Resumo

O pressuposto da normalidade é uma conveniência matemática. Em muitos casos, a distribuição de características quantitativas não é normal, apresentando caudas pesadas que por seu turno fazem com que observações regulares aparentem ser outliers. Entretanto, é também sabido que os métodos clássicos que se baseiam na verosimilhança da distribuição normal têm um fraco desempenho aquando da violação deste pressuposto. Tal é igualmente o caso do teste F da ANOVA cuja potência fica comprometida nestas circunstâncias. Os métodos robustos por seu lado, são desenhados por forma a serem fiáveis quando não se verificam alguns dos pressupostos subjacentes aos métodos clássicos, por exemplo a violação da condição da normalidade dos erros num modelo de regressão. Neste sentido, pretendemos comparar a abordagem clássica com duas alternativas não paramétricas e a regressão-M no contexto dos estudos de associação genética de características quantitativas em que temos marcadores genéticos como variáveis explicativas e se usam os modelos de regressão para aferir possíveis associações. Os resultados mostrarão a vantagem da metodologia robusta proposta bem como os perigos que podem advir da utilização da abordagem clássica. Adicionalmente, proporemos um coeficiente de determinação robusto para aproximar a proporção de variabilidade fenotípica que é explicada pela componente genética do modelo e apresentaremos uma proposta robusta para detecção de outliers onde são avaliadas várias medidas de correcção para testes de multiplicidade.

Palavras Chave: não-normalidade; regressão-M; estudos de associação genéticos; polimorfismo de nucleótido único (SNP); herdabilidade; outliers; testes de outliers; medidas de correcção em testes de hipóteses múltiplos (MTPs); taxa de falsas descobertas (FDR); coeficiente de determinação.

Abstract

Data normality is just a mathematical convenience. For many real-life data sets, the distribution of continuous variables is not normal, often showing heavy tails, which in turn tend to make regular observations look like outliers. In such scenarios, the classical approach, whose likelihood-based inference leans on the normality assumption and is known to be non-robust to small model deviations, may be inappropriate. Also, contrary to some statements in the literature, the ANOVA F-test is not robust against non-normality often seeing its power seriously compromised. Robust statistical methods are designed to be resistant to influent factors such as outlying observations, non-normality and other model misspecifications, allowing for reliable results. We analyze the case of statistical tests to detect associations between genomic individual variations and quantitative traits when deviations from the normality assumption are observed. We consider the classical ANOVA tests for the parameters of the appropriate linear model and a robust version of those tests based on M-regression. We compare both approaches with two non-parametric alternatives, in terms of empirical power and level using simulated data with several degrees of contamination. Results will show that the robust methodology can be more powerful and thus more adequate for association studies than the classical approach. Additionally, we will propose a robust coefficient of determination to approximate the phenotypic variation that is explained by the models' genetic component and a robust outlier test where several multiple correction measures are evaluated.

Keywords: non-normality; M-regression; genetic association studies; single nucleotide polymorphism (SNP); heritability; outliers; outlier tests; multiple testing procedures (MTPs); false discovery rate (FDR); coefficient of determination.

Acknowledgements

I would like to express my deep gratitude to my advisor, Professor Ana Maria Pires, for her support and guidance throughout this academic laborious path. I thank her as well for being such an inspiration and most of all, for her continuous belief in my work even on those times of back and forth that usually characterize investigation.

To CEMAT I thank the financial support that allowed me to attend important conferences indoors and abroad.

To my children Carolina and Rafael, I leave my full appreciation for their unconditional support and encouragement, especially in the past year when work seriously volumed. I hereby promise to catch up on all those weekends and holidays that were lost.

I thank also António, my friend and companion, for his constant encouragement. For being the pillar that held up my world from the very beginning and above all things, for loving me.

To my grandfather João, who is now almost 88 years old, I must leave a no less big thanks for putting his faith on me and my work. This accomplishment is a dream come true for both of us. To my grandmother Amélia, no longer among us, I miss you..

I leave my full recognition to my mother Emilia, for her priceless help during the past years. She has undoubtedly been a vital part of this fulfillment and so, thanks mom! To my dad Arlindo as well, thank you for being deprived of mother so many times!

For last but not least, I thank the rest of the family and most dear friends for sharing this journey with me without complaining for the times I just couldn't be there!

Publications

Lourenço V.M., Pires A.M. and Kirst M. (2011). Robust linear regression methods in association studies. *Bioinformatics* **27(6)**, 815-821. doi: 10.1093/bioinformatics/btr006.

Lourenço V.M. and Pires A.M. Outlier detection in genetic association studies of quantitative traits: integrating M-regression and false discovery rates. *In progress.*

Abbreviations

ANOVA	**AN**alysis **O**f **VA**riance
BH	**B**enjamini and **H**ochberg
BKY	**B**enjamini, **K**rieger and **Y**ekutieli
BLUE	**B**est **L**inear **U**nbiased **E**stimator
bp	**b**ase **p**air
BP	**B**reakdown **P**oint
BY	**B**enjamini and **Y**ekutieli
CWER	**C**omparison **W**ise **E**rror **R**ate
DFFITS	**DiF**erence in **Fit S**tandardized
DNA	**D**eoxyribo**N**ucleic **A**cid
EM	**E**xpectation **M**aximization
FDR	**F**alse **D**iscovery **R**ate
FERL	**F**emale **EaR L**ength
FWER	**F**amily **W**ise **E**rror **R**ate
F1	First **F**ilial
F2	Second **F**ilial
HWE	**H**ardy-**W**einberg **E**quilibrium
IBD	**I**dentity **B**y **D**escent
IF	**I**nfluence **F**unction
IRWLS	**I**teratively **R**e**W**eighted **L**east **S**quares
KR	**K**enward-**R**oger
LD	**L**inkage **D**isequilibrium
LMS	**L**east **M**edian of **S**quares
LRT	**L**ikelihood **R**atio **T**est
LS	**L**east **S**quares
LTS	**L**east **T**rimmed **S**quares
ML	**M**aximum **L**ikelihood
MLE	**M**aximum **L**ikelihood **E**stimator/**E**stimate
mRNA	**m**essenger **R**ibo**N**ucleic **A**cid
MSE	**M**ean **S**quare **E**rror
MTP	**M**ultiple **T**esting **P**rocedure
NR	**N**ewton-**R**aphson

OLS	**O**rdinary **L**east **S**quares
PCER	**P**er **C**omparison **E**rror **R**ate
pFDR	**p**ositive **F**alse **D**iscovery **R**ate
PFER	**P**er **F**amily **E**rror **R**ate
QQ-Plots	**Q**uantile-**Q**uantile **P**lots
REML	**RE**stricted **M**aximum **L**ikelihood
RNA	**R**ibo**N**ucleic **A**cid
RT	**R**ank-**T**ransform
SF	**S**hapiro-**F**rancia
SNP	**S**ingle **N**ucleotide **P**olymorphism
SSE	**S**um of **S**quared **E**rrors
SST	**T**otal **S**um of **S**quares
ST	**S**torey and **T**ibshriani
STP	**S**imultaneous **T**est **P**rocedure
tRNA	**t**ransfer **R**ibo**N**ucleic **A**cid
UMVUE	**U**niform **M**inimum **V**ariance **U**nbiased **E**stimator
WIL	**WIL**coxon
WY	**W**estfall and **Y**oung

Contents

Resumo i

Abstract ii

Acknowledgements iv

Publications vi

Abbreviations viii

1 Introduction 1

1.1 Robust statistics . . . 2

1.2 Genetic association studies of quantitative traits . . . 4

1.3 The multiple testing problem . . . 9

1.4 Thesis goal and outline . . . 10

2 A short insight on the basics of genetics 13

2.1 General concepts . . . 13

2.2 Mendel's experiment, findings and laws . . . 19

2.3 Population genetics . . . 24

2.3.1 Hardy-Weinberg equilibrium . . . 24

2.3.2 Recombination and linkage desiquilibrium. . . . 27

2.4 Quantitative genetics . . . 30

2.4.1 Genetic models . . . 31

2.4.2 Heritability . . . 32

3 Multiple hypothesis testing 35

3.1 Controlling the family wise error rate . . . 36

3.2 Controlling the false discovery rate . . . 49

4 Multiple linear regression 61
4.1 The regression model . . . 62
4.2 Least squares and maximum likelihood estimators . . . 62
4.3 M-estimators . . . 64
4.4 Testing the models parameters . . . 69
4.4.1 The classical approach . . . 69
4.4.2 The robust approach . . . 70
4.5 Coefficient of determination . . . 71
4.6 Outlier detection . . . 72
4.7 Estimation and tests in the mixed linear model setup . . . 81
4.7.1 Using maximum likelihood for the estimation of Φ^* . . . 83
4.7.2 Hypothesis testing . . . 83
4.7.3 Coefficient of determination . . . 85

5 Simulation study I 87
5.1 The simulation model . . . 88
5.2 Choosing the simulation model's parameters . . . 90
5.3 Simulation settings . . . 93
5.4 Simulation results . . . 94
5.5 Discussion . . . 101

6 Simulation study II 103
6.1 False discovery rate control in outlier tests: a robust approach. . . . 104
6.2 Coefficient of determination and trait heritability . . . 120

7 Example 125
7.1 Example data and methodology . . . 126
7.2 Single-SNP analysis . . . 129
7.3 Multiple-SNP analysis . . . 131

8 Discussion 137

A APPENDIX A: Some extra tables 143

B APPENDIX B: Some extra figures 147

C APPENDIX C: R routines 167
C.1 R code for the phones data example on Section 4.6 . . . 168
C.2 Description of the R `lmekin` routine . . . 169
C.3 Description of the R robust version of the R `lmekin` routine . . . 173

C.4 R routine for the example robust single-SNP analysis 177
C.5 R routine for the example robust multiple-SNP analysis 180
C.6 R routine for the example classic single-SNP analysis 182
C.7 R routine for the example classic multiple-SNP analysis 186

References **191**

List of Figures

2.1 Double Helix. *Image from Wikipedia.* 14

2.2 DNA strand. *Image from Wikipedia.* 16

2.3 Crossover scheme: 2 homologous chromosomes crossover during meiosis producing 4 distinct gametes. *Image from* `www.bio3400.nicertutor.com`. 18

2.4 Mutation types I: deletion (1), duplication (2) and inversion (3). *Image from Wikipedia.* 18

2.5 Mutation types II: insertion (1) and translocation (2). *Image from Wikipedia.* 18

2.6 Single Nucleotide Polymorphism (SNP): the same chromosome is genotyped for several individuals in a population and variation is observed in 3 distinct sites. *Image from The Individualist HapMap.* 19

4.1 Huber and LS function plots; $b = 1.345$. 68

4.2 Plot of *year* versus *calls* (black dots), with q-value rescaled plot superimposed (blue dots), plus the regression line (black dashed line) and FDR 10% rescaled threshold (green dashed line) 80

5.1 P-value histogram of the global tests for association under H_0 in the 10SNP model. Recall that we only have 1000 observations in the Wilcoxon approach 94

5.2 Power comparison between classical (in blue) and robust analysis (in black) from simulation under H1, with and without a % of smooth and gross data contamination, for the 2, 5 and 10 SNPs models. 99

5.3 Q-Q Plots for the 10 SNP model under H_1 with 2% Gross, Intermediate and Smooth contamination, from one out of the 10000 simulations, for the classic and robust analysis only. 100

6.1 Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.3$, a percentage of $2, 5$ and 10% of smooth data contamination and robust parameter of scale s. Benjamini and Hochberg (1995) correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust standardized residuals (4.32). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1. 105

6.2 Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.3$, a percentage of $2, 5$ and 10% of smooth data contamination and robust parameter of scale Q_n. Benjamini and Hochberg (1995) correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust standardized residuals (4.32). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1. 105

6.3 Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.3$, a percentage of $2, 5$ and 10% of smooth data contamination and robust parameter of scale s. Benjamini and Hochberg (1995) correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust studentized residuals (6.1). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1. 107

6.4 Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.3$, a percentage of $2, 5$ and 10% of smooth data contamination and robust parameter of scale Q_n. Benjamini and Hochberg (1995) correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust studentized residuals (6.1). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1. 107

6.5 Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.3$, a percentage of $2, 5$ and 10% of smooth data contamination and robust parameter of scale s. Bonferroni correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust standardized residuals (4.32). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1. 108

6.6 Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.3$, a percentage of $2, 5$ and 10% of smooth data contamination and robust parameter of scale Q_n. Bonferroni correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust standardized residuals (4.32). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1. 108

6.7 Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.5$, 2% of smooth data contamination and robust parameter of scale Q_n. Benjamini and Hochberg (1995) correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust standardized residuals (4.32). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1. 110

6.8 Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.5$, 2% of smooth data contamination and robust parameter of scale Q_n. Bonferroni correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust standardized residuals (4.32). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1. 110

6.9 Observed FDR for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.75$, 2% of smooth data contamination and robust parameter of scale Q_n. Benjamini and Hochberg (1995) correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust standardized residuals (4.32). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1. 111

6.10 Observed FDR for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.75$, 2% of smooth data contamination and robust parameter of scale Q_n. Bonferroni correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust standardized residuals (4.32). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1. 111

6.11 Observed FDR in models with up to 10 SNPs for several alternatives to the BH-1995 FDR correction. Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1. 114

6.12 Observed number of True Positives in models with up to 10 SNPs for several alternatives to the BH-1995 FDR correction. Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1. 115

6.13 FDR observed values for nominal levels of $\alpha = 0.01, 0.05, 0.1$ in models with 10 SNPs. BH-2000 FDR correction was used. Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1. 117

6.14 Number of True Outliers observed for nominal levels of $\alpha = 0.01, 0.05, 0.1$ in models with 10 SNPs. BH-2000 FDR correction was used. Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1. Black dots refer to the total amount of outliers in the model. 118

6.15 Number of True Outliers observed for nominal level of $\alpha = 0.1$ in models with 10 SNPs. BH-2000 FDR correction was used. 119

7.1 Histograms of the 498 p-values for both simple and full models, in the classical and robust analysis. 127

7.2 Cumulative distribution of the 498 p-values for both simple and full models, classical method. 128

7.3 Conditional residual plots and QQ-normal plots from the multiple-SNP analysis. 132

B.1 Q-Q Plots for the 10 SNP model with 2% gross, intermediate and smooth contamination. 148

B.2 Q-Q Plots for the 10 SNP model with 5% gross, intermediate and smooth contamination. 149

B.3 Q-Q Plots for the 10 SNP model with 10% gross, intermediate and smooth contamination. 150

B.4 Q-Q Normal plots of residuals from the 10 SNP model association analysis with 2% smooth contamination, with SF p-values of 0.00001, 0.00003 and 0.00045, respectively, corresponding to Pearson p-values of 0.134, 0.256 and 0.379. 151

B.5 Q-Q Normal plots of residuals from the 10 SNP model association analysis with 2% smooth contamination, with Pearson p-values of 0.0249 and 0.0352 corresponding to SF p-values of 0.000001 and 0.0009, respectively. 151

B.6 Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.3$, a percentage of $2, 5$ and 10% of smooth data contamination and robust parameter of scale s. BH-1995 correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust standardized residuals (4.32). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1. 152

B.7 Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.3$, a percentage of $2, 5$ and 10% of smooth data contamination and robust parameter of scale Q_n. BH-1995 correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust standardized residuals (4.32). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1. 153

B.8 Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.3$, a percentage of $2, 5$ and 10% of smooth data contamination and robust parameter of scale s. BH-1995 correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust studentized residuals (6.1). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1. 154

B.9 Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.3$, a percentage of $2, 5$ and 10% of smooth data contamination and robust parameter of scale Q_n. BH-1995 correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust studentized residuals (6.1). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1. 155

B.10 Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.3$, a percentage of $2, 5$ and 10% of smooth data contamination and robust parameter of scale s. Bonferroni correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust standardized residuals (4.32). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1. 156

B.11 Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.3$, a percentage of $2, 5$ and 10% of smooth data contamination and robust parameter of scale Q_n. Bonferroni correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust standardized residuals (4.32). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1. 157

B.12 Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.3$, a percentage of $2, 5$ and 10% of smooth data contamination and robust parameter of scale s. Bonferroni correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust studentized residuals (6.1). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1. 158

B.13 Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.3$, a percentage of $2, 5$ and 10% of smooth data contamination and robust parameter of scale Q_n. Bonferroni correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust studentized residuals (6.1). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1. 159

B.14 Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.5$, 2% of smooth data contamination and robust parameter of scale s. BH-1995 correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust standardized residuals (4.32). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1. 160
B.15 Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.5$, 2% of smooth data contamination and robust parameter of scale s. Bonferroni correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust standardized residuals (4.32). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1. 160
B.16 Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.75$, 2% of smooth data contamination and robust parameter of scale s. BH-1995 correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust standardized residuals (4.32). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1. 161
B.17 Observed FDR True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.75$, 2% of smooth data contamination and robust parameter of scale s. Bonferroni correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust standardized residuals (4.32). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1. 161
B.18 Number of True Outliers observed for nominal level of $\alpha = 0.05$ in models with 10 SNPs. BH-2000 FDR correction was used. 162
B.19 Cumulative distribution of the 498 p-values for both simple and full models, in the classical and robust analysis. 163
B.20 Cumulative distribution of the 498 p-values for both simple and full models - robust methodology. 164
B.21 QQ-normal residual plots from the classic single-SNP analysis. 165
B.22 QQ-normal residual plots from the robust single-SNP analysis. 166

List of Tables

2.1 Codon/amino-acid table. Codon `AUG` coding for *methionine* (Met) also serves as the start codon if it is preceded by a string of alphabets, much longer than a single codon, called the *promoter* sequence. 15

2.2 Crossing between F_1 `Aa` hybrids. 21

2.3 Mendel's experiment results in the crosses between plants differing on one of the seven characters. 22

2.4 Results of the Pearson's Chi-square test on Mendel's data. 22

2.5 Hybrid crossing scheme under the principle of independent assortment . . . 23

2.6 Resulting genotypes and phenotypes from the previous hybrid crossing scheme. 23

2.7 Punnet square for HWE: shows the allelic frequencies from random mating with two alleles. 25

2.8 Factors that condition LD behaviour in a population (Rafalski and Morgante, 2003; Pritchard and Przeworski, 2001). 28

2.9 Expected frequencies of gene `A` genotypes for an F2 population 31

2.10 Expected frequencies of genotypes for gene `B` conditional on gene `A` when the two are linked with recombination fraction r for an $F2$ population . . . 32

3.1 Number of errors committed when testing m null hypothesis 51

3.2 Summary table of the discussed multiple testing procedures which have strong control of either the FWER or the FDR - some properties. 59

5.1 Classic POWER (%) obtained from simulation under H_1 without outliers, for the global test of association in the 2 SNP simulation model, for several values of the recombination fraction, r, relative additive and dominance effects, k, and number of individuals, n. 91

5.2 FWER (in %) obtained from simulation* under H_0 with a percentage of contaminated traits - (iii) setting under H_0 ** 96

5.3 POWER (in %) obtained from simulation under H_1 without contamination and with a percentage of grossly (G), intermediate (I) and smoothly (S) contaminated traits. (truncated minimum power observed to detect every SNP in the simulation model) 97

5.4 Truncated, average number of SNP detected in the simulation under H_1 without contamination and with a percentage of grossly (G), intermediate (I) and smoothly (S) contaminated traits. 98

5.5 POWER (%) obtained from simulation under H_1, without contamination and with a percentage of contaminated traits for the global test of association in the 10 SNP simulation model. 101

6.1 Average number of detected true positives corresponding to Figures 6.1 and 6.2 - for the 2% contamination setting only, which amounts to a total of 10 contaminant observations. 106

6.2 Average number of detected true positives corresponding to Figures 6.5 and 6.6 - for the 2% contamination setting only, which amounts to a total of 10 contaminant observations. 106

6.3 Values of the observed FDR obtained using Bonferroni correction at the nominal levels of $\alpha = 0.01,\ 0.05$ and 0.1, with the robust estimate of scale Q_n and $H^2 = 0.3$, for models with $n/5/2$ independent SNPs. 112

6.4 Values of the observed FDR obtained using Bonferroni correction at the nominal levels of $\alpha = 0.01,\ 0.05$ and 0.1, with the robust estimate of scale Q_n and $H^2 = 0.3$, for models with $\lceil n/10/2 \rceil$ independent SNPs. 113

6.5 Number of detected true positives corresponding to Figure 6.12 for the three FDR controlling approaches, where there is a total of 10 outliers. . . 116

6.6 Mean values of the adjusted coefficient of correlation R_a^2 for all the models considered in the 1000 simulations with varying heritabilities and percentage of data smooth contamination - $n = 500$ 121

6.7 Mean values of the adjusted robust coefficient of correlation R_a^2 for all the models considered in the 1000 simulations with varying heritabilities and percentage of data smooth contamination. Only 20% of the SNPs in the models are in association with the trait. 122

6.8 Mean values of the adjusted robust coefficient of correlation R_a^2 for all the models considered in the 1000 simulations with varying heritabilities and percentage of data smooth contamination. Only 10% of the SNPs in the models are in association with the trait. 123

6.9 Mean values of the adjusted robust coefficient of correlation R_a^2 for all the models considered in the 1000 simulations with varying heritabilities and percentage of data smooth contamination. Not all the SNPs in the model were used to generate the quantitative trait. 124

7.1 Significant SNPs detected at the 5% level and with FDR under 10% in the single-SNP analysis. 129
7.2 Results from the Shapiro-Francia test on the models' residuals. 131
7.3 Diagnostic measures for the influence of the eight plants identified as outliers by the rule of thumb. Last column refers to the weights given by the robust analysis. 133
7.4 Multiple-SNP analysis results of the SNPs declared significant in the single-SNP analysis. P-values are raw. 135

A.1 Candidade genes for the quantitative phenotype FERL. 144
A.2 Pairwise LD r^2 for the 9 SNPs detected in the single-SNP analysis. 145

The greater our knowledge increases the more our ignorance unfolds.

John F. Kennedy

1 Introduction

Man has been interested in genetics since FOREVER! Whether concerned with animal breeding, farming or even his own ancestry, questions like "Why do progeny resemble their parents?" or "Why sometimes they do not?" occupied the mind of men since the beginning of men-hood, giving rise to the most various theories. Early stock breeders and farmers knew from experience that offsprings inherited characteristics from their parents using that empirical knowledge to their advantage. Also, and from experience alone, they knew that sometimes variation within breeding groups occurred and although they could not answer the question "Why?", still the struggle with the laws of inheritance continued.

It was not until the 19^{th} century, with the discovery of the gene, that those questions started being answered. Gregor-Mendel ($1822 - 1884$), an Austrian Augustinian monk, while breeding different cultivars of peas, faced himself with those questions and set up to find some answers. He was the first known person to have experimented with the purpose of understanding variation.

By that time, science was worrying and discussing the theory of evolution and hence was not prepared to embrace Mendel's amazing results. Its importance was only recognized at the beginning of the 20^{th} century, many years after his death. Mendel's work was rediscovered by three botanic scientists, namely, Hugo de Vries, Erich Von Tschermak

and Karl Correns. These men tripped on Mendel's work by chance when searching the scientific libraries for possible precursors on their own work. They promptly announced Gregor Mendel's work to the world and science finally saw the light!

Nonetheless, in due time another headache soon began: how to appropriately address and conduct genetic association studies and how could statistics fit in helping to uncover patterns, relationships and variability in the data? At this point we must refer to the works of Fisher (1930), Wright (1931) and Haldane (1932) who, among others not less important, brought together evolutionary biology and genetics into a whole that could be quantitatively modeled.

1.1 Robust statistics

In the last 40/50 years there has been a major effort in developing statistical procedures that are resistant to small model's assumptions deviations. Namely, procedures that are robust to outliers and stable with respect to deviations from the normality assumption, in the case of the classic parametric model (see Tukey, 1960; Huber, 1964 and 1972; Hampel, 1968; Huber and Ronchetti, 1981; Rousseeuw and Leroy, 1987, for the foundations of robust statistics and Maronna *et al.*, 2006; Ronchetti, 2006; Dell'Aquila and Ronchetti, 2006, for an overview). This interest is motivated, among others, by the fact that the removal of outlying observations is, first of all, not an easy task and secondly it may not be the best way to proceed because

i. true outliers are not always visible due to masking and swamping effects;

ii. gross outliers are easily identifiable but that is not usually the case of mild outliers, which sometimes cannot be differentiated from regular data;

iii. outliers may contain relevant information for the analysis;

iv. removing outliers reduces sample size, may affect the distribution theory and variances may be underestimated from the cleaned data.

Also, data transformation may not always be sufficient to solve all the problems caused by non-normality frequently raising interpretation issues (Pires and Rodrigues, 2007).

Robust methods are designed to accommodate for certain data deficiencies, allowing for reliable results under various conditions. They ought to be resistant to influent factors such as outlying observations, non-normality and other model misspecifications (Huber *et al.*, 1972; Hampel *et al.*, 1986; Rousseeuw and Leroy, 1987; Maronna *et al.*, 2006;

Daszykowski *et al.*, 2007). Moreover, if the model verifies the classical assumptions, the robust methods provide results close to the classical ones. Therefore, the use of robust methods has been advocated for inference in the linear model setup (Copt and Feser, 2006; Pires and Rodrigues, 2007; Daszykowski *et al.*, 2007).

For many real-life data sets, the distribution of continuous variables is not normal often showing heavy tails, which in turn tend to make regular observations look like outliers. This is mainly the reason why non-normality and outlier presence are usually associated.

In such scenarios, the classical approach, whose likelihood-based inference leans on the normality assumption and is known to be non-robust to small model deviations (Tuckey, 1960; Huber, 1964/1967), may be inappropriate, having low statistical efficiency if the tails are symmetric and large bias if the tails are asymmetric. After all, the ordinary least squares (OLS) estimator is known to be quite sensitive to outliers (Jajo and Hussain, 1989; Jajo, 1993) and long-tailed distributions having poor efficiency relative to many robust estimators when the errors are not normally distributed.

This evidently leads to tests with unreliable level and low power and to confidence intervals with also unreliable level and large expected interval length. We emphasize the fact that, contrary to some statements in the literature dating back to Box (1953), the ANOVA F-test is not robust against non-normality (Schrader and Hettmansperger, 1980; Tan, 1982; Hampel *et al.*, 1986; Ronchetti, 1987; Zimmerman, 1994; de Haan *et al.*, 2009) and consequently efforts have been made in order to enhance its performance (Ronchetti, 1982; Copt and Heritier, 2007).

So, we have seen that data transformation comes with interpretation problems and should therefore, if possible, be avoided. Outlier removal, as well, comes at the cost of losing potential important information and should thus not be an alternative to the use of methodologies that are able to accommodate for them. However, outlier identification is of the utmost importance to the understanding of the underlying mechanisms of the system described by the data because they are not always a result of measurement error. In fact, they sometimes follow from intrinsic characteristics of the individual being measured. As an example, in the context of environmental sciences, an outlier could very well represent an unusual event that took place (Daszykowski *et al.*, 2007) and whose comprehension could be helpful for the study. And this happens not only in environmental sciences but in many other areas of research like e.g, in genetic association studies (Tzeng *et al.*, 2003).

It is not therefore surprising the amount of literature dedicated to outlier detection, especially with application to multivariate and high dimensional data, where the search for more effective outlier methods/tests that are not subject to masking comprehends the search for measures of location and scale that are not themselves affected by outliers (Wilcox, 2010). We mention only the three more cited books on the subject and a more recent publication: the books of Barnett and Lewis (1978); Hawkins (1980); Rousseeuw and Leroy (1987) and Hettmansperger and McKean (2001). And also the papers of Chatterjee and Hadi (1986); Boente *et al.* (2002); Oh and Gao (2009).

Another very important point is the effect of outliers in the coefficient of determination in linear regression models. The coefficient of determination should reflect the suitability of the explanatory variables in predicting the response variable, i.e, the amount of variation of the response variable that is explained by the variation among the explanatory variables. In the classical setting however, it can become misleading in the presence of a single outlier. There are many alternatives to the least squares coefficient of determination in the literature, amongst which some robust proposals (Huber, 1981; Rousseeuw, 1984; Rousseeuw and Yohai, 1984; Greene, 1997; Croux and Dehon, 2003; Maronna *et al.*, 2006; Heritier *et al.*, 2009; Renaud and Feser, 2010). A recent comparative study of the most familiar measures of R^2 can be found in Yarmohammadi and Mahmoudvand (2010).

For an overview of robust statistics in all its dimensions, the literature already mentioned, plus the one that will be mentioned throughout this work, is more than enough for an "outsider" to be acquainted with this particular area of statistics that has been around since the sixties, and which finds in our time, with the advent of technology, a new breakthrough. For those who prefer Portuguese, we have Pires and Branco (2007).

1.2 Genetic association studies of quantitative traits

Genetic association studies aim to identify genetic polymorphisms that cause phenotypic variation for a trait of interest, or that are in linkage disequilibrium (LD) with the causative genetic variant. If association is present, a particular allele, genotype or haplotype of a polymorphism or polymorphism(s) will be seen more often than expected by chance in an individual carrying the trait. For example, an individual carrying one or two copies of a high-risk variant is at increased risk of developing the associated disease or having the associated trait.

A quantitative trait is a measurable trait that shows continuous variation through a

population, such as height, weight or other. Association studies of quantitative traits can be performed using an unrelated population sample or family trios in which the quantitative trait is measured in the offspring. However, in population-based studies, associations found are valid population wide and not only for the group of individuals analyzed, which is often the case in family based-studies. Some other advantages of population-based studies over family-based are, now in the context of plant genetics: (i) the use of natural populations circumvents the need to raise large controlled crosses; (ii) the costs associated to finding natural populations as opposed to generating sourcing pedigrees are low.

In population-based genetic association studies of quantitative traits, the unit of analysis is the single nucleotide polymorphism (SNP). The strategy behind these studies is to explore linkage disequilibrium (LD) between SNPs in natural populations that have been mating for generations. The idea is that recombinations that occur through generations of random mating break down every possible association of phenotype with all but the most tightly linked SNPs. This LD can be statistically detected and has been widely applied to map and eventually clone a number of genes underlying complex genetic traits in humans (Abdurakhmonov and Abdukarimov, 2008).

In particular, we are interested in using a number of genotyped SNPs in a gene, or region, to detect the genetic factors underlying a quantitative trait of interest that does not follow simple Mendelian patterns of inheritance. The most straightforward and still more favoured approach in association studies, though raising multiple testing problems (Nyholt, 2004), is to perform a single SNP test for every genotyped SNP via regression or ANOVA methods: Tao and Boulding (2003), used a linear model to test the association between SNPs in 8 candidate genes and age-specific growth rate in the Artic charr; Martínez *et al.* (2007), used mixed linear models to test for association between 57 SNPs from 20 candidate genes and some wood properties in *Pinus taeda*; Weber *et al.* (2007; 2008), used a mixed random effects linear model to test for the association between a collection of SNPs, from a set of candidate genes, and several Teosinte traits of interest; Moe *et al.* (2009) used a linear model to test for association between 151 SNPs from 57 candidate genes and several traits of boar related to taint and reproduction; Jiang *et al.* (2010) used a mixed linear model to test for association between 39163 SNPs and five milk production traits in chinese holstein population. Many more examples can be easily found in the literature.

Though this approach may be considered if we are looking for a single causal variant,

it is not very efficient when the SNPs have limited LD with that causal variant, meaning smaller power. Plus, a single-SNP analysis may neglect information in the SNPs joint distribution, situation that may be of little consequence if SNPs are widely spaced in the genome as to have little or no LD between them or when we have all SNPs genotyped so that the causal variant is likely to be among them (Balding, 2006). Unfortunately, or not, most studies have SNP densities lying in between these two situations.

When it comes to quantitative traits, we know that they are usually controlled by several and sometimes many genes. Thus, a joint analysis of SNPs may be more adequate, being much more informative than single-SNP analysis (Jannot *et al.*, 2003). However, it also may loose power due to the usually large number of degrees of freedom involved. Ideally, one should make use of the information provided by multiple SNPs, capturing as much of the genetic variance as possible, without raising the degrees of freedom too much (Bureau *et al.*, 2005; Wang and Elston, 2006; Chapman and Whittaker, 2008; Xiang *et al.*, 2008/2009; Kwee *et al.*, 2008; Li *et al.*, 2009) and thus not compromising power. Note that the joint analysis of SNPs (multiple-SNP approach) can only be applied to situations where the number of explanatory variables is much smaller than the number of individuals, therefore implying that in a genome-wide association study context a preliminary step of dimension reduction is necessary.

Two major problems that can arise when testing for association between a quantitative trait and a set of candidate SNPs in a population-based study, and for which there is extensive literature, are: LD, i.e, there may be SNPs whose alleles are in moderate to strong LD which may produce, due to the induced correlation among SNPs, misleading results in standard multiple regression analysis that do not account for this problem; population structure/stratification (PS; Pritchard *et al.*, 2000a and 2000b; Bacanu *et al.*, 2002; Cardon and Palmer, 2003; Freedman *et al.*, 2004; Yu *et al.*, 2006) and within-population kinship among individuals, inflating the number of false positives and failing to detect true associations, due to a confounding effect on the tests.

Several methods have also been proposed in literature to accommodate for LD and also to correct, and therefore reduce, the impact of PS on population-based association studies: Ridge regression (Malo *et al.*, 2008), tagSNP selection (Zhang *et al.*, 2002; Eskin *et al.*, 2003; Cousin *et al.*, 2003; Carlson *et al.*, 2004; Takeuchi *et al.*, 2005) structured association (SA) (Pritchard *et al.*, 2000a and 2000b; Li *et al.*, 2008), genomic control (GC) (Devlin and Roeder, 1999/2001; Bacanu *et al.*, 2000; Zheng *et al.*, 2006) and principal component analysis (PCA) (Price *et al.*, 2006; Wang and Abbot, 2008). The three latter

methodologies are compared in Zhang *et al.* (2008).

Recently, Yu *et al.* (2006) introduced a unified mixed model approach to account for multiple levels of relatedness, where population structure is modeled simultaneously as a fixed effect, through a Q matrix (derived using STRUCTURE: Pritchard *et al.*, 2000a; Falush *al*, 2003), and as a random effect, using a matrix K of pairwise kinship coefficients. Price *et al.* (2006) suggest that Q can be replaced by a matrix derived from PCA, due to STRUCTURE demanding algorithm, and Zhao *et al.* (2007) show that both approaches perform similarly.

Another frequent data problem, which may have the same sort of undesirable effects, is non-normality and/or presence of outliers in the phenotypic data. This problem is far less studied than LD or PS. For instance, the review paper by Balding (2006) treats non-normality in one sentence where the only mentioned remedy is a transformation of the original trait values. Also, in the literature, data transformation seems to be the tip of the day: Tao and Boulding (2003) used a cubic root transformation; Weber *et al.* (2008) a square root transformation; Kardia *et al.* (2008) a logarithmic transformation; Smith *et al.* (2009) a natural logarithmic transformation; ... and so forth.

Some alternatives have been proposed in the literature to be used as a way of circumventing violations of model assumptions. E.g: non-parametric methods (Conover and Iman, 1981; Wang and Huang, 2002; Hicks *et al.*, 2004; McKean, 2009; Beasley and Erickson, 2009; Gerhard and Hothorn, 2010; Wilcox, 2010); Bayesian methods (Rosa *et al.*, 2003; Pereira *et al.*, 2007); semi-parametric methods (Li *et al.*, 2008); *et cetera*. At this point we must mention that the semi-parametric model proposed by Li *et al.* (2008), with the aim of dealing with error distribution non-normality, was compared in their work with the classical approach where a log-transformation of the data was made. They concluded that their proposed method performed better than the classical plus data transformation approach.

Also, we already see some applications of the robust methodology in genetic association studies (Xu *et al.*, 2010; Tan *et al.*, 2010; Gudbjartsson *et al.*, 2010) which show increasing concern with the violation of model assumptions and growing interest in using methods that are capable of coping with them.

In the classical analysis, when the residuals' distribution fails normality tests (i.e, if normality tests are performed, which they usually are not) or looks non-normal, instead of

transforming the data, we frequently encounter in the literature another approach to the problem: identification and removal of outlying observations from the study. By doing this, the investigator usually gets the residuals' distribution to be approximately normal and thus the opportunity of performing the analysis via the classical approach without "fear". One can easily find in the literature examples of genetic association studies where outlying observations were eliminated from the bulk of the data after some rule of thumb: Pachou *et al.* (2008); Arshadi *et al.* (2009); Inouye *et al.* (2010). However we do not agree with this approach and our reluctance with it will be fully justified by the results observed in Chapter 5.

An important population parameter in genetic association studies of quantitative traits is heritability, which corresponds to the total phenotypic variation that is explained by genetic factors. The estimation of this parameter is of the utmost importance since response to artificial and natural selection and the degree of resemblance between relatives are functions of it. Knowledge of heritability is therefore very useful for e.g, choosing an efficient breeding scheme, estimating the gain to be expected under mass selection and constructing a selection index (Kempthorne and Tandon, 1953). In a plant genetic context for example, good knowledge of heritability in different yield parameters is a prerequisite for effective plant improvement (Waqar-Ul-Haq *et al.*, 2008). A good review on heritability estimation for plant breeding can be found in Holland and Martinez (2003). Usual approaches to the estimation of heritability involve moment and likelihood estimators that are used in variance components analysis (Devlin and Roeder, 1997; Thomas, 2005; Visscher *et al.*, 2006). Ritland (2000) proposes a linear regression marker-based method for estimating heritability of quantitative traits in natural populations.

In genetic association studies of quantitative traits, where we know that the quantitative trait is controlled by several to many genes/SNPs, at the candidate gene level our genotyped SNP data usually comprehend only a portion of those SNPs. The model used for those studies should therefore reflect as accurately as possible that proportion of heritability that is explained by the genotypic data. Additionally, should we have an idea in advance of the heritability of the trait in hand, from previous within family studies or other, that model information would be valuable for the search of the un-genotyped SNP markers responsible for the remaining unexplained heritability.

To end this brief incursion in genetic association studies of quantitative traits, we must refer to another very important aspect of the analysis, which is data quality control (do not mistake it with genomic control - Devlin and Roeder, 1999/2001). We will not go

through this point here but we find it mandatory reading for anyone engaging in these studies and therefore leave some starting useful literature: Gomes *et al.* (1999); Balding (2006); Sebastiani *et al.* (2009).

Much more is to say on the subject of genetic association studies of quantitative traits, so if the reader feels the need to learn more about it we strongly advise the books of Lynch and Walsh (1997), Balding *et al.* (2003) and of Wu, Ma and Casella (2007) as good starting points.

1.3 The multiple testing problem

When we assign for example a probability of $\alpha = 0.05$ to a type-I error in a set of tests, it means that virtually one out of every twenty such tests will show a false positive. So, in a set of 100 tests we expect 5 tests to be declared significant if we use that α threshold for each test. The problem with this is that we would like to control the false positive rate not only for one test but for all family of tests comprehending our study. And this is what is called the multiple testing problem.

Multiple testing, being concerned with testing several hypotheses simultaneously, is a thorny issue in many areas of research. Citing Westfall and Young (1993), a major drawback of multiple testing is the increased probability of declaring statistically significant associations where none exist. Researchers need therefore to use multiple testing procedures (MTPs) in order to be able to control for false inferences as well as maintaining power in detecting real effects.

Genetic association studies of quantitative traits, as many other studies in other areas of research, are prone to the multiple testing problem due to the usual large amount of hypotheses being simultaneously tested, situation that escalates in genome-wide genetic association studies, e.g, when tens to hundreds of thousands of SNPs are being tested for association with a given trait. Balding (2006) even called it *the bane of statistical genetics*.

This problem has received a lot of attention from people working in the field of genetics (Westfall *et al.*, 2001; Storey and Tibshirani 2001/2003; Reiner *et al.*, 2003; Oord, 2005; Tabangin *et al.*, 2007; Dudoit and van der Laan, 2008; Edwin, 2008; Zhang *et al*, 2008; Foulkes, 2009; Lee *et al.*, 2009; Stephens and Balding, 2009) where false discovery rate (FDR) control is primarily preferred over family wise error rate (FWER) control.

However, the decision on what multiple testing correction to be used while performing genetic association studies is not consensual: Tao and Boulding (2003) used a permutation derived threshold; Martínez *et al.* (2007) a positive FDR (pFDR) at 10% level; Weber *et al.* (2007/2008) a FDR at 10% level; Kardia *et al.* (2008) and Smith *et al.* (2009) a FDR at 30% level; Jiang *et al.* (2010) used Bonferroni correction at 5% level. So it really depends on how much false positive associations we are willing to tolerate in exchange for a few more true positive associations. If you want to go conservative then Bonferroni and Bonferroni-type corrections (Bonferroni, 1935; Sidák, 1968/1971; Holm, 1979; Simes, 1986; Hochberg, 1988; Hommel, 1988; etc.) are the way to go, but if you are willing to me more permissive, then FDRs (Benjamini and Hochberg, 1995/2000; Benjamini and Yekutieli, 2001; Storey, 2002; Storey and Tibshriani, 2003; etc.) should be your choice.

The problem of multiple testing extends itself as well, in genetic association studies, when we need to perform outlier tests on a usually large number of individuals. This situation represents therefore a good opportunity for integrating robust multiple linear regression and outlier tests with multiple testing corrections, which we have not seen done before.

1.4 Thesis goal and outline

We found in the literature a few recent examples of the use of robust regression in the context of genetic association studies. However, no simulations are yet reported that assess these methods properties in this area of research.

Since in genetics the traits studied are mainly complex, having non-normal distributions and often showing heavy tails, our concern that the use of available methods in the literature should be explored is strongly sustained. Especially the robust methodology which has given proofs of being resistant to those model misspecifications, giving reliable results where others fail.

As normality violation comes together with outliers and outliers that are not a result of mere error measurement may even be the most important observations, methods that are able to correctly identify them, thus minimizing the swamping effect, and also powerful enough as to avoid masking, must be made known to researchers working in genetic association studies. This issue obviously comes together with the multiple testing problem

and therefore implies that multiple testing corrections that control the FDR are evaluated. Additionally, we have still another issue to unravel: outlier tests in the context of a multiple linear regression model, which is the one we advocate for these studies.

Furthermore, regression models that accurately reflect the proportion of phenotypic variation explained by genotypic variation among individuals are of crucial importance to prediction. And again, there are several proposals in robust literature that can be handy when using robust multiple linear regression in genetic association studies of quantitative traits.

Having said, the outline of this thesis is as follows:

In Chapter 2 we make a small incursion into the basic concepts of genetics to make the reader acquainted with some important concepts, giving as well an insight on the work of Gregor-Mendel, the father of genetics.

Chapter 3 takes us through the problematic of multiple testing, where we review some state of the art methods used whenever facing simultaneous testing of multiple hypotheses.

In Chapter 4 we present in detail both the usual classical approach to genetic association studies of quantitative traits and our proposed robust methodology for those same studies. More precisely, Huber's M-estimation together with adapted Wald-type tests to assess SNP/trait associations. We also propose a robust coefficient of determination as an approximation to the proportion of phenotypic variation that is explained by the genetic variation among individuals. And additionally review the theory of mixed linear models proposing the robustification of the correspondent F-test and coefficient of determination by the appropriate plug-in of robust estimates of scale and location.

The performance of the proposed approach is compared with both the classical and two other non-parametric methods in terms of type-I error rate and power in a simulation study under several contamination settings in Chapter 5, after a detailed description and full specification of a simulation model from the literature. The two non-parametric methods selected for comparison are rank-transform (RT; Conover and Iman, 1981), which was recently used in genetics (de Haan et al., 2009) and a Wilcoxon based (WIL; McKean *et al.*, 2009), used in quantitative trait locus (QTL) analysis (Zou *et al.*, 2003). An appropriate correction measure from Chapter 3 will be used.

In Chapter 6 we present a robust outlier test whose performance is assessed via a simulation study in the context of robust multiple linear regression where several multiple correction measures are also evaluated. This Chapter comprehends as well a simulation study that appraises the adequacy of the robust coefficient of determination as an approximation of the true phenotypic variation that is explained by the models' genetic component.

Finally, in Chapter 7 we present a real data example from the literature. Results of that paper are replicated and compared with the ones obtained from the robust analysis proposed in Chapter 4. Differences in both results are then discussed.

Chapter 8 being the final discussion, summarizes and pinpoints the most relevant aspects of the study as well as making some important alert notes. Also, directions for future work are laid down.

As a final goal, we aim at the publication of the obtained results in a distinguished scientific journal of the area therefore bringing robust methodologies to the knowledge of people working in genetic association studies of quantitative traits to whom they can definitely be a valuable asset.

I have called this principle, by which each slight variation, if useful, is preserved, by the term of Natural Selection.

Charles Darwin

2

A short insight on the basics of genetics

The principal aim of this chapter is to make the reader acquainted with some genetic basic notions without going into deep detail. It is therefore intended that the reader understands the concepts introduced and the subsequent issues related with quantitative genetics. For further details on genetics, the books of Gonick and Wheelis (1991) and Robinson (2010) can be quite useful, providing an easy and joyful trip into the underlying concepts of genetics. For deep detail see Russel (1992) and Lewin (2006). Animations of some of the processes described below can be found online, e.g, at
`http://highered.mcgraw-hill.com/sites/dl/free/0072437316/120060/ravenanimation.html`
or at `http://vcell.ndsu.nodak.edu/animations` and the reader is advised to peep them throughout the text.

2.1 General concepts

In nature there are two kinds of living organisms: the *Prokaryotes* and the *Eukaryotes*. Each cell of every living eucaryotic organism (e.g, plants and animals) is characterized by a membrane, a nucleus (DNA housing region) and the cytoplasm and its subcell parts (*organelles*) whereas in procaryotes (e.g, *bacteria* and *Cyanophytes*), cells are organized into complex structures enclosed within membranes, being the nucleus the membrane-bound structure. I.e, basically procaryotes lack a nucleus having DNA floating around in the cell.

From now on we will refer only to the eukaryotes case.

What is DNA?

DNA stands for deoxyribonucleic acid and is a large double stranded molecule[1], also known as *double helix* (see Figure 2.1), that is packaged in *chromosomes*, which are nothing more than an organized structures of super coiled DNA. It actually contains the genes that direct the production of *proteins* and therefore, genes are no more than DNA segments coding for a specific protein. In turn, proteins are molecules that play a critical role in structure, function and regulation of body cells, tissues and organs.

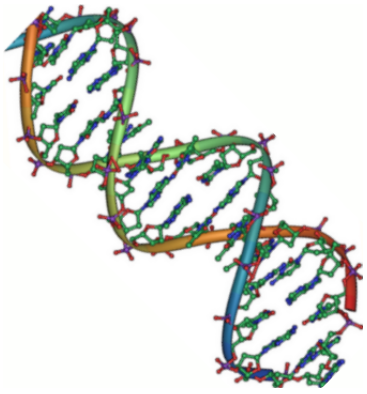


Figure 2.1: Double Helix. *Image from Wikipedia.*

At the molecular level, DNA information is stored as a code made up of four chemical bases (nitrogen bases): adenine (`A`), guanine (`G`), cytosine (`C`), and thymine (`T`), also called *chemical building blocks*. Those bases are placed in a unique order to code for all of the genes in all living organisms[2]. In the coding sequences[3] (*exons*), the bases are organized in groups of three (*triplets**codons**words*) forming the genetic code[4]. Each triplet may be translated into an *amino-acid*[5] of a protein, which is no more than a collection of hundreds or even thousands of amino-acids. The non-coding sequences are known as *introns*. Table 2.1 lists all the amino-acids one can get with the three letter words. We may see that not always a base change changes the final amino-acid and this is why there are only 20

[1]Stable electrically neutral group of at least two atoms in a definite arrangement held together by very strong chemical bonds;

[2]Human DNA consists of about 3 billion bases (3×10^9), and more than 99 percent of those bases are the same in all people;

[3]There are sequences of bases that do not code for proteins, known as junk DNA or introns which are removed from mRNA, in eucaryotes, only after transcription;

[4]Triplets of letters make words, words make sentences and sentences make genes;

[5]There are 20 standard amino-acids which can be placed in many different orders to form a wide variety of proteins;

Table 2.1: Codon/amino-acid table. Codon `AUG` coding for *methionine* (Met) also serves as the start codon if it is preceded by a string of alphabets, much longer than a single codon, called the *promoter* sequence.

1^{st} position 5'	2^{nd} position U	2^{nd} position C	2^{nd} position A	2^{nd} position G	3^{rd} position 3'
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA *Stop*	UGA *Stop*	A
	UUG Leu	UCG Ser	UAG *Stop*	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

standard amino-acids instead of $4^3 = 64$.

DNA bases pair up with each other, `A` with `T` and `C` with `G`, to form units called *base pairs*.Each base is attached to a sugar molecule and a phosphate molecule. Together, a base, sugar, and phosphate are called a *nucleotide*. Nucleotides are arranged, with phosphate and sugar molecules alternating, in two long strands that form a spiral and thus the name double helix. The leading end of each strand is called $5'$ *end* and the tail end $3'$ *end* (Figure 2.2).

One should be aware that variation in a gene, even at a single nucleotide, may cause a protein's structure and function to change. That variation can be either insignificant, having small or null effect on the final protein, or even cause serious disease. It is therefore important to understand how genes are translated into proteins. This path is known as the *Central Dogma of Molecular Biology* and consists of three major steps: DNA replication, transcription and translation.

So, how is DNA replicated?

DNA has a unique structure that enables the molecule to copy itself during cell division. In this way, when a cell prepares to divide, the DNA helix splits down the middle becoming two separate single strands. These single strands now serve as templates for building two new, double-stranded DNA molecules - each a exact copy of the original

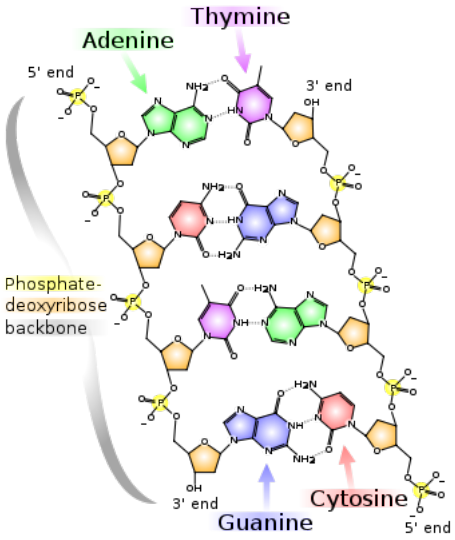


Figure 2.2: DNA strand. *Image from Wikipedia.*

DNA molecule. The process is quite simple: an `A` base is added wherever there is a `T`, a `C` where there is a `G`, and so on until all the bases are again paired up (see animation at `http://highered.mcgraw-hill.com/olc/dl/120076/bio23.swf`).

And, how are DNA sequences used to make proteins?

The process of going from gene to protein is complex and tightly controlled within each cell. It consists of two major steps: transcription and translation. Together, transcription and translation are known as *gene expression*. During the process of transcription (see animation at `http://vcell.ndsu.nodak.edu/animations/transcription/index.htm`), the information stored in a gene's DNA is transferred to a similar molecule called RNA[6] (*ribonucleic acid*) in the cell nucleus (this is done by RNA polymerase *enzyme*). The type of RNA that contains the information for making a protein is called *messenger RNA* (mRNA). This is because it carries the information, or message, from the DNA out of the cell nucleus into the cytoplasm. After transcription, introns in the sequence are removed by a process called *RNA-splicing* (see animation at `http://vcell.ndsu.nodak.edu/animations/mrnasplicing/index.htm`).

The second step, translation (see animation at `http://vcell.ndsu.nodak.edu/animations/translation/index.htm`), takes place in the cytoplasm. The mRNA interacts with a specialized complex called *ribosome*, a protein assembly machinery, which reads the sequence

[6]Both RNA and DNA are made up of a chain of nucleotide bases, but they have slightly different chemical properties. For instance, the ribose sugar component of RNA is slightly different chemically than that of DNA. RNA has a 2' oxygen atom that is not present in DNA. Another fundamental structural difference is for example, the fact that uracil takes the place of the thymine nucleotide found in DNA, and one more is that RNA is, for the most part, a single-stranded molecule.

of mRNA bases and uses that information to synthesize the exact protein coded for by the gene. The ribosome identifies the start codon `AUG` and begins the reading from that point forward. A type of RNA molecule called *transfer RNA* (tRNA)[7] assembles the protein, one amino acid at a time. Protein assembly continues until the ribosome encounters a stop codon: `AUG`, `UAA` or `UGA`. The mRNA product is called an *mRNA transcript*.

What's gene regulation?

Gene regulation is the process of turning genes on and off. Each cell expresses only a fraction of its genes and this is what makes cells in particular parts of the organism different from others in the same organism. E.g., a liver cell and a muscle cell, although having the same chromosomes, look and act quite differently because they both turn on only the genes that make them what they are, leaving all the other genes turned off. Gene regulation also allows cells to react to environmental changes. Signals from the environment or other cells activate proteins that bind to regulatory regions of a gene and increase or decrease the level of transcription.

One gene - one protein or not exactly?

During the RNA splicing process, where introns are removed, exons may be reconnected, in a process called *alternative splicing*, in many different ways resulting in different mRNAs and thus translating into different protein isoforms. It is believed that around 60% of genes are affected by alternative splicing. This means of course that genes may code for multiple proteins and not just one. But alternative splicing is not the only reason why genes sometimes code for different proteins. Genetic variations in the DNA sequence of a gene may also play its part. We call *allele* to the various forms a gene or DNA sequence can present.

How many sources of genetic variation are there?

There are several forms of genetic variation. These include chromosome crossing over, mutation and genetic drift, among others. Chromosome crossing-over occurs during *Meiosis* (see animation at `http://www.sumanasinc.com/webcontent/animations/content/meiosis.html`), which is an essential process of reproduction, when *homologous* chromosomes, after duplicating themselves, sometimes touch each other and exchange genetic material (Figure 2.3).

Mutations on the other hand, are changes in the DNA sequence that can translate in a large effect on how the organism looks like, i.e, its *phenotype*, the way it behaves and its physiology. They can be either neutral, beneficial or harmful and they happen at random.

[7] Small molecules that translate the genetic code, they attach to and activate specific amino-acids

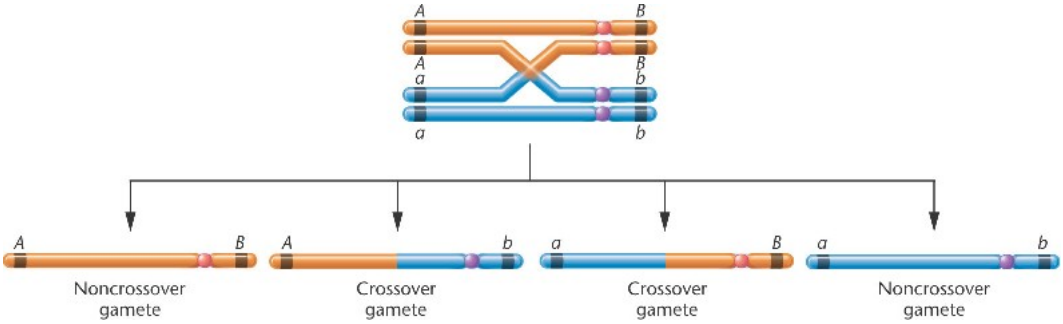


Figure 2.3: Crossover scheme: 2 homologous chromosomes crossover during meiosis producing 4 distinct gametes. *Image from* `www.bio3400.nicertutor.com`.

However, not all mutations matter to a species evolution. Some of them are just *somatic mutations*, i.e, they occur in non-reproductive cells and are'nt therefore passed on to offsprings. The mutations that are passed on are called *germ line mutations*. Several types of mutations are illustrated in Figures 2.4 and 2.5. The causes to this phenomena are

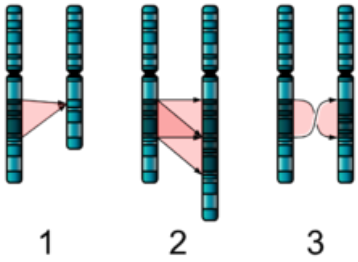


Figure 2.4: Mutation types I: deletion (1), duplication (2) and inversion (3). *Image from Wikipedia.*

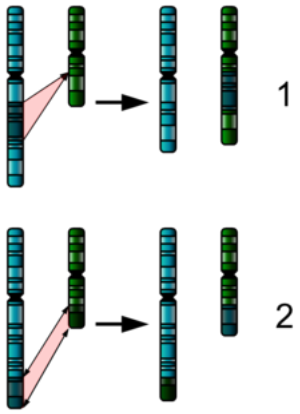


Figure 2.5: Mutation types II: insertion (1) and translocation (2). *Image from Wikipedia.*

DNA failure to copy and/or repair accurately and external influences, like e.g., exposure to radiation or specific chemical substances. If we analyze a population on a specific DNA locus, the *minor allele* is the allele that occurs less often in the population and the *major allele* the one that occurs more often. If the *minor allele* has a frequency larger than 1%

then the variant is called *polymorphism*[8]. This condition helps distinguishing between inherited variation and spontaneous mutations.

Genetic drift is the fluctuation of allelic frequencies in a population from generation to generation and plays an important role in the evolutionary process, leading to allelic frequency changes over time and sometimes causing some gene variants to completely disappear reducing genetic variability. Its effect is larger in small populations rather than in large populations.

What are SNPs?

The most abundant source of genetic variation is the *single nucleotide polymorphism* (SNP), which is every site on the genome that shows a base-pair (bp) change that is variable across the general population of a species. The SNP can either be an *anonymous* or *functional* variant with respect to protein coding or gene function. From a statistical perspective, at the nucleotide level and in the case of bi-allelic SNPs, this unit of analysis can be regarded as a three-category variable. For instance, a SNP with alleles `A` (adenine) and `G` (guanine) has categories `AA`, `AG` and `GG`.

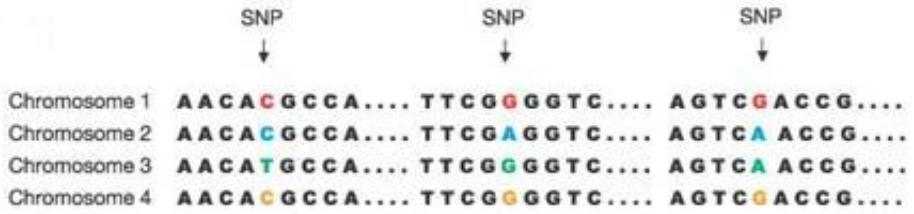


Figure 2.6: Single Nucleotide Polymorphism (SNP): the same chromosome is genotyped for several individuals in a population and variation is observed in 3 distinct sites. *Image from The Individualist HapMap.*

2.2 Mendel's experiment, findings and laws

Gregor Mendel, also known as the father of genetics, was an Austrian augustinian monk who lived between the years of 1822 and 1884. He was a curious amateur gardener who happened to notice, while walking through the monastery gardens, that his, self-pollinating, pea plants (*Pisum sativum*) were different from each other in several ways, like:

[8]This cut-off point between a mutation and polymorphism is arbitrary - see Twyman, R. (2003) on "Mutation or polymorphism?" on the Wellcome Trust website at `http://genome.wellcome.ac.uk/doc_wtd020780.html`.

i. seed shape: round & wrinkled;

ii. seed color: yellow & green;

iii. seed coat color: grey & white with correspondent flower color violet & white;

iv. stem flower position: axial & terminal;

v. stem size: tall & short;

vi. pod form: inflated & constricted;

vii. pod color: green & yellow.

In order to understand why pea plant offsprings resembled their parents, Mendel decided to make some experiments. He started out with tall and short plants observing that both plants gave rise to also tall and short plants thus maintaining their parents height characteristic. These plants are known as *true-breedings*. He then wondered what would happen if he crossed a tall plant with a short one, referring to the two parental plants as the *parent* generation (P generation). How would the hybrids resulting plants, which were the first filial generation (F1 generation), look like? Medium-height? He soon found out that the resulting hybrids were always tall plants! This indicated that the tall trait somehow dominated over the short trait. Hummm?.. Since one trait remained in the next generation and the other disappeared, Mendel decided to name those traits as *dominant* and *recessive* respectively.

Next, he just let the hybrids self-pollinate. He referred to these offsprings as the second filial or F2 generation. He was quite surprised to observe that there were short plants coming up again in that generation. More precisely, the recessive trait reappeared and in a proportion close `3:1`, i.e, for each 3 tall plants there was a short one! Moreover, taking these hybrids, he observed that the short plants continued to reproduce in short plants and the tall ones in tall and short plants with the same ratio has before. By this point Mendel was already convinced that there must be something in plant egg and pollen that determined plant height. He referred to this abstract entity as *factor* or *unit of inheritance*, later to be called GENE.

Mathematically speaking, Mendel's reasoning was:

i. each plant egg and pollen grain have one height gene;

ii. the height gene may have two *forms*, which we now call *alleles*, e.g, `A` for tallness (dominant allele) and `a` for shortness (recessive allele);

iii. each plant will therefore have two forms of the height gene, i.e *genotype*, which may be `AA`, `Aa` or `aa`. Plants with genotypes `AA` and `aa` are said *homozygous* and the ones with genotype `Aa` are *heterozygous*.

Summing up, for a plant to reproduce, the two pollen gene alleles must separate in two separate gametes each one having one allele. The same must happen with the egg gene alleles. This process was Mendel's first law and is known as the *Principle of segregation*. However it was not proved until the *meiosis* process was discovered. Its modern version is displayed in the box below.

Principle of segregation. (modern version) *The alleles in homologous chromosomes segregate from each other in the formation of gametes. Half the gametes carry one allele and the other half carry the other allele.*

Mendel's experiment showed that plants with `AA` and `Aa` were tall plants and the ones with `aa` were short, with a phenotypic ratio close `3:1` and a genotypic ratio of `1:2:1`, agreeing with the crossing square in Table 2.2. He also verified these results for the other traits mentioned and therefore concluded that heredity is controlled by genes. His results are displayed in Table 2.3 and may be verified by the Pearson's Chi-square test:

Table 2.2: Crossing between F_1 `Aa` hybrids.

×	A	a
A	AA	Aa
a	aA	aa

Pearson's Chi-square test. To test the classification of n objects on k distinct classes we consider the test statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - np_i)^2}{np_i}$$

where O_i is the number of observed counts for class i and p_i is the probability of an observation belonging to class i. Under the null hypothesis

$$H_0: \text{ the } i^{th} \text{ class has probability } p_i, \ i = 1, ..., k$$

the test statistic has a chi-square distribution with $k - 1$ d.f..

In this case we have two classes, the dominant class and the recessive class with theoretical probabilities of 3/4 and 1/4, respectively for the theoretical ratio of 3:1. Results

Table 2.3: Mendel's experiment results in the crosses between plants differing on one of the seven characters.

Character	F_1			Ratio
		Dominant	Recessive	
seeds: smooth x wrinkled	all smooth	5474	1850	2.96:1
seeds: yellow x green	all yellow	6022	2001	3.01:1
seed coats & flower: grey & purple x white & white	all grey & purple	705	224	3.15:1
flowers: axial x terminal	all axial	651	207	3.14:1
pods: inflated x pinched	all inflated	882	299	2.95:1
pods: green x yellow	all green	428	152	2.82:1
stem: tall x short	all tall	787	277	2.84:1

Table 2.4: Results of the Pearson's Chi-square test on Mendel's data.

Experiment	χ^2	p-value
1	0.26288	0.6081
2	0.01500	0.9025
3	0.39074	0.5319
4	0.34965	0.5543
5	0.06351	0.8010
6	0.63391	0.4259
7	0.60651	0.4361

of the Pearson's test are displayed in Table 2.4 where we observe evidence to accept H_0 in all Mendel's experiments.

With the purpose of accessing the independency of the traits studied, Mendel continued his experiments. He now crossed plants differing in two characteristics. E.g, a tall plant with round seeds (`AARR`, where `R` is the allele for roundness) with a short plant with wrinkled seeds (`aarr`, where `r` is the allele for wrinkleness), resulting in hybrids of the form `AaRr` tall and round due to the dominance of the traits tallness and roundness.

If the height and smoothness genes sort out independently of each other, then the four allele combinations `AR, Ar, aR` and `aa` for the next hybrid generation will be equally likely, in which case the crossing scheme can be seen in Table 2.5. It can also be easily observed (see Table 2.6) a ratio of `9:3:3:1` for, respectively, tall and round (`AARR`, `AARr`, `AaRR` or `AaRr`), tall and wrinkled (`AArr` or `Aarr`), short and round (`aaRR` or `aaRr`) and short and wrinkled (`aarr`) plants.

Having the F1 hybrids and letting them self-pollinate Mendel observed close ratios in his F2 generation thus stating that different traits are inherited independently of each other, i.e, so that there is no relation between one another. And this is Mendel's second law, the *Principle of independent assortment*, whose modern version is displayed in the

Table 2.5: Hybrid crossing scheme under the principle of independent assortment

×	AR	Ar	aR	ar
AR	AARR	AARr	AaRR	AaRr
Ar	AArR	AArr	AarR	Aarr
aR	aARR	aARr	aaRR	aaRr
ar	aArR	aArr	aarR	aarr

Table 2.6: Resulting genotypes and phenotypes from the previous hybrid crossing scheme.

N	Genotype	Phenotype
1	AARR	
2	AARr	*Tall &*
2	AaRR	*Round*
4	AaRr	
1	AArr	*Tall & Wrinkled*
2	Aarr	
1	aaRR	*Short & Round*
2	aaRr	
1	aarr	*Short & Wrinkled*

box below.

Principle of independent assortment. (modern version) *Genes lying on different chromosomes are distributed independently of one another during the production of gametes.*

However, this is not always the case, dominant alleles don't always dominate (we may have *incomplete dominance* resulting in blending phenotypes or *codominance* when alleles share equally in the expression of their phenotypes) and genes aren't always inherited independently. E.g, they may be close of one another on the same chromosome (*linked* genes) and be inherited together from generation to generation. Mendel did not know about chromosomes but still was lucky enough as to have 7 qualitative pea traits, each one controlled by a single gene, having each gene lying in a different chromosome of the 7 total pea chromosomes! Coincidence or publication bias?!.. One really has to wonder and Fisher (1936) even dared say that either Mendel was exceptionally lucky or he cheated, conveniently throwing away the results he did not understand, the ones that make inheritance messy like *epistasis* and *linkage*, and that did not agree to the laws of segregation and independent assortment. This controversy between Fisher and Mendel's results and who's right or wrong, has been a topic of discussion in literature (Froggatt and Levin, 1971; Farrall, 1975; Campbell, 1976; Pilgrim, 1984; Weiling, 1986 and 1989; Van Valen, 1987; Fairbanks and Rytting, 2001; Novitski, 2004*a* and 2004*b*; Pilpel, 2007). Recently, Franklin *et al.* (2008) ended this discussion arguing that Mendel was not guilty

of fraud and that Fisher's conclusions may be explained without invoking fraud, and Pires and Branco (2010) presented a statistical model to end this controversy once and for all.

Despite the controversy, Mendel's work is undoubtedly valuable being the percussor of modern genetics. Traits following Mendel's rules are said to be *Mendelian* traits. Moreover, the study of how different traits are inherited and expressed is called *Mendelian genetics*.

Mendel's work was first published in 1865 in *Verhandlungen des Naturforschenden Vereines in Brünn* and a translation of his work can be found in Mendel (1866).

2.3 Population genetics

Allelic frequencies are important components for genomic analysis in natural populations and the number of alleles and their distribution are indicators of genetic diversity. Population genetics studies the distribution of allele frequency and the way it evolves under the processes of natural selection, gene flow, genetic drift and mutation. This field of research was born between the years of 1920 and early 30s by the hands of Fisher (1918;1930), Haldane (1932) and Wright (1931). These men integrated the principles of Mendelian genetics with those of the Darwinian natural selection.

In this section we will only focus on a few but important concepts of population genetics theory. For a deeper insight on the subject, literature is available, e.g, see Lynch and Walsh (1997).

2.3.1 Hardy-Weinberg equilibrium

The Hardy-Weinberg equilibrium (HWE) denotes the independence of alleles at a single site between two homologous chromosomes. It was discovered independently in 1908 by Godfrey H. Hardy (1877-1947), an English prominent mathematician, and Wilhem Weinberg (1862-1937), a German physician, and is one of the simplest and most important principles in population genetics. It can easily be shown that it actually is a consequence of Mendel's law of segregation.

Take a bi-allelic gene with alleles `A` and `a` segregating in a diploid population with relative frequencies of p and q, respectively. If the genotypic frequencies of `AA`, `Aa` and `aa` observed in the population agree with the square in Table 2.7, i.e, if $f(\texttt{AA}) = p^2$,

$f(\texttt{Aa}) = 2pq$ and $f(\texttt{aa}) = q^2$, then the population is said to be in HWE and we have

$$(p+q)^2 = p^2 + 2pq + q^2 = f(\texttt{AA}) + f(\texttt{Aa}) + f(\texttt{aa}) = 1.$$

Table 2.7: Punnet square for HWE: shows the allelic frequencies from random mating with two alleles.

×		*sperm*	
		A (p)	a (q)
eggs	A (p)	AA (p^2)	Aa (pq)
	a (q)	aA (pq)	aa (q^2)

Deviations from HWE may occur and usually indicate evolution of a species. It may also indicate inbreeding, population stratification, eventual problems in genotyping, or small population size. Moreover, in case-control studies, in samples of affected individuals it may also indicate evidence of association.

Hardy-Weinberg equilibrium law. *Both allelic and genotypic frequencies in a population remain constant from generation to generation unless specific disturbance influences (e.g, non-random mating, mutation, natural selection, small population size, genetic drift and gene flow) are introduced..*

It is important to underline that whatever the initial population one-locus genotypic frequencies are, random-mating will automatically produce an offspring population in HWE. This means that if generations are non-overlapping, i.e, parents die as soon as they reproduce, HWE will be achieved in one generation only, whereas if they overlap more than one generation of random-mating may be needed to achieve HWE. Once it is achieved, under random-mating, HWE will be maintained generation after generation.

HWE is often used as a quality control measure. As such, deviations from this relationship may indicate quality problems in the genotyping procedure. However, this may not always be the case since the HWE principle is based in several assumptions and deviations from it may occur not because of genotyping error but due to selection or assortative mating. In case control studies for example, it is a usual procedure to test the control group SNPs for HWE, excluding the ones failing the test from further analysis. Nevertheless it has been criticized in the literature and due care should be taken before making a decision. Wittke *et al.* (2005) have a very nice discussion on this topic that we strongly advise reading before engaging in case-control genetic association studies. Other

interesting readings are the works of Nielsen *et al.* (1998), Gomes *et al.* (1999), Xu *et al.* (2002), Hosking *et al.* (2004) and Györffy *et al.* (2004), among others.

The departure of HWE is known as *disequilibrium* and may be quantified by a coefficient $D_{\mathtt{A}}$ such that:

$$\begin{cases} f(\mathtt{AA}) = p^2 + D_{\mathtt{A}} \\ f(\mathtt{Aa}) = 2pq - 2D_{\mathtt{A}} \\ f(\mathtt{aa}) = q^2 + D_{\mathtt{A}} \end{cases} .$$

Bounds for the disequilibrium coefficient $D_{\mathtt{A}}$ are $-\max(p^2, q^2) \leq D_{\mathtt{A}} \leq pq$ (Weir, 1996). When $D_A = 0$ we have HWE.

Estimating the HWE parameter p. Let's assume we have two alleles `A` and `a` segregating in a size n diploid population in HWE where $p = f(\mathtt{A})$ is unknown. Take, X_i, $i = 1, 2, 3$ to represent the population genotypes `AA`, `Aa` and `aa` counts, respectively. Then $(X_1, X_2, X_3) \sim \mathcal{M}\text{ultinomial}(n, p^2, 2p(1-p), (1-p)^2)$. Assuming that we now have a realization (N_1, N_2, N_3) of (X_1, X_2, X_3), the likelihood function is

$$L(p, \mathbf{x}) = \frac{n!}{N_1!N_2!N_3!} p_1^{N_1} p_2^{N_2} p_3^{N_3} = \frac{n!2^{N_2}}{N_1!N_2!N_3!} p^{2N_1+N_2}(1-p)^{N_2+2N_3}$$

and the maximum likelihood estimate of p, which is UMVUE (uniform minimum variance unbiased estimate), is given by $\widehat{p} = \dfrac{2N_1 + N_2}{2n}$. Also $E(X_i) = np_i$.

Chi-square test for HWE. The usual test to assess HWE is the goodness of fit chi-square test. Being the null hypothesis

$$H_0 : \text{ the population is in HWE,}$$

the test statistic is calculated as

$$\chi^2 = \sum_{i=1}^{3} \frac{(N_i - np_i)^2}{np_i}$$

where N_i, $i = 1, 2, 3$ are the observed genotypic counts and np_i, $i = 1, 2, 3$ the expected ones. Under the null, the test statistic follows a chi-square distribution with 1 d.f. and so we reject H_0 at the 5% significance level if p-value $= P(\chi^2_{(1)} > \chi^2_{obs}) < 0.05$.

This test however can have inflated type-I error rates and exact tests (e.g, Fisher's exact test) may be preferred. This happens for example in many genetic case-control studies, where the individuals are not randomly sampled from a general population but rather ascertained by affection status (Li and Li, 2008). To an insight into exact tests see Emig (1980) and Wigginton *et al.* (2005). The Wigginton proposed exact test is available in package *SNPassoc* via instruction `tableHWE()` (González *et al.*, 2008). The usual chi-square test can be performed for example with instruction `HWE.chisq()` from the R package *genetics* or with the instruction `HWChisq()` from the R package *HardyWeiberg*, using continuity correction `cc` $= 0$. This package also performs the HWE likelihood ratio test (LRT) via `HWLratio()` and the Wigginton exact test via `HWExact()`. Fisher's exact test for HWE is available also in the R package *genetics* via instruction `HWE.exact()`.

2.3.2 Recombination and linkage desiquilibrium.

In the meiosis process, when homologous chromosomes pair up (*maternal* and *paternal* chromosomes), they often intertwine and exchange sections or fragments of chromosome, i.e, they crossover. In this way, the gametes produced by this process will present new combination of genes that differs from the combination supplied by the parents (see Figure 2.3) and thus the name *recombination*. Gene recombination is thus one way that organisms have to produce offspring with new combinations of maternal and paternal traits that may contribute to or enhance survival.

Also, as a result of recombination, alleles previously on the same chromosome may end up in different gametes, i.e, the further the two alleles are apart in the same chromosome, the greater the chance that a crossover event may occur between them, possibly separating the alleles. Conversely, the closer they are of one another, the smaller the chance a crossover event will occur and therefore they will tend to be inherited together. This is exactly the exception to Mendel's *principle of independent assortment*. The modern version of Mendel's 2^{nd} law (Section 2.2) already accounts for this possibility.

The distance between alleles on the same chromosome can be measured either in terms of recombination events between them or in terms of the number of base pairs in between. As to the recombination approach, quite simply, the greater the recombination frequency between alleles, the farther apart they are in the chromosome. The lower the recombination frequency, the smaller the physical distance between alleles. The percentage of recombination between alleles is called *recombination fraction*. It can not exceed 50%, i.e, the recombination fraction between two alleles is less or equal than 0.5. Moreover, when it is 0.5, it means that both alleles are on opposite ends of the chromosome.

Linkage disequilibrium (LD) reflects the non-random association of alleles at two or more *loci* (specific sites on the genome) not necessarily on the same chromosome. If two loci/genes are said to be in linkage disequilibrium, it means that there is lack of historical recombination events in between. So basically, LD refers to a historically reduced (non-equilibrium) level of the recombination of specific alleles at different loci controlling particular genetic variations in a population. This LD can be detected statistically, and has been widely applied to map and eventually clone a number of genes underlying complex genetic traits in humans. However, significant LD does not necessarily indicate causal associations since it can also be influenced by population history.

One should not mistake LD with linkage, the latter meaning genetic proximity. In fact, if two loci are tightly linked, i.e, close to one another on the same chromosome, they are in LD, however, if two loci are in LD it does not necessarily mean that they are linked because LD may be influenced by many other factors such as mutation rate, random drift or non-random mating and population structure (Table 2.8). Therefore, LD needs to be interpreted cautiously.

Factor	Effect on LD behaviour
Recombination rate	higher rec. lowers LD
Mating system: Selfing species	high LD
Mating system: Outcrossing species	low LD
Natural and artificial selection	locally increases LD
Population size	small populations have more LD
Mutation rate	high mutation rate decreases overall LD but LD around newly created mutated allele remains high until dissipated by recombination
Bottlenecks	inflates LD
Population structure	inflates LD

Table 2.8: Factors that condition LD behaviour in a population (Rafalski and Morgante, 2003; Pritchard and Przeworski, 2001).

LD provides a mean of examining genetic polymorphisms in different genetic backgrounds, taking advantage of generations of recombination. It has a couple of advantages over QTL analysis (see Liu, 1997 and Wu *et al.*, 2007), namely, the fact that it uses natural populations instead of breeding schemes, making experimental design straightforward and thus saving considerable time. Also, traits may be mapped to very small regions allowing the discovery of the underlying gene(s) and/or application for selection across a wider range of germplasm[9].

[9]A germplasm is a collection of genetic resources for an organism. For plants, the germplasm may be

In natural populations we can find LD between loci for which recombination has not had enough time to dissipate the initial disequilibrium. However, when a natural population mates at random, disequilibrium reduces progressively in each generation, being the degree of LD therefore dependent on recombination fraction θ, and time in generations t, since the origin of a new mutation at time $t = 0$ and according to the formula (Falconer and Mackay, 1996):

$$D_t = (1 - \theta)^t D_0, \tag{2.1}$$

where, D_0 is the value of LD at generation 0 and D_t the value of LD at generation t.

Rewriting the above equation, knowing the values of D_0 and θ it is possible to determine the number of generations t needed to achieve a pre-established value D_t:

$$t = \frac{\ln \frac{D_t}{D_0}}{\ln(1 - \theta)}. \tag{2.2}$$

However, the time LD takes to decay differs between species. For example, LD is known to decay faster in outbred species like maize, eucalyptus globulus and humans than in inbred species like soybean and *Arabidopsis Thaliana* (see Garcia *et al.*, 2003, in "Structure of LD in plants" for more detail). In these two latter species it is known to persist over tens to hundreds of kilobases. In most human populations however, LD is known to extend for small distances in most genomic regions although it may extend for larger distances when population sizes are small. Also, LD in humans is highly dependent on the population in which it is measured (Reich *et al.*, 2001). In some populatins of maize as well, LD is known to decay rapidly over a few hundred base pairs (Tenaillon *et al.*, 2001).

There are several measures of LD on literature and a good review can be found in Weir (1996). The basis of most of those measures is the difference between observed and expected allelic frequencies under independence and, in this way, if both frequencies match then the loci are in equilibrium and so there is no LD. The preferred measures of LD quantify the degree of association between pairs of markers (SNPs or other) and they differ only in the way in which they depend on the marginal allele frequencies. Amongst those, the most popular is r^2 or Δ^2 and will be described below.

Consider two loci A and B on the same chromosome, with alleles `A` and `a` and `B`

stored as a seed collection or, for trees, in a nursery;

and b respectively, and correspondent allelic frequencies $p_{\mathtt{A}}$, $p_{\mathtt{a}}$, $p_{\mathtt{B}}$ and $p_{\mathtt{b}}$. Let the four haplotype frequencies be $p_{\mathtt{AB}}$, $p_{\mathtt{Ab}}$, $p_{\mathtt{aB}}$ and $p_{\mathtt{ab}}$. Then, the squared correlation coefficient between pairs of bi-allelic markers (Hill and Robertson, 1968) is given by

$$r^2 = \frac{(p_{\mathtt{AB}} - p_{\mathtt{A}}p_{\mathtt{B}})^2}{p_{\mathtt{A}}p_{\mathtt{B}}p_{\mathtt{a}}p_{\mathtt{b}}}. \tag{2.3}$$

Estimates of r^2 are usually obtained by plugging in sample frequencies $\widehat{p}_{\mathtt{A}}$, $\widehat{p}_{\mathtt{a}}$, $\widehat{p}_{\mathtt{B}}$ and $\widehat{p}_{\mathtt{b}}$, $\widehat{p}_{\mathtt{AB}}$ into equation (2.3).

Different allele frequencies at loci A and B imply $r^2 < 1$ and we have $r^2 = 1$ if and only if the loci are in complete disequilibrium, in which case loci A and B are fully predictive of each other.

The study of patterns of LD is crucial in genome-wide association studies making possible to have carefully selected sets of SNPs that provide information on much larger sets of unobserved SNPs. Even in the study of complex traits there is evidence that this approach will often lead to detectable associations which are not in themselves causal but rather in LD with the causal variant (Abecassis *et al.*, 2005). Knowledge of these patterns of LD makes even possible to have study designs based only on randomly genotyped SNPs, implying that the costs associated with those studies may drop significantly. Therefore SNP subset selection via LD is a hot topic in the literature of genetic association studies (Byng *et al.*, 2003; Carlson *et al.*, 2004; Morkûnienë and Kuèinskas, 2005; Howie *et al.*, 2006).

In R there are several genetic packages one may use to calculate, among other things, pairwise LD. One such package is package *genetics* where instruction `LD()` calculates not only r^2 but other common LD measures (namely D and D' not discussed in this work). It can either be used to calculate LD between two SNPs or all the pairwise LD for all the SNPs in the data.

2.4 Quantitative genetics

Quantitative genetics usually refers to the study of continuous traits and their underlying mechanisms. It can be regarded as an extension of simple Mendelian inheritance where the combined effect of many genes results in phenotype values with continuous distribution. However, it is not limited to continuous traits alone, but rather all traits that are controlled by several genes, the so-called *polygenic traits*.

Like the height of an individual, that is a trait of both genetic and environmental influences, there are many others whose genetic variation we would like to fully understand. These traits are not Mendelian and do not verify the HWE since they are influenced by a multitude of genes.

The aim of quantitative genetics is to study the degree of dependence between parental traits and their offsprings traits. This does not necessarily mean that there is a genetic basis for this dependence since it is possible that environmental effects also play its part.

The mathematical entity related to the dependence between parent and offspring characteristics is *heritability* and will be described in more detail just ahead (Section 2.4.2). However, when we talk about disease then we interpret heritability as *familial risk*.

A good review on the genetics and analysis of quantitative traits can be seen in Lynch and Walsh (1997). In this small section we will only describe the one-gene and multiple-gene models for an $F2$ population and briefly discuss the definition of the important notion of heritability of a trait.

2.4.1 Genetic models

In this sub-section we will see how to simulate one or more bi-allelic genes from an $F2$ population (Section 2.2).

One-gene model. Let us assume we want to simulate one bi-allelic gene `A` having alleles `A`, `a` for n individuals. Then, the genotypic data is simulated based on the expected cumulative frequencies presented on Table 2.9.

Gene `A`	$p_{\mathtt{A}}$	Cumulative Frequency
`AA`	0.25	0.25
`Aa`	0.5	0.75
`aa`	0.25	1

Table 2.9: Expected frequencies of gene `A` genotypes for an F2 population

Two-gene model. If we want to simulate not one but two bi-allelic genes `A` and `B`, having alleles `A`, `a` and `B`, `b` respectively, linked with a certain recombination fraction r, for n individuals, we may start by simulating gene `A` as previously. Gene `B` is then simulated based on gene `A` genotypes obtained and the recombination fraction r between them, according to the expected cumulative frequencies presented on Table 2.10.

Gene A	Gene B	$p(\texttt{B}\|\texttt{A})\Big\|_r$	Cumulative Frequency
AA	BB	$(1-r)^2$	$(1-r)^2$
	Bb	$2r(1-r)$	$1-r^2$
	bb	r^2	1
Aa	BB	$r(1-r)$	$r(1-r)$
	Bb	$(1-r)^2+r^2$	r^2-r+1
	bb	$r(1-r)$	1
aa	BB	r^2	r^2
	Bb	$2r(1-r)$	$-r^2+2r$
	bb	$(1-r)^2$	1

Table 2.10: Expected frequencies of genotypes for gene B conditional on gene A when the two are linked with recombination fraction r for an $F2$ population

Multiple-gene model. If we intend to simulate N genes for n individuals, the usual approach to do it, which avoids considering all the 3^N possible genotypes, is to consider the simulation of one gene at a time conditional only on the previous simulated gene. Hence, should we want to simulate for instance another gene C with alleles C and c, we would consider only the genotypes observed for gene B and the frequencies on Table 2.10. This procedure can be repeated for the number of genes desired. We must underline that this scheme assumes no crossover interference is present. When it is, at least three genes have to be simulated simultaneously.

2.4.2 Heritability

Heritability is the proportion of phenotypic variation in a population that is attributable to genetic variation among individuals. It therefore quantifies the importance of the genetic effects to the trait value and is defined as the ratio of the genotypic to the phenotypic variance

$$H^2 = \frac{\sigma_G^2}{\sigma_P^2} = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_e^2} \tag{2.4}$$

where σ_G^2 and σ_e^2 are the variance components associated with the genetic effects and the residual error, which may include undetected genetic effects, environmental effects, genetic-environmental effects and the random effects. Heritability, as defined in Equation (2.4) is also known as *broad sense heritability*.

The genetic variance component can still be partitioned into other components arising from different types of interactions between genes:

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2 \tag{2.5}$$

where σ_A^2 and σ_D^2 are the additive and dominance variances respectively, representing the

genetic variation among individuals that are due to the additive and dominance effects of genes, and σ_I^2, which refers to another source of genetic variation, epistasis that accounts for the interactions between genes. If we assume no epistatic effects between genes then equation (2.5) simplifies to

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2. \tag{2.6}$$

The heritability of a trait can range from 0 to 1. Having $H^2 = 1$ suggests that all the phenotypic variance is genetically based whereas having $H^2 = 0.5$ means that 50% of the phenotypic variation is attributable to genetic differences among individuals. Note that if $H^2 = 0$ then no phenotypic variation is due to genetic variation among individuals. However one must be extra careful when interpreting heritability. As Russel (1992) opportunely points out, heritability does not indicate the extent to which a trait is genetic. E.g, we can have a genetic trait but still a population genetic variance of zero leading to $H^2 = 0$, and this fact does not indicate what proportion of an individual's phenotype is genetic since it is a population characteristic and not an individual's characteristic and is not fixed for a given trait but rather depends upon the genetic makeup and the specific environment of the population.

The estimation of H and thus of σ_A^2 and σ_D^2 is based upon the expected proportion of genes shared between different types of relatives (Fisher, 1918). However, nowadays, in the context of the genome-wide era, it is also possible to estimate such parameters within families using the identity-by-descent sharing degree between individuals (Ritland, 2000; Thomas, 2005; Visscher *et al.*, 2006).

Repeated confirmation is an essential element. Otherwise, a very specific association may be a reflection of the multiple comparison problem; If enough contrasts are created by fractionation of a single data set, some apparent significant result is likely to appear.

Mantel and Haenszel (1959)

3

Multiple hypothesis testing

Multiple testing is a challenging issue in many areas of research where there is need for the simultaneous testing of many hypotheses. In particular, in genetic association studies where there are usually a large number of genotyped SNP markers, many of which in LD. Failure to adjust for this problem may produce excessive false positive associations as well as overlook true positive associations.

The usual approach to the multiplicity problem accounts for the control of the experiment-wise or family-wise error rate (FWER). Another state of the art approach accounts for controlling the expected proportion of falsely rejected hypotheses, i.e, the false discovery rate (FDR). It should be highlighted that when all hypotheses are true this error rate is equivalent to the FWER. Should not this be the case, then it is known to be smaller than the FWER. FDR is thus potentially more powerful.

There are situations however, where FWER control is required (e.g, when rejecting the hypothesis necessarily implies an action like, for example, the approval of a drug, or when a conclusion from the various individual inferences is likely to be erroneous when at least one of them is) and others where FDR control is sufficient (e.g, the analysis of gene expression where the goal is to extract genes that are potential candidates for further studies).

Either way, multiple testing procedures aim to achieve a balance between type-I and type-II errors. The usual principle is, once type-I errors are controlled at a pre-specified level, it is preferable to make more type-I errors than type-II errors.

In the next two sections we will present and discuss some of the usual methods available in literature for each of these approaches. Additional information on multiple testing issues, aside from the ones provided throughout the text, can be found in Westfall and Young (1993), Pollard and Van der Laan (2003) and Dudoit and Van der Laan (2008).

3.1 Controlling the family wise error rate

In large-scale association studies where a large number of SNPs are tested for association with a given trait, any true association result will be masked in the lot of spurious associations that will be falsely declared.

A widely used approach to deal with this problem is multiple testing correction. Multiple testing corrections adjust p-values derived from multiple statistical tests to correct for the occurrence of false positives, i.e, wrong rejections of the null hypothesis (type-I error). That adjustment keeps the overall FWER below a pre-specified p-value cutoff.

Another approach for the control of the type-I error is the re-sampling permutation test. Permutation based tests do not use individual association scores based on family-wise corrections, they estimate statistical significance directly from the data being analyzed instead, i.e, they draw the threshold directly from the experimental data. For this reason they have become widely accepted and recommended in studies that involve multiple statistical testing (e.g, Belmonte and Yurgelun, 2001; Dudoit *et al.*, 2003; Nakagawa, 2004; Kimmel *et al.*, 2007) especially when the tests are correlated and despite of being sometimes computationally time consuming.

In this section we will make a review on procedures designed to control the FWER. More detail on this subject can be found in Hochberg and Tamahne (1987) and Westfall and Young (1993). Also, in Dudoit *et al.* (2003), procedures for the control of the FWER and other error rates with application to genomics and DNA microarray data are reviewed and compared. Another review with application to microarray data is the work of Sebastiani *et al.* (2003).

The methods discussed below can be categorized as one-step (or single-step), step-

down and step-up methods (or stepwise)[1]:

One-step methods means that all the p-values will be compared to a single value calculated from the desired FWER threshold α and from the number of tests m and therefore each null hypothesis is tested "independently" from the outcome of the others, i.e, equivalent adjustments are performed for all hypotheses.

Step-down methods have the p-values examined from the smallest to the largest. At each step, the p-value is compared to its correspondent threshold, which depends on the p-value's rank. If smaller, then the correspondent hypothesis is rejected. Once a p-value is greater than its cut-off, the correspondent hypothesis and all others referring to higher p-values are not rejected.

Step-up are similar to step-downs. The p-values are now examined from the largest to the smallest. At each step, the p-value is compared to its correspondent threshold, which depends on the p-value's rank. If greater than its cut-off then the correspondent hypothesis is not rejected. Once a p-value is smaller than its cut-off, the correspondent hypothesis is rejected and so are the remaining hypotheses with smaller p-values.

As examples of one-step procedures we have those of Bonferroni (Algorithm 1) and Sidák (Algorithm 2). Holms' method (Algorithm 3) is an example of a step-down method and Hochbergs' (Algorithm 7) is an example of a step-up method. Both step-down and step-up methods are said to be multi-step or stepwise procedures since the tests are made in sequential order depending on the outcome of the previous tests. Stepwise procedures for controlling FWER are usually less conservative and more powerful than single-step methods.

In this section we will present some criteria for the rejection of an overall test $H_0 : H_1, ..., H_m$ where the hypotheses may or may not be independent. When, by any of the methods discussed, H_0 is rejected, the question to which individual H_i's should be rejected sometimes remains unclear. Moreover, as multiple testing corrections depend on the amount of hypotheses being tested, the larger the list of tests is, the more conservative the correction will be (Bonferroni is quite a good example of this).

One solution to solving the multiplicity issue is to make each individual test more conservative, i.e, more difficult to reject each individual hypothesis H_i, $i = 1, ..., m$. Such

[1]See Westfall and Young (1993) for a general method for constructing step-down methods from one-step methods.

methodology is often called a Simultaneous Test Procedure (STP) and STPs are commonly used to control the FWER. Bonferroni procedure is an example of an STP and is the most stringent amongst those.

Bonferroni-type corrections are the usual approaches for controlling the FWER at level α, by specifying the individual tests' α' values (comparison-wise error rate - CWER).

Suppose we want to test for association between a trait and m independent SNPs, i.e, we want to perform m independent tests, each with a pre-specified common type-I error α'. In this case, the experiment-wise error rate is:

$$\begin{aligned}
\text{FWER} &= P(\text{at least one test causes rejection} \mid H_0 \text{ true}) \\
&= 1 - P(\text{all test do not cause rejection} \mid H_0 \text{ true}) \\
&= 1 - P(\text{one test does not cause rejection} \mid H_0 \text{ true})^m \quad (\text{because of independence}) \\
&= 1 - (1-\alpha')^m \simeq m\alpha'.
\end{aligned}$$

Definition 3.1 If one wishes to control the FWER at a pre-specified level α, the adjusted p-value for hypothesis H_i is given by

$$p_i^* = \inf\Big\{\alpha \in [0,1] : H_i \text{ is rejected at nominal level FWER} = \alpha\Big\}. \tag{3.1}$$

Adjusted p-values are a way of better comparing different procedures allowing direct interpretation against a pre-specified α-value since they already take into account that multiple tests are being conducted (see Wright, 1992, for a discussion on the topic). Moreover, they are a way of reporting the result of a single test using p-values, with the advantage that the level of the test does not have to be determined in advance. We illustrate this with a simple example: let's assume we have an adjusted ordered p-value of $p^*_{(i)} = 0.0111$ obtained from any FWER controlling procedure. If we meant to control the FWER at level 0.01 then in this case the hypothesis $H_{(i)}$ would be accepted. On the other hand, if we relaxed a bit the significance level, e.g., to $\alpha = 0.015$, $H_{(i)}$ would be rejected.

One FWER controlling procedure is (see Bonferroni (1935, 1936) and Dunn (1961)):

Algorithm 1 - Bonferroni (1935) -

i. Calculate the m individual test p-values, say, $p_1, ..., p_m$ of the test statistics $T_1, ..., T_m$ of the hypotheses $H_1, ..., H_m$.

ii. Should there be a $p_i \leq \frac{\alpha}{m}$ for some $i = 1, ..., m$, then the null hypothesis, $H_0 : H_1, ..., H_m$, is rejected. In particular, each specific hypothesis H_i is rejected if $p_i \leq \frac{\alpha}{m}$.

The adjusted p-values are $p_i^* = \min(mp_i, 1)$ (follows directly from (3.1)).

Definition 3.2 One procedure is said to control FWER in:

1. the **weak sense**, if it controls the FWER under the complete null, i.e, when all null hypotheses are actually true.

2. the **strong sense**, if it controls the FWER under any combination of true and false hypotheses, i.e, under any $H_0^{'}$ where $H_0^{'}$ is a subset of H_0 of $m' \leq m$ individual hypotheses.

In general, weak control of the FWER without other safeguards is unsatisfactory. This is because the complete null is in general not realistic (Dudoit *et al.*, 2003).

It is easy to see that strong control of the FWER implies weak control. When weak control is all we have, one hopes that the error rate under the true distribution is bounded by the error rate under the complete null (Farcomeni, 2008).

The Bonferroni inequality

$$P\Big(\bigcup_{i=1}^{m}(P_i \leq \frac{\alpha}{m})\Big) \leq \alpha, \quad (0 \leq \alpha \leq 1) \tag{3.2}$$

ensures that the probability of rejecting at least one hypothesis when all are true is no greater than α, i.e, FWER $\leq \alpha$, and so Bonferroni STP is said to control de FWER in the weak sense. Hocheberg (1988) shows that it also controls the FWER in the strong sense. If the number of true hypothesis m_0 were known, then the Bonferroni procedure with cut-off value α/m_0 would also control the FWER in the strong sense. This procedure is known in the literature as the *oracle* Bonferroni. Should we use this adjustment having m_0 much, much smaller than m, the level of α/m_0 would be higher than α/m and we would thus have more power.

Although being a simple procedure to use, requiring no distributional assumptions and enabling individual alternative hypotheses to be identified, it lacks power if several tests are highly correlated (e.g, when there are several or many SNPs in LD[2]). Moreover, it offers the most conservative approach to control for false positives, especially with increasing number of tests, and when allowing possible unacceptable levels of type-II errors it may contribute to publication bias and the exclusion of potentially relevant hypotheses (e.g, significant SNP-phenotype associations).

This method however, when combined with strategies to reduce the number of tests can sometimes be a useful approach. For example, if the number of independent tests among the m tests can be correctly inferred (Cheverud (2001), Nyholt (2004, 2005), Li and Ji (2005), Salyakina *et al.* (2005), Gao *et al.* (2008)), we can still use the standard Bonferroni to rapidly adjust for multiple testing.

Although Bonferroni correction is the most common method for type-I error correction, less conservative methods exist. One example is the Dunn-Sidák method. This method, comes from the independence assumption and the equality

$$\alpha = 1 - (1 - \alpha')^m.$$

Solving for α' we obtain $\alpha' = 1 - (1 - \alpha)^{1/m}$. We thus have,

---

Algorithm 2 - Sidák (1968/1971) -

i. Calculate the m individual test p-values, say, $p_1, ..., p_m$ of the test statistics $T_1, ..., T_m$ of the hypotheses $H_1, ..., H_m$.

ii. Should there be a $p_i \leq 1 - (1 - \alpha)^{1/m}$ for some $i = 1, ..., m$, then the null hypothesis, $H_0 : H_1, ..., H_m$, is rejected. In particular, each specific hypothesis H_i is rejected if $p_i \leq 1 - (1 - \alpha)^{1/m}$.

The adjusted p-values are $p_i^* = \min(1 - (1 - p_i)^m, 1)$.

The Sidák method gives the exact control of the FWER under the null hypothesis, where the unadjusted p-values are independently $U(0, 1)$ distributed, but like the Bonferroni method, it is also conservative. In fact, it actually is less conservative than the former, although for very small p-values there is apparently little difference between the

[2]Since closely spaced SNPs are usually in LD when association studies are performed with many SNPs, the tests performed depend on the correlation structure among SNPs.

two methods. It does not however provide FWER control for arbitrary distributions of the test statistics. In these circumstances control of the FWER can be guaranteed if the test statistics verify the Sidák's inequality:

$$P(|T_1| \leq c_1, ..., |T_m| \leq c_m) \geq \prod_{j=1}^{m} P(|T_j| \leq c_j). \tag{3.3}$$

This inequality is also known as the *positive orthant dependence property* and was initially derived by Dunn (1961) in a particular setting.

If some of the hypotheses are true, step-wise procedures are generally preferable. Additionally, one should also note, that small values of α reduce the power for any single test. Thus, if we perform a huge number of simultaneous tests one may end with no significant declared associations. In this case as well, another correction method should be preferred.

Holm (1979), proposed a simple sequential rejective multiple test procedure based on the ordered test individual p-values. This step-down method is based on the Bonferroni inequality and is therefore valid regardless of the joint distribution of the test statistics. It also controls the FWER in the strong sense. This method was made quite popular in ecological studies by Rice (1989).

Considering m tests at an overall significance level α, Holms' proposal consists of the following steps:

Algorithm 3 - Holm (1979) -

i. Calculate the m individual test p-values, say, $p_1, ..., p_m$ of the test statistics $T_1, ..., T_m$ of the hypotheses $H_1, ..., H_m$ and rank them.

ii. Should there be a $p_{(i)} \leq \frac{\alpha}{m-(i-1)}$ for some $i = 1, ..., m$, then the null hypothesis, $H_0 : H_1, ..., H_m$, is rejected. In particular, each specific hypothesis $H_{(i)}$ is rejected if $p_{(i)} \leq \frac{\alpha}{m-(i-1)}$ and if the previous hypothesis $H_{(1)}, ..., H_{(i-1)}$ have been rejected. Acceptance of $H_{(k)}$ implies acceptance of all $H_{(l)}$ with $l > k$, $l \leq m$.

The adjusted p-values are $p^*_{(i)} = \max_{k=1,...,i} \Big[\min \Big((m-k+1)p_{(k)}, 1 \Big) \Big]$.

This is a little less conservative correction procedure compared to Bonferroni and Sidák corrections, providing a non negligible increase in power of the tests while keeping all the properties of the formers. This modification cannot however shorten Bonferroni's confidence intervals for individual (scalar) parameters. And, like Bonferroni, it does not as well account for possible dependencies between the tests.

A major assumption of Bonferroni, Sidák and Holm's methods is, as mentioned previously, that the tests are uncorrelated, otherwise corrections would be too extreme. It so happens that in genome-wide association studies this assumption is often violated when, for example, LD introduces correlation between tests. But still, if the number of independent tests among the total of tests can be correctly inferred, these techniques can still be used.

Another improvement on the Bonferroni procedure was made by Simes (1986). Like Holm's procedure, it is also based on the ordered individual test p-values. Considering m tests at an overall significance level α, it consists of the following steps:

Algorithm 4 - Simes (1986) -

i. Calculate the m individual test p-values, say, $p_1, ..., p_m$ of the test statistics $T_1, ..., T_m$ of the hypotheses $H_1, ..., H_m$ and rank them.

ii. Should there be a $p_{(i)} \leq i\frac{\alpha}{m}$ for some $i = 1, ..., m$, then the null hypothesis, $H_0 : H_1, ..., H_m$, is rejected. In an exploratory sense, individual hypotheses $H_{(1)}, ..., H_{(j)}$ can be rejected for every $j = \max\{k : p_{(k)} \leq k\frac{\alpha}{m}\}$.

Simes method is known to be less conservative than the classical Bonferroni procedure, thus having more power, and quite simple to apply. Simes (1986) shows in is work, via a simulation study, that the probability of a type-I error does not exceed the family-wise threshold α for a variety of multivariate normal and gamma test statistics (Simes' inequality - equality holds when the p-values P_i are independently uniformly distributed on $[0, 1]$):

$$P\Big(\bigcup_{i=1}^{m}(P_{(i)} \leq i\frac{\alpha}{m})\Big) \leq \alpha, \quad (0 \leq \alpha \leq 1). \tag{3.4}$$

Moreover, he shows that his procedure has level α under the null when the tests are independent, meaning that in this case we have weak control of the FWER. Additionally, his simulation study supports the advantage of the method over the Bonferronis' when

there are several highly-correlated test statistics. Sarkar (1998) proved Simes inequality for general positive dependence.

This criterion however, isn't always satisfactory as Simes (1986) carefully pointed out (an example of this can be found in Hommel (1988), pp. 384). Also, there is a problem of making statements on individual hypotheses for which a solution is not provided - we only have an exploratory decision criteria. This problem was overcome by Hommel (1988) (see Algorithm 8) who extended Simes' procedure for making statements on individual hypotheses. He also recommended the use of the Simes' method whenever possible since it is strictly more powerful than Bonferroni's algorithm (Algorithm 1) and Ruger's (1978) and Hommel's (1983) procedures.

For situations where the truth of some hypotheses implies the truth of others, Shaffer (1986) proposed a modification on Holm's procedure in order to increase power, however with a complexity cost. Considering m dependent tests at an overall significance level α, we have the following steps:

Algorithm 5 - Shaffer (1986) -

i. Calculate the m individual test p-values, say, $p_1, ..., p_m$ of the test statistics $T_1, ..., T_m$ of the hypotheses $H_1, ..., H_m$ and rank them.

ii. Calculate the values

$$t_i = \text{ maximum number of possible true hypothesis given that } H_{(1)}, ..., H_{(i-1)} \text{ are false}$$

for $i = 1, ..., m$. Obviously $t_i \in \{1, ..., m - i + 1\}$ and $t_1 = m$.

iii. Let $i^* \in \{1, ..., m\}$ be the first index to verify the inequality $p_{(i^*)} > \frac{\alpha}{t_{i^*}}$. Then reject $H_{(1)}, ..., H_{(i^*-1)}$ and accept $H_{(i^*)}, ..., H_{(m)}$. Should there be a $p_{(i)} \leq \frac{\alpha}{t_i}$ for some $i = 1, ..., m$, then the null hypothesis, $H_0 : H_1, ..., H_m$, is rejected.

When there are no logical implications among the m hypotheses, i.e, when the m tests are independent, $t_i = m - i + 1$ and, thus, this method coincides with Holm's. Moreover, since we have the inequalities $\frac{\alpha}{m} \leq \frac{\alpha}{m - i + 1} \leq \frac{\alpha}{t_i}$ it is clear that any hypothesis rejected by Bonferroni procedure is subsequently rejected by Holm's and Shaffer's procedures. The complexity cost issue mentioned, arises from the need to compute the values $t_1, ..., t_m$, which will require some effort especially when m is large.

Holland and Conpenhaver (1987) made a small but uniform improvement on the Shaffer and Holm procedures for positively orthant dependent test statistics:

Algorithm 6 - Holland & Conpenhaver (1987) -

i. Calculate the m individual test p-values, say, $p_1, ..., p_m$ of the test statistics $T_1, ..., T_m$ of the hypotheses $H_1, ..., H_m$ and rank them.

ii. Calculate the values

$$t_i = \text{ maximum number of possible true hypothesis given that } H_{(1)}, ..., H_{(i-1)} \text{ are false}$$

for $i = 1, ..., m$. Obviously $t_i \in \{1, ..., m - i + 1\}$ and $t_1 = m$.

iii. Let $i^* \in \{1, ..., m\}$ be the first index to verify the inequality $p_{(i^*)} > C(t_{i^*})$, where $C(x) = 1-(1-\alpha)^{1/x}$. Then reject $H_{(1)}, ..., H_{(i^*-1)}$ and accept $H_{(i^*)}, ..., H_{(m)}$. Should there be a $p_{(i)} \leq \frac{\alpha}{t_i}$ for some $i = 1, ..., m$, then the null hypothesis, $H_0 : H_1, ..., H_m$, is rejected.

This procedure keeps the overall type-I error probability below the threshold α and is at least as powerful as Shaffer and Holm's procedures. Holm (1979) himself had already recognized that his rule could be improved with $p_{(i^*)} > C(m - i^* + 1)$ for the case of independent statistics, although he did not mention it was also applicable in certain dependent situations as is the case of positively orthant dependence.

In their paper, Holland and Copenhaver suggest their method to be used in all circumstances where Bonferroni is preferred whenever the positive orthant dependence holds.

Power gains are still possible with the sequential approach by using a step-up instead of a step-down procedure. Hochberg (1988), presented what he called "a sharper Bonferroni procedure", which is an extension of Sime's procedure that now has strong control of the FWER.

For m independent tests at an overall significance level α, this step-up method has the following steps:

Algorithm 7 - Hochberg (1988) -

i. Calculate the m individual test p-values, say, $p_1, ..., p_m$ of the test statistics $T_1, ..., T_m$ of the hypotheses $H_1, ..., H_m$ and rank them.

ii. Take $i = 1$.

iii. If $p_{(m-i+1)} \leq \frac{\alpha}{i}$ then reject $H_{(1)}, ..., H_{(m-i+1)}$ and accept the others. Go to step v.

iv. If $p_{(m-i+1)} > \frac{\alpha}{i}$, increment i, i.e, make $i = i + 1$ and go back to iii.

v. H_0 is rejected if at least one $H_{(i)}$ as been rejected. H_i is rejected if $p_j \leq \frac{\alpha}{m-j+1}$ for any $j \geq i$.

The adjusted ordered p-values are $p^*_{(i)} = \min_{k=i,...,m} \left\{ \min \left\{ (m-k+1)p_{(k)}, 1 \right\} \right\}$.

This step-up method, that can be viewed as an analog of Holm's step down, is uniformly more powerful than the latter, but, unlike Holm's method, you can only use it in independent tests. Step-up methods are not always better than step-downs' (see example in Huang and Hsu, 2007). Both methods are special cases of partition testing (Stefansson *et al.* (1988), Hayter and Hsu (1994), Finner and Strassburger (2002)). The difference is that, Holm's method tests each partition hypothesis using the largest order statistic, setting a critical value based on the Bonferroni inequality (Equation 3.2) while, Hochberg's tests each partition hypothesis using all the order statistics, setting a series of critical values based on Sime's inequality (Equation 3.4). See Huang and Hsu (2007) for more detail and also Lehman *et al.* (2005) for a discussion on step-up and step-down procedures optimality.

The Hochberg procedure, is perhaps the most popular improved Bonferroni-type procedure. Not only it is easier to understand and implement but also only slightly less powerful than other FWER controlling Bonferroni-type procedures. Sarkar and Chang (1997) proved that it also controls FWER when the distribution of the test statistics has exchangeable positive dependence.

Bonferroni, Hochberg, Holm, and Sidak's procedures are implemented in the R package *multtest* via instruction `mt.rawp2adjp()`.

Related procedures to Hochberg's are the ones of Hommel (1988) and Rom (1990), both being more powerful than the former. We will describe below only Hommel's procedure since it is a generalization of the Simes' procedure that gives strong control of the FWER whenever Simes' gives weak control (e.g., with independent tests).

Algorithm 8 - Hommel (1988) -

i. Calculate the m individual test p-values, say, $p_1, ..., p_m$ of the test statistics $T_1, ..., T_m$ of the hypotheses $H_1, ..., H_m$ and rank them.

ii. Calculate the value $k_i = \max_i p_{(m-i+j)} > \alpha \frac{j}{i}$, with $j = 1, ..., i$.

iii. Should there be $p_{(i)} \leq \frac{\alpha}{k_i}$ for some $i = 1, ..., m$, then the null hypothesis, $H_0 : H_1, ..., H_m$, is rejected. In particular, each specific hypothesis $H_{(i)}$ is rejected if $p_{(i)} \leq \frac{\alpha}{k_i}$.

The ordered adjusted p-values are $p^*_{(i)} = \min\left\{k_i p_{(i)}, 1\right\}$.

Hommel's method, although a bit more complicated when compared to Hochberg's, is more powerful than the Hochberg correction and, thus, less conservative. It can be viewed as an extension of Simes' procedure as it now enables inferences on individual hypotheses. An example of application can be found in Wright (1992).

A popular approach in genetic association studies for the control of the type-I error rate, where Bonferroni-type methods may be too conservative, is the use of a permutation procedure. Permutation tests provide valid adjustments if the data are permuted as to simulate the null hypothesis while keeping the data original structure. In genetic association studies, it means retaining the genotype data, so that the LD structure between SNPs is preserved, and randomizing the phenotype over individuals. In this way the resulting data set will satisfy the null hypothesis of no association between the SNPs and the phenotype. The analysis of many such data sets (the more the better), comparing the permutation-based results to the original ones, allows the false-positive rate to be approximated, based on the probability of observing a p-value as small as the original minimum, given the correlation between tests. The draw back of such an approach in the context of large association studies is, of course, computational burden.

Several permutation methods have been proposed in the literature to test the significance of partial regression coefficients in multiple regression: (1) permute the raw data,

(2) permute the residuals of a null model (2 variants), or (3) permute the residuals of a full model (see Westfall and Young, 1993; Legendre and Legendre, 1998; Anderson and Legendre, 1999; for more information).

Wesfall and Young (1993) discuss several FWER controlling procedures and resampling techinques[3] to obtain the adjusted p-values when the distribution of the original p-values is unknown. We will only mention below the proposed minP and maxT procedures, which are both step-downs, take into account the joint distribution of the test statistics, are less conservative than Bonferroni, Holm or Hocheberg's procedures, and whose p-values can be estimated by resampling, making successively smaller adjustments at each step.

Algorithm 9 `- single-step minP -`
`Calculate the adjusted p-values` $p_i^* = P(\min_{1 \le j \le m} P_j \le p_i \mid H_0^C)$, `where` H_0^C `is the complete null hypothesis and` P_j `denotes the random variable for the unadjusted p-value of the` j`-th hypothesis.`

Procedures based on minP adjusted p-values provide weak control of the FWER. Strong control is achieved under the assumption of subset pivotality[4] (see Westfall and Young, 1993, pp. 42-43, for more detail). These two last considerations are also valid for maxT based procedures.

We list some more considerations on this procedure:

1. If the unadjusted minP p-values are independent and P_j is $U(0,1)$ distributed under H_j, then the minP adjusted p-values coincide with Sidák's p-values:

$$p_i^* = P(\min_{1 \le j \le m} P_j \le p_j \mid H_0^C) = 1 - P(P_j > p_i \text{ for all } j \mid H_0^C) \underset{P_j \sim U(0,1)}{=} 1 - (1 - p_i)^k.$$

2. When the Sidák's inequality (Equation 3.3) holds, the minP adjusted p-values are less than or equal to Sidák's adjusted p-values;

[3] Could either be through permutation (resamplig without replacement) or bootstrap (resampling with replacement) - see Westfall and Young (1993), chapters 4 and 5, for more detail.

[4] The distribution of unadjusted p-values $\{P_1, ..., P_m\}$ is said to have the subset pivotality property, if the joint distribution of the random vector $\{P_j : j \in \Omega\}$ is identical for distributions satisfying the composite null hypothesis $\bigcap_{j \in \Omega} H_j$ and $H_0^C = \bigcap_{i=1}^{m} H_j$, for all subsets Ω of $\{1, ..., m\}$.

3. Under Sidák's inequality, procedures based on minP adjusted p-values are less conservative than Bonferroni or Sidák procedures.

Algorithm 10 `- single-step maxT -`
`Calculate the adjusted p-values` $p_i^* = P(\max_{1 \le j \le m} |T_j| \ge |t_i| \mid H_0^C)$, `where` H_0^C `is the complete null hypothesis and` T_j `and` t_i `denote the random and the observed test statistics.`

Usually the p-value distribution under the null is unknown and therefore resampling techniques such as bootstrap or permutation, need to be used to estimate the adjusted p-values. Westfall and Young (1993) suggested several ways of deriving this distribution, including permutation and bootstrap methods, using either p-values or test statistics. It is clear that when using permutation methods to obtain the adjusted p-values for the above procedures, they both become computationally slow, although maxT is known to be less demanding than the minP. Ge *et al.* (2003) suggested a fast procedure for the computation of the minP adjusted p-values - a recent review can be found in Farcomeni (2008). The adjusted p-values from minP and maxT are the same when the test statistics are identically distributed. When they are not, maxT adjustments may be unbalanced in the sense that not all tests contribute equally to the adjustment. Also, maxT can be a more powerful approach in small "n" large "m" situations.

As an example of application of the use of the minP procedure in a genetic case-control study context, we have the work of Chen *et al.* (2006). In this paper they investigate whether the gain in power, obtained by using permutation resampling to account for correlations between polymorphisms and test statistics, is big enough as to justify the associated computational burden. Therefore, they compared the resampling-based minP procedure with Simes' procedure and with haplotype analysis (see Mander, 2001, for a quick review) under several scenarios via simulation, where empirical haplotype structures in 15 genes were incorporated. Recall that inference with the minP test, relies on the permutation distribution of the minimum of the ordered p-values, which takes the correlations between SNPs into account whereas the Simes' global test uses a non-iterative procedure to adjust the minimum observed p-value for multiplicity. In their work, Chen *et al.* found minP to be only slightly more powerful than Simes procedure (around 2%) and observed a few situations where minP under-performed haplotype-analysis. They thus proposed an omnibus test procedure combining both haplotype analysis and minP to obtain a more effective gain in power while still controlling the FWER.

Other application examples of minP can be found in the work of Hosgood *et al.* (2008), also in a case-control associtation context, and Schwartzbaum *et al.* (2010), now in a gene expression study context. Chapman and Whittaker (2008) showed that the minP approach performs quite well under scenarios like e.g., the presence of high and low LD betweens SNPs, and recommended its use in genetic association studies. Pollard *et al.* (2005) argue that approaches that use a data generating distribution, such as a permutation or bootstrap distribution, that satisfies the complete null, typically rely on the subset pivotality condition to justify strong control of the FWER, which in turn is violated in many relevant biomedical and genomic data analysis testing problems, e.g., tests concerning correlation coefficients and tests concerning regression coefficients.

Both minP and maxT procedures are implemented in the R package *multtest* where correspondent adjusted p-values can be obtained through instructions `mt.minP()` and `mt.maxT()`. The permutation algorithm used to compute these p-values is the one of Ge *et al.* (2003).

Despite all the improvements made on the Bonferroni correction procedure (see Garcia, 2004, for a review), they all focus on limiting the chance of making a single type-I error which, with an increasing number of tests comes at a cost: more type-II errors. I.e, low power, translating into failing to detect a true effect. In genome-wide genetic association studies the Bonferroni correction is therefore undesirable, since only a few very strong effects will be detected among the many many tests performed. And this brings us to the next section.

3.2 Controlling the false discovery rate

In genetic association studies the risk of a false discovery (indication of association when there is none) is very high because only few among all markers/genes/SNPs that can be tested will have an effect on the phenotype and/or disease. See for example the work of Colhoun *et al.* (2003) where speculation is made that 19 out of 20 marker-disease associations reported in the literature are false. To this respect, population stratification and failure to exclude chance can be argued. Moreover, since it has recently become possible to screen hundreds of thousands or even millions of SNPs across the genome of plants and animals to test for association with a disease and/or phenotype of interest, without the proper control, researchers will be flooded with false discoveries, those being proportional to the number of tests performed. It is therefore compelling to develop

proper methods for the adequate control of false discoveries so that a lot of time and resources are not spent in pursuing dead-end leads.

Motivated by the paper of Soriç (1989), Benjamini and Hochberg (1995) introduced the FDR as the expected proportion of the true null hypotheses that are wrongly rejected, the so-called false positives:

$$\text{FDR} = E\Big[\frac{\text{number of false positives}}{\text{total number of rejections of the null hypothesis}}\Big]$$

as long as the denominator is different from zero in which case FDR = 0. Contrary to the FWER, this approach is meant to control the proportion of significant results that are in fact type-I errors. The reasoning is that in genome-wide genetic association studies, a few false positives are acceptable if in small proportion to the total number of identified SNPs, i.e, rejected null hypotheses. To this respect the FWER may be too conservative and hence, this error rate seems to be appropriate on those problems where the control of the FWER is not necessarily implied.

This criterion seems to be the most satisfactory approach for dealing with tests' multiplicity and has been advocated for ecological studies involving a large number of tests in which several or even many individual hypotheses have been rejected and in situations where the most significant p-value is relatively large (Waite and Campbell, 2006). Benjamini and Hochberg thus proposed a sensible and powerful way of addressing the multiple testing problem.

Garcia (2004) summarized the following advantages of the FDR approach over the FWER:

1. it enables controlling the proportion of false positives among the rejected null hypotheses;

2. it avoids performing individual tests at very low p-levels in large problems;

3. it is more powerful than the sequential Bonferroni procedures (e.g, Holm, 1979; Hochberg, 1988; Hommel, 1988);

4. when no actual true positive findings exist (i.e, all the null hypotheses are true), the FDR method has the same control as the previous methods, i.e, it has a so-called weak control of the family-wise type-I error;

5. the FDR threshold may be determined from the observed p-value distribution, and

hence is adaptive to the amount of signal in the data (Genovese and Wasserman, 2002);

6. FDR is family-wise robust, i.e, it tends to be far more consistent than procedures controlling the FWE in terms of whether a particular hypothesis is rejected, as the family in which this hypothesis is located changes in size[5]. Additionally, FDR may account for the exact dependence structure of the data, via resampling-based procedures (Yekutieli and Benjamini, 1999) or under the normal assumption (Troendle, 2000).

The FDR is less stringent than the FWER and so any procedure controlling the FWER also controls the FDR. This allows FDR controlling procedures to have greater power than FWER controlling procedures, which may lead more researchers to actually use multiplicity adjustment. Also notice that the gain in using the FDR compared to the FWER increases when the number of false null hypotheses is large. Consequently, it seems that the FDR is well suited for testing situations in which there are many different treatments or outcomes for which there is little or no prior knowledge, and the objective is to identify a smaller set of them to study further.

To better illustrate the differences between FWER and FDR control, let us consider the possible outcomes of m tests as in Table 3.1, where V is the number of type-I errors, T the number of type-II errors, U the number of true negatives, S the number of true positives, m_0 the number of true null hypothesis, $m - m_0$ the number of false null hypothesis. Note that m, R and m-R are observed and the rest of the variables are unknown.

Table 3.1: Number of errors committed when testing m null hypothesis

	Accept H_0	Reject H_0	Σ
H_0 true	U	V	m_0
H_0 false	T	S	m-m_0
Σ	m-R	R	m

From the notation in Table 3.1, we have that

$$\text{FWER} = P(V \geq 1) \tag{3.5}$$

[5]See Holland and Cheung (2002) for family-wise robustness criteria for multiple comparisons procedures.

and (as defined by Benjamini and Hocheberg, 1995)

$$\text{FDR} = \begin{cases} E\left(\frac{V}{R}\right) & \text{if } R > 0 \\ 0 & \text{if } R = 0 \end{cases} \tag{3.6}$$

where the ratio V/R is called the false discovery proportion (Genovese and Wasserman, 2004).

Other frequently used rates in the literature are the per-family error rate (PFER), given by $\text{PFER} = E[V]$, and the per-comparison error rate (PCER) given by $\text{PCER} = \frac{E[V]}{m}$. The value of the FDR is in between the PCER and the FWER, i.e,

$$\text{PCER} \leq \text{FDR} \leq \text{FWER}.$$

When $m_0 = m$, FDR=FWER (Benjamini and Hochberg, 1995). When $m_0 < m$, a FDR control may be more powerful than a FWER control at the same level. The proposal of Benjamini and Hochberg (1995) (Algorithm 11) thus controls the FWER in the weak sense. Moreover, the FDR is adaptive in the sense that, the more hypotheses are not true, the smaller the FDR.

Also, we have

$$\text{PCER} \leq \text{FWER} \leq \text{PFER}.$$

From the inequalities above, one can easily see that for a fixed threshold, procedures that control the PFER are generally more conservative, that is, lead to fewer rejections, than those that control either the FWER or the PCER, and procedures that control the FWER are more conservative than those that control the PCER. These results can be better acknowledged in a comparison study example in Dudoit *et al.* (2003)[6].

We will now present a couple of FDR control methods based in the availability of the p-values for each of the individual hypotheses. When such is not the case, a generalized p-value approach is a possibility (see Tsui and Weerahandi, 1989; Tang and Tsui, 2007). The rational behind the p-value adjustment for FDR controlling procedures is the same as seen for FWER controlling procedures:

Definition 3.3 If one wishes to control the FDR at a pre-specified level α, the adjusted

[6]In this paper different approaches to large-scale multiple hypotheses testing in the context of DNA microarray experiments are discussed.

p-value for hypothesis H_i is given by

$$p_i^* = \inf\Big\{\alpha \in [0,1] : H_i \text{ is rejected at nominal level FDR} = \alpha\Big\}. \tag{3.7}$$

We start with the Simes-type proposal of Benjamini and Hochberg (1995) who introduced a distribution-free method that controls the expected fraction of falsely rejected null hypotheses among those rejected, with the understanding that if no null hypotheses are rejected the proportion is zero.

Algorithm 11 - Benjamini & Hocheberg (1995) - (BH-1995)
Consider m independent tests.

i. Calculate the m individual test p-values, say, $p_1, ..., p_m$ of the test statistics $T_1, ..., T_m$ of the null hypotheses $H_1, ..., H_m$ and rank them in ascending order $p_{(1)} \leq ... \leq p_{(m)}$.

ii. Denote by $H_{(i)}$ the hypothesis corresponding to $p_{(i)}$.

iii. Let k be the largest i for which $p_{(i)} \leq \alpha\dfrac{i}{m} = \alpha_i$.

iv. Reject all null hypothesis $H_{(1)}, ..., H_{(k)}$. The adjusted p-values are $p^*_{(i)} = \min\limits_{k=i,...,m}\Big[\min\Big(\dfrac{mp_{(k)}}{k}, 1\Big)\Big]$.

This procedure, also referred to in literature as the step-up linear procedure, was shown by Benjamini and Hochberg (1995) to control the FDR at the level $\alpha m_0/m$ and thus at level α (Storey *et al.*, 2004 also has an elegant proof of this) under independence. However, at the beginning it did not deserve much attention since it only provided weak control of the FWER. Proof against strong control follows from Seeger (1968) and Hommel (1988). Sarkar (2004) proved that the BH-1995 procedure is unbiased, being a special case of FDR-controlling step-up-step-down procedures proposed by Sarkar (2002).

Benjamini and Yekutieli (2001) studied the BH-1995 procedure under dependence and showed that the it controlled the FDR with the same upper bound $\alpha m_0/m$ when tests are positively correlated. In a more general dependence structure it controls the FDR at level

$$\alpha(1 + \frac{1}{2} + ... + \frac{1}{m}). \tag{3.8}$$

When tests are negatively correlated, or have a more complex structure, they demonstrated that replacing m in the Algorithm 11(iii) with $m\sum_{i=1}^{m}\frac{1}{i}$ also provides FDR control (we will refer to this modified procedure as BY). The adjusted p-values for this modified step-up version of the previous procedure are given by

$$p^*_{(i)} = \min_{k=i,\ldots,m}\Big[\min\Big(\frac{mp_{(k)}}{k}\sum_{i=1}^{m}\frac{1}{i}, 1\Big)\Big]. \tag{3.9}$$

In that paper they also proposed an FDR controlling procedure that does not make any assumption on the joint distribution of the test statistics. Some asymptotic results and plug-in procedures for the BH-1995 procedure can be found in Genovese and Wasserman (2002, 2006). Note that BH-1995 and BY procedures differ only in the multiplicity penalty attributed to the unadjusted p-values: in BH-1995 the applied penalty is of m/k whereas in BY we have a penalty of $(m\sum_{i=1}^{m}\frac{1}{i})/k$.

Both BH-1995 and BY procedures are implemented in the R package *multtest* and can be calculated with instruction `mt.rawp2adjp()`. We also have the BH-1995 procedure available via instructions `qvaluebh95()` and `p.adjust()` from the R packages *GenABEL* and *stats*, respectively. We note that the two last functions return the adjusted p-values in the initial order whereas the first returns them ordered.

We emphasize the fact that positive dependence is a very common situation in genetic association studies. E.g., it may arise in SNP/trait association studies when there are SNPs in LD. Since this correction becomes more stringent as the p-value decreases, i.e, as we go from $p_{(m)}$ to $p_{(1)}$ the rejection significance levels of the individual tests decrease, in the context of genetic association studies, this method provides a good balance between discovery of statistically significant genes (i.e, associations) and limitation of false positive occurrences thus improving power. It so provides a most satisfactory approach for coping with multiplicity when a large number of tests are performed simultaneously, which is usually the case in this specific setting.

Knowledge on m_0 can be very useful as a tool to improve the power of BH-1995 procedure. Even with the Bonferroni procedure, taking α/m_0 instead of α/m would enhance its performance. Hochberg and Benjamini (1990) used this idea to construct more powerful versions of the Holm and Hochberg's procedures. There is therefore much interest in constructing procedures that firstly estimate m_0 and then use its estimate in the BH-1995 procedure. These plug-in procedures are usually referred to in literature as

adaptive procedures. We must underline however, that once an estimator of m_0 is used in the BH-1995 procedure, there is no guarantee that the new procedure assures FDR control at the desired level. For that purpose additional adjustments in the estimator may be needed (Storey *et al.*, 2004; Benjamini *et al.*, 2006; Gravilov *et al.*, 2009; Blanchard and Roquain, 2009). These adjustments are critical under dependence and when $m_0/m \sim 1$ or even $m_0/m = 1$.

Schweder and Spjϕtvoll (1982) were the first ones to propose an informal way of estimating m_0 from inspecting the quantile-plot of the p-values versus their ranks. The idea is quite simple: since the plot will tend to show a linear behaviour for the largest p-values which are the ones most likely to correspond to the true null hypotheses, one chooses the k-largest p-values and takes $\widehat{m}_0$ as the slope of the line that passes through the $\{(1-p_1), ..., (1-p_k)\}$ k p-values. Still, it is known that this method tends to overestimate the number of nulls (Walsh, 2004). Hochberg and Benjamini (1990) referred to this methodology proposing a straight line to be fit through the largest p-values by OLS estimation. Heuristic methods on how many largest p-values should be used to fit this line are also discussed.

Following the idea of m_0 estimation with subsequent plug-in in the BH-1995 procedure, Benjamini and Hochberg (2000) introduced a sharpening procedure for their FDR method showing that power over the initial method can be substantially improved. Basically, this procedure uses the data to provide a 'prior' estimate of the number of true nulls, m_0, among the m null hypotheses:

Algorithm 12 - `Benjamini & Hochberg (2000) - (BH-2000)`

i. `Use the linear step-up procedure BH-1995 at level` α `and if no hypothesis is rejected STOP.`

ii. `Calculate` $S_k = (m + 1 - k)/(1 - p_{(k)})$.

iii. `Starting with` $k = 2$ `STOP when for the first time` $S_k > S_{k-1}$.

iv. `Estimate` $\widehat{m}_0 = min\{S_k, m\}$ `rounding up to the next highest integer.`

v. `Use the linear step-up procedure BH-1995 with` $\alpha_i = \alpha \dfrac{i}{\widehat{m}_0}$.

This adaptive procedure is implemented in the R package *multtest* via instruction `mt.rawp2adjp()`, which returns the adjusted ordered p-values.

Storey (2002/2003) introduced a different concept of FDR, the positive FDR, pFDR, defined as

$$\text{pFDR} = E\Big[\frac{V}{R} \mid R > 0\Big]$$

and the concept of q-value of a particular test which gives the expected proportion of false discoveries when the cutoff point for H_0 is at that test's p-value.

Although similar to the notion of the p-value, these two significance measures should not be confused. The p-value is a measure of significance in terms of the false positive rate, whereas the q-value is a measure in terms of the FDR. E.g., a false positive rate of 5% means that on average 5% of true nulls in a study will be declared significant. A FDR of 5% means that among all features called significant, 5% of these are true nulls on average.

To control the introduced pFDR he proposed the following adaptive linear step-up procedure:

Algorithm 13 `- Storey (2002) -`
`Let the tuning parameter` λ `be estimated by resampling techniques after fixing the rejection area.`

`i. Take` $r(\lambda) = \#\{p_{(i)} \leq \lambda\}$.

`ii. Estimate` m_0 `by` $\widehat{m}_0(\lambda) = (m - r(\lambda))/(1 - \lambda)$.

`iii. Use the linear step-up procedure BH-1995 with` $\alpha_i = \alpha \dfrac{i}{\widehat{m}_0(\lambda)}$.

There is an inherent bias/variance trade-off in the choice of λ: usually, when λ gets smaller, the bias of $\widehat{m}_0(\lambda)$ gets larger and the variance gets smaller. A balance of this trade-off can be achieved with an adequate choice of λ. Storey (2002) and Storey and Tibshriani (2003) recommended the use of $\lambda = 0.5$ for the estimation of m_0.

Storey *et al.* (2004) modified this procedure by replacing in Algorithm 13(ii) $(m - r(\lambda))$ by $(m + 1 - r(\lambda))$ and additionally requiring that $p_{(i)} \leq \lambda$ for a hypothesis to be rejected.

Storey and Tibshriani (2003) provide a way of transferring a list of p-values to q-values as an approach to measure statistical significance in genome-wide studies. These ideas were incorporated in function `qvalue()` from the R package *qvalue*, that returns the

adjusted p-values without changing their initial order. They also presented a Bayesian interpretation of the pFDR and gave an algorithm for the estimation of the prior probability, $\pi_0 = m_0/m$, that a gene is not differentially expressed (in a genome-wide study context):

Algorithm 14 - Storey and Tibshriani (2003) - (ST)
Let p_i be the raw p-value of gene i, with $i = 1, ..., m$.

i. For $\tau = 0, 0.01, 0.02, ..., 0.95$ compute $\widehat{\pi}_0(\tau) = \#\{p_i > \tau\}/((1-\tau)m)$.

ii. Fit a natural cubic spline, *spl*, with three degrees of freedom through the data points $(\tau, \widehat{\pi}_0(\tau))$.

iii. Estimate π_0 as $\widehat{\pi}_0 = \min\{spl(1), 1\}$.

Benjamini *et al.* (2006), makes a review of some FDR controlling procedures that estimate m_0 and propose, among others, the following two-stage linear step-up procedure:

Algorithm 15 - Benjamini, Krieger & Yekutieli (2006) - (BKY)

i. Use the linear step-up procedure of BH-1995 at level $\alpha/(1+\alpha)$ and let r_1 be the number of rejected hypotheses. If $r_1 = 0$ do not reject any hypothesis and STOP. If $r_1 = m$ reject all hypotheses and STOP, otherwise move on;

ii. Take $\widehat{m}_0 = m - r_1$;

iii. use the linear step-up procedure of BH-1995 with $\alpha_i = \dfrac{\alpha}{1+\alpha}\dfrac{i}{\widehat{m}_0}$.

They show this procedure to be more powerful than the BH-1995 and to control the FDR under independence. Gravilov *et al.* additionally show it to control the FDR also under positive dependence. This procedure is also implemented in the R package *multtest* via instruction `mt.rawp2adjp()`.

Some more adaptive procedures to control the FDR have been proposed in the literature, e.g., Storey *et al.* (2004), Genovese and Wasserman (2004), Sarkar (2006), Benjamini and Heller (2007), Blanchard and Roquain (2009), Gravilov *et al.* (2009), Sarkar and Guo

(2009). A common drawback among all is that asymptotically they have been shown to control the FDR only under independence. Under dependence, Benjamini *et al.* (2006) showed that some adaptive procedures sometimes may fail to control the FDR.

In situations where the joint distribution of the test statistics is unknown, resampling methods (e.g., boostrap, permutation) can be used to estimate unadjusted and adjusted p-values, the same way as in FWER controlling procedures (Yekutieli and Benjamini, 1999; Tusher et al., 2001; Reiner et al., 2003). In this way, parametric assumptions on the test statistics distributions are avoided.

With the development of high-throughput biotechnologies we now have the ability of screening hundreds of thousands of SNPs and tens of thousands of gene expression profiles. This is why we are said to be facing the Era of genomics where such an enormous amount of genetic information certainly poses a huge and obvious problem to data handling and statistical analysis. To the crucial question "How can one assess the significance of biological findings over families of thousands or even millions of hypothesis tests?" there is extensive debate in the literature. Throughout this chapter we have reviewed several of the many proposed procedures, some able to control the FWER and others the FDR, or both. The choice of one correction method over another will depend of course on the context of the investigator's research. In genetic association studies it is usually better to allow false associations to occur as long as we can find a few positive associations rather than being over conservative and incur in a zero findings study, usually accompanied with time and money losses. As far as we have seen from several genetic association studies, the preference goes to the use of BH-1995 at the 10% level, although there is actually no consensus in the literature as to what the threshold should be.

We would now like to finish this chapter by leaving the reader with a small summary table (Table 3.2) where we briefly review some of the FWER and FDR procedures discussed and their properties.

Some of these measures will be compared in Chapter 6 in the context of outlier tests.

More recent developments on the subject of FDRs can be found, e.g, in Benjamini and Gravilov (2009) and Benjamini *et al.* (2009). We also have a recent and nice state of the art in Benjamini (2010).

Table 3.2: Summary table of the discussed multiple testing procedures which have strong control of either the FWER or the FDR - some properties.

Procedure	type-I error rate	Strong/weak control	Stepwise structure	Dependence structure
Bonferroni	FWER	Strong	Single	*
Sidák	FWER	Strong	Single	Positive orthant
Holm	FWER	Strong	Down	*
Hochberg	FWER	Strong	Up	Some dependence
minP	FWER	Strong	Single	Subset pivotality
maxT	FWER	Strong	Single	Subset pivotality
Benjamini and Hochberg	FDR	Strong	Up	Positive regression dep.
Benjamini and Yekutieli	FDR	Strong	Up	*
Yekutieli and Benjamini	FDR	Strong	Up	Some dependence
Storey	pFDR	Strong	Up	Independence

* means that the procedure controls the claimed type-I error rate for general dependence structures, but does not explicitly take into account the joint distribution of the test statistics.

Everyone believes in the [normal] law of errors, the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact.

Henri Poincaré

4

Multiple linear regression

The linear regression model is the simplest and most widely used model to describe the relationship between variables in many areas of research. More precisely, between a set of explanatory variables (one or more) and a response variable (or more if we refer to the multivariate case). Usually, models containing more than one explanatory variable can be more helpful in predicting values for the response variable, especially on those cases where the response variable is believed to be controlled by multiple predictor variables, as in the case of genetic quantitative traits which are known to be ruled by sometimes many genes.

Linear regression methodology however, is affected by the presence of highly discrepant observations in the data, which may arise from heavy-tailed distributions, mixture of distributions or simply from errors in data collection and/or recording. Robust linear regression on the other hand, presents an alternative to the classical methodology, being resistant to some of those model misspecifications, in particular to regression outliers/normality deviations.

In this Chapter we will review both the classic and the proposed robust approaches in a linear regression context (Sections 4.1-4.3), going through hypothesis testing (Section 4.4), coefficient of determination (Section 4.5), outlier detection methods (Section 4.6)

and robust plug-in estimators in mixed linear models (Section 4.7).

4.1 The regression model

We describe the general multiple-linear regression model as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{ip-1} + \varepsilon_i, \quad i = 1, \ldots, n \tag{4.1}$$

where $n > p$ is the number of observed individuals. This model appropriately rewrites to

$$Y = X\beta + \varepsilon \tag{4.2}$$

where $Y = (Y_1, \ldots, Y_n)^T$ is the $(n \times 1)$ vector of the response variable, X is the $(n \times p)$ design matrix, $\beta = (\beta_0, \ldots, \beta_{p-1})^T$ are the unknown parameters and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is a vector of non-observable independent errors with expectation $E(\varepsilon) = \mathbf{0}$ and covariance matrix $\text{var}(\varepsilon) = \sigma^2 I_n$, the latter equality meaning the errors are *homoscedastic*. Hence, $E(Y) = X\beta$ and $\text{var}(Y) = \sigma^2 I_n$.

In the literature it is also usual to see model (4.1) written as

$$Y_i = x_i \beta + \varepsilon_i \tag{4.3}$$

where $x_i = (1 \; X_{i1} \; ... \; X_{ip-1})$.

For documentation on heteroscedastic errors please see the work of Park (1966), Rutmiller and Bowers (1968), Harvey (1976), Smyth *et al.* (2001), Smyth (2002) and Yuan and Wahba (2004), among others, in the case of the classical approach. Check also the work of Giltinan and Carroll (1986), Bianco and Rienzo (2000), Carroll and Ruppert (1982) for the robust approach.

4.2 Least squares and maximum likelihood estimators

From equation (4.1), taking $X_{i\bullet} = (1 \; X_{i1} \; \ldots \; X_{ip-1})$ and writing $\varepsilon_i = Y_i - X_{i\bullet}\beta$, the least squares (LS) or ordinary least squares (OLS) estimate of β is obtained by minimizing the

error sum of squares $\sum_{i=1}^{n} \varepsilon_i^2$, i.e,

$$\widehat{\beta}_{LS} = \arg\min_{\beta} \sum_{i=1}^{n} \varepsilon_i^2 = \arg\min_{\beta} \varepsilon^T \varepsilon = \arg\min_{\beta} (Y - X\beta)^T (Y - X\beta). \tag{4.4}$$

If X has rank $p \leq n$ the problem has one unique solution $\widehat{\beta}_{LS} = (X^T X)^{-1} X^T Y$, with covariance $\Sigma_{\widehat{\beta}_{LS}} = \sigma^2 (X^T X)^{-1}$ and $\widehat{\sigma}^2_{LS} = MSE = \dfrac{\widehat{\varepsilon}^T \widehat{\varepsilon}}{n-p} = \dfrac{SSE}{n-p}$.

In the classical linear multiple regression model we additionally have $\varepsilon \sim N(\mathbf{0}, \sigma^2 I_n)$ and consequently $\widehat{\beta}_{LS}$ coincides with the maximum likelihood estimate (MLE) of β. Moreover, the *Gauss-Markov* theorem, that ensures the OLS estimator is the best linear unbiased estimator (BLUE), with the normality assumption further ensures that this estimator is the uniform minimum variance unbiased estimator (UMVUE). Plus, we have

$$\widehat{\beta}_{LS} | X \sim N(\beta, \Sigma_{\widehat{\beta}_{LS}})$$

where $\Sigma_{\widehat{\beta}_{LS}}$ can be unbiasedly estimated by $\widehat{\sigma}^2_{LS}(X^T X)^{-1}$.

The classic estimates of the mean response and the covariance have thus desirable statistical properties under the true model. The thorny issue is that we usually do not know the underlying model.

Having the estimates $\widehat{\beta}$, the predicted value of Y by that estimation is given by $\widehat{Y} = X\widehat{\beta}$. The difference $Y - \widehat{Y}$ is called *residual*. We define:

Definition 4.1 Being Y_i the ith observed value and $\widehat{Y}_i$ its predicted value via model (4.1), the residual of the ith observed value is defined defined as $Y_i - \widehat{Y}_i = r_i(\widehat{\beta})$.

With the residual notation, having $\varepsilon_i = r_i(\beta)$, we re-write equation (4.4) to

$$\widehat{\beta}_{LS} = \arg\min_{\beta} \sum_{i=1}^{n} r_i^2(\beta). \tag{4.5}$$

We must note at this point that, it is known that even a single corrupted observation is enough to disrupt the estimate $\widehat{\beta}_{LS}$. We will not go into detail here but, the largest fraction of data that can arbitrarily turned into bad values without perturbing the estimator greatly is known as the *breakdown point* of that estimator. More information on the subject can be found for example in Hampel (1968, 1971) and Donoho and Huber (1983).

4.3 M-estimators

M-estimators were introduced in regression by Huber (1973) and are a class of flexible estimators which played a central role in the development of robust statistics, the "M" standing for *maximum likelihood type*.

In the robust approach, the normality condition on the error distribution is relaxed to a quasi-normality condition and the estimators are obtained by methods other than ML. In the M-regression approach the estimates are the solutions to the following minimization problem,

$$\widehat{\beta}_R = \arg\min_{\beta} \sum_{i=1}^{n} \rho\Big(\frac{\varepsilon_i}{\widehat{\sigma}}\Big) = \arg\min_{\beta} \sum_{i=1}^{n} \rho\Big(\frac{r_i(\beta)}{\widehat{\sigma}}\Big) \tag{4.6}$$

where $\widehat{\sigma}$ is a robust scale estimate of σ and ρ is a function, also called *objective function*, with the following properties:

(i) ρ is even;

(ii) $\rho(0) = 0$;

(iii) $\rho(x)$ is increasing for $x > 0$ such that $\rho(x) < \rho(\infty)$;

(iv) if ρ is bounded then it is also assumed that $\rho(\infty) = 1$.

It is easy to verify that when $\rho(x) = x^2$ we have the LS/ML situation described above, since

$$\arg\min_{\beta} \sum_{i=1}^{n} \rho\Big(\frac{r_i(\beta)}{\widehat{\sigma}}\Big) = \arg\min_{\beta} \sum_{i=1}^{n} \frac{r_i^2(\beta)}{\widehat{\sigma}^2} = \arg\min_{\beta} \sum_{i=1}^{n} r_i^2(\beta) = \widehat{\beta}_{LS}.$$

The LS estimator is therefore an M-estimator.

There are several proposals in the literature for the robust scale parameter and a nice review can be found in Rousseeuw and Croux (1993) or more recently in Daszykowski *et al.* (2007). We have for example the re-scaled MAD (Hampel, 1974)

$$\text{MAD} = 1.4826 \times \text{median}|r_i - \text{median}(r_i)| \tag{4.7}$$

where the value of 1.4826 is to assure consistency at the gaussian distribution. This estimator although having a high breakdown point (50%) lacks efficiency at the normal model (37%). For this reason, among others, Croux and Rousseeuw (1992) proposed the estimator

$$Q_n = d \times \{|r_i - r_j| : i < j\}_{(k)} \tag{4.8}$$

where d is a constant and $k = \begin{pmatrix} h \\ 2 \end{pmatrix}$ with $h = \lfloor n/2 \rfloor + 1$, which is also a high breakdown estimator (50%) but now with 83% efficiency at the normal when we take $d = 2.2219$. Note that here (k) refers to the k-th order statistic of the $\begin{pmatrix} h \\ 2 \end{pmatrix}$ interpoint distances. Both robust scale measures can be calculated in the R language via instructions `mad()` and `Qn()` from packages *MASS* and *robustbase* (Maechler, 2006), respectively.

Differentiating (4.6) for every β_j, and equating to zero we get the p equations

$$\sum_{i=1}^{n} \psi\Big(\frac{r_i(\beta)}{\widehat{\sigma}}\Big) X_{ij} = 0, \;\; j = 0, \ldots, p-1 \tag{4.9}$$

where $\psi = \rho'$ and $X_{i0} = 1$, $i = 1, ..., n$. ψ is said to be redescending if it tends to zero at infinity, i.e,

$$\lim_{z \to \pm\infty} \psi(z) = 0.$$

In particular, if ρ is bounded, then ψ is redescending. Also, redescending ψ functions exclude observations with large residuals. One example is the Tukey biweight ψ function with tuning parameter b:

$$\psi(x) = \begin{cases} x(b^2 - x^2)^2 & \text{if } \; |x| \leq b \\ 0 & \text{if } |x| > b \end{cases} \quad . \tag{4.10}$$

Although (4.6) and (4.9) are not always equivalent, (4.9) is useful in the search of solutions to (4.6).

Estimates that are obtained through this process are regression, scale and affine equivariant (Rousseeuw and Leroy, 1987, pp. 116; Maronna *et al.*, 2006, pp.92), which are desirable properties. Solutions of (4.9) with monotone ψ are called monotone regression M-estimates and have the major advantage that all solutions to (4.9) are solutions of (4.6).

Considering the weights

$$W_i = \psi\left(\frac{r_i(\beta)}{\widehat{\sigma}}\right) \Big/ r_i(\beta), \quad i = 1, \ldots, n \tag{4.11}$$

the p equations (4.9) rewrite to

$$\sum_{i=1}^{n} W_i \times r_i(\beta) \times X_{ij} = 0,\ j = 0, \ldots, p-1 \tag{4.12}$$

which are the *weighted normal equations* that lead to $\widehat{\beta}_R = (X^T W X)^{-1} X^T W Y$ when the W_i's are known, where W is a diagonal matrix with elements W_i. When the W_i's are unknown (4.9) can be solved by Iteratively Reweighted Least Squares (IRWLS), which consists of the steps:

Step 0.

$$\begin{cases} \widehat{\beta}^{(0)} = \widehat{\beta}_{LS} \\ e_i^{(0)} = Y_i - \widehat{Y}_i^{(0)} = Y_i - X\widehat{\beta}^{(0)},\ i = 1, ..., n \\ \widehat{\sigma}^{(0)} = \text{MAD}(e_1^{(0)}, ..., e_n^{(0)}) \\ W^{(0)}:\ W_i^{(0)} = \dfrac{\psi\left(\frac{r_i(\widehat{\beta}^{(0)})}{\widehat{\sigma}^{(0)}}\right)}{r_i(\widehat{\beta}^{(0)})},\ i = 1, ..., n \end{cases} \tag{4.13}$$

Step k+1. With $k \geq 0$

$$\begin{cases} \widehat{\beta}^{(k+1)} = (X^T W^{(k)} X)^{-1} X^T W^{(k)} Y \\ e_i^{(k+1)} = Y_i - \widehat{Y}_i^{(k+1)} = Y_i - X\widehat{\beta}^{(k+1)},\ i = 1, ..., n \\ \widehat{\sigma}^{(k+1)} = \text{MAD}(e_1^{(k+1)}, ..., e_n^{(k+1)}) \\ W^{(k+1)}:\ W_i^{(k+1)} = \dfrac{\psi\left(\frac{r_i(\widehat{\beta}^{(k+1)})}{\widehat{\sigma}^{(k+1)}}\right)}{r_i(\widehat{\beta}^{(k+1)})},\ i = 1, ..., n \end{cases} . \tag{4.14}$$

The iterative procedure stops when a stoping criterion is attained, converging to the solution if ρ is convex[1].

Choosing a robust estimator within the class of M-estimators is not always an easy task. The objective function ρ proposed by Huber (1964) is known to lead to efficient estimators under ideal conditions. It is given by

$$\rho(x) = \begin{cases} x^2/2 & \text{if } |x| \leq b \\ b(|x| - b/2) & \text{if } |x| > b \end{cases} \tag{4.15}$$

[1]A real-valued function f(x) defined on an interval is called convex if for any two points x_1, x_2 of its domain, and any $t \in [0, 1]$ we have that $f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$.

where b is a tuning parameter. The correspondent ψ function is

$$\psi(x) = \begin{cases} -b & \text{if } x < -b \\ x & \text{if } |x| \leq b \\ b & \text{if } x > b \end{cases} . \tag{4.16}$$

Smaller values of b give more resistance to outliers in the data but at the expense of lower efficiency when the errors are normally distributed. Therefore, it is generally chosen as to give high efficiency at the normal. Huber (1964) suggested that the tuning parameter b should take the value of 1.5 if there are no outliers in the data and 1.345 if there are. In particular, a value of $b = 1.345$ produces 95% efficiency relative to the sample mean when the data is normal, and provides substantial resistance to outliers when it is not.

Efficiency and robustness however are inversely related and thus a good compromise between both should be considered.

Below, in Figure 4.1, we can see the plots of the Huber and LS ρ functions, as well as their correspondent ψ and weight functions. Although both objective functions ρ increase without bound as the residual departs from zero, the LS objective function increases faster than Huber's ρ.

Some other differences are quite evident from the plot. For instance, Huber's ψ function is limited whereas the LS is not and it additionally accounts for a normal model in a linear way. We know that in general a ψ function that has a linear behavior in the middle results in better efficiency at the normal distribution (Tukey, 1960). Also, Huber's weight function weights the observations differently giving less weight to more extreme observations, and actually weighting all observations within the bounds of $[-b, b]$ with weight 1, whereas the LS weight function corresponds to weighting all observations equally to 1.

Other choices are available in the literature like the already mentioned Tukey's biweight ψ function, but do not always guarantee a unique solution to (4.9) and one needs good initial estimates of the parameters to assure the convergence of the corresponding algorithm to the optimal solution. Using Huber's ρ, the resulting M-estimator of β is efficient (for both normal and non-normal data) and it is robust against outliers in the response variable, which is precisely the situation we may find in genetic association studies of quantitative traits. Library MASS in R has a model-fitting function `rlm()` in the

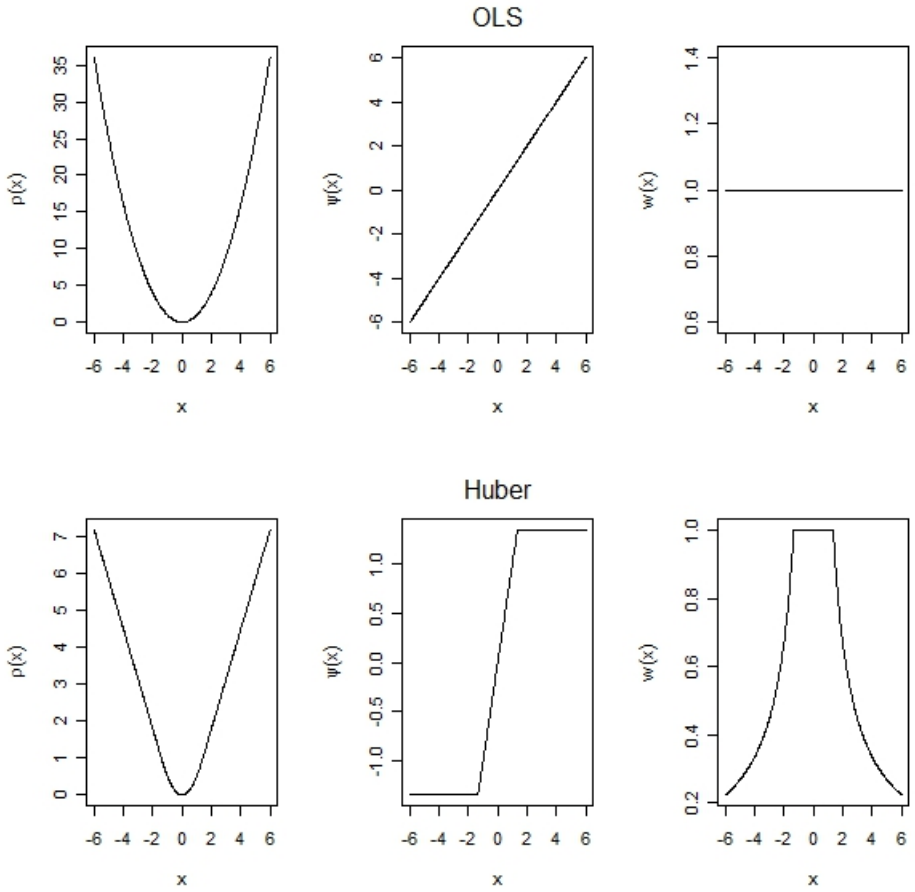


Figure 4.1: Huber and LS function plots; $b = 1.345$.

conditions described above. By default it uses ρ-Huber with tuning constant $b = 1.345$ and both β and the scale parameter σ are estimated by the IRWLS procedure with initial estimates of β and σ given by the LS estimate and the rescaled MAD, respectively.

When there are observations with extreme values in the predictor space (*leverage points*), M-regression is not robust and even shows no advantage over the least squares approach. In such situations, other alternatives exist, e.g, least trimmed squares (LTS) estimation, S-estimation and MM-estimation, among others, the latter attempting to retain the robustness and resistance of S-estimation while still keeping the efficiency of M-estimation (Rousseeuw and Leroy, 1987; Hubert and Rousseeuw, 1997; Maronna and Yohai, 2000; Ryan, 2008).

4.4 Testing the models parameters

The first step in data analysis is parameter estimation and we are usually interested in testing if a sub-vector of the vector of unknown parameters equals zero, i.e, if a subset of the predictor variables associate or not with the response variable.

4.4.1 The classical approach

A general linear hypothesis concerning β is of the form $H_0: A\beta = \mathbf{0}$, where A is a known $q \times p$ matrix with $q \leq p$. The general test for testing this hypothesis, which is derived from the LRT[2], is to reject H_0, at the level α, if $F \geq F_\alpha$, where $P_{H_0}(F \geq F_\alpha) = \alpha$ and

$$F = \frac{(Y - X\widehat{\widehat{\beta}})^T(Y - X\widehat{\widehat{\beta}}) - (Y - X\widehat{\beta})^T(Y - X\widehat{\beta})}{(Y - X\widehat{\beta})^T(Y - X\widehat{\beta})}. \tag{4.17}$$

$\widehat{\beta}$ and $\widehat{\widehat{\beta}}$ are the unrestricted and restricted MLE of β, respectively. We also know that,

$$\frac{n-p}{q}F \underset{H_0}{\sim} F_{q,n-p} \underset{a}{\sim} \chi^2_{(q)} \quad \text{and} \quad \frac{n-p}{q}F \underset{H_1}{\sim} F_{q,n-p,\delta} \tag{4.18}$$

where δ is a non-centrality parameter (see Rohatgi and Saleh, 2001, for further detail).

In the context of genetic association studies of quantitative traits, we are primarily interested in the following two testing situations:

(i) $H_0: \beta_1 = \cdots = \beta_{p-1} = 0 \equiv A\beta = \mathbf{0}$ where $A = [\mathbf{0}\ I_{p-1}]$, referring to the global test of association, i.e,

$H_0:$ there is no association between the SNPs in the model and the quantitative trait Y

(ii) $\{H_{0k}: \beta_k = 0\}_{k=1,\ldots,p-1} \equiv \{A\beta = 0\}$ where now A is a $1 \times p$ vector of nulls with 1 at position $k+1$ referring to individual tests of association, i.e, to the individual tests

$H_{0i}:$ there is no association between SNP_i and the quantitative trait Y

In (i), the number of parameters to be tested is $q = p - 1$. In this case we have $\widehat{\widehat{\beta}}_1 = \cdots = \widehat{\widehat{\beta}}_{p-1} = 0$ and $\widehat{\widehat{\beta}}_0 = \overline{Y}$, thus the F statistic of (4.17) rewrites to

$$F = \frac{\sum_{i=1}^n (Y_i - \overline{Y})^2}{\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2} - 1 = \frac{SST}{SSE} - 1 \tag{4.19}$$

[2] And therefore assumes the normality of the errors.

and, under H_0, from (4.18) we have $(n-p)F/(p-1) \sim F_{p-1,n-p} \underset{a}{\sim} \chi^2_{(p-1)}$.

In (ii), for each H_{0k} test, we have $q = 1$ and so, again under H_0 and from (4.18), $(n-p)F \sim F_{1,n-p}$. Furthermore, we know that $\sqrt{(n-p)F} \underset{H_0}{\sim} |t_{n-p}|$ and thus, taking

$$T = \frac{\widehat{\beta}_k}{\sqrt{\text{var}(\widehat{\beta}_k)}} = \frac{\widehat{\beta}_k}{\sqrt{(\text{var}(\widehat{\beta}))_{kk}}} = \frac{\widehat{\beta}_k}{\sqrt{(\sigma^2(X^TX)^{-1})_{kk}}} \tag{4.20}$$

we therefore reject H_{0k} at significance level α if $|T| \geq t_{1-\alpha/2,n-p}$.

In practice, in the simulation study of Chapter 5, for each bi-allelic SNP, there will be two dummy variables in the regression model, that is, $q = 2$ and therefore, under H_0, $(n-p)F/2 \sim F_{2,n-p} \underset{a}{\sim} \chi^2_{(2)}$.

The classic multiple linear regression model is available in R via instruction `lm()` of package *stats*. The global tests of association (4.19) are provided in the summary table of the `lm()` adjusted model. The individual tests (4.20), are available in the `Anova()` table from package *car* and in the `anova()` table from package *stats*. However, we must underline the fact that in the simulation study of Chapter 5, the design is unbalanced (Shaw, 1993), thus the use of instruction `Anova()` of package *car* and not `anova()` from package *stats*. We must also add that we used the default type-II tests though type-III are also available. A nice discussion on the differences between these two types of tests can be found in Lansgrud (2003).

4.4.2 The robust approach

As already mentioned in the introduction, contrary to some statements in the literature dating back to Box (1953), the ANOVA F-test is not robust against non-normality (Schrader and Hettmansperger, 1980; Ronchetti, 1987; de Haan *et al.*, 2009). What happens is that the use of the word robust in the literature is usually associated with a good control of the type-I error rather than with the robustness property of the method (see for example Taplin, 1999). For this reason we have considered for the robust tests for the general linear hypothesis, taking $\gamma = A\beta$, the robust Wald-type statistic (see Maronna, 2006, pp. 107)

$$T_W = \frac{\widehat{\gamma}^T \widehat{\Sigma}_\gamma^{-1} \widehat{\gamma}}{q} \underset{H_0}{\sim} F_{q,n-p} \tag{4.21}$$

where $\widehat{\Sigma}_\gamma$ is an estimate of the covariance matrix of γ. Since $\widehat{\Sigma}_{\beta_R} = \widehat{\upsilon}(X^TX)^{-1}$, with $\widehat{\upsilon} = \widehat{\sigma}^2 \frac{\text{ave}_i(\psi(r_i/\widehat{\sigma})^2)}{(\text{ave}_i(\psi'(r_i/\widehat{\sigma})))^2} \times \frac{n}{n-p}$, it follows that $\widehat{\Sigma}_\gamma = \widehat{\upsilon} A(X^TX)^{-1}A^T$.

For the global test for association, whether the predictor variables in the model are continuous or categorical, we can obtain the test p-value via the R function `f.robftest()` of package *sfsmisc*. However, if the predictor variables are continuous or categorical with up to two categories (thus one dummy variable in the regression model), to perform the individual association tests we have to recode function `f.robftest()` with due care to the unbalanced data design. Again, in the simulation study (Chapter 5) the T_W test statistic will never simplify to the case where $q = 1$ since the simulated SNPs are three-category variables and thus there will be two estimated dummies per covariate. Nevertheless, in the real data set study (Chapter 7), we had a couple of SNPs with only two categories and so the situation where $q = 1$ did happen and may therefore happen as well in other studies.

4.5 Coefficient of determination

The coefficient of determination is a statistical tool that is commonly used as a summary measure of predictive power of the selected regression model. It is defined as the percentage of variance of the response variable that is explained by the regression model. More precisely,

$$R^2 = 1 - \frac{\sigma^2}{\sigma_Y^2}. \tag{4.22}$$

Thus, the larger the R^2, the larger the proportion of variation of Y that is explained by the explanatory variables. When $R^2 = 1$ we say that the regression model perfectly fits the data.

In the classical approach, the coefficient of determination is estimated as

$$R^2 = 1 - \frac{\sum_{i=1}^{n} r_i(\widehat{\beta}_{LS})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \tag{4.23}$$

where $r_i(\widehat{\beta}_{LS})$ are the residuals obtained by the LS fit. There are however several other estimates of R^2 in the literature and a good review can be found in Kvalseth (1985). One major drawback of (4.23) is that this value increases as the number of covariates in the

model increase[3] and it is therefore usual to use the adjusted version

$$R_a^2 = 1 - R^2 \frac{n-1}{n-p} \tag{4.24}$$

which may decrease if we enter variables that do not add significantly to the model.

As the LS estimator, this measure is vulnerable to data contamination, i.e, the presence of outliers. An example of how drastically it may decrease when a single outlier is introduced in the data set can be seen e.g, in Croux and Dehon (2003). Several robust proposals have been placed in the literature (Roussseuw, 1984; Anderson-Sprecher, 1994; Croux and Dehon, 2003; Maronna *et al.*, 2006; Renaud and Vitoria-Feser, 2010). The parameters σ^2 and σ_Y^2 may be robustly estimated, for instance, taking $\widehat{\sigma}^2 = Q_n(r_i(\beta_R))^2$ and $\widehat{\sigma_Y}^2 = Q_n(Y)^2$ and then plugged-in (4.22). The estimate of R^2 thus obtained, can be adjusted in the same way as the classical one but using the weights sum derived from the robust analysis instead of the number of individuals n used in the regression model.

4.6 Outlier detection

Statistics obtained by fitting a regression model can be highly influenced by a small set of the data that is different from the bulk of the data. In particular, extremely deviant outliers reduce the power of parametric tests[4] (Barnett and Lewis, 1978; Hampel *et al.*, 1986).

The identification of outlying observations is thus of the utmost importance and so is the assessment of their effects in the analysis of the data. Therefore, it is natural to look for such observations after the model has been adjusted and there are several diagnostic measures available in the literature that one can use to do it. Some of them were planned to identify individual observations, and others to detect groups of individuals. It may actually happen that an observation is not influential in itself but a group of observations of which it is part of is (Rousseeuw and Leroy, 1987).

In this thesis we are mainly interested in regression outliers since that will be the situation encountered in genetic association studies of quantitative traits where the response variable is continuous, and thus prone to outlying observations - the explanatory variables are categorical and therefore there will be no outlying observations to this respect.

[3] It is known to reach its maximum value of 1 for any saturated model even when the predictors and outcome are independent of each other.

[4] For a note on the influence of outliers on non-parametric tests please see Zimmerman (1994).

In this section we will briefly review some of the usual methodology of outlier identification. We would like to underline though, that outlier pinpointing is not the end of the line in what concerns outliers, but rather the beginning of something far more complex as some pertinent questions then arise:

i. What should we do with outliers? Should we ignore them? Discard them? Incorporate them in the model?

ii. Why are these points different from the bulk of the data? Are they measurement errors? Do they represent rare events?

iii. How do these points affect the conclusions of our analysis? Are they influent?

One thing is for sure though, no observation should ever be removed from the data on statistical grounds alone! Here, statistical grounds include, for example, outlier testing.

We start by defining the hat matrix, which plays a major role in the identification of outliers:

Definition 4.2 We define the hat-matrix as $H = X(X^tX)^{-1}X^T$ and h_{ii} as the i-th element of the diagonal of H.

The hat-matrix verifies $H = H^T$ (H is symmetric), $H = H^2$ (H is idempotent) and we have $tr(H) = p$. More detailed properties of H can be found for example in Hoaglin and Welsch (1978) and Cook and Weisberg (1982). This matrix plays as well an important role in the identification of leverage points. More precisely, the diagonal elements h_{ii} of the hat matrix can reveal the leverage points in the data. When h_{ii} is near 1, the variance of the i-th residual is near 0 which means that observation i has a large influence in the regression estimates. More precisely, if $h_{ii} \geq 2p/n$ then observation i should be considered a leverage point. In simple regression hat values h_{ii} measure distance from the mean of X, whereas in multiple linear regression h_{ii} measures the distance from the centroid point of all of the X's (mean vector).

We write a few results in terms of the hat-matrix, which can easily be derived and that will be handy ahead (see Myers *et al.*, 2001, for more detail):

i. $\widehat{Y} = X\widehat{\beta}_{LS} = HY$; H is sometimes called in the literature as *prediction matrix* because of this equality; usually we know it by *projection matrix* since it projects the Y values onto their predicted values;

ii. $\text{var}(\widehat{Y}) = \sigma^2 H$;

iii. $\widehat{\varepsilon} = r(\widehat{\beta}_{LS}) = Y - \widehat{Y} = (I - H)Y$ where $I - H$ is called the residual matrix;

iv. $\text{var}(\widehat{\varepsilon}) = \text{var}(r(\widehat{\beta}_{LS})) = \sigma^2(I - H)$ and thus, the variance of the i-th residual is $\text{var}(\widehat{\varepsilon}_i) = \text{var}(r_i(\widehat{\beta}_{LS})) = \sigma^2(1 - h_{ii})$, which is estimated by $MSE(1 - h_{ii})$;

v. $\text{cov}(\widehat{\varepsilon}_i, \widehat{\varepsilon}_j) = -h_{ij}\sigma^2$ for $i \neq j$, which is estimated by $-h_{ij} \times MSE$.

Since regression outliers are obviously not contemplated by the hat matrix, they have to be acknowledged separately. The analysis of the least squares residuals is one common way of detecting model failures. In multiple regression, an outlier can thus be defined as an observation with large residuals. As a general rule of thumb, 95% of the standardized residuals should lie between -2 and 2. However, this approach could be misleading since outliers' residuals often look like normal ones, e.g., when an outlier is also a high-leverage point, it pulls the fitted least squares line towards it and the resulting residual often does not indicate that it is actually an outlier. Plus, there are two additional problems in outlier diagnostic methods through residual analysis: the masking effect problem, where outliers mask the effect of others leaving those undetected, and the swamping effect, when regular observations are flagged as outliers when they are not.

In (iii) above, we can see that the LS estimated residuals $\widehat{\varepsilon} = r(\widehat{\beta}_{LS})$ and the related unknown true residuals $\varepsilon = r(\beta)$ are typically correlated since it can be showed that $(I - H)Y = (I - H)\varepsilon$ and have different variances (iv). For diagnostic purposes this variances should be equalized. As such, several transformations of the residuals have been suggested and the hat matrix can be used to easily studentize the regression residuals:

i. *Studentized residuals*

$$r_i^* = \frac{\widehat{\varepsilon}_i}{\sqrt{\sigma^2(1 - h_{ii})}} \tag{4.25}$$

where $E(r_i^*) = 0$ and $\text{var}(r_i^*) = 1$.

ii. *Studentized deleted residuals*

$$r_i^\circ = \frac{\widehat{\varepsilon}_i}{\sqrt{\widehat{\sigma}^2_{(-i)}(1 - h_{ii})}} \tag{4.26}$$

where $\widehat{\sigma}_{(-i)}$ is the residual mean squares estimate of σ when the i-th observation is omitted from the data, i.e, $\widehat{\sigma}^2_{(-i)} = MSE_{(-i)}$. Studentized residuals, have mean near 0 and a variance slightly larger than 1. They also follow a t-distribution with $n - 1 - p$ degrees of freedom[5].

[5]Note that we are always assuming $\varepsilon_i \underset{i.i.d}{\sim} N(0, \sigma^2)$.

In large data sets, the values of the standardized and studentized residuals should not differ greatly. When compared with standardized residuals, studentized residuals have the particularity of overcoming the problem of a single large outlier inflating $\widehat{\sigma}$ and therefore masking itself.

The value $MSE_{(-i)}$ that appears in the studentized deleted residuals is also known in the literature as the external MSE. To calculate its value, one needs not to run the regressions over and over again, dropping observation i, $i = 1, ..., n$, in each turn. Instead we can use the hat matrix to estimate $MSE_{(-i)}$ as follows:

$$\widehat{\sigma}^2_{(-i)} = \frac{(n-p)\widehat{\sigma}^2 - \dfrac{\widehat{\varepsilon}_i^2}{1-h_{ii}}}{n-1-p} \ . \tag{4.27}$$

With a little algebra this equation rewrites to

$$\widehat{\sigma}^2_{(-i)} = \frac{1}{n-1-p}\left(\widehat{\varepsilon}^T\widehat{\varepsilon} - \frac{\widehat{\varepsilon}_i^2}{1-h_{ii}}\right) \tag{4.28}$$

which is just the SSE subtracted by a correction element and divided by the degrees of freedom. We can as well acknowledge the important role of the hat matrix in determining the magnitude of a studentized residual and hence in the identification of outlying Y observations.

Observations that have a studentized residual outside the ± 2 standard deviations range are considered statistically significant at the 95% level, i.e, are flagged as outliers. We have reasons to be even more concerned and therefore alert, with observations whose residuals exceed ± 3 standard deviations.

Another way to assess if any of the n observations are outliers or not, is to perform a test using the studentized deleted residuals statistic r_i° for every observation, adjusting for multiple testing via, e.g., Bonferroni correction (see Netter *et al.*, 1996, for more detail). The hypothesis

$$H_{0i} : \text{observation } i \text{ is not an outlier}$$

will be rejected if $p_i^* < \alpha$, where p_i^* is the i-th adjusted p-value and α a pre-specified threshold. The Bonferroni critical value for the test is $t_{(1-\alpha/2;n-1-p)}$. The R function `outlier.test()` from the *car* package gives the Bonferroni corrected p-value for the largest absolute studentized residual. Instruction `ls.diag()` from package *stats* provides the studentized residuals. With these values the tests can easily be constructed and p-

values corrected via Bonferroni with instruction `mt.rawp2adjp()` from the R package *multtest*.

At this point we must bring to the reader's attention the fact that this correction may be too conservative if n is large, in which case other alternatives should be considered (see Chapter 3 for more detail).

Another technique would be to plot the studentized residuals from the regression against the t-distibution. Observations straying outside the 95% confidence interval should be flagged out as outliers. This can be done with instruction `qq.plot()` from the R package *car*. Graphical methods are important tools in outlier detection, since the patterns of residuals are often more important than their magnitudes (Atkinson, 1985). Plotting the fitted values versus standardized residuals is also an usual approach to regression outliers identification.

Having an outlier identified does not mean that it necessarily influences the regression line. However it will tend to increase the standard error, and this is why in particular it messes up the F test. Be aware that, as a rule of thumb, if more than 20% of the observations are identified as outliers, one should begin questioning oneself about the distribution assumed by the model, the model itself and/or about the quality of the data collected.

Next to identifying outliers, the following natural step is to ascertain whether an outlier is influential or not. There are three popular measures of influence, each one based on the omission of a single observation to measure its influence. They are briefly described below.

Cook's distance (Cook, 1977) is a diagnostic method that can be used to identify influent points. It measures the influence of the i-th observation on all n fitted values of the outcome. One calculates the distance:

$$D_i = \frac{(\widehat{Y} - \widehat{Y}_{(-i)})^T(\widehat{Y} - \widehat{Y}_{(-i)})}{p \times MSE} = \frac{r_i^{\circ 2}}{p}\frac{h_{ii}}{1-h_{ii}} = \frac{\widehat{\varepsilon}_i^2}{p \times MSE_{(-i)}}\frac{h_{ii}}{(1-h_{ii})^2} \tag{4.29}$$

where $\widehat{Y}$ is the predicted value of Y from the whole model, i.e, the model with all n observations, and $\widehat{Y}_{(-i)}$ is the predicted value of Y based on the parameters estimated after deleting observation i from the data. The last equality shows the Cook's distance D_i to depend on the size of the residual $\widehat{\varepsilon}_i$ and the leverage value h_{ii}. The larger either one

of these values is, the larger D_i is. In this way, an observation i is considered influential if:

i. it has a large residual $\widehat{\varepsilon}_i$ and a moderate leverage h_{ii};

ii. it has a large leverage h_{ii} and a moderately sized residual $\widehat{\varepsilon}_i$;

iii. it has both large residual $\widehat{\varepsilon}_i$ and leverage value h_{ii}.

One rule of thumb is to consider point i influential when $D_i > \dfrac{4}{n-p}$. However, since under the null the D_i's are distributed as an $F_{n,n-p}$, another possibility is to reject for values of $D_i > F_{0.5,n,n-p}$. The R library *MASS* as an instruction `cooks.distance()` that allows for the calculation of the D_i's once a regression model has been adjusted. It can also be done via instruction `ls.diag()` from package *stats*, which, among other statistics, gives Cook's distances, or with `cookd()` from package *car*. Inspecting the plot of Cook's distances versus observations index is also a useful tool of analysis. Another useful plot is the influence plot where hat values are plotted against the studentized residuals and where there are circles plotted as well for each observation representing the relative size of Cook's D. This can be done in R via instruction `influenceplot()` from package *car*.

Contrary to Cook's distance, DFFITS (difference in fit standardized) distance, a popular influence statistic, measures the influence of the i-th observation solely on the fitted value $\widehat{Y}_i$:

$$\text{DFFITS}_i = \frac{\widehat{Y}_i - \widehat{Y}_{i(-i)}}{\sqrt{MSE_{(-i)}h_{ii}}} \tag{4.30}$$

where again, $\widehat{Y}_i$ is the predicted value of Y_i from the whole model and $\widehat{Y}_{i(-i)}$ is the predicted value of Y_i based on the parameters estimated after deleting observation i from the data.

If DFFITS_i is large, it means that the i-th observation is influential on the model's predicted value for the i-th observation meaning the model does not fit observation i well. In general, any observation i with $|\text{DFFITS}_i| > 2\sqrt{p/n}$ is set aside for further investigation, but only if n is large. For small values of n, the cut-off is $|\text{DFFITS}_i| > 1$.

The two methods discussed above measure the influence of an observation in the fitted values of the regression. To assess the influence of such an observation on the coefficients' estimates we have the DFBETAS distance, which measures how much the regression

coefficients change when an observation is omitted:

$$\text{DFBETAS}_{k,i} = \frac{\widehat{\beta}_k - \widehat{\beta}_{k(-i)}}{\sqrt{((X^TX)^{-1})_{kk} MSE_{(-i)}}}, \tag{4.31}$$

where $\widehat{\beta}_k$ is the estimate of parameter β_k obtained from the whole model, i.e, the model with all n observations, and $\widehat{\beta}_{k(-i)}$ is the estimate of β_k obtained when observation i is deleted from the data. If $\text{DFBETAS}_{k,i}$ is positive then the effect of the i-th observation is to increase the estimate of β_k, whilst when it is negative it has the contrary effect.

A large value of $|\text{DFBETAS}_{k,i}|$ indicates that observation i has a large impact on the k-th regression coefficient. In particular, for large data sets, an observation i is considered influential in the estimation of parameter β_k if $|\text{DFBETAS}_{k,i}| > 2/\sqrt{n}$ (Belsley *et al.*, 1980). For small n we should look for values of $|\text{DFBETAS}_{k,i}| > 1$.

If after investigating whether an outlier is or not an influential observation we conclude negatively, then we have no reason to worry about that observation. At least, in what concerns the regression estimates. As to the tests of association, that is totally a different matter, since we have already pointed out that this observation will tend to increase MSE thus messing up the F test. The simulation study in Chapter 5 will show how bad the F test may become in the presence of contamination, while giving evidence that the alternative proposed may be more adequate for the association studies of quantitative traits.

Robust estimation and outlier detection are intimately related. McKean *et al.* (1993) showed that with monotone ψ, M-regression residual plots can be interpreted the same way as LS residual plots. That will be the case when we take Huber's ψ function. In this way, given such robust estimates of the true values of an assumed null model, analyzing the robust residuals[6] plotted against the fitted values, can indicate which, if any, of the observations are outliers.

Robust methods are particularly important in situations where there are multiple outliers whose presence is masked when diagnostic methods that delete one observation at a time are used. In M-regression (with monotone ψ), observations with final small weights are usually outlying. Another important aspect is that robust regression may be a useful tool for confirming the reasonability of OLS results, e.g., when the residuals are similar, one is reassured that OLS is not being influenced by outlying observations. Jajo

[6] I.e, the residuals obtained from a robust fit.

(2005) makes a succinct review of robust regression and diagnostic measures in linear regression.

Rice and Spiegelhalter (2006) proposed a simple outlier diagnostic plot assembling M-regression, residual plots and false discovery rates, that will be briefly described below.

To test the hypothesis H_{0i} : observation i is not an outlier proceed as follows:

i. take the standardized robust residuals

$$r_i^R = \widehat{\varepsilon}_i^R / \widehat{\sigma}_R \tag{4.32}$$

where $\widehat{\varepsilon}_i^R$ are the robust residuals and $\widehat{\sigma}_R$ a robust estimate of scale (e.g, equations (4.7) and (4.8));

ii. calculate $p_i = 2 \times \Phi\Big(- |r_i^R|\Big)$ for each observation $i = 1, ..., n$;

iii. adjust for multiple testing with the BH-1995 FDR (Chapter 3, Algorithm 11), i.e, consider $q_i = n \times p_{(i)}/i$

iv. reject the null hypothesis at a given FDR threshold, e.g, 10%.

They showed through simulation for various levels of the FDR (more precisely, 0.01, 0.02 and 0.1) and different number of individuals (namely, $50, 100, 150, ..., 1000$) that, when the data is contaminated by a symmetrical distribution, the procedure works well with an excellent agreement between sample and nominal FDR levels. When an asymmetric contaminant distribution was considered, the level of FDR control was not as accurate as in the previous setting. However it did not amount to much concern since the FDR level was never exceeded by more than 1%. We must underline that in this study the regression model considered only the intercept.

Next, to better illustrate the performance of the proposed method, they presented an example of application with the "phones" data from Rousseeuw and Leroy (1987), which is available in the R library *MASS*. This data set contains information on the total number of calls, in millions, made in Belgium from the year of 1950 to 1973 (24 observations in total). It is also known to have misrecorded number of calls between the years of 1964 to 1969. Here, the residuals used in the outlier tests were based on the regression of *calls* on *years*.

We have replicated this study and present the diagnost plot of Rice and Spiegelhalter (the R code is available in routine C.1) together with the correspondent plot that would be obtained from the classical residual analysis (using r_i^*):

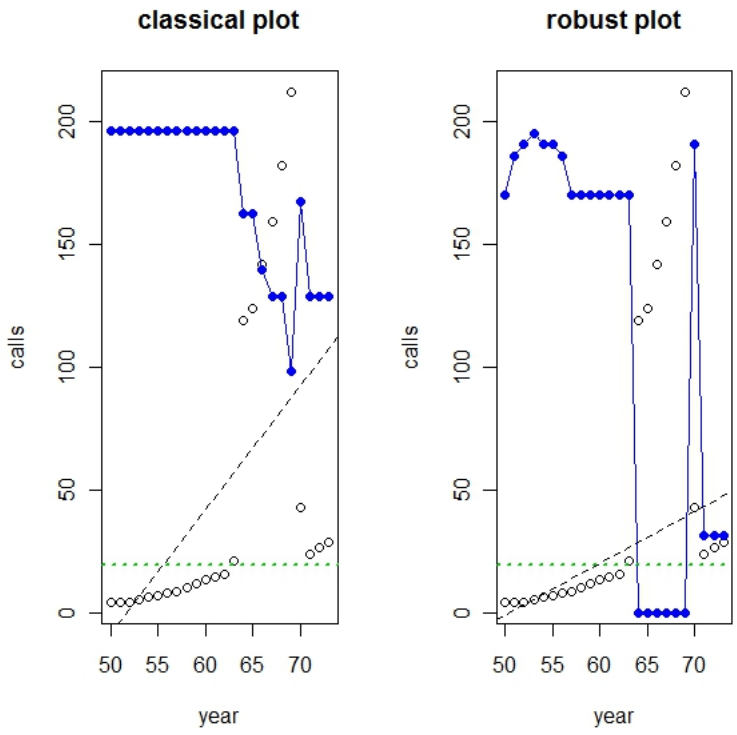


Figure 4.2: Plot of *year* versus *calls* (black dots), with q-value rescaled plot superimposed (blue dots), plus the regression line (black dashed line) and FDR 10% rescaled threshold (green dashed line)

Clearly, the robust methodology identifies the 6, known to be, outliers whereas the classical approach does not identify any, and this diagnostic plot is very useful in observing just that. However, the use of it is only viable in situations where we have a single explanatory variable. We must also add that in this particular case, the FDR threshold level of 10% used could and should be lowered to at least 1% without fear of losing significance of the 6 identified outliers. In general, to avoid the known swamping effect, the FDR level should be as low as possible.

4.7 Estimation and tests in the mixed linear model setup

The general linear mixed model can be written as

$$\mathbf{y} = X\beta + Z\mathbf{u} + \epsilon, \tag{4.33}$$

where $\mathbf{y} \in \mathcal{M}_{n\times 1}$ is the vector of observed values, $X \in \mathcal{M}_{n\times p}$ and $Z \in \mathcal{M}_{n\times n}$ are the incidence matrices for the fixed and random effects, respectively, $\beta \in \mathcal{M}_{p\times 1}$ and $\mathbf{u} \in \mathcal{M}_{n\times 1}$ are the correspondent fixed and random effects with $\mathbf{u} \sim N(\mathbf{0}, G)$ and $\epsilon \in \mathcal{M}_{n\times 1}$ is a vector of residual deviations with $\epsilon \sim N(\mathbf{0}, R)$. Both $\mathbf{u}$ and ϵ are assumed to be independent vectors and so $\text{cov}(\mathbf{u}, \epsilon) = \mathbf{0}$ and $\text{var}(\mathbf{y}) = ZGZ' + R = \Phi \in \mathcal{M}_{n\times n}$.

In a genetics association context, $\mathbf{y}$ is the phenotype for each of the n individuals, X the individual genotype and $\mathbf{u}$ usually quantifies the genetic effects.

For the model described we have the following likelihood function

$$L(\beta, \sigma^2|\mathbf{y}) = (2\pi)^{-n/2}|\Phi|^{-1/2} \exp\Big[-\frac{1}{2}(\mathbf{y} - X\beta)'\Phi^{-1}(\mathbf{y} - X\beta)\Big] \tag{4.34}$$

and the correspondent log-likelihood

$$l(\beta, \sigma^2|\mathbf{y}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Phi| - \frac{1}{2}(\mathbf{y} - X\beta)'\Phi^{-1}(\mathbf{y} - X\beta) \tag{4.35}$$

where $\sigma^2 = [\sigma^2_{ij}]_{i,j=1,\dots,n}$ are parameters of Φ.

We are interested in studying the model where $G = \sigma^2_A A$ and $R = \sigma^2_E I$, where A is a relationship matrix between the n individuals, e.g., a kinship matrix in a genetics context, also referred to as an *identity-by-descent* (IBD) matrix. Moreover, we consider that we only have one observation per individual, i.e, $Z = I$ and that Φ is a positive definite-matrix. Therefore the model (4.33) rewrites to

$$\mathbf{y} = X\beta + I\mathbf{u} + \epsilon \tag{4.36}$$

and we have the following derivation for the variance of $\mathbf{y}$

$$\text{var}(\mathbf{y}) = \sigma^2_A A + \sigma^2_E I = \Phi = \sigma^2_E\Big(\frac{\sigma^2_A}{\sigma^2_E}A + I\Big) = \sigma^2_E\Big(\sigma^2_* A + I\Big) = \sigma^2_E \Phi^*.$$

The vector of parameters of Φ thus becomes $\sigma^2 = (\sigma_*^2, \sigma_E^2)$.

If Φ^* is known, the generalized least squares estimate of β, that is a best linear unbiased estimate (BLUE) of β, and its covariance are given by

$$\hat{\beta} = (X'\Phi^{-1}X)^{-1}X'\Phi^{-1}\mathbf{y} = (X'\Phi^{*^{-1}}X)^{-1}X'\Phi^{*^{-1}}\mathbf{y} \tag{4.37}$$

and

$$\text{var}(\hat{\beta}) = (X'\Phi^{-1}X)^{-1} = \hat{\sigma}_E^2(X'\Phi^{*^{-1}}X)^{-1} = \hat{\Sigma} \quad , \tag{4.38}$$

the best linear unbiased predictor (BLUP) of $\mathbf{u}$ is

$$\hat{\mathbf{u}} = \hat{\sigma}_E^2 GZ'\Phi^{*^{-1}}(\mathbf{y} - X\hat{\beta}) = \hat{\sigma}_E^6\sigma_*^2 A\Phi^{*^{-1}}(\mathbf{y} - X\hat{\beta}) \tag{4.39}$$

where

$$\hat{\sigma}_E^2 = \frac{1}{n}(\mathbf{y} - X\hat{\beta})'\Phi^{*^{-1}}(\mathbf{y} - X\hat{\beta}). \tag{4.40}$$

Substituting $\hat{\beta}$ and $\hat{\sigma}_E^2$ in (4.35) we get the profile log-likelihood function

$$l_P(\sigma_*^2|\beta = \hat{\beta}, \sigma_E^2 = \hat{\sigma}_E^2, \mathbf{y}) = -\frac{n}{2}\Big(\log(2\pi) + \log(\hat{\sigma}_E^2) + 1\Big) - \frac{1}{2}\log|\Phi^*| \tag{4.41}$$

$$= \text{constant} - \frac{n}{2}\log(\hat{\sigma}_E^2) - \frac{1}{2}\log|\Phi^*| = l_P(\sigma_*^2, \hat{\sigma}_E^2). \tag{4.42}$$

Being Φ^* a semi-definite positive matrix, taking its Cholesky decomposition, model (4.36) rewrites to

$$Y^* = X^*\beta + \varepsilon^* \tag{4.43}$$

where $Y^* = \text{chol}(\Phi^*)^{-1}Y$, $X^* = \text{chol}(\Phi^*)^{-1}X$, $\varepsilon^* = \text{chol}(\Phi^*)^{-1}\varepsilon$. Moreover, now we have $\text{var}(\varepsilon^*) = \sigma^2 I$ and we are thus in the multiple linear regression setup described in Section 4.1 allowing us to easily obtain the LS and robust estimates needed.

When Φ^* is unknown, an estimate $\hat{\Phi}^*$ of the covariance matrix Φ^* has to be considered. It is usually obtained through the ML or REML methodology, although there is some criticism around the former since the ML estimators of the variance components do not account for the loss in the degrees of freedom in the calculation of β and are therefore biased downward. REML, on the other hand, corrects for this. For an insight on this approach we refer to the works of Harville (1974), Searl (1979), Searl *et al.* (1992), Netter *et al.* (1996) and LaMotte (2007), and in a genetic context to Meyer (1983/1985) and Lynch and Walsh (1997). As to the ML approach, we will briefly describe below a

derivative-free methodology for the estimation of Φ^*, which is easy to understand and quite straightforward.

4.7.1 Using maximum likelihood for the estimation of Φ^*

A derivative-free approach for the estimation of Φ^*, i.e, the estimation of σ_*^2, consists on taking the profile log-likelihood function (4.42) and an initial value for σ_*^2, say $\sigma_*^{2^{(0)}}$. For this value $\sigma_*^{2^{(0)}}$, we calculate $\sigma_E^{2\,(0)}$ from (4.43) and evaluate $l_P(\sigma_*^2 = \sigma_*^{2^{(0)}}, \widehat{\sigma}_E^2 = \sigma_E^{2\,(0)})$ in (4.42). We iterate in the same way for other possible values of σ_*^2 until (4.42) is maximized. With the final value of σ_*^2 obtained from the iterative process, say $\sigma_*^{2^{(k)}}$ for some k, we calculate Φ^* and proceed as in model (4.43) to obtain the estimates of β and σ_E^2.

Another possibility would be to write the log-likelihood equations and use some numerical method, like the Newton-Raphson (NR) or the expectation-maximization (EM) algorithms, to calculate an approximation of their solution.

The function `lmekin()` from the R package *kinship* takes the former approach, using the `optim()` function to minimize minus the log-likelihood (4.42).

In order to obtain an estimate of Φ^* for the robust approach, we suggest plugging-in the robust estimates in (4.42) and proceed in analogous way. We have recoded the R function `lmekin()` as to do so and named it `lmekinR()`. Its code can be found in the Appendix C, Section C.3.

4.7.2 Hypothesis testing

Inference for the fixed-effects terms, i.e, the tests $H_0 : L\beta = \mathbf{0}$ of $q \leq p$ fixed-effects are conducted by the F statistic

$$F = \frac{\hat{\beta}' L' (L\hat{C}L')^{-1} L\hat{\beta}}{q} \tag{4.44}$$

where

$$\hat{C} = (X'\hat{\Phi}^{-1}X)^- = (X'(\hat{\sigma}_E^2\hat{\Phi}^*)^{-1}X)^- = \hat{\sigma}_E^2(X'\hat{\Phi}^{*^{-1}}X)^- \tag{4.45}$$

pointing out that we now consider instead of (4.40), the centered estimate

$$\hat{\sigma}_E^2 = \frac{1}{n - \text{rank}(X)}(\mathbf{y} - X\hat{\beta})'\Phi^{*^{-1}}(\mathbf{y} - X\hat{\beta}). \tag{4.46}$$

Under the null[7], the sampling distribution of F is asymptotically distributed as an

[7]Where we consider as well the normal distribution of the errors.

$F_{q,\nu}$ where ν has to be appropriately estimated. There are several proposed methods in literature to the estimation of ν. E.g., the Satterthwaite approximation, the Kenward-Roger (KR) method[8] (available in SAS) and others. When we are dealing with unbalanced data though, these two methods are usually preferred over the remaining.

A full theoretical derivation of the KR approximation can be found in Alnosaier (2007) and evidence that the KR method outperforms the Satterthwaite at the nominal type-I error rate control can be found in Spike *et al.* (2005) and Schaalje *et al.* (2001).

The Satterthwaite approximation for ν is obtained as follows:

i. Compute the spectral decomposition of $L\hat{C}L'$, where $P \in \mathcal{M}_{q\times q}$ is an orthogonal matrix of $L\hat{C}L'$ eigenvectors and $D \in \mathcal{M}_{q\times q}$ a diagonal matrix of $L\hat{C}L'$ eigenvalues.

ii. Let $l_m = PL[m,]$, i.e, the mth row of PL.

iii. Compute the variance/covariance matrix of the unknown parameters in Φ, $\sigma^2 = (\sigma_E^2, \sigma_*^2)$, $B_{ij} = (S^{-1})_{ij}$ where $S_{ij} = \frac{1}{2}tr(\Phi^{-1}\Phi_i\Phi^{-1}\Phi_j)$ and

$$\Phi_i = \begin{cases} I & \text{when } \sigma_i^2 = \sigma_E^2 \\ A & \text{otherwise} \end{cases} \tag{4.47}$$

iv. Let $\nu_m = \dfrac{2(D_m)^2}{g_m' B g_m}$ where D_m is the mth diagonal element of D and g_m the gradient of $l_m C l_m'$ with respect to σ^2 evaluated at $\hat{\sigma^2}$, where $\hat{\sigma^2}$ is the ML/REML estimate of σ^2.

v. Let $E = \sum_{m=1}^{q} \dfrac{\nu_m}{\nu_m - 2} I\{\nu_m > 2\}$.

vi. Compute $\nu = \dfrac{2E}{E-q}$ provided $E > q$, otherwise set $\nu = 0$.

Inference for the fixed-effects terms in the robust analysis can be performed using the Wald-type F statistic above (4.44) plugging in the robust estimates of β, σ^2 and Φ^* obtained as described in the previous sections.

[8] F does not necessarily have an F-distribution and so KR approximates the distribution of F by choosing λ and m so that $\lambda F \sim F(q, m)$.

4.7.3 Coefficient of determination

Generalizing the concept of R^2 from the general linear multiple regression model to the context of a mixed model is not straightforward and there is not yet a consensus in the literature as to which of the proposed measures should be used to assess the goodness of fit of a linear mixed model. One of those measures is the one of Magee (1990)

$$R^2_{LR} = 1 - \exp\Big(-\frac{2}{n}(\log L_M - \log L_0)\Big) \tag{4.48}$$

where $\log L_M$ is the maximum log-likelihood of the model of interest M and $\log L_0$ is the maximum log-likelihood of the null model, i.e, the model with only the intercept.

This measure was compared with several others by Sun (2006) in the context of also several linear mixed models, one of which the genetic association model proposed by Yu *et al.* (2006). They used a maize and an Arabidopsis data set in their comparisons and concluded that (4.48) was the most useful R^2-type statistic and the one that made more sense amongst all the R^2 proposals studied. Kramer (2005), also advised its use as a measure of goodness of fit of the model to the data since it is available in most software and is also easily calculated. A recent review on the theme can be seen e.g, in Edwards *et al.* (2008) where, besides presenting a new approach, he also reviews, among others, the proposals of Snijders and Bosker (1994), Vonesh and Chinchilli (1997), Zheng (2000), Xu (2003), Kramer (2005), Gelman and Pardoe (2006).

A robust version of this measure may be taken plugging in (4.48) the robust estimates of β, σ^2 and Φ^* obtained as described in the previous sections and replacing n in the formula by the weight sum obtained from the robust linear mixed model regression.

As far as the laws of mathematics refer to reality, they are not certain, and as far as they are certain, they do not refer to reality.

Albert Einstein

5

Simulation study I

The main goal of this chapter is to compare the robust approach with the classic methodology, both described in Chapter 4, the latter being commonly used in genetic association studies, and two other non-parametric methods. Namely, the non-parametric method proposed by McKean *et al.* (2009) and the usual rank-transform (RT) method which consists on transforming the observed phenotypes to their ranks and consider these values as the response variable in the classic linear regression model.

It is our intention to show that, free from the simulation burden they once were victims of, the use of robust methods now have their own place also in the context of genetic association studies to which they can be a much valuable asset.

This chapter consists of 5 sections. First, the simulation model is introduced (Section 5.1). Next the simulation model's parameters are chosen (Section 5.2). We then discuss the simulation settings (Section 5.3), present the results (Section 5.4) and end with a brief discussion (Section 5.5).

5.1 The simulation model

The model from which the data to use in the comparative analysis was simulated, is the one proposed by Liu (1997). This model is not completely specified by Liu in his book but one can find the necessary information to complete the model's specification in Wu *et al.* (2007). The simulation model will be briefly described below.

We simulate N biallelic genes on one pair of chromosomes of an F2 population (see Section 2.4.1). We start by simulating the genotype for the first gene. Afterwards, each new gene genotype is simulated based on the previous gene genotype, assuming no crossover interference and a recombination fraction r between both genes, randomly taken from the uniform distribution, $U(0, 0.5)$ (see Liu, 1997 and Section 2.4.1 for further details on the simulation of an F2 population genotype). This one gene at a time type of data simulation is easier than the joint approach where there may be a large number of possible genotypes to consider. E.g, for N biallelic genes there would be 3^N possible genotypes.

If a quantitative trait is assumed to be controlled by a number N of genes, without epistatic interactions[1], it can be simulated by

$$y_j = \mu + \sum_{i=1}^{N} \Big(a_i x_{(2i-1)j} + d_i x_{(2i)j} \Big) + \varepsilon_j, \tag{5.1}$$

where y_j is the trait value for the j-th individual in the population, $(x_{(2i-1)j}, x_{(2i)j})$ are the dummy variables associated to the additive and dominance effects for the N genes, respectively, (a_i, d_i) are the additive and dominance effects for each gene, μ is the overall mean for the trait and ε_j is the random error for the j-th individual, $j = 1, \ldots, n$. The dummy variables are coded $(x_{(2i-1)j}, x_{(2i)j}) = (1, 0)$, $(x_{(2i-1)j}, x_{(2i)j}) = (0, 1)$, and $(x_{(2i-1)j}, x_{(2i)j}) = (-1, 0)$, for genotypes AA, Aa and aa, respectively. In order to use (5.1), we need to specify some parameters underlying the simulation model. These are:

Heritability (*broad sense*). See Section 2.4.2 for more detail. From equation (2.4) and for a certain value of H^2 $(\neq 0, 1)$ we have the relations

$$\sigma_G^2 = \frac{H^2 \sigma_e^2}{1 - H^2} \quad \text{and} \quad \sigma_e^2 = \frac{(1 - H^2)\sigma_G^2}{H^2} \quad . \tag{5.2}$$

Genetic, additive and dominance variances. The genetic variance can be de-

[1]Without epistatic interactions means that the effect of one gene is not conditioned by the effect of another. This is the same as "no interaction" in statistical terminology.

composed as $\sigma_G^2 = \sigma_A^2 + \sigma_D^2$, where σ_A^2 and σ_D^2 are the additive and dominance variances respectively. Under model (5.1) the additive and dominance variances between genes are given by (see Wu *et al.*, 2007)

$$\begin{cases} \sigma_A^2 = \frac{1}{2}\sum_{i=1}^{N} a_i^2 + \frac{1}{2}\sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N}(1-2r_{ij})a_i a_j \\ \sigma_D^2 = \frac{1}{4}\sum_{i=1}^{N} d_i^2 + \frac{1}{4}\sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N}(1-2r_{ij})^2 d_i d_j \end{cases} \tag{5.3}$$

where r_{ij} represents the recombination fraction between genes i and j.

From the genotype data simulation, the values of $r_1, \ldots, r_{N-1}$ coincide with the values of $r_{12}, r_{23}, \ldots, r_{(N-1)N}$. We also have $r_{ii} = 0 \ \forall_i$.

In order to obtain the remaining values of the recombination fractions between genes, and since recombination fractions are not additive, $r_1, \ldots, r_{N-1}$ can be converted to map distances $d_1^*, \ldots, d_{N-1}^*$ via Kosambi's or Haldane's map function. Map distances are additive so, with the values of $d_1^*, \ldots, d_{N-1}^*$, the distances between all genes can be calculated and through the inverse process the recombination fractions between all genes can thus be obtained (see Liu, 1997, for further detail).

The relative importance of the additive and dominance variances can be quantified by their ratio and so we may consider, for simulation purposes, $\sigma_D^2/\sigma_A^2 = t$ and write

$$\begin{cases} \sigma_A^2 = \sigma_G^2 \dfrac{1}{t+1} \\ \sigma_D^2 = \sigma_G^2 \dfrac{t}{t+1} \end{cases} \tag{5.4}$$

Additive and dominance effects. Assuming the relative sizes of the effects to be given by $k_i = a_i/a_1 = d_i/d_1$, $i = 1, \ldots, N$, the additive and dominance effects can be obtained from (5.3) in the following way (derivation for σ_D^2 is analogous):

$$\begin{aligned} \sigma_A^2 &= \frac{1}{2}\sum_{i=1}^{N} a_i^2 + \frac{1}{2}\sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N}(1-2r_{ij})a_i a_j = \frac{a_1^2}{a_1^2}\Big[\frac{1}{2}\sum_{i=1}^{N} a_i^2 + \frac{1}{2}\sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N}(1-2r_{ij})a_i a_j\Big] = \\ &= a_1^2\Big[\frac{1}{2}\sum_{i=1}^{N}\frac{a_i^2}{a_1^2} + \frac{1}{2}\sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N}(1-2r_{ij})\frac{a_i a_j}{a_1^2}\Big] \underset{k_i=\frac{a_i}{a_1}}{=} a_1^2\Big[\frac{1}{2}\sum_{i=1}^{N} k_i^2 + \frac{1}{2}\sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N}(1-2r_{ij})k_i k_j\Big] \end{aligned}$$

$$\Rightarrow a_1 = \sqrt{\sigma_A^2 / \Big(\frac{1}{2} \sum_{i=1}^{N} k_i^2 + \frac{1}{2} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} (1 - 2r_{ij}) k_i k_j \Big)}$$

and therefore we have the equalities

$$\begin{cases} a_1 = \sqrt{\sigma_A^2 / \Big(\frac{1}{2} \sum_{i=1}^{N} k_i^2 + \frac{1}{2} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} (1 - 2r_{ij}) k_i k_j \Big)} \\ a_i = a_1 k_i, \quad i = 2, \ldots, N \\ d_1 = \sqrt{\sigma_D^2 / \Big(\frac{1}{4} \sum_{i=1}^{N} k_i^2 + \frac{1}{4} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} (1 - 2r_{ij})^2 k_i k_j \Big)} \\ d_i = d_1 k_i, \quad i = 2, \ldots, N \end{cases} . \tag{5.5}$$

For a fixed value of H and $\sigma_G^2 = 1$, and using equation (5.2), we have $\varepsilon_j = z\sigma_e = z\sqrt{(1 - H^2)/H^2}$ for each individual $j = 1, \ldots, n$, where z is a random observation from the $N(0, 1)$ distribution. Finally, for a fixed value of t, the additive and dominance variances can be calculated from (5.4) and from there all the additive and dominance effects (5.5). The quantitative trait is then simulated for the n individuals.

The `R` language was used to implement the model.

5.2 Choosing the simulation model's parameters

The aim of this simulation study is to compare the performance of classic, robust and two non-parametric approaches in genetic association studies, i.e, in association tests between a quantitative trait and several genetic markers, more precisely, SNPs. At this point we should underline the fact that SNPs are biallelic with alleles `A` and `T` or `C` and `G` and this is the reason why the model described in the previous section can be used to simulate a quantitative trait from SNP data.

For that purpose we first need to simulate the data sets with which to perform the comparative analysis. We must therefore, decide which parameters to use in model (5.1).

The idea is to have simulation settings with appropriate size, where the classic tests have the best behavior possible, i.e, uniform p-value distribution under the null and high

power to detect global and individual SNP/trait associations when such exist.

We started out by making a few experiments with the parameters on a 2 SNP model, observing p-value distribution under the null and test power under the alternative hypothesis. We randomly fixed $\mu = 50$ and $t = 0.6$. Also, we considered $H^2 = 0.3$ since quantitative traits variation is only shortly explained by genetic factors. We varied the number of individuals n, the recombination fraction between SNPs r and the relative sizes of the effects k. Association tests were run 10000 times and the conservative Bonferroni correction was used. The goal is to choose fixed values for these parameters.

Table 5.1: Classic POWER (%) obtained from simulation under H_1 without outliers, for the global test of association in the 2 SNP simulation model, for several values of the recombination fraction, r, relative additive and dominance effects, k, and number of individuals, n.

r	k	SNP1			SNP2		
		n=200	n=350	n=500	n=200	n=350	n=500
0.1	$k_1 = 1,\ k_2 = 1/1$	> 70	> 90	$\simeq 100$	> 70	> 90	$\simeq 100$
	$k_1 = 1,\ k_2 = 1/2$	> 90	$\simeq 100$	100	> 30	> 55	> 75
	$k_1 = 1,\ k_2 = 1/3$	> 95	$\simeq 100$	100	> 15	> 30	> 45
	$k_1 = 1,\ k_2 = 1/5$	$\simeq 100$	100	100	> 5	> 10	> 15
0.2	$k_1 = 1,\ k_2 = 1/1$	> 95	$\simeq 100$	100	> 95	$\simeq 100$	100
	$k_1 = 1,\ k_2 = 1/2$	$\simeq 100$	100	100	> 60	> 85	> 95
	$k_1 = 1,\ k_2 = 1/3$	$\simeq 100$	100	100	> 30	> 60	> 75
	$k_1 = 1,\ k_2 = 1/5$	$\simeq 100$	100	100	> 10	> 25	> 35
0.3	$k_1 = 1,\ k_2 = 1/1$	100	100	100	$\simeq 100$	100	100
	$k_1 = 1,\ k_2 = 1/2$	100	100	100	> 75	> 95	100
	$k_1 = 1,\ k_2 = 1/3$	100	100	100	> 45	> 75	> 90
	$k_1 = 1,\ k_2 = 1/5$	100	100	100	> 15	> 30	> 45
0.4	$k_1 = 1,\ k_2 = 1/1$	100	100	100	100	100	100
	$k_1 = 1,\ k_2 = 1/2$	100	100	100	> 85	$\simeq 100$	$\simeq 100$
	$k_1 = 1,\ k_2 = 1/3$	100	100	100	> 55	> 80	> 90
	$k_1 = 1,\ k_2 = 1/5$	100	100	100	> 20	> 35	> 55
0.5	$k_1 = 1,\ k_2 = 1/1$	100	100	100	100	100	100
	$k_1 = 1,\ k_2 = 1/2$	100	100	100	> 90	$\simeq 100$	100
	$k_1 = 1,\ k_2 = 1/3$	100	100	100	> 55	> 85	> 95
	$k_1 = 1,\ k_2 = 1/5$	100	100	100	> 20	> 40	> 55

We observed that p-value distribution under the null was uniform for all parameter settings tested for both global and individual SNP/trait tests. As to the tests under the alternative hypothesis, results concerning the power of the tests can be seen in Table 5.1. From that table it follows, as expected, that power increases:

1. when the number of individuals, n, increases;

2. with the increase of the recombination fraction r;

3. when $k_2 \to k_1 = 1$.

Other interesting observations are:

1. The gain in power with the increase of the number of individuals is greater when we increase from 200 to 350 individuals than when we increase from 350 to 500. This is not strange since we can increase power by increasing the number of individuals but only up to a limit, i.e, up to a point from where increasing the number of individuals in the experiment is then unjustified.

2. When both SNPs have the same effect on trait variation ($k_1 = k_2$) and recombination fraction is greater or equal than 0.3, the association detection between them and the quantitative trait is 100%. When $r < 0.3$ power is still good enough, i.e, it stays over 70%.

3. Despite the number of individuals and the value of the recombination fraction, power drops dramatically in the detection of association regarding SNP2 when its contribution to the trait value is 1/5 of that of SNP1. Even in the best scenario, i.e, with $r = 0.5$, it stays merely above 55%.

Since we are interested in selecting the settings that provided an uniform distribution under the null (they all did) and gave more power under the alternative hypothesis, it seems appropriate to take $r = 0.5$, which assures the independence of the SNPs thus avoiding any collinearity issues regarding the specification matrix of the regression model, and $n = 500$. As to k, although the situation $k_1 = k_2 = 1$ presented itself as the most powerful, it is most unlikely that two genes/SNPs contribute exactly in the same amount to a trait value and so we have decided to take $k_1 = 1$ and k_i out of an uniform distribution $U(0.8, 1)$, with $i = 2, ..., N$, N being the number of SNPs in the model. In this way we assure that the SNPs have an important though unequal contribution to the trait value.

Finally, with these fixed parameters, we enlarged our experiment to the situation where we have $N = 2, 3, 4, 5$ and 10 SNPs in order to asses the decrease in power with the increase in the number of SNPs to be simultaneously tested. The global tests detected association with 100% power in all settings. Moreover, the minimum individual SNP/trait test power observed was in the 10 SNP model with a reasonably value of over 75%.

5.3 Simulation settings

From the exploratory analysis made in the previous section we have the following fixed parameters: $H^2 = 0.3$, $t = 0.6$, $\mu = 50$, $n = 500$. Also, we considered the relative additive and dominance effects of the genes $k_i \sim U(0.8, 1)$, $i = 2, \ldots, N$ so that in this way every SNP in the model would have an important, though unequal, contribution to the trait value.

Additionally, for the purpose of the simulation, we considered independent SNPs, that is $r_{ij} = 0.5 \; \forall_{i \neq j}$. This assures that the SNPs are in HWE and that there is no LD between pairs of SNPs.

As to the number of SNPs in the model, we took $N = 2, 3, 4, 5$ and 10. A percentage, $2, 5$ and 10%, of data contamination (outliers) was also considered. The contamination was generated from a normal distribution $N(\mu_c, \sigma_c^2)$ where σ_c was obtained from a uniform distribution $U(1, 5)$ and μ_c from $U(80, 90)$, $U(60, 70)$ and $U(55, 60)$, corresponding, respectively, to *gross*, *intermediate* or *smooth* contamination. Association tests were run 10000 times.

In order to compare the modeling approaches, one needs first to investigate how they control the FWER at a pre-specified level. On the other hand a better performance of one over another means POWER. Hence, we have the following two testing settings:

Under the null hypothesis we looked at three distinct situations:

i. traits were simulated independently from the SNPs genotypes out of a $N(50, 1)$ distribution;

ii. traits were simulated independently from the SNPs genotypes out of a $N(50, 1)$ and a percentage of outliers (2, 5 and 10%) was then introduced in the trait values;

iii. a percentage of the traits (98, 95 and 90%) were simulated independently from the SNPs genotypes, that is under H_0, but using the normal contaminant distributions, whereas the remaining traits (2, 5 and 10%) were generated from model (5.1), that is under H_1.

Under the alternative hypothesis two situations were considered:

i. the traits were simulated according to model (5.1);

ii. the traits were simulated from model (5.1) and a percentage (2, 5 and 10%) of outliers, obtained from the normal distributions $N(\mu_c, \sigma_c^2)$ previously described, was introduced.

5.4 Simulation results

We adjusted in R the additive model described in (4.1) via the usual linear regression (`lm()`) and the robust linear regression (`rlm()`) with the M-Huber estimator, in the two simulation settings described. Function `lm()` was also used for the RT method, i.e, the original quantitative traits were replaced by their ranks.

As to the Wilcoxon methodology (WIL) we used the R function `wwest()` from McKean *et al.* (2009).

We then tested for association using the ANOVA table for the usual linear regression model and RT method, and robust Wald-type tests for the robust model as described in Chapter 4.

Tests for the WIL approach were obtained from function `wwest()` also via Wald-type tests.

SNPs were always tested in a multiple regression framework. We must also underline that due to the computational effort of the `wwest()` function, simulations for the WIL approach ran only 1000 times instead of the 10000.

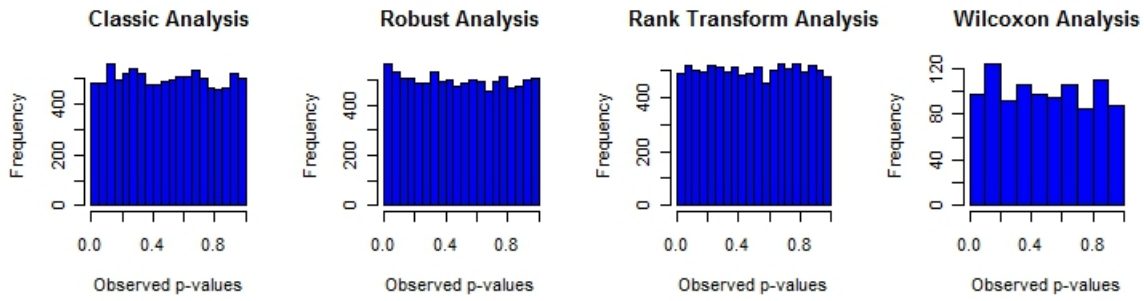


Figure 5.1: P-value histogram of the global tests for association under H_0 in the 10SNP model. Recall that we only have 1000 observations in the Wilcoxon approach

We concluded that under the null without contamination, the distribution of the p-values is, in all cases, approximately $U(0,1)$[2] (see Figure 5.1) and thus all methods are comparable, allowing for the use of the same multiple testing corrections.

In the (ii) situation under H_0, there were no surprises, i.e., all methods were able to control the FWER at the 5% level whatever the contamination, gross, intermediate or smooth. However, in the (iii) situation under H_0, we noticed that in the smooth contamination setting, the FWER of the classic approach surpasses 5% most of the times, while the robust approach always keeps it around that threshold (Table 5.2). This tendency of the classic approach towards inflated type-I error rates was also observed in the intermediate contamination setting and accompanied by the WIL approach also in the smooth and intermediate contamination settings. As to the RT approach, it only failed to control for the FWER at the desired level once.

Under H_1 (Table 5.3), all methods show very good power to detect association between the SNPs and the quantitative trait, in the following order: Classical>WIL>Robust>RT, and with relatively small differences. Even using Bonferroni correction, we had over 99% power in all methods, up to 5 SNPs, and still over 76% power when we took the model with 10 SNPs (0% contamination).

With the introduction of contamination, as expected, the power of all methods decreases as the contamination level increases. However, the robust method shows much higher power to detect associations than the classic method (Figure 5.2) and higher power than the RT approach, being neck to neck with the WIL approach. At the worst scenario, 10 SNPs in association with the trait and 10% gross outliers, the robust method, with RT and WIL close behind, has a power over 52% to detect those associations, whilst the usual model stays under 1%. Even in the smooth contamination case, the power of the classical method is only 23.5% versus over 48% for all the other approaches.

Moreover, we must stress out that there are relevant power losses even in cases where the residual deviations from normality are not evident. See for example in Table 5.3 the 10 SNP model with 2% smooth contamination, where there is 13% power loss from the classic approach relative to the robust one but whose residuals do not look different from normal (see the correspondent residual Q-Q plots in Figure 5.3). There are even situations where the Shapiro-Francia (SF) normality test[3] p-values may go as low as 0.000001 and still the

[2]Uniformity of the p-value distribution was tested with the R instruction `ks.test()`.
[3]Available in the R *nortest* package via instruction `sf.test()`.

Table 5.2: FWER (in %) obtained from simulation* under H_0 with a percentage of contaminated traits - (iii) setting under H_0 **

	Cont	*Model*	2SNP	3SNP	4SNP	5SNP	10SNP
	2%	Classical	4.7	4.8	4.6	4.7	4.5
		Robust	5.0	4.8	5.0	5.0	4.8
		RT	4.3	5.0	4.9	5.0	4.8
G		WIL	5.3	5.4	5.2	5.8	4.3
R	5%	Classical	5.0	4.9	5.1	4.7	4.9
O		Robust	5.3	4.9	4.8	4.8	5.0
S		RT	5.2	4.8	4.7	5.3	5.3
S		WIL	5.9	6.1	5.8	4.3	5.2
	10%	Classical	5.2	5.4	5.2	5.3	5.2
		Robust	4.9	4.9	5.1	4.8	4.7
		RT	5.3	5.1	4.9	5.1	5.1
		WIL	4.9	4.5	5.0	5.3	5.1
	2%	Classical	5.1	4.9	4.9	4.8	4.8
		Robust	5.0	4.8	4.9	4.8	5.0
		RT	4.9	4.7	5.0	5.2	4.8
I		WIL	5.3	5.7	3.9	4.5	4.3
N	5%	Classical	**5.6**	5.3	5.4	5.3	5.0
T		Robust	5.0	4.9	5.0	4.9	4.7
E		RT	4.9	5.3	4.8	5.0	5.3
R		WIL	5.8	4.6	4.4	4.6	4.5
.	10%	Classical	**6.2**	**5.8**	**5.6**	**5.7**	**5.5**
		Robust	4.8	4.8	5.0	5.0	5.0
		RT	5.3	5.2	5.1	4.7	4.8
		WIL	**6.9**	4.5	6.0	5.4	5.7
	2%	Classical	5.3	5.1	5.1	4.9	5.0
		Robust	5.2	5.1	4.9	4.9	5.1
		RT	5.1	**5.6**	5.1	4.8	4.8
S		WIL	5.3	5.4	5.1	5.1	4.7
M	5%	Classical	**5.9**	**5.6**	**5.6**	**5.6**	5.0
O		Robust	5.1	4.8	5.0	5.1	4.8
O		RT	5.1	5.1	5.0	4.9	5.1
T		WIL	5.5	4.1	4.4	5.1	5.2
H	10%	Classical	**7.6**	**6.9**	**6.5**	**6.6**	**6.0**
		Robust	5.3	5.1	5.2	5.4	5.2
		RT	5.4	5.1	5.3	4.8	4.9
		WIL	**7.1**	5.7	6.3	**7.2**	5.0

* Results obtained from 10000 simulations except for the WIL method which ran only a 1000 times;

** Results in bold correspond to cases significantly different from 5%, according to the binomial exact test at the 5% level.

Q-Q normal plot gives indication of approximate normality (Figure B.4). We emphasize the fact that this SF p-value of 0.000001 corresponds to a Pearson's goodness-of-fit[4] test p-value of 0.134 giving no evidence against H_0 should this be the normality test used. We may thus be tempted either to lower the SF threshold upon which we reject the null hypothesis or to opt by the Pearson's test at the usual threshold of 5% (or lower!.. see

[4]Available in the R *nortest* package via instruction `pearson.test()`.

Table 5.3: POWER (in %) obtained from simulation under H_1 without contamination and with a percentage of grossly (G), intermediate (I) and smoothly (S) contaminated traits. (truncated minimum power observed to detect every SNP in the simulation model)

	Cont.	*Model*	2SNP	3SNP	4SNP	5SNP	10SNP
	None	Classical	100	100	99.9	99.7	81.2
		Robust	100	100	99.9	99.7	78.2
		RT	100	100	100	99.6	76.2
		WIL	100	100	99.9	99.5	79.5
	2%	Classical	62.6	41.1	26.6	15.2	3.4
		Robust	100	100	99.9	99.3	75.0
		RT	100	100	99.8	98.5	71.0
G		WIL	100	100	100	99.7	74.0
R	5%	Classical	25.1	12.4	8.5	4.7	1.2
O		Robust	100	100	99.9	98.6	61.3
S		RT	100	99.9	99.6	97.4	57.6
S		WIL	100	100	99.6	97.6	67.7
	10%	Classical	11.9	5.6	4.0	2.3	0.8
		Robust	100	99.7	99.1	92.6	52.6
		RT	100	99.3	96.5	92.4	44.5
		WIL	100	100	98.7	94.3	51.0
	2%	Classical	99.1	94.7	86.6	72.8	30.6
		Robust	100	100	99.9	99.4	74.6
		RT	100	100	99.8	98.5	71.3
I		WIL	100	100	100	99.4	74.3
N	5%	Classical	86.8	65.2	55.4	38.7	11.0
T		Robust	100	99.9	99.8	98.4	62.5
E		RT	100	99.9	99.6	97.6	58.1
R		WIL	100	99.9	99.8	98.2	65.8
.	10%	Classical	59.1	36.2	27.33	16.33	4.9
		Robust	100	99.7	99.0	92.4	53.0
		RT	100	99.3	97.0	92.0	44.2
		WIL	100	100	98.6	93.7	52.2
	2%	Classical	100	99.9	99.3	95.9	62.2
		Robust	100	100	99.9	99.5	75.5
		RT	100	100	99.9	98.6	71.9
S		WIL	100	99.8	99.9	99.8	73.5
M	5%	Classical	99.6	96.8	93.1	82.9	37.1
O		Robust	100	99.9	99.8	98.5	63.3
O		RT	100	99.9	99.6	97.8	59.6
T		WIL	100	100	99.7	98.4	69.1
H	10%	Classical	95.6	82.8	75.2	55.4	23.5
		Robust	100	99.8	99.2	93.4	55.3
		RT	100	99.5	97.3	93.5	48.3
		WIL	100	99.8	99.3	95.2	56.9

* Results obtained from 10000 simulations except for the WIL method which ran only a 1000 times.

Figure B.5). In the latter situation, in 10000 simulations of association tests performed in the 10 SNP model with 2% smooth contamination, where only models with Pearson's p-values between 0.05 and 0.1 were considered, the power loss from the classic approach relative to the robust counterpart amounted to 11%. This difference in power is likely to

Table 5.4: Truncated, average number of SNP detected in the simulation under H_1 without contamination and with a percentage of grossly (G), intermediate (I) and smoothly (S) contaminated traits.

	Cont.	*Model*	2SNP	3SNP	4SNP	5SNP	10SNP
	None	Classical	2	3	3.99	4.99	9.01
		Robust	2	3	3.99	4.99	8.80
		RT	2	3	4	4.99	8.70
		WIL	2	3	3.99	4.99	8.83
	2%	Classical	1.22	1.04	0.84	0.61	0.17
		Robust	2	3	3.99	4.98	8.39
		RT	2	3	3.99	4.98	8.20
G		WIL	2	3	4	4.99	8.47
R	5%	Classical	0.45	0.28	0.18	0.12	0.03
O		Robust	2	3	3.99	4.96	7.76
S		RT	2	2.99	3.99	4.92	7.20
S		WIL	2	3	3.99	4.96	7.88
	10%	Classical	0.20	0.10	0.06	0.04	0.01
		Robust	2	2.99	3.97	4.85	6.30
		RT	2	2.99	3.93	4.74	5.46
		WIL	2	3	3.97	4.84	6.38
	2%	Classical	1.98	2.85	3.53	3.92	3.61
		Robust	2	3	3.99	4.99	8.40
		RT	2	3	3.99	4.98	8.23
I		WIL	2	3	3.99	4.99	8.49
N	5%	Classical	1.77	2.12	2.20	2.04	1.16
T		Robust	2	2.99	3.99	4.96	7.82
E		RT	2	2.99	3.99	4.93	7.24
R		WIL	2	2.99	3.99	4.97	7.90
.	10%	Classical	1.24	1.15	0.93	0.78	0.32
		Robust	2	2.99	3.97	4.86	6.33
		RT	2	2.99	3.94	4.74	5.46
		WIL	2	3	3.98	4.85	6.39
	2%	Classical	1.99	2.99	3.98	4.89	7.21
		Robust	2	3	3.99	4.99	8.42
		RT	2	3	3.99	4.98	8.29
S		WIL	2	2.99	3.99	4.99	8.50
M	5%	Classical	1.99	2.95	3.78	4.41	4.94
O		Robust	2	2.99	3.99	4.97	7.89
O		RT	2	2.99	3.99	4.94	7.43
T		WIL	2	3	3.99	4.97	7.93
H	10%	Classical	1.94	2.65	3.09	3.25	2.65
		Robust	2	2.99	3.98	4.87	6.61
		RT	2	2.99	3.95	4.79	5.81
		WIL	2	2.99	3.98	4.87	6.65

* Results obtained from 10000 simulations except for the WIL method which ran only a 1000 times.

increase as one loosens the threshold for the tests.

In real data example analysis, the authors rarely report a normality test p-value. We think this happens mostly because there are several available criteria in the literature

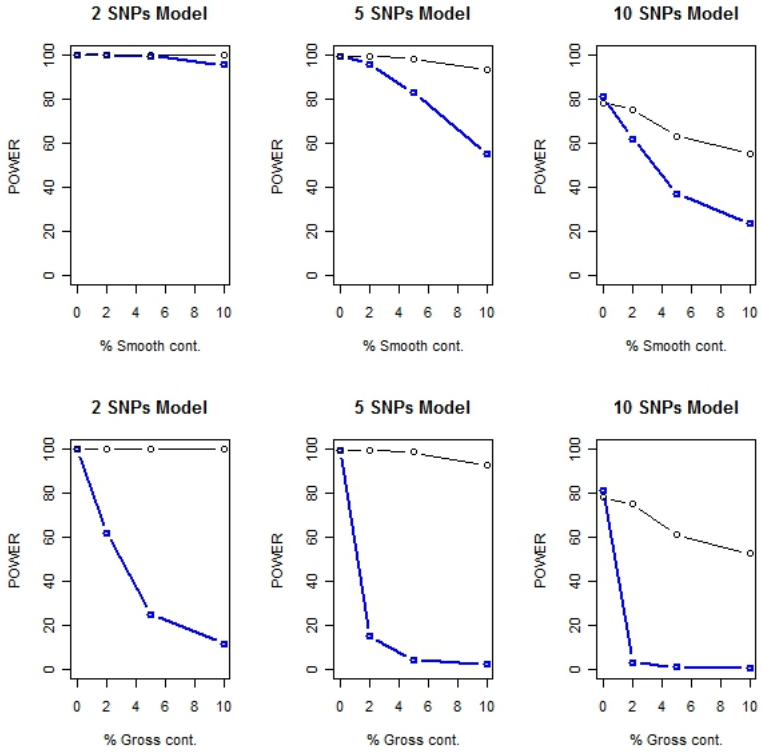


Figure 5.2: Power comparison between classical (in blue) and robust analysis (in black) from simulation under H1, with and without a % of smooth and gross data contamination, for the 2, 5 and 10 SNPs models.

and no consensus about the most adequate to use. Most researchers opt to look at the normal Q-Q plots. In the real data example analyzed in Chapter 7 this procedure was no exception. Quoting Weber *et al.* (2008): "Residual plots were examined to determine if there were any patterns indicating that a transformation was necessary". In theory, when such is the case, you may find a normality test that gives you the p-value you need to accept H_0, i.e, evidence not to reject the normality assumption. In the 10000 simulations for the 10 SNP model with 2% smooth contamination, the SF normality test at the 5% threshold gave us indication of approximate normality only 5% of the times, fact that justifies the loss in power from the classical approach to the robust one. Pearson's test gave that same evidence 56% of the times. When using a 1% threshold, that evidence escalated to 72%. This means of course, that one may be confident of one's analysis results, and still be incurring in power loss in the detection of possible associations.

To end this line of thought, we must add that, being the SF test the least conservative of the two normality tests mentioned[5], rejecting the normality hypothesis 95% of the times

[5]We also compared it with Shapiro-Wilks and Lilliefors tests concluding the SF test to be the least conservative of them all.

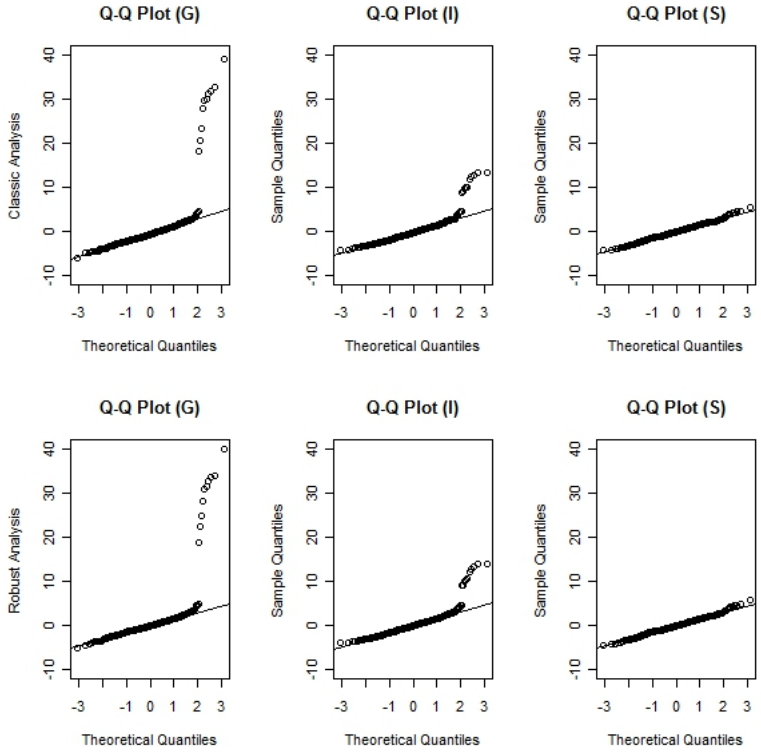


Figure 5.3: Q-Q Plots for the 10 SNP model under H_1 with 2% Gross, Intermediate and Smooth contamination, from one out of the 10000 simulations, for the classic and robust analysis only.

at a 5% level (in 10000 simulations for the 10 SNP model with 2% smooth contamination), this test is, if nothing else, a good indicator of data contamination.

If we now analyze Table 5.4, we see that the robust method detects in general more SNPs than the classical and RT procedures. Although that difference may not look substantial in the smooth contamination setting, it is quite evident in the 10 SNP simulation for 5 and 10% gross contamination. If compared to the WIL approach, in the 10 SNP simulation WIL comes off better than the robust approach but with a maximum difference of only 0.08.

Table 5.5 shows the results for the 10 SNP simulation scenario. Note that the global robust test, as well as the RT and WIL global tests, keep a 100% power in all simulation settings whereas the classical power falls down to 15.21% at the 10% gross contamination setting.

Table 5.5: POWER (%) obtained from simulation under H_1, without contamination and with a percentage of contaminated traits for the global test of association in the 10 SNP simulation model.

Cont.	Model	Gross	Interm.	Smooth
0%	Classical	100	100	100
	Robust	100	100	100
	RT	100	100	100
	WIL	100	100	100
2%	Classical	76.71	100	100
	Robust	100	100	100
	RT	100	100	100
	WIL	100	100	100
5%	Classical	32.48	96.54	100
	Robust	100	100	100
	RT	100	100	100
	WIL	100	100	100
10%	Classical	15.21	77.24	99.69
	Robust	100	100	100
	RT	100	100	100
	WIL	100	100	100

5.5 Discussion

In the previous Sections we compared the performance of classic, robust and two non-parametric methodologies in genetic association studies in a particular simulation frame.

We showed that under the null, without contamination, the four methods considered have control of the FWER at the desired level. We acknowledged not only the tendency of both the classic and WIL approaches towards inflated type-I error rates, but also that the robust approach proposed (Huber M-regression plus Wald-type tests to assess association) is not as sensitive to outlier contamination as the classical approach and is more powerful than the RT approach to detect SNP/trait associations.

Despite the fact that the WIL approach kept close to the robust methodology in terms of power, its tendency to inflated type-I error rates and computational issues indicate that for an association study involving a small number of independent SNPs to be tested (as in the simulation study), the robust multiple-SNP linear model is preferable over the remaining approaches.

We conclude this chapter with the belief that the robust methodology is much promising to genetic association studies of quantitative traits. Therefore we would suggest it should be used in this context in a regular basis together with the classical methodology in order to assess and discuss possible arising differences in the results.

Our energy is in proportion to the resistance it meets. We attempt nothing great but from a sense of the difficulties we have to encounter, we persevere in nothing great but from a pride in overcoming them.

William Hazlitt

6

Simulation study II

In Section 4.6 we reviewed a proposal of Rice and Spiegelhalter (2006) for outlier detection via robust residuals, Benjamini and Hochberg FDR and a simple diagnostic plot. They only studied the case when a single explanatory variable is considered. Here we propose to study the case of regression outliers in the context of association studies and thus multiple regression.We do not have in this context problems with outlying observations in the explanatory variables since they are categorical and will therefore not be concerned with leverage points, although caution is needed with regression outliers that may be influential to the analysis on the account of the unbalanced design and/or the complexity of the model, i.e, the sometimes large number of explanatory variables. Another important concern will be with the allowed level of swamping. As opposed to association studies where an FDR of 10% is usually acceptable, in outlier detection, depending on what we intend to do after detecting the outlying observations, a 10% FDR threshold may be exaggerate.

The goal of the simulation study in the first section of this Chapter will therefore be to assess if the control of the FDR in outlier detection, via the proposed approach, is possible in the context of increasing model complexity and in the presence of smooth data contamination. Several FDR correction measures proposed in the literature will also be studied to see if power in detecting true outliers can be improved.

In Section 6.2 we will study how well the robust coefficient of determination reflects the heritability of a trait, also in the context of a multiple regression model with independent SNPs.

6.1 False discovery rate control in outlier tests: a robust approach.

For this study we considered:

1. $n = 250,\ 500$ and 750 individuals
2. $N = 2, 3, 4, 5, 10, 50$ and 100 independent SNPs
3. BH-1995, BH-2000, BY, ST, BKY and Bonferroni corrections (see Algorithms in Chapter 3) at nominal levels of 0.01, 0.05 and 0.1;
4. data generated from model (5.1) in Chapter 5;
5. trait heritabilities of $H^2 = 0.3, 0.5$ and 0.75;
6. for the outlier tests we used the standardized residuals in equation (4.32) and the studentized residuals
$$r_i^R = \frac{\widehat{\varepsilon}_i^R}{\widehat{\sigma}_R\sqrt{1-h_{ii}}} \tag{6.1}$$
in a robustified version of (4.25);
7. for the robust estimate $\widehat{\sigma}_R$ in equations (4.32) and (6.1) we considered the robust $s = \text{MAD}$ and Q_n (see equations (4.7) and (4.8) respectively);
8. $B = 1000$ simulations.

Some results of the simulations will only be displayed in the Appendix B.

I. Simulations considering $H^2 = 0.3$; $n = 500$.

From Figures 6.1 and 6.2 (also B.6 and B.7) we can see that as the % of contamination increases the observed values of the FDR decrease, a trend that was also observed in all other simulations. However it is quite clear the inability of the BH-1995 FDR to control the FDR at the nominal levels established. Nevertheless, using Q_n as a robust estimate of scale instead of s allows us to grasp that control for models with up to 10 SNPs at

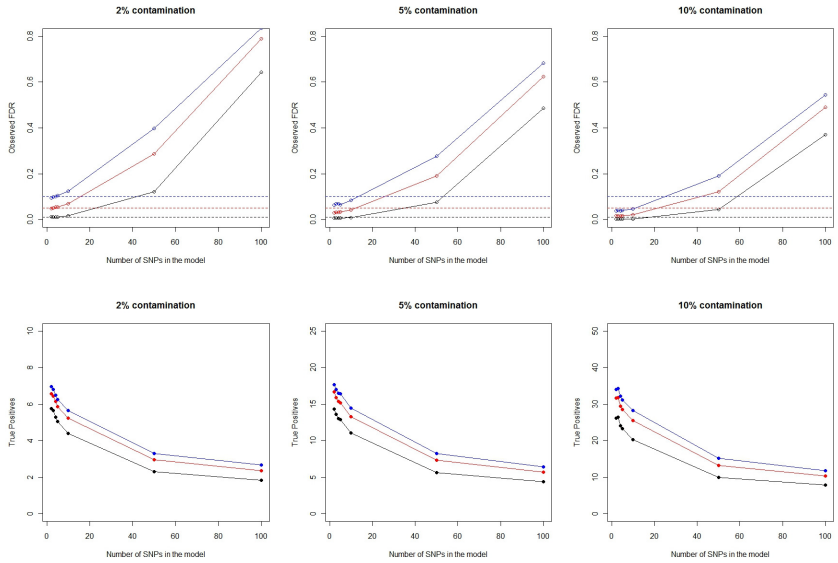


Figure 6.1: Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.3$, a percentage of $2, 5$ and 10% of smooth data contamination and robust parameter of scale s. Benjamini and Hochberg (1995) correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust standardized residuals (4.32). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1.

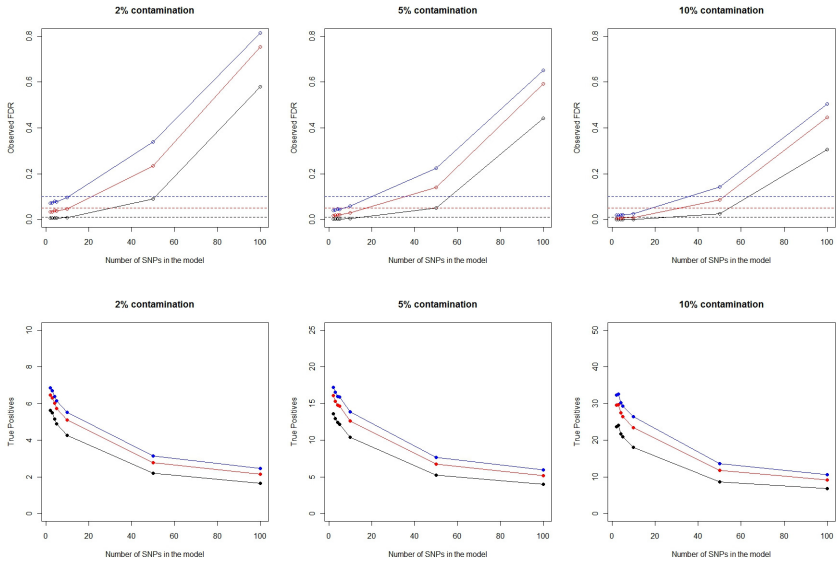


Figure 6.2: Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.3$, a percentage of $2, 5$ and 10% of smooth data contamination and robust parameter of scale Q_n. Benjamini and Hochberg (1995) correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust standardized residuals (4.32). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1.

Table 6.1: Average number of detected true positives corresponding to Figures 6.1 and 6.2 - for the 2% contamination setting only, which amounts to a total of 10 contaminant observations.

	True Outliers					
nSNPs	s			Q_n		
	0.01	0.05	0.1	0.01	0.05	0.1
2 SNPs	5.763	6.575	6.964	5.633	6.464	6.864
3 SNPs	5.656	6.434	6.803	5.493	6.302	6.699
4 SNPs	5.288	6.137	6.487	5.160	6.021	6.373
5 SNPs	5.056	5.857	6.258	4.904	5.737	6.144
10 SNPs	4.409	5.232	5.638	4.277	5.103	5.517

Table 6.2: Average number of detected true positives corresponding to Figures 6.5 and 6.6 - for the 2% contamination setting only, which amounts to a total of 10 contaminant observations.

	True Outliers					
nSNPs	s			Q_n		
	0.01	0.05	0.1	0.01	0.05	0.1
2 SNPs	5.097	5.738	6.019	4.949	5.599	5.909
3 SNPs	4.930	5.598	5.924	4.749	5.478	5.789
4 SNPs	4.550	5.262	5.602	4.410	5.128	5.484
5 SNPs	4.358	5.065	5.375	4.220	4.916	5.237
10 SNPs	3.795	4.439	4.744	3.670	4.301	4.644

a cost of a decrease in the number of true positives detected as shown in Table 6.1. If we think of the level of equation (3.8) given for models with a more general dependence structure, then with Q_n we achieve that control in models up to 50 SNPs but this is not an interesting situation given the high number of false positives.

When using the stundentized robust residuals (6.1) instead of the standardized ones (4.32) the results are pretty much the same (Figures 6.3/B.8 and 6.4/B.9) and therefore this choice does not appear to compensate the effort of a few more calculations.

When we used the Bonferroni correction the values of the observed FDR dropped when compared to the ones resulting from the BH-1995 FDR correction (Figures 6.5/B.10 and 6.6/B.11). However, it looks that that control will not be maintained much farther away, in which case further simulations will be needed to assess the exact point where it is actually lost. Any way, from Tables 6.1 and 6.2 we can acknowledge that the number of true outliers detected using Bonferroni correction is lower than the one we obtain with BH1995. Nevertheless, from the plots we can see that the proportion of observed false positives in this case, up to a 10 SNP model is near zero, which is a good thing meaning no swamping effect.

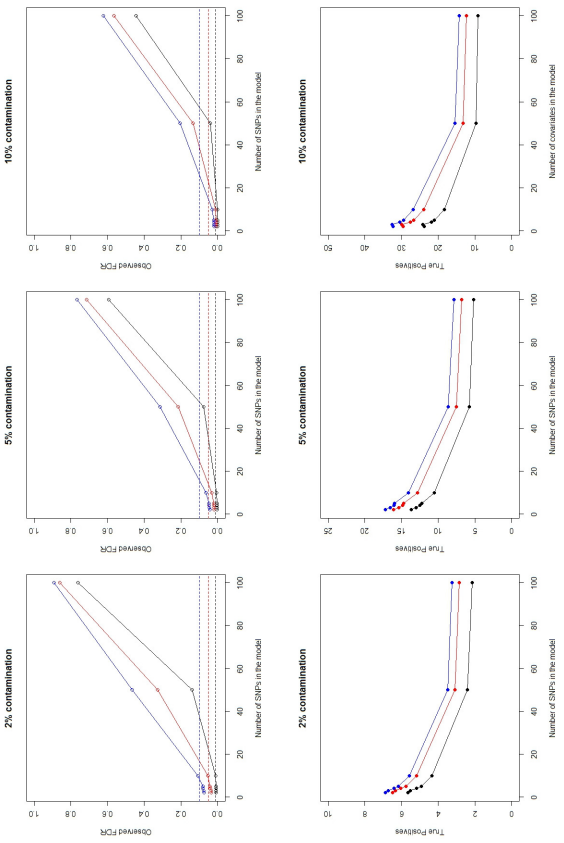


Figure 6.3: Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.3$, a percentage of $2, 5$ and 10% of smooth data contamination and robust parameter of scale s. Benjamini and Hochberg (1995) correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust studentized residuals (6.1). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1.

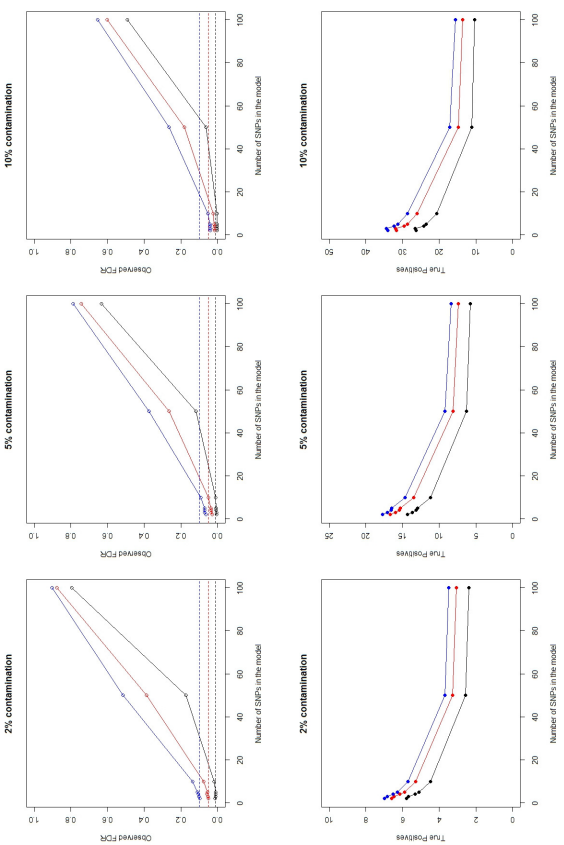


Figure 6.4: Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.3$, a percentage of $2, 5$ and 10% of smooth data contamination and robust parameter of scale Q_n. Benjamini and Hochberg (1995) correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust studentized residuals (6.1). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1.

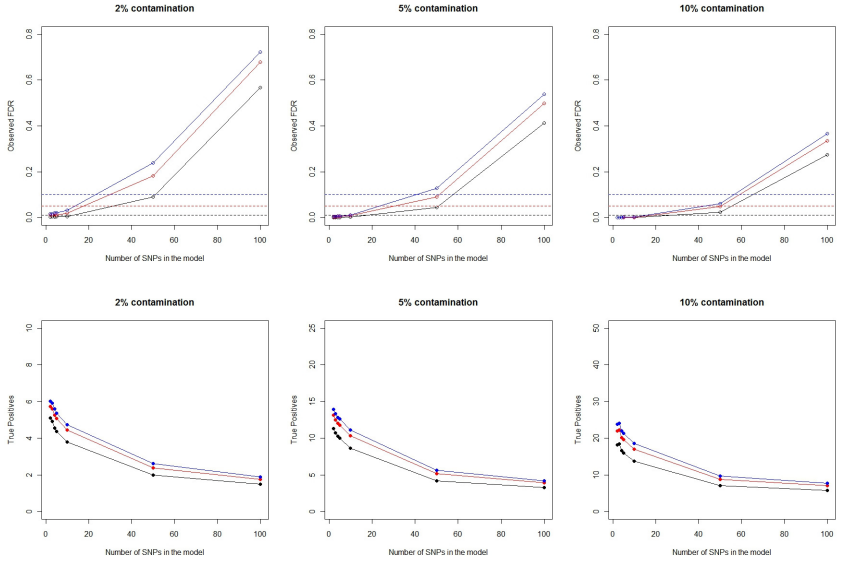


Figure 6.5: Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.3$, a percentage of $2, 5$ and 10% of smooth data contamination and robust parameter of scale s. Bonferroni correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust standardized residuals (4.32). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1.

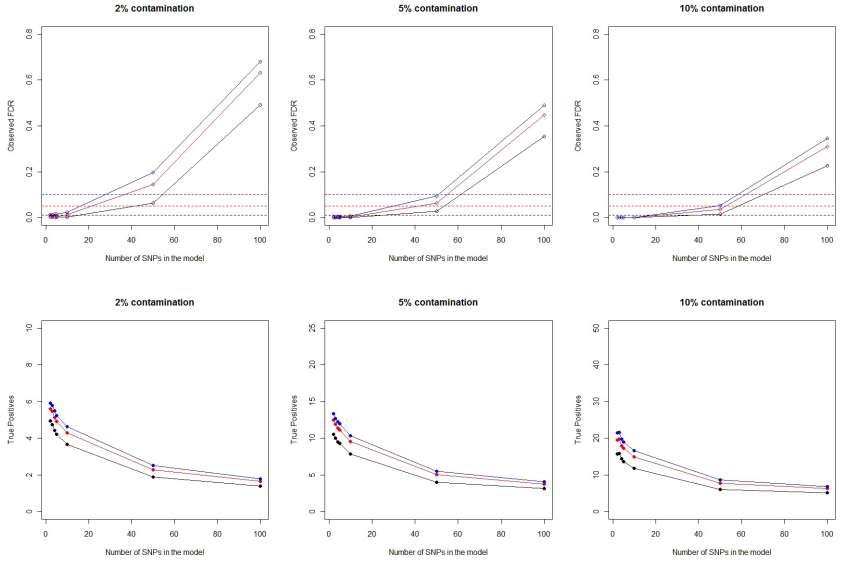


Figure 6.6: Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.3$, a percentage of $2, 5$ and 10% of smooth data contamination and robust parameter of scale Q_n. Bonferroni correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust standardized residuals (4.32). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1.

Results using the robust studentized residuals for the Bonferroni correction (6.1) can be seen in Figures B.12 and B.13 in Appendix B. As before, there seems to be no particular advantage in using these residuals for the outlier tests.

First considerations.

Ending this first set of simulations we're inclined to advise the use of standardized robust residuals together with the robust estimate of scale Q_n for outlier detection since with this choice, independently of using BH-1995 FDR correction or Bonferroni's, we were able to control the FDR at the nominal levels in models up to 10 SNPs, and we also acknowledged the control of the FDR at levels $\alpha(1 + \frac{1}{2} + ... + \frac{1}{n})$ (equation 3.8) for models with up to 50 SNPs when using Bonferroni correction. With BH-1995 this was only verified for $\alpha = 0.05$ and 0.10.

It also seems evident from these results that if we can improve both the control of the FDR and the detection of true outliers in models with 2% contamination, then those improvements will show even better in the other contamination settings of 5 and 10%. Hence, hereafter we will only worry about the 2% contamination setting.

II. Simulations considering $H^2 = 0.5,\ 0.75$; $n = 500$; Q_n only; 2% contamination only.

Let us see what happens to the FDR and number of true positives when H^2 varies. Results for values of $H^2 = 0.5$ and 0.75 are displayed in Figures 6.7-6.10, and also B.14-B.17. As expected, the control of the FDR at the nominal α levels improves and so does the number of true positives as the trait heritability increases.

More considerations.

In the simulations we acknowledged that in general it is not possible to control the FDR in models above 10 SNPs, unless of course we are dealing with a high heritable trait (Tables 6.8 and 6.10) which is usually not the case of quantitative traits. In models with 10 or less SNPs, using Bonferroni corrected p-values at the nominal levels considered, we were able to control the FDR at those same levels independently of the trait heritability, the robust estimate of scale used and of the correction made to the robust residuals.

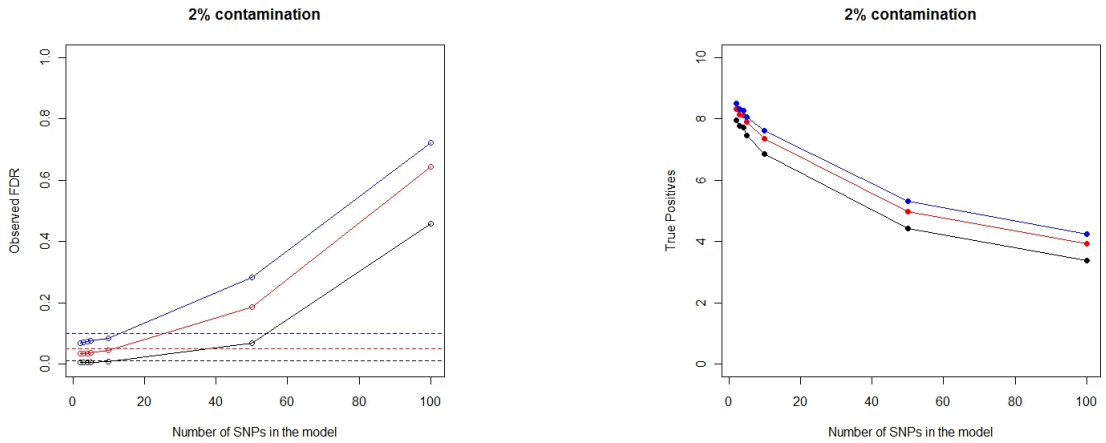


Figure 6.7: Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.5$, 2% of smooth data contamination and robust parameter of scale Q_n. Benjamini and Hochberg (1995) correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust standardized residuals (4.32). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1.

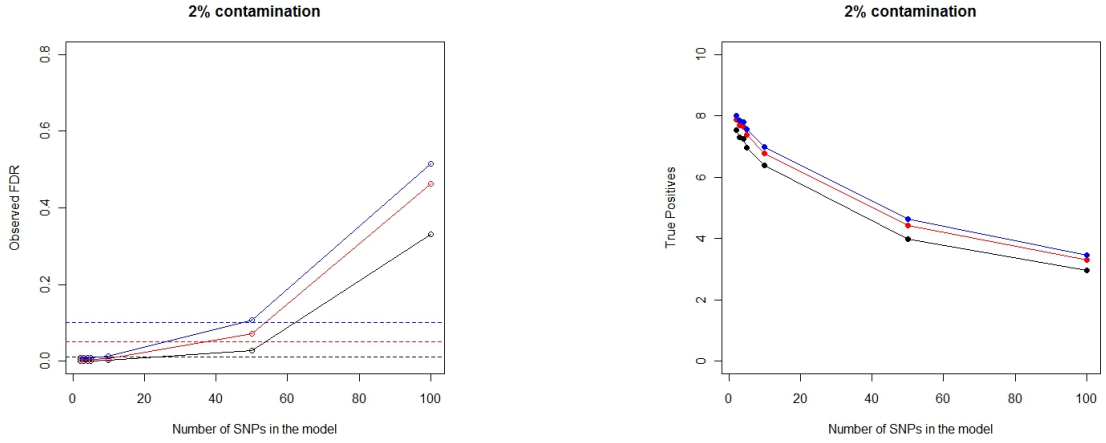


Figure 6.8: Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.5$, 2% of smooth data contamination and robust parameter of scale Q_n. Bonferroni correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust standardized residuals (4.32). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1.

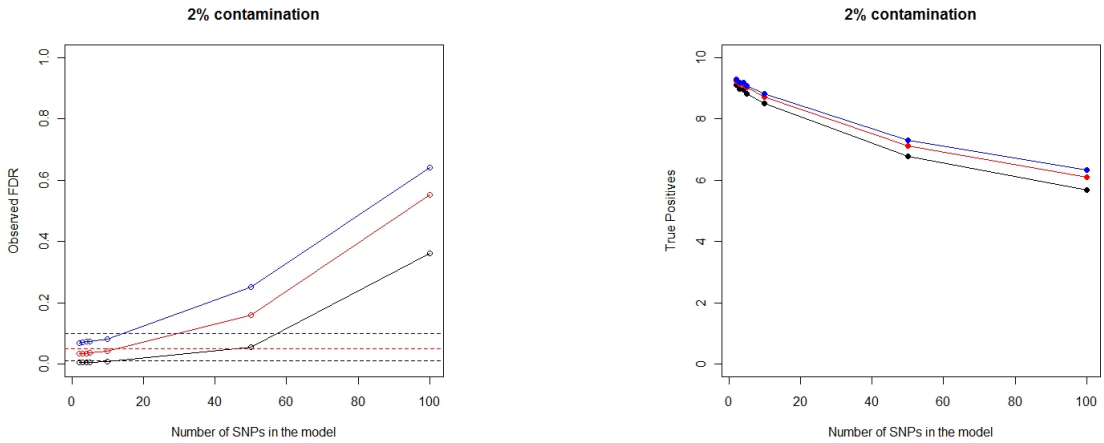


Figure 6.9: Observed FDR for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.75$, 2% of smooth data contamination and robust parameter of scale Q_n. Benjamini and Hochberg (1995) correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust standardized residuals (4.32). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1.

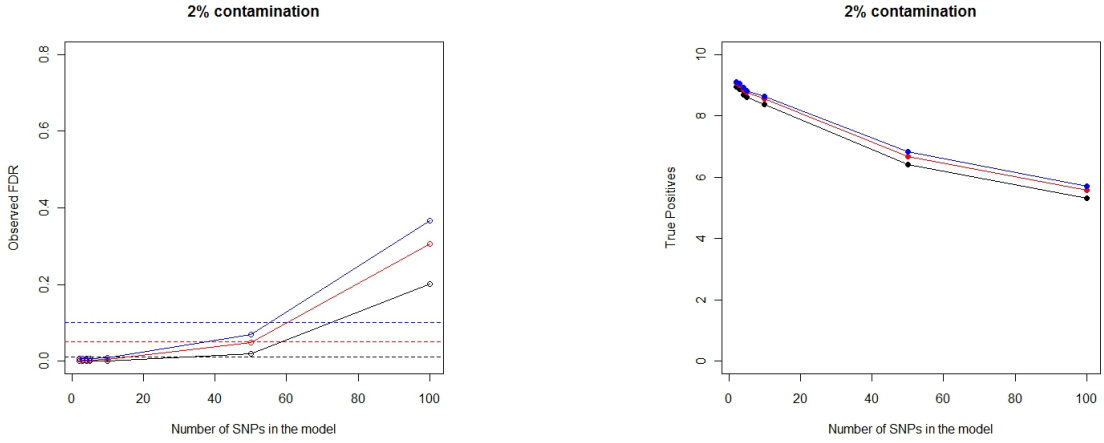


Figure 6.10: Observed FDR for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.75$, 2% of smooth data contamination and robust parameter of scale Q_n. Bonferroni correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust standardized residuals (4.32). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1.

Table 6.3: Values of the observed FDR obtained using Bonferroni correction at the nominal levels of $\alpha = 0.01,\ 0.05$ and 0.1, with the robust estimate of scale Q_n and $H^2 = 0.3$, for models with $n/5/2$ independent SNPs.

n	$n/5$	nSNPs	$\alpha \sum_{i=1}^{n} 1/i$			Observed FDR		
			0.01	0.05	0.1	0.01	0.05	0.1
250	50	25	0.0610	0.3050	0.6101	0.0535	0.1143	0.1713
500	100	50	0.0679	0.3396	0.6793	0.0636	0.1450	0.1967
750	150	75	0.0720	0.3599	0.7198	0.0777	0.1556	0.2166

We have seen this approach to also control the FDR at the level $\alpha(1 + \frac{1}{2} + ... + \frac{1}{n})$ in all simulations with $H^2 = 0.3$, $n = 500$ and up to 50 SNPs which is good in a way but not a solution since this threshold is much higher than the nominal levels of α. Any way, when we performed the outlier tests with $n = 250$ and 750 and nSNPs= 25 and 75, respectively, that control was not achieved simultaneously for all nominal levels of α considered (see Table 6.3 for different values of n) and thus, the hope that we could control the FDR at that level for models with up to nSNPs$\leq n/5/2$ did not succeed. We saw it to be possible only when we take $\alpha \geq 0.05$ in which case the observed FDR stayed over 10%, a percentage not at all acceptable for swamping. When we considered instead nSNPs$\leq n/10/2$ then we were able to control the FDR at that threshold simultaneously for all the α values (see Table 6.4), meaning that if we want to perform outlier tests in models over 10 and below $n/10/2$ independent SNPs, using Bonferroni correction with $\alpha = 0.01$ could be an option, having the FDR controlled at level $0.01(1 + \frac{1}{2} + ... + \frac{1}{n})$.

With the Benjamini and Hochberg FDR we were also able to control the FDR at the nominal α levels but only when using Q_n as the robust scale estimate. Control of the FDR at levels $\alpha(1 + \frac{1}{2} + ... + \frac{1}{n})$ as well, was not attained simulaneously for all considered α values in the simulations with $n = 500$ for models with 50 SNPs or more (results not shown) and therefore we cannot assure it in general. Nevertheless, since this approach got the highest rate of true positives detected it is therefore the one we advise should be used in these tests when the total of SNPs in the model does not go over 10 and, for the time being, when ≥ 500 individuals are used in the study.

III. Simulations considering $H^2 = 0.3$; nSNPs= $2, 3, 4, 5, 10$ only; Q_n only; 2% contamination; BH-2000, BY, ST, BKY.

It is now time to see if we can improve power while still keeping FDR control at the nominal levels of α. According to the FDR procedures discussed in Chapter 3, Section 3.2, we have several alternatives to BH-1995 that allow us to increase power, in this case

Table 6.4: Values of the observed FDR obtained using Bonferroni correction at the nominal levels of $\alpha = 0.01,\ 0.05$ and 0.1, with the robust estimate of scale Q_n and $H^2 = 0.3$, for models with $\lceil n/10/2 \rceil$ independent SNPs.

n	$n/10$	nSNPs	$\alpha \sum_{i=1}^{n} 1/i$			Observed FDR		
			0.01	0.05	0.1	0.01	0.05	0.1
250	25	13	0.0610	0.3050	0.6101	0.0118	0.0488	0.0751
500	50	25	0.0679	0.3396	0.6793	0.0089	0.0374	0.0618
750	75	38	0.0720	0.3599	0.7198	0.0145	0.0420	0.0677

the number of true positives detected. Four such methods are e.g, the ones of the BH-2000, BY, ST and BKY. Simulations were made to assess the control of the FDR at the pre-specified α levels in models with up to 10 SNPs. The number of true outliers was also recorded in order to provide a comparison of the methods.

From Figure 6.11 we can see that not all methods allowed for the control of the FDR at the nominal levels of α considered: ST procedure failed its control at the 5% level in the 10 SNP model; BH-2000 and BKY procedures, both reached the 5% level in the 10 SNP model and BY procedure controlled the FDR way below the nominal levels of α considered. This could of course be a good thing if the number of true detected outliers was big, which was not the case. It is not even better than the correspondent results obtained using Bonferroni correction (compare Tables 6.2 and 6.5).

Among the three FDR controlling methods, the ones of BH-2000 and BKY come up with more true outliers detected (Figure 6.12). Looking at Table 6.5, it is not clear which one is best, for their results are quite similar to one another. As to the number of false negatives observed in each one, they're also quite close (results not shown). So, we cannot say which one is actually best either from the number of true outliers detected or the number of false positives detected point of view. However they both look slightly superior when compared to the true outliers detected with BH-1995 FDR correction (compare Tables 6.1 and Table 6.5).

IV. Simulations considering $H^2 = 0.3$; nSNPs= 10; Q_n only; $2, 5$ and 10% contamination; BH-2000 correction and $n = 150, 200, 250, ..., 1000$.

We will now perform new simulations with BH-2000 correction varying the number of individuals and the percentage of contamination. Note that 1000 individuals in a quantitative trait association study is usually too much, i.e, would be to good to be true! We usually do not have much more than 500 individuals after quality control has been made, e.g, individuals with insufficient genotype call rates are usually eliminated from

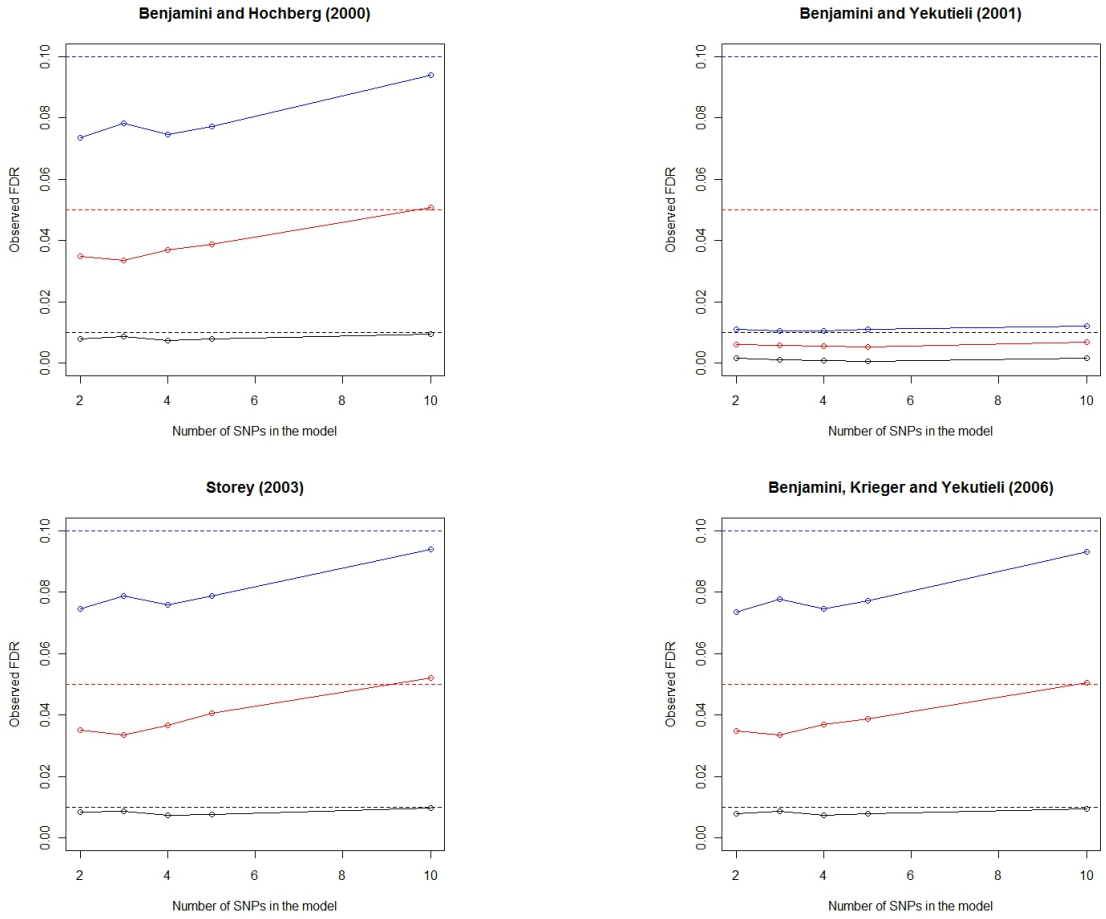


Figure 6.11: Observed FDR in models with up to 10 SNPs for several alternatives to the BH-1995 FDR correction. Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1.

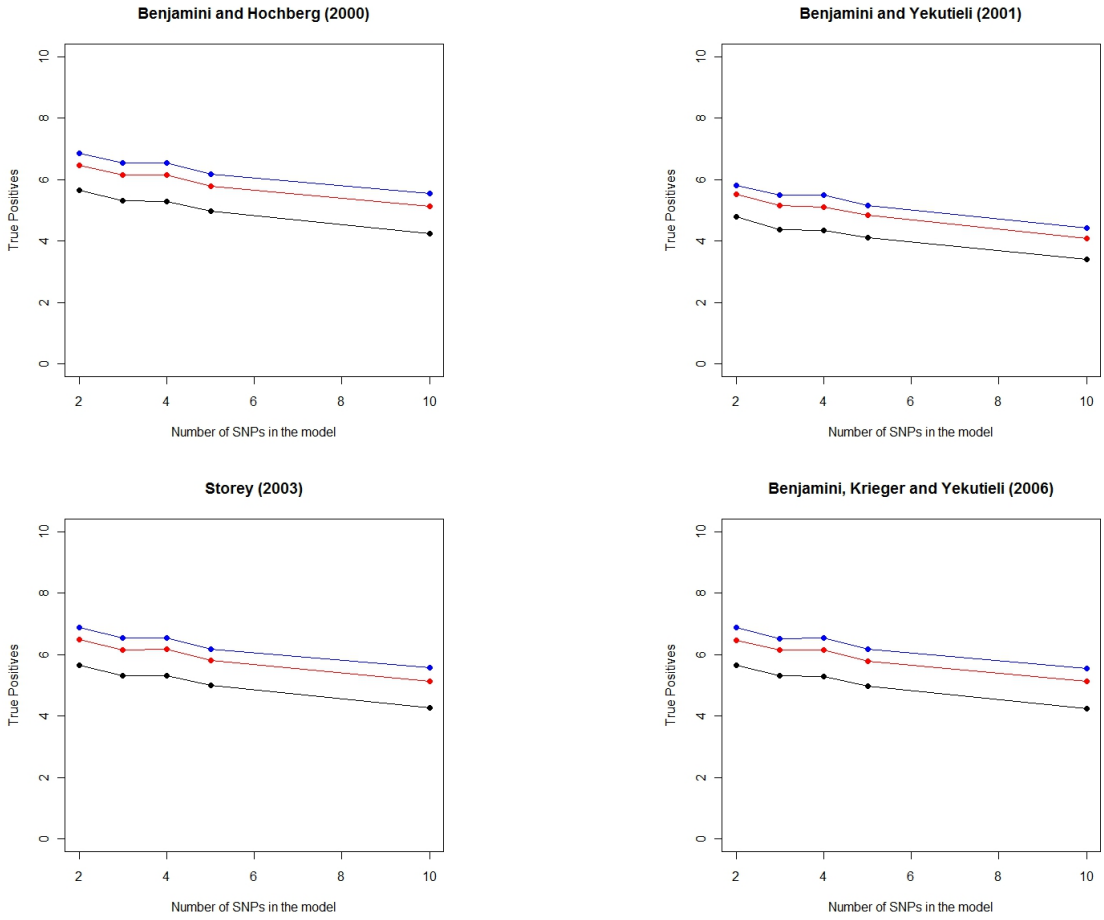


Figure 6.12: Observed number of True Positives in models with up to 10 SNPs for several alternatives to the BH-1995 FDR correction. Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1.

Table 6.5: Number of detected true positives corresponding to Figure 6.12 for the three FDR controlling approaches, where there is a total of 10 outliers.

	True Positives			
	nSNPs	0.01	0.05	0.1
BH-2000	2 SNPs	5.644	6.471	6.865
	3 SNPs	5.300	6.135	6.528
	4 SNPs	5.287	6.146	6.531
	5 SNPs	4.974	5.786	6.164
	10 SNPs	4.242	5.125	5.553
BY	2 SNPs	4.797	5.507	5.820
	3 SNPs	4.370	5.151	5.503
	4 SNPs	4.352	5.108	5.503
	5 SNPs	4.119	4.871	5.159
	10 SNPs	3.392	4.093	4.421
BKY	2 SNPs	5.642	6.470	6.866
	3 SNPs	5.298	6.134	6.524
	4 SNPs	5.287	6.146	6.527
	5 SNPs	4.975	5.785	6.164
	10 SNPs	4.242	5.123	5.552

the study (Devlin and Roeder, 1999/2001; Zheng *et al.*, 2006). Also, we have started the simulations at $n = 150$ because below that value for a 10 SNP model with 2% contamination it is not possible to control the FDR at the nominal levels (results not shown).

From the plots in Figure 6.13, we acknowledge the control of the FDR to be possible only in models with over $n = 400$ individuals, where an exceedance of less of one percentage point is sometimes observed for the nominal levels of $\alpha = 0.01$ and 0.05. As to the number of true outliers detected (Figure 6.14) they always keep way below the total number of outliers in the model. However, from Figure 6.15 we see that when considering $\alpha = 0.1$ we manage to detect over half of that total. This situation is not verified for $\alpha = 0.05$ as can be seen in the 10% contamination setting in Figure B.18, nor for $\alpha = 0.01$ (results nor shown). The plots also show that as the level of contamination increases the control of the FDR begins to be attained earlier in terms of the number of individuals in the study. At the same time, the number of true detections sees a slight fall but still keeps above the half number detection for the nominal level $\alpha = 0.1$.

Final considerations.

To conclude, we advise outlier tests to be conducted in the context of robust multiple linear regression. Being such, if the total number of SNPs in the model still exceeds 10 independent SNPs, we suggest that outlier tests are performed solely with the SNPs that

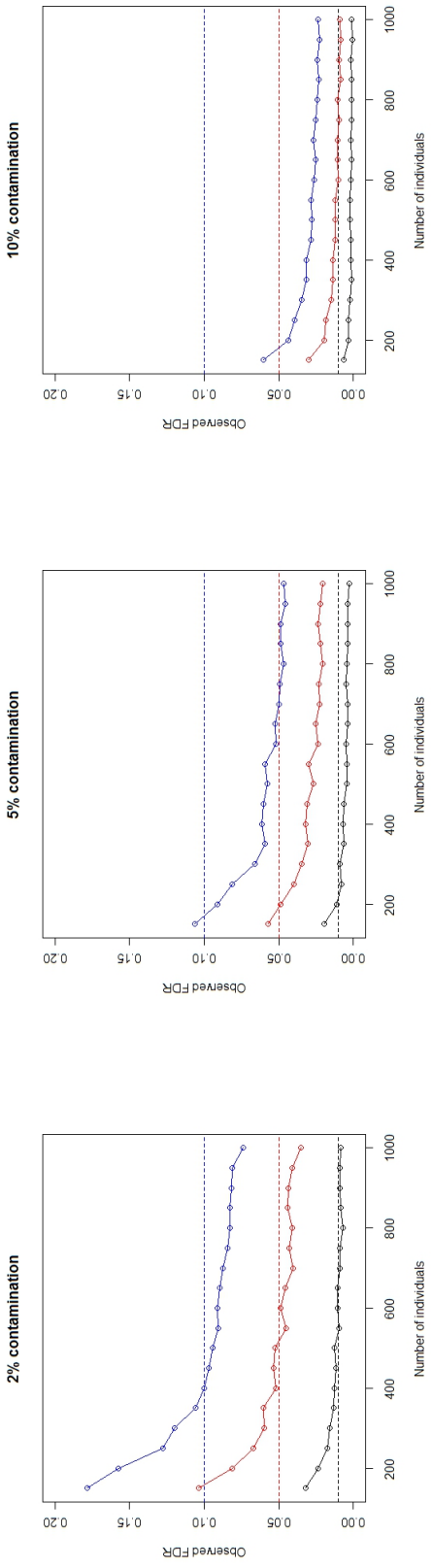


Figure 6.13: FDR observed values for nominal levels of $\alpha = 0.01, 0.05, 0.1$ in models with 10 SNPs. BH-2000 FDR correction was used. Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1.

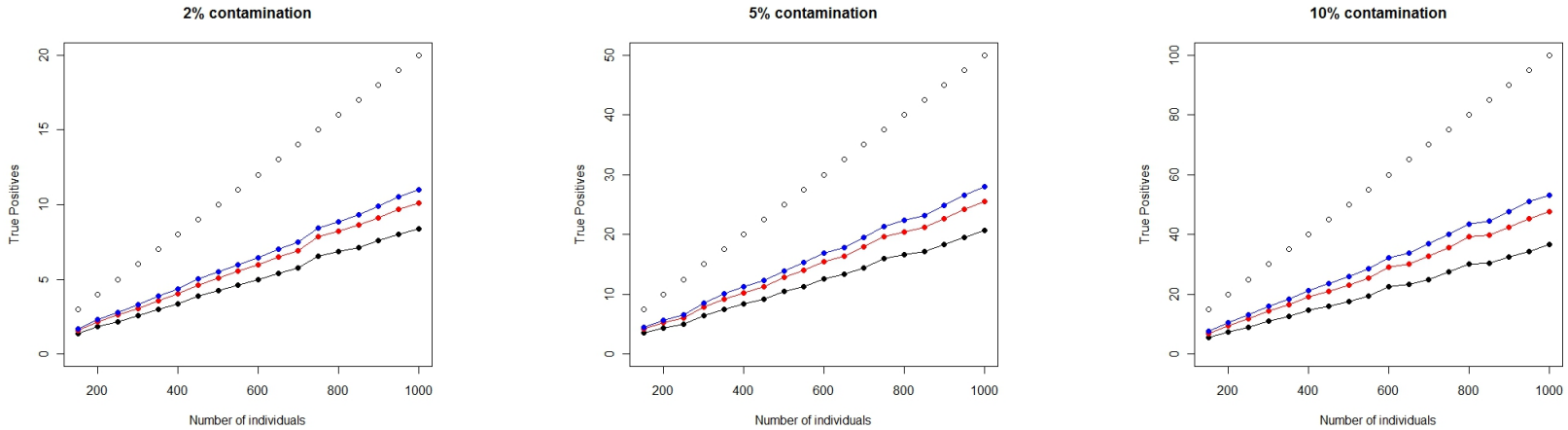


Figure 6.14: Number of True Outliers observed for nominal levels of $\alpha = 0.01, 0.05, 0.1$ in models with 10 SNPs. BH-2000 FDR correction was used. Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1. Black dots refer to the total amount of outliers in the model.

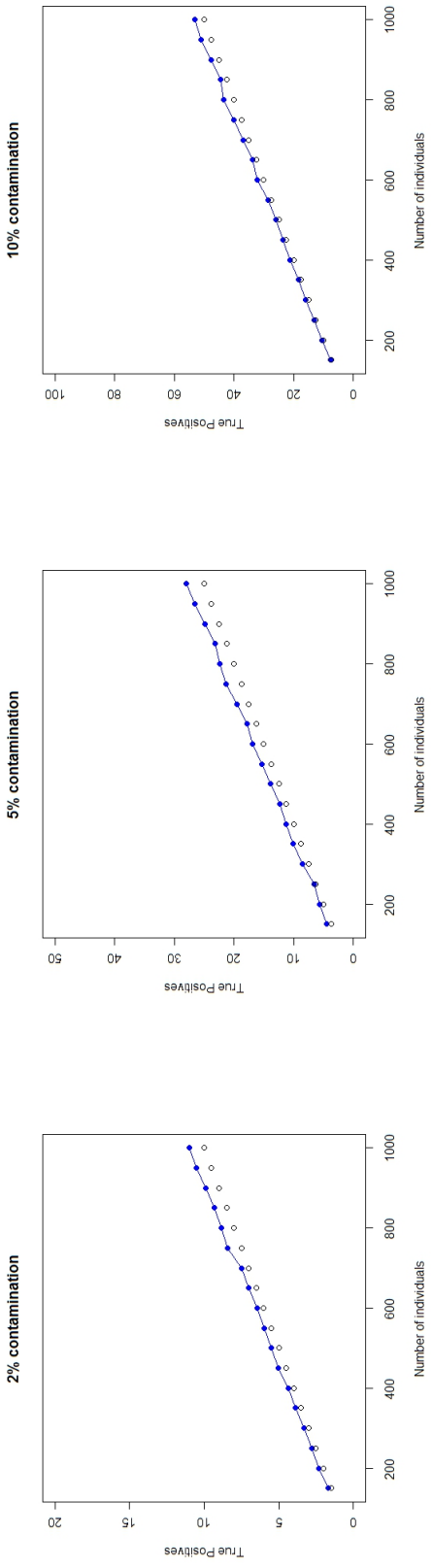


Figure 6.15: Number of True Outliers observed for nominal level of $\alpha = 0.1$ in models with 10 SNPs. BH-2000 FDR correction was used.

were declared significant in the association tests. This recommendation has to do with the inability of the evaluated correction measures in controlling the FDR in models above 10 SNPs.

When the total number of significantly declared SNPs from the multiple-SNP analysis stays above 10, then the user should opt for the Bonferroni correction together with the robust estimate of scale Q_n for the standardized robust residuals, knowing that up to 50 SNPs, for a nominal FDR of α, the FDR is controlled at level $\alpha(1 + \frac{1}{2} + ... + \frac{1}{n})$. This is obviously a high price to pay in terms of swamping but is nonetheless an alternative.

Please remember that this results were obtained for a smooth contamination of the data, which is usually the case where outlier tests have difficulties in distinguishing true outliers from regular observations. Should the data have been contaminated more drastically, the true number of detections would most certainly be higher. Do not forget as well that both FDR control and true detection improve as the trait heritability increases, and that brings us to the next section..

6.2 Coefficient of determination and trait heritability

In the previous section we have established that it is possible to perform outliers tests with control of the FDR at the usual nominal levels, in models with up to 10 independent SNPs, with a minimum number of 400 individuals using BH-2000 FDR correction as long as the trait heritability is $\geq 30\%$ (simulations not made for lower values of H^2). We also saw in this case, that fixing the FDR threshold at 10% we could identify over half the total of contaminant observations. However we do not always know the value of the trait heritability and therefore, in this section we will study how well the robust coefficient of determination reflects the heritability of a trait.

In order to answer that question, we therefore simulated several models with $N = 2, 3, 4, 5, 10, 50$ and 100 independent SNPs under $0, 2, 5$ and 10% smooth contamination scenarios for trait heritabilities of $H^2 = 0.1, 0.3, 0.5$ and 0.75.

Results presented in Table 6.6 show that under no contamination both measures (see Section 4.5) approximate the real value of H^2 up to the 10 SNP model. Above that, we see a tendency of the robust measure to inflate the value of the adjusted R^2. See for example in the $H^2 = 0.1$ case the 100 SNP model where the robust R^2 overestimates H^2 in approximately 14%, meaning that an adaptive correction should be used in this

Table 6.6: Mean values of the adjusted coefficient of correlation R_a^2 for all the models considered in the 1000 simulations with varying heritabilities and percentage of data smooth contamination - $n = 500$

	Contamination							
H^2=0.1	**0**%		**2**%		**5**%		**10**%	
	classic	robust	classic	robust	classic	robust	classic	robust
2 SNPs	0.1012	0.1031	0.0873	0.0991	0.0737	0.0971	0.0581	0.0916
3 SNPs	0.0986	0.1005	0.0870	0.1001	0.0748	0.0961	0.0575	0.0883
4 SNPs	0.0985	0.1011	0.0845	0.0992	0.0742	0.0955	0.0595	0.0886
5 SNPs	0.1018	0.1043	0.0894	0.1001	0.0760	0.0974	0.0596	0.0878
10 SNPs	0.0985	0.1065	0.0905	0.1050	0.0778	0.0999	0.0622	0.0910
50 SNPs	0.1015	0.1482	0.0934	0.1418	0.0850	0.1358	0.0739	0.1283
100 SNPs	0.0997	0.2361	0.0982	0.2343	0.0883	0.2261	0.0790	0.2132
H^2=0.3								
2 SNPs	0.3015	0.3043	0.2206	0.2995	0.1572	0.2971	0.1040	0.2917
3 SNPs	0.2976	0.2997	0.2208	0.2990	0.1623	0.2954	0.1063	0.2855
4 SNPs	0.2979	0.3001	0.2263	0.2988	0.1612	0.2947	0.1096	0.2854
5 SNPs	0.3021	0.3048	0.2289	0.2997	0.1668	0.2950	0.1120	0.2826
10 SNPs	0.2975	0.3040	0.2344	0.3026	0.1763	0.2960	0.1194	0.2827
50 SNPs	0.2984	0.3369	0.2589	0.3260	0.2145	0.3131	0.1644	0.2892
100 SNPs	0.3008	0.4057	0.2725	0.3960	0.2329	0.3779	0.1879	0.3469
H^2=0.5								
2 SNPs	0.5011	0.5032	0.3189	0.4987	0.2049	0.4964	0.1244	0.4914
3 SNPs	0.4975	0.4997	0.3204	0.4991	0.2134	0.4965	0.1289	0.4872
4 SNPs	0.4981	0.5006	0.3298	0.4993	0.2119	0.4958	0.1329	0.4875
5 SNPs	0.5015	0.5046	0.3348	0.5003	0.2210	0.4956	0.1369	0.4848
10 SNPs	0.4973	0.5022	0.3465	0.5013	0.2378	0.4959	0.1477	0.4841
50 SNPs	0.5001	0.5260	0.4020	0.5155	0.3156	0.5010	0.2243	0.4727
100 SNPs	0.4988	0.5751	0.4327	0.5624	0.3556	0.5407	0.2707	0.4886
H^2=0.75								
2 SNPs	0.7504	0.7345	0.4119	0.7322	0.2425	0.7312	0.1383	0.7288
3 SNPs	0.7484	0.7473	0.4154	0.7470	0.2542	0.7458	0.1447	0.7399
4 SNPs	0.7489	0.7508	0.4291	0.7494	0.2525	0.7473	0.1491	0.7421
5 SNPs	0.7505	0.7528	0.4375	0.7503	0.2651	0.7473	0.1544	0.7408
10 SNPs	0.7481	0.7508	0.4576	0.7502	0.2896	0.7475	0.1683	0.7399
50 SNPs	0.7497	0.7626	0.5577	0.7559	0.4119	0.7456	0.2753	0.7240
100 SNPs	0.7492	0.7873	0.6138	0.7774	0.4801	0.7601	0.3465	0.7211

case, whereas the classic adjusted R^2 still gives a good approximation (within ± 0.0003). With the introduction of contamination in the data, the classic adjusted coefficient of determination no longer approximates the trait heritability H^2 (see in Table 6.6 the 10% setting). The robust measure on the other hand, although decreasing with the increase of contamination, up to 50 SNPs kept around $H^2 \pm 0.05$, the worst situation being in the $H^2 = 0.1$ case with 50 SNPs. That difference gets higher with high dimensionality as seen in the no contamination setting. All in all, remembering that a model with 50 SNPs implies the estimation of $100 + 1$ parameters (one for the intercept), results show us that up to a ratio of $n/5$ parameters (intercept not included), the robust adjusted R_a^2 is a good indicator of the trait heritability.

Table 6.7: Mean values of the adjusted robust coefficient of correlation R_a^2 for all the models considered in the 1000 simulations with varying heritabilities and percentage of data smooth contamination. Only 20% of the SNPs in the models are in association with the trait.

	$n = 500$		**Contamination**			
	Total SNPs	20% SNPs	**0**%	**2**%	**5**%	**10**%
$H^2 = 0.1$	10 SNPs	2 SNPs	0.1077	0.1043	0.1026	0.0933
	15 SNPs	3 SNPs	0.1109	0.1081	0.1035	0.0948
	20 SNPs	4 SNPs	0.1154	0.1125	0.1096	0.0094
	25 SNPs	5 SNPs	0.1226	0.1175	0.1092	0.1008
	50 SNPs	10 SNPs	0.1465	0.1474	0.1341	0.1176
	100 SNPs	20 SNPs	0.2293	0.2297	0.2223	0.2128
$H^2 = 0.3$	10 SNPs	2 SNPs	0.3084	0.3011	0.2976	0.2860
	15 SNPs	3 SNPs	0.3104	0.3048	0.3010	0.2848
	20 SNPs	4 SNPs	0.3123	0.3098	0.2985	0.2834
	25 SNPs	5 SNPs	0.3169	0.3129	0.3024	0.2804
	50 SNPs	10 SNPs	0.3351	0.3273	0.3093	0.2759
	100 SNPs	20 SNPs	0.4019	0.3896	0.3613	0.3023
$H^2 = 0.5$	10 SNPs	2 SNPs	0.5051	0.4999	0.4967	0.4853
	15 SNPs	3 SNPs	0.5087	0.5039	0.4967	0.4839
	20 SNPs	4 SNPs	0.5098	0.5049	0.4966	0.4825
	25 SNPs	5 SNPs	0.5114	0.5087	0.4993	0.4781
	50 SNPs	10 SNPs	0.5252	0.5170	0.5021	0.4694
	100 SNPs	20 SNPs	0.5728	0.5580	0.5321	0.4464
$H^2 = 0.75$	10 SNPs	2 SNPs	0.7353	0.7328	0.7314	0.7258
	15 SNPs	3 SNPs	0.7522	0.7498	0.7454	0.7381
	20 SNPs	4 SNPs	0.7551	0.7520	0.7475	0.7388
	25 SNPs	5 SNPs	0.7561	0.7545	0.7494	0.7363
	50 SNPs	10 SNPs	0.7632	0.7578	0.7490	0.7283
	100 SNPs	20 SNPs	0.7864	0.7771	0.7597	0.6977

We must not forget though, that the SNP models considered contained all and only the SNPs responsible for the trait heritability. But what happens if those SNPs are among a bigger set of SNPs, e.g, if they represent only a total of $10, 20\%$ of the total number of SNPs in the model? Do we still get a robust coefficient of determination that approximates H^2?

Tables 6.7 and 6.8 tell us that up to a total of 50 independent SNPs in a model, if all the SNPs responsible for the trait are amongst them, then the robust coefficient of determination still approximates the real value of H^2, overestimating this parameter with no more than 5% error (the worst case being the 0% contamination setting in $H^2 = 0.1$).

From Tables 6.6, 6.7 and 6.8 it is also clear that the accuracy of the robust adjusted R^2 improves with the increase of the real trait heritability, being a good approximation

Table 6.8: Mean values of the adjusted robust coefficient of correlation R_a^2 for all the models considered in the 1000 simulations with varying heritabilities and percentage of data smooth contamination. Only 10% of the SNPs in the models are in association with the trait.

	$n = 500$		**Contamination**			
	Total SNPs	10% SNPs	**0**%	**2**%	**5**%	**10**%
$H^2 = 0.1$	20 SNPs	2 SNPs	0.1154	0.1104	0.1061	0.0957
	30 SNPs	3 SNPs	0.1245	0.1240	0.1157	0.1031
	40 SNPs	4 SNPs	0.1344	0.1310	0.1218	0.1083
	50 SNPs	5 SNPs	0.1453	0.1422	0.1337	0.1204
	100 SNPs	10 SNPs	0.2351	0.2254	0.2188	0.2015
$H^2 = 0.3$	20 SNPs	2 SNPs	0.3122	0.3072	0.2989	0.2822
	30 SNPs	3 SNPs	0.3190	0.3160	0.3048	0.2800
	40 SNPs	4 SNPs	0.3271	0.3211	0.3054	0.2753
	50 SNPs	5 SNPs	0.3356	0.3276	0.3115	0.2752
	100 SNPs	10 SNPs	0.4050	0.3860	0.3579	0.2845
$H^2 = 0.5$	20 SNPs	2 SNPs	0.5078	0.5041	0.4971	0.4816
	30 SNPs	3 SNPs	0.5138	0.5112	0.5017	0.4793
	40 SNPs	4 SNPs	0.5199	0.5149	0.5015	0.4734
	50 SNPs	5 SNPs	0.5265	0.5187	0.5043	0.4682
	100 SNPs	10 SNPs	0.5749	0.5577	0.5286	0.4393
$H^2 = 0.75$	20 SNPs	2 SNPs	0.7353	0.7328	0.7314	0.7258
	30 SNPs	3 SNPs	0.7522	0.7498	0.7454	0.7381
	40 SNPs	4 SNPs	0.7551	0.7520	0.7475	0.7388
	50 SNPs	5 SNPs	0.7561	0.7545	0.7494	0.7363
	100 SNPs	10 SNPs	0.7632	0.7578	0.7490	0.7283

even in the 100 SNP model (see values of R_a^2 in all tables when $H^2 = 0.75$).

Another interesting question would be, what if among a set of SNPs only a few are responsible for a proportion of the trait variation? Will the robust coefficient of determination drop by that proportion?

We considered the situation where only half of the SNPs responsible for the trait variation were present. Results in Table 6.9, as expected, show the coefficient of determination to drop, but only to a bit less than a half thus still reflecting the amount of genetic variation that is explained by the causative SNPs that remained in the model. Let's not forget that in our simulation model we have considered all the SNPs used to generate the quantitative trait to contribute unequally but approximately in the same amount to the trait variation. And this fact explains why we don't exactly get half the value of the robust coefficient of determination when we take from the model half of the responsible SNPs.

Table 6.9: Mean values of the adjusted robust coefficient of correlation R_a^2 for all the models considered in the 1000 simulations with varying heritabilities and percentage of data smooth contamination. Not all the SNPs in the model were used to generate the quantitative trait.

$n = 500$		**Contamination**			
	SNPs*	**0**%	**2**%	**5**%	**10**%
$H^2 = 0.1$	1 SNPs	0.0433	0.0429	0.0414	0.0395
	2 SNPs	0.0457	0.0453	0.0433	0.0403
	5 SNPs	0.0527	0.0519	0.0491	0.0452
	25 SNPs	0.0714	0.0667	0.0638	0.0608
$H^2 = 0.3$	1 SNPs	0.1288	0.1281	0.1262	0.1241
	2 SNPs	0.1349	0.1344	0.1316	0.1280
	5 SNPs	0.1503	0.1487	0.1451	0.1395
	25 SNPs	0.1677	0.1599	0.1530	0.1427
$H^2 = 0.5$	1 SNPs	0.2179	0.2171	0.2150	0.2127
	2 SNPs	0.2371	0.2364	0.2309	0.2270
	5 SNPs	0.2426	0.2418	0.2381	0.2328
	25 SNPs	0.2652	0.2601	0.2521	0.2295
$H^2 = 0.75$	1 SNPs	0.3746	0.3728	0.3704	0.3680
	2 SNPs	0.3614	0.3601	0.3556	0.3506
	5 SNPs	0.3626	0.3630	0.3580	0.3522
	25 SNPs	0.3859	0.3805	0.3729	0.3478

* The quantitative trait was generated with twice the number of SNPs in the model.

We have thus showed that in the context of robust multiple linear regression where we have a set of $N \leq 50$ independent SNPs, the robust adjusted coefficient of determination proposed and studied is a good approximation of the proportion of phenotypic variation that is explained by the genetic variation among individuals. Should one have before hand, via family based studies or other, an idea as to what is the heritability of the trait under study, then the difference between the value of the robust coefficient of determination and the hypothesized trait heritability gives an indication of how much of the trait heritability was left unexplained, giving us a motif to look for that variability somewhere else in the species' genome.

Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.

John Tukey

7 Example

Zea mays ssp. mays, commonly addressed to as maize or, in many English-speaking countries, simply as corn, is an important crop species, if not the most important, for besides feeding the world's population and livestock, it also provides raw materials for several industries. One can hardly imagine the use of corn in industry. Not only corn is in most processed food like hamburgers, soda, doughnuts but also in farmed fish, fruit wax, sweeteners, ethanol (ethyl alcohol), corn plastic and so forth[1]. It definitely has the highest production of any crop in the world and in particular, it plays a central role in all of US agriculture and food production.

At the diversity level, maize is the crop species presenting more variation in morphological traits and showing greater extension of DNA polymorphisms: modern maize lines usually have a SNP every $\sim$ 100 bp (Tenaillon *et al.*, 2001). Understanding the mechanisms underlying both morphological and genetic diversity of maize is critical for future plant breeding and of the utmost interest to scientists and plant breeders.

Teosintes are wild growing relatives of maize. Amongst those, *Zea mays ssp. parviglumis*, or Balsas teosinte (because the largest populations grow in the river Balsas basin

[1]Out of curiosity, do watch the interview at `http://abcnews.go.com/GMA/OnCall/story?id=4439943page=1` and also the documentary "King corn" easily found on YouTube.

in southwestern Mexico), is believed to be maize's wild ancestor (White and Doebley, 1998; Galinat, 2001; Doebley, 2004). The domestication of maize around 9000 years ago and the bottlenecks it experienced somewhere in between, first due to the limited pool of founder plants and afterwards because of artificial selection, reduced the genetic diversity of maize relative to its ancestor Balsas teosinte (Eyre-Walker *et al.*, 1998; Wright *et al.*, 2005; Buckler and Stevens, 2006). Two major key questions therefore naturally follow

How natural and artificial selection have shaped the molecular diversity of maize?

and

How does this molecular diversity relate to functional trait variation?

To this purpose, among the certainly many researches in the field, we highlight the ongoing project "The Genetic Architecture of Maize and Teosinte" (formerly entitled "Molecular and Functional Diversity in the Maize Genome")[2], funded by the National Science Foundation, where one of the five proposed goals is to identify genes that control domestication traits and three key agronomic traits: flowering time, plant height, and kernel quality. To achieve this goal, populations are being genotyped for SNPs in genes that putatively control the traits of interest and association genetic studies are being performed to assess if those associations are valid or not and eventually find new associations worthwhile investigating further. Altogether, data and results obtained from this project are being released at the project website `http://www.panzea.org` and are freely available for download.

In Chapter 5, we saw that the robust methodology presented itself quite promising for genetic association studies of quantitative traits, although in a very specific setting. It thus makes sense to now apply it to some real data example, preferably one that has already been studied in the literature. This is where the "The Genetic Architecture of Maize and Teosinte" project comes in: we searched the project site for some published work with correspondent available data set. This Chapter is therefore dedicated to the replication of the results obtained in the selected paper and confrontation of those results with the ones obtained when analyzing the data with a robust regression model approach.

7.1 Example data and methodology

As an example of application we downloaded the data of Weber *et al.* (2008) (see Zhao *et al.*, 2006, and `http://www.panzea.org`) on 493 maize plants, with respect to the

[2]Detailed information on this project can be found at `http://www.panzea.org` and in Zhao *et al.* (2006).

quantitative trait FERL (length of the female and hermaphroditic portions of the basal-most ear on the lateral branch), getting information on 61 SNPs from 22 candidate genes (see data description on Table A.1), chosen based on their possible effects on the trait under study given their known mutant phenotype in maize or other plants. The goal of Weber's work was to determine if major regulatory genes of maize contribute to natural variation in its wild ancestor Balsas teosinte and follows from previous work (Weber *et al.*, 2007). We will however restrict ourselves only to the analysis of the quantitative phenotype FERL.

In that paper, Weber studied the association between these SNPs and this Teosinte trait, among others, by adjusting the mixed linear model, as presented by Yu *et al.* (2006),

$$y = Pv + S\alpha + Iu + e, \tag{7.1}$$

where y is the vector of phenotypic values[3], v is a vector of fixed effects regarding population structure (inferred via PowerMarker - Liu and Muse, 2005), α is the fixed effect for the candidate SNP, u is a vector of random effects relative to recent coancestry, e a vector of residuals, P is a matrix of the 10 significant principal components (determined via EIGENSTRAT - Price *et al.*, 2006), S a vector of the SNP genotypes and I an identity matrix. The structure assumed for the variances is: $\text{var}(u) = 2KV_g$ and $\text{var}(e) = IV_R$, where K is the Kinship matrix (generated via PowerMarker), which quantifies the proportion of shared alleles, $V_g = \theta$ is the genetic variance and $V_R = \sigma^2$ is the residual variance. The pertinence of this mixed linear model is justified by the systematic sources for spurious associations found under the simple model,

$$y = S\alpha + e, \tag{7.2}$$

non uniformity of the p-values and high type-I error rates (Figure 7.1; the correspondent Q-Q normal plots can be seen in Figure B.19). The adequacy of the full model to test under the null hypothesis is also clear, with only $\approx 5.4\%$ by chance significant associations in both the classical and the robust methodologies.

Here, we configured SNP as a numerical covariate, which is the proper way to analyze this kind of data when testing under the null hypothesis (see Figure 7.2 for the classical analysis and the correspondent Figure B.20 for the robust analysis).

[3]In his study, Weber transformed, with a square root, 2 of the 31 phenotypic traits studied after examining residual plots; FERL was not one of them.

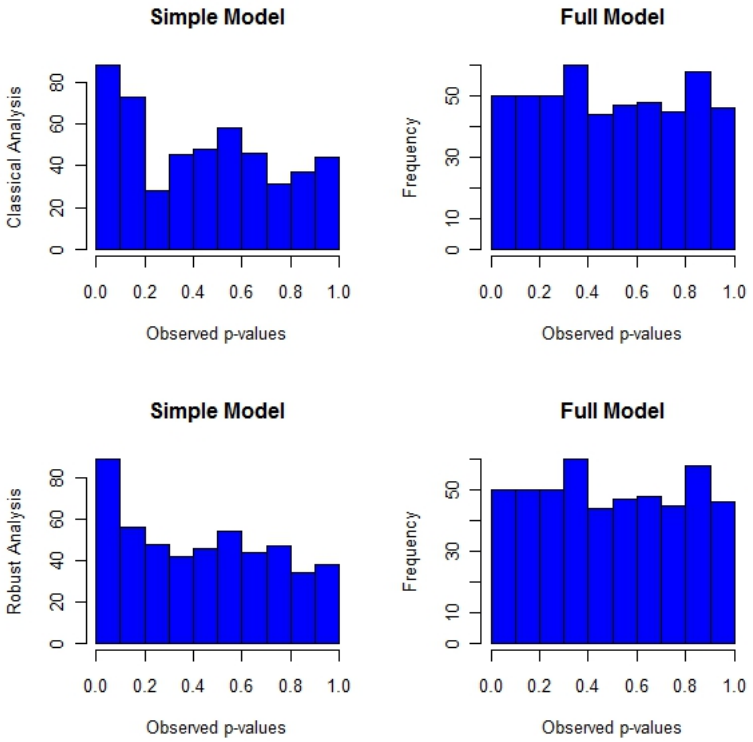


Figure 7.1: Histograms of the 498 p-values for both simple and full models, in the classical and robust analysis.

However, when we are not under the null, the codification of SNP as factor is more adequate in order to capture eventual dominance effects, unless of course the effects in the model are only additive, which may not always be the case. In these circumstances, we will therein consider SNP coded as factor. We started by rewriting model (7.1) as

$$Y^* = X^*\beta + \varepsilon^* \tag{7.3}$$

where $Y^* = y$, $X^* = Pv + S\alpha$ and $\varepsilon^* = Iu + e$, with $\text{var}(\varepsilon^*) = \sigma^2 V$, $V = K\frac{\theta}{\sigma^2} + I = K\theta_0 + I$ (i.e., we consider $\theta_0 = \theta/\sigma^2$). When V is known, and is positive semi-definite, taking its Cholesky decomposition, leads to

$$Y = X\beta + \varepsilon \tag{7.4}$$

where $Y = \text{chol}(V)^{-1}Y^*$, $X = \text{chol}(V)^{-1}X^*$, $\varepsilon = \text{chol}(V)^{-1}\varepsilon^*$ and now $\text{var}(\varepsilon) = \sigma^2 I$. The ML estimates of β and σ^2 may then easily be obtained.

Usually, the matrix V is unknown and θ_0 is estimated by ML, REML or some other method. We should note that Weber used the SAS Proc Mixed routine with REML, and we used ML from the R function `lmekin()` in package *kinship*, which may justify the slight differences between both results. Once an estimate of V, $\widehat{V}$, has been obtained, it

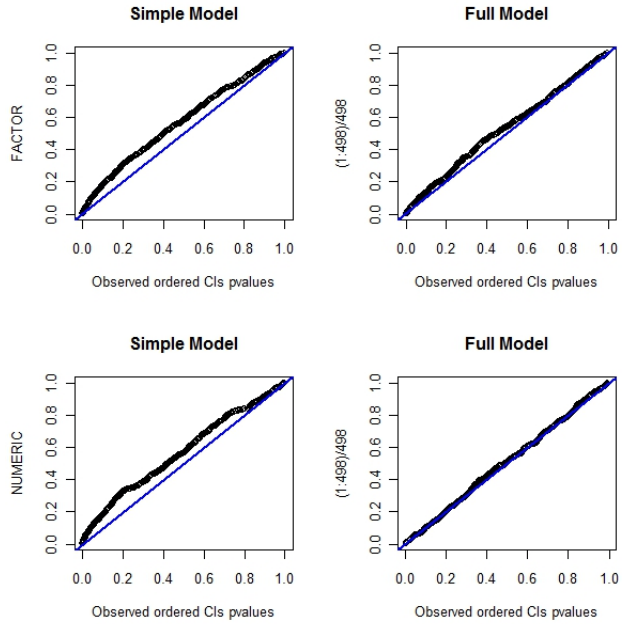


Figure 7.2: Cumulative distribution of the 498 p-values for both simple and full models, classical method.

is used in (7.4) so that the estimates of β and σ^2 can be calculated.

Inference for the fixed-effects terms, i.e., the tests $H_0 : H\beta = \mathbf{0}$ of $q \leq p$ fixed-effects, were conducted by the usual F statistic with degrees of freedom estimated by the Satterthwaite approximation (as in Weber *et al.*, 2008; see Sections 4.4.1; 4.4.2 and 4.7.2).

7.2 Single-SNP analysis

Following Weber's analysis, we performed the classical analysis described above for each of the 61 candidate SNPs with routine `lmekin()` from the R package *kinship* (see routine C.2). The robust analysis was performed with a robustified version of this instruction (see routine `lmekinR()` in C.3). This single-SNP analysis allowed for the detection of 9 associations in the classic analysis and 8 in the robust analysis at the 5% level (i.e, $p \leq 0.05$), all associations being significant after correction for multiple testing via FDR ($q \leq 0.1$) - Table 7.1.

In both analysis we have the 6 SNPs identified by Weber. Moreover, with the exception of zagl1.6, all other detected SNP are common between approaches. Also,

i. from the identified SNP, SNP PZD00073.5 and PZD00073.8 are not independent since they are in high LD ($r^2 = 0.6770$); all other observed pair-wise-LD is low ($r^2 \leq$

Table 7.1: Significant SNPs detected at the 5% level and with FDR under 10% in the single-SNP analysis.

	SNP	N	R^2_*	R^2_{LR}	$R^2_{LR_a}$	$2a/\sigma^2_P$	d/a	$2a$	d	p-value	q-value
	zagl1.1	487	0.021	0.103	0.080	0.176	-2.55	1.90	-2.43	0.0057	0.0212
C	PZD00073.5	474	0.015	0.107	0.084	0.839	-1.01	9.05	-4.57	0.0069	0.0212
L	PZD00006.1	465	0.006	0.114	0.091	0.301	-0.11	3.26	-0.17	0.0083	0.0212
A	PZD00022.3	455	0.021	0.112	0.088	0.436	0.87	4.72	2.06	0.0102	0.0212
S	PZD00073.8	457	0.017	0.110	0.086	0.934	-0.94	10.20	-4.82	0.0112	0.0212
S	te1.3	489	0.010	0.106	0.083	0.422	1.01	4.55	2.29	0.0208	0.0312
I	ba1.9	453	0.027	0.104	0.079	0.984	-0.86	10.60	-4.57	0.0267	0.0344
C	PZB00049.7	453	0.016	0.100	0.075	0.094	5.08	1.01	2.56	0.0336	0.0378
	zagl1.6	490	0.008	0.101	0.079	0.109	3.07	1.18	1.82	0.0499	0.0499
	zagl1.1	487	0.120	0.200	0.179	0.139	-3.15	1.50	-2.37	0.0067	0.0169
R	PZD00073.5	474	0.030	0.179	0.157	0.834	-1.03	9.00	-4.65	0.0090	0.0169
O	PZD00006.1	465	0.057	0.256	0.235	0.279	0.06	3.03	0.09	0.0083	0.0169
B	PZD00022.3	455	0.226	0.229	0.207	0.430	0.61	4.66	1.42	0.0092	0.0169
U	PZD00073.8	457	0.127	0.206	0.184	0.936	-1.05	10.22	-5.37	0.0106	0.0169
S	te1.3	489	0.014	0.281	0.261	0.273	0.96	2.95	1.41	0.0442	0.0201
T	ba1.9	453	0.002	0.210	0.187	0.985	-0.79	10.61	-4.21	0.0151	0.0450
	PZD00049.7	453	0.046	0.213	0.191	0.052	8.17	0.56	2.72	0.0450	0.0450

N is the number of individuals used in each analysis;
R^2_* are the R^2 values obtained via simple model (7.2) and as in Section 4.5;
R^2_{LR} and $R^2_{LR_a}$ are the unadjusted and adjusted R^2 values proposed for the mixed linear model in Section 4.7.3.

0.02) - we used `LD()` instruction from the R package *genetics* for the calculations (see Table A.2);

ii. only SNP PZD00006.1 shows an additive mode of inheritance ($-0.5 < d/a < 0.5$); zagl1.1, zagl1.6 and PZB00049.7 show overdominance ($|d/a| > 1.25$) and the remaining SNPs show partial or complete dominance ($0.50 < |d/a| < 1.25$);

iii. from the R^2_* values, which were the ones calculated by Weber *et al.*, we acknowledge that all individual effects are small, ranging from 0.8% to 2.7% in the classic analysis. Being true this could mean the marker assayed is not the causative site but is in LD with the causative site giving an underestimate R^2 of the real effect, the trait may have low heritability or the associations may be due to alleles of small effect. The robust counterpart in turn, signals SNPs zagl1.1, PZD00022.3 and PZD00073.8 with significant individual effects;

iv. if we look instead at the values of the unadjusted and adjusted R^2_{LR}, we see that all the SNPs detected in both analysis have larger values of R^2.

We additionally performed the Shapiro-Francia (SF) normality test, via the R instruction `sf.test()` from package *nortest*, on both the residuals from the classic and robust approaches and observed that they all failed the normality assumption (see Table 7.2 and

residual QQ-plots in Figures B.21 and B.22) showing heavy tails, specially on the left side of the distribution. Although it is not a pre-requisite in the robust analysis, residual normality is one of the classical assumptions. This violation indicates that either the classical analysis normality assumption we have made is not realistic or the model adjusted is not good. Either way, one should take care with false association detection (possibly the case of SNP zagl1.6 detected in the classical methodology but not in the robust) and possible reduction in power.

Table 7.2: Results from the Shapiro-Francia test on the models' residuals.

SNP	Classic	Robust
zagl1.1	0.0015	0.0002
PZD00073.5	0.0027	0.0006
PZD00006.1	0.0007	0.0002
PZD00022.3	0.0013	0.0002
PZD00073.8	0.0029	0.0006
te1.3	0.0051	0.0007
ba1.9	0.0014	0.0003
zagl1.6	0.0037	—
PZB00049.7	0.0020	0.0004

To end, we must underline that the quantitative trait "FERL" was the only that was not transformed to normality by Weber *et al.*. They argued that the normal Q-Q plots were approximately normal, therefore not justifying data transformation. If we had performed the Pearson chi-square normality test instead of the SF test, we too could not reject the normality assumption for each of these single-SNP analysis. And this happens because both normality tests differ mainly in the importance they give to the tails of the distribution, which is higher in the SF test as compared to the Pearson's.

7.3 Multiple-SNP analysis

It is known that a joint analysis of SNPs may be more adequate whenever we believe the trait of interest to be controlled by many genes, which is usually the case of quantitative traits, being much more informative than single-SNP analysis (Jannot *et al.*, 2003). Multiple regression can accommodate many explanatory variables that may somehow be correlated, allowing us to infer causality where single-SNP analysis did not. It has however some drawbacks and/or limitations, e.g, the number of explanatory variables must not be greater than the number of individuals in the sample and the large number of degrees of freedom usually involved inevitably implicates power loss.

In this particular case, we have a total of 61 SNPs each with either 2 or 3 categories, and all with missing observations ranging from 0.2% and 9.3%. This means that if we

consider the model with the 61 SNPs altogether, after eliminating from the model all observations for which the SNP genotype was not observed, we will end with a $p >> n$ situation which will derail the regression analysis as it is. One possible solution is to do data imputation where missing genotypes are replaced by predicted genotypes based on the genotypes of nearby SNPs (Little and Rubin, 2002; Souverein *et al.*, 2006; Balding, 2006). Another way to circumvent the dimensionality problem would be to do some previous TagSNP selection (Byng *et al.*, 2003; Stram, 2004; Carlson *et al.*, 2004; Halperin *et al.*, 2005; Gopalakrishnan and Qin, 2006; He and Zelikovsky, 2006; Nicolas *et al.*, 2006; Xu *et al.*, 2007) based on some criteria like e.g. LD, before doing the regression analysis, thus reducing the number of SNPs on the model hopefully enough as not to still have $p >> n$.

Since data imputation and TagSNP selection are not the scope of our work, we decided only to further investigate the SNPs detected in the single-SNP analysis now in a joint analysis. This procedure is in itself another way of avoiding the $p >> n$ issue, i.e, we consider for the multiple-SNP analysis only the SNPs that were significant in the single-SNP analysis.

The multiple-SNP analysis regarding the previously detected SNPs (Table 7.4), after correcting for multiple testing with the conservative Bonferroni correction (bold p-values), left us with only two significant SNPs in the classic analysis (PZD0006.1 and PZD00022.3) and three significant SNPs in the robust analysis (PZD0006.1, PZD00022.3 and $ba1.9$). If we consider the FDR at level 10% (bold q-values), as in Weber, then both methods detect SNPs PZD00006.1, PZD00022.3, ba1.9 and PZB00049.7. The classic analysis additionally detects SNP zagl1.6. In order to evaluate the models adequacy we plotted the conditional residuals plots (Figure 7.3), and performed the SF test of normality, verifying, again, the non-normality of both the classical ($p \simeq 0.034$) and the robust residuals ($p \simeq 0.003$) at the 5% level.

The plots in Figure 7.3 also show heavier tails on the left of the distributions (normal QQ-plots) and the presence of possible outlying observations (conditional residuals plots).

To conclude, we removed the outlying observations identified by the robust analysis either using a rule of thumb (in this case standardized robust residuals that deviated more than 3 standard deviations from the mean) or the outlier robust test described in Section 4.6, where the robust residuals of equation (4.32) were considered with the robust estimate of scale Q_n (equation (4.8)) and the multiple test corrections of Bonferroni and

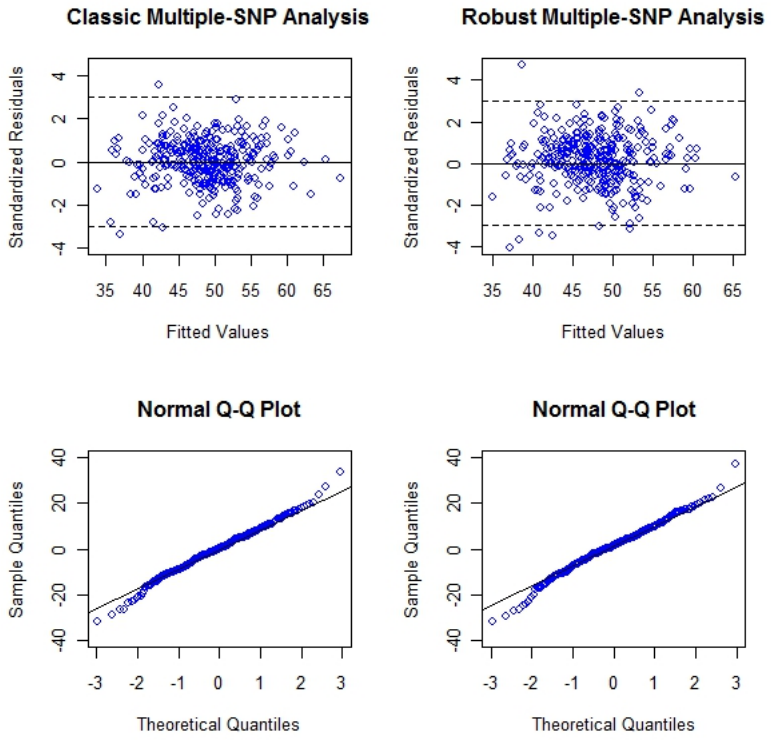


Figure 7.3: Conditional residual plots and QQ-normal plots from the multiple-SNP analysis.

BH-2000 at the 10% level (these choices are supported by the results obtained in the simulation study of Chapter 6).

In the first situation we identified a total of 8 outliers, namely TAMex0344/0719/0775/0802/0805/0807/0821/1534. On the other hand, via the robust outlier test, we identified as outliers only one plant from the previous outlier list using Bonferroni correction and two using BH-2000. More precisely, TAMex0344 in the first situation and TAMex0344/0775 in the second. This is not a surprising result since we have seen in Chapter 6 that the BH-2000 has more power in detecting outlying observations while controlling the FDR at the same time, relative to the use of the Bonferroni correction. We must note however that the simulations' conditions are not quite the same as the ones of this study. In particular, the SNPs in this model are not independent. As mentioned before, SNPs PZD00073.5 and PZD00073.8 are in high LD.

According to the influence diagnostic measures discussed in Chapter 4 we have for the eight plants the results displayed in Table 7.3. There we observe that plants TAMex0344/0775/0719 are the only ones among the total of eight that are declared influential by the three influence measures simultaneously. Also, TAMex0344/0775 are the ones with the

Table 7.3: Diagnostic measures for the influence of the eight plants identified as outliers by the rule of thumb. Last column refers to the weights given by the robust analysis.

Plant ID	Cook's D_i	|DFFITS|	|DFBETAS|	Weights
TAMex0344	**0.0185**	**0.7193**	YES	0.3672
TAMex0719	**0.0190**	**0.7262**	YES	0.4309
TAMex0775	**0.0170**	**0.6903**	YES	0.3340
TAMex0802	0.0018	0.2208	NO	0.5108
TAMex0805	0.0036	0.3165	YES	0.5211
TAMex0807	0.0103	0.5345	YES	0.4172
TAMex0821	0.0116	0.5648	YES	0.5448
TAMex1534	0.0091	0.5021	YES	0.4251

bold Cook values and |DFFITS| and YES |DFBETAS|, are the ones that are influent according to the thresholds $4/(n^* - p)$, $2\sqrt{p/n^*}$ and $2/\sqrt{n^*}$, respectively; p is the number of estimated parameters; n^* is the number of plants used in the multiple analysis.

smallest weights given by the robust analysis, meaning our robust outlier test with BH-2000 correction was able to identify among the plants whose residuals deviated the most from the bulk, the ones that were highly influencing simultaneously the fitted values, the estimated parameters and who had the smallest weights. We also observed through the DFBETAS values, that TAMex0344/0775 are influencing the estimated parameters of 3 SNPs each, whereas TAMex0719 is influencing only 2 SNPs (results not shown). The weight given by the robust analysis to TAMex0719 is not the third smallest of all. These two observations may be the reason why TAMex0719 was not declared an outlier in the robust test.

We then re-ran the classical analysis without those three sets of outliers in each time (Table 7.4,*, ** and ***). As an immediate result, we now had $p \simeq 0.9325$, $p \simeq 0.0895$ and $p \simeq 0.3873$ for the SF normality test in the three analysis, respectively. Also, we can see that SNP zagl1.6 no longer appears as a significant association in any of the new classic tests of association reinforcing the idea that it was in fact a false positive in the first classical analysis. Both ** and *** analysis provided the same results in terms of the SNPs detected, in a total of four.

It was expected though, that the classical analysis without the outlying observations would produce the same results as the robust analysis with all observations. However, in *, SNP PZB00049.7 now misses significance by merely 0.0092. This happens because the robust analysis gave the outlying observations weights between 0.36 and 0.55, whereas the classical approach without the outliers actually corresponds to giving zero weight to the outliers and one to the remaining observations. Moreover, we have seen in the simulation study of Chapter 5 that it is sometimes possible to have a residual distribution

approximately normal, i.e, passing the SF test, and still the classic approach losing power relative to the robust approach. We thus have reason to believe at this point that SNP PZB00049.7 may be a false negative of the classical analysis * without outliers, idea that is corroborated by the results of ** and ***. Furthermore, this situation clearly alerts us to the danger of removing outlying observations from the data. After all, outliers may contain some relevant information on the model that should not be disregarded.

To end, if we now take a look at the adjusted R^2 values of the multiple-SNP models in Table 7.4, we see that the robust approach has the highest value of all. So, from this observation and based on what was said before, we therefore have reasons to believe that the SNPs identified in the robust analysis should be the ones further investigated for association with FERL.

Table 7.4: Multiple-SNP analysis results of the SNPs declared significant in the single-SNP analysis. P-values are raw.

$R^2_{LR_a}$	0.1727		0.2918		0.1698		0.1929		0.1834	
R^2_{LR}	0.2418		0.3494		0.2409		0.2606		0.2521	
SNP	**Classic**		**Robust**		**Classic***		**Classic****		**Classic*****	
	p-value	q-value	p-value	q-value	p-value	q-value	p-value	q-value	p-value	q-value
zagl1.1	0.1510	0.1699	0.1656	0.2208	0.4125	0.4366	0.2169	0.2440	0.3636	0.4379
PZD00073.5	0.1075	0.1612	0.2432	0.2780	0.0607	0.1092	0.0954	0.1431	0.0825	0.1357
PZD00073.8	0.5895	0.5895	0.6252	0.6252	0.2994	0.3849	0.4285	0.4285	0.3939	0.4379
PZD00006.1	**0.0008**	**0.0076**	**0.00004**	**0.0003**	**0.0004**	**0.0032**	**0.0001**	**0.0011**	0.0004	**0.0019**
PZD00022.3	**0.0042**	**0.0193**	**0.0007**	**0.0029**	**0.0040**	**0.0178**	**0.0003**	**0.0014**	0.0002	**0.0019**
te1.3	0.1356	0.1699	0.0910	0.1456	0.4366	0.4366	0.1316	0.1692	0.4379	0.4379
ba1.9	0.0138	**0.0413**	**0.0055**	**0.0147**	0.0090	**0.0272**	0.0081	**0.0245**	0.0135	**0.0405**
zagl1.6	0.0327	**0.0736**	——	——	0.1849	0.2774	0.0771	0.1387	0.0905	0.1357
PZB00049.7	0.0535	**0.0962**	0.0243	**0.0487**	0.0537	0.1092	0.0294	**0.0661**	0.0389	**0.0875**

***** After removing the 8 outliers identified with the rule of thumb;

****** After removing the only outlier identified with the robust outlier tests using Bonferroni correction with robust Q_n;

******* After removing the two outliers identified with the robust outlier tests using BH-2000 correction with robust Q_n;

Bold p-values refer to those that pass Bonferroni correction at the 5% level;

Bold q-values refer to those that pass a FDR correction of 10%;

R^2_{LR} and $R^2_{LR_a}$ values are the unadjusted and adjusted R^2 described in Section 4.7.3.

It is not the strongest of the species that survives, nor the most intelligent that survives. It is the one that is the most adaptable to change.

Charles Darwin

8 Discussion

In this thesis we have reviewed some basics of genetics, the multiple testing issue arising from the simultaneous test of a high number of null hypothesis, the classical methodology approach used in genetic association studies of quantitative traits and the robust counterpart we proposed to be used in this particular context (Chapters 2, 3 and 4).

In Chapter 5 we compared the performance of the usual approach with the robust methodology proposed and two other literature suggestions, more precisely, two non-parametric methods, in genetic association studies of quantitative traits under a specific frame, when the normality assumption is violated. We acknowledged that the robust approach was not as sensitive to outlier contamination as the classical one, having greater power to detect SNP/trait associations while still controlling the FWER. Moreover, we alerted to the danger of conducting the classical analysis even when the residuals distribution looks approximately normal, since in one of the simulation settings studied, results revealed a 13% power loss of the classical approach relative to the robust. For one reason or another, the other two studied methodologies did not compete with the performance of the robust suggested methodology.

The results obtained in this chapter clearly indicate that for an association study involving a not very large number of independent SNPs to be tested ($p << n$), the

robust multiple-SNP linear model is preferable over the classical linear model, assuring the investigator that the outliers that may or not be in the data, and therefore disturbing or not the residuals' distribution, will not translate into false negatives and biased estimates for the model's parameters.

Chapter 6 took us to another much discussed problem in the literature which concerns the identification of regression outliers. In the literature, we have seen outlier detection being made, among other strategies, via robust residuals but always in models with only one explanatory variable. In genetic association studies of quantitative traits, I would dare say, we never have models with just a single explanatory variable! It seemed therefore compelling that an exploratory study was made in this area to assess the ability of the robust outlier tests in detecting outlying observations when model complexity increases. Moreover, since there are usually hundreds of individuals in an association study, adds to this problem the one of multiple testing, presenting therefore a wonderful opportunity of integrating robust outlier tests with multiple testing corrections with the goal of FDR control and power increase. This intent was not accomplished in full, although we saw Benjamini and Hochberg adaptive correction to control the FDR in models with up to 10 independent SNPs while warranting over half of the outliers in the data detected when a 10% threshold was used.

In this chapter we also acknowledged the ability of the robust coefficient of determination proposed in reflecting the proportion of phenotypic variation that is actually explained by the genetic variation among individuals, i.e, in reflecting the trait heritability. This aspect of the analysis is quite important, among other things, for prediction purposes, although this it was not part of the thesis scope. In the ambit of our work, it can be of paramount importance for the outlier tests, since we have seen FDR control and true outlier detection increase with the increase of the trait heritability.

In Chapter 7 we applied both the classical and the robust methodologies to a published real data set. This data set was not in the conditions of the simulation study and was therefore treated accordingly. Results from both analysis showed the presence of outliers and therefore the non-normality of residuals which is known to compromise the results of the classical approach. Having these outliers removed using three different criteria and re-running the classical analysis for those new three data sets, given the variability of the classical approach in detecting and non-detecting SNP/trait associations, the robust methodology again proved to be more adequate, a fact that strengthens our belief against removing outlying observations on statistical grounds alone. These results

together with the simulation study of Chapter 5, were recently published by Lourenço, Pires and Kirst (2011) which certifies the importance of bringing to the area of research of genetic association studies, methods able to circumvent the handicaps of traditional methodologies.

Log-transforming the data, or any other suitable transformation, should not be preferred over a good robust model since it may be difficult to find and/or raise interpretation problems.

Most genetic association studies of quantitative traits in plants and animals are on economically important traits. For this reason, too many false positive associations and few or no true associations detected incur in time and money losses. It is therefore unquestionable the need for a good compromise between the number of true and spurious associations. That compromise can only be successfully achieved if we indulge ourselves in the search for more and more robust methods and more effective multiple testing correction procedures and we believe that this work, away from being finished, enlightens a new pathway in achieving that goal.

Future/Ongoing Work.

In Chapter 4 we briefly reviewed the coefficient of determination in the general multiple regression model and in the context of a mixed model. We became aware of the difficulties in extending the valuable R^2 statistic from the linear univariate model to the linear mixed model, in part because the new version of R^2 now should account not only for the proportion of variation in the response variable that is explained by the fixed effects but also by the random effects. The bottom line is that there seems to be no general definition for R^2 for mixed models that will cover every model.

In the work of Edwards *et al.* (2008) a proposal based on the F-test in equation (4.44) has been made to measure association for fixed effects in the linear mixed model. They compared their R^2 proposal for different methods of estimation of the degrees of freedom ν, in terms of intercept, fixed effects and covariance estimation, concluding the Kenward-Roger approximation to be the best choice in this case. It would thus be interesting to see the performance of this proposal when robust estimates of β, σ^2 and Φ^* are plugged in, making this an open issue for future work since it clearly stays in the interest of genetic association studies where mixed models are currently being used to test for association between a set of genotyped SNPs and a quantitative trait.

In Chapter 6 we have discussed multiple testing corrections for robust outlier tests. We have seen that, in general, we are able to control the FDR at the usual nominal levels for several correction measures in models up to 10 SNPs, therefore minimizing swamping. However, the number of true associations detected stayed a bit over half the total of outliers introduced in the data, when using a 10% threshold, leaving us with an obviously high level of masking. In our opinion, a new more powerful correction measure not prejudicing FDR control and which accounts for the complexity of the model and not just the total number of tests being made should be investigated further. There is for example the recent work of Hekimoglu and Erenoglu (2009) where the performance of Huber's M-estimation for the detection of outliers is increased via reducing the weights of some observations. They showed the performance of their proposed approach to be high especially for small magnitude outliers, situation that is prone to masking and swamping. For that reason, this line of thought certainly deserves our attention. Outliers that are not mere errors are far too important to be simply disregarded. Either way, given the importance of the results already obtained in this chapter a paper is already underway.

In the data analysis example (Chapter 7) we did not use REML but a derivative-free ML methodology to estimate the variance-covariance matrix V, via our robustified version of `lmekin()` function from the R package *kinship*. This was due to the impossibility of incorporating the kinship matrix K in the available R functions for mixed linear models, e.g, `lme()` and `lmer()` (Pinheiro and Bates, 2000). We did find an R package that allows for this and that uses REML, the *ASReml* package. However this R package is not free. We believe results in the estimation of V can be improved with the use of a robustified REML methodology together with the Kenward-Roger approximation for the degrees of freedom of the F test statistic (Kenward and Roger, 1997) in detriment of the Satterthwaite approximation used by Weber (Schaalje *et al*, 2001; Padilla and Algina, 2004; Spike *et al*, 2004 and 2005; Alnosaier, 2007) unless of-course the data is balanced in which case both approximations yield similar results, the Satterthwaite approximation being computationally less demanding. This clearly leaves us with an open topic for future work where the analysis of the work of Graser *et al.* (1987), where a derivative-free REML approach for the estimation of variance components in animal models is proposed, could be a good starting point together with the book of Searle *et al.* (1992) on variance components, where both ML and REML methodologies are discussed for balanced and unbalanced data. Also in this chapter, we have used the BH-1995 FDR at 10% level to assess significance of associations as done in Weber *et al.* (2008). However, in the context of genetic association studies permutation/bootstrap p-values could be investigated. There are already some robust approaches in the literature, e.g,

the robust bootstrap: Salibian-Barrera and Zamar, 2002/2004; Salibian-Barrera, 2005. In the latter paper, Salibian-Barrera studies the estimation of p-values for robust tests in the linear regression model but focusing only on score-type tests. The author states that his approach can be extended to e.g, Wald-type tests and so this leaves us with another window of opportunity.

There are suggested semi-parametric methods in literature for association studies (e.g, Zhang *et al.*, 2003; Li *et al.*, 2008). In a near future we are sure many more will come up. A comparison between these and robust methods is always of high interest. However, in our case it will make more sense after we have improved the methodology discussed in this work with, for example, the ideas presented in the previous paragraphs.

> Now this is not the end. It is not even the beginning of the end. But it is, perhaps, the end of the beginning.
>
> Winston Churchill

A

APPENDIX A: Some extra tables

Table A.1: Candidade genes for the quantitative phenotype FERL.

Gene	Gene symbol	Num SNPs	Genotyped SNPs
another ear	an1	1	an1.4
teosinte branched	tb1	4	PZD00062.2 tb1.18 tb1.19 tb1.5
indeterminate growth	id1	2	PZD00068.1 PZD00068.3
indeterminate spikelet1	ids	3	PZD00069.2 PZD00069.4 PZD00069.5
zea floricaula leafy2	zfl2	1	zfl2.6
aberrant phyllotaxy	abph1	1	abph1.20
zea apetala homolog1	zap1	4	zap1.2 PZD00022.1 PZD00022.3 PZD00022.4
barren stalk	ba1	3	ba1.6 ba1.9 PZD00078.2
terminal ear1	te1	5	PZD00006.1 PZD00007.1 PZD00008.3 te1.3 te1.4
ramosa2	ra2	3	Ra2−ORF.2 Ra2−ORF.4 Ra2−promoter.3
zea mays MADS19	zmm19	1	PZD00030.5
fasciated ear2	fea2	4	Fea2.5 PZD00067 PZD00067.2 PZD00067.3
teosinte glume architecture1	tga1	1	PZD00009.5
thick tassel dwarf1	td1	2	PZB01115.5 PZB01115.6
ramosa1	ra1	4	PZD00073.5 PZD00073.8 PZB00175.1 PZB00175.3
ramosa3	ra3	1	PZB01238.5
zea mays MADS1	zagl1	8	PZD00019.1 PZD00020.2 PZD00020.3 PZD00020.4 PZD00021.2 PZD00021.5 zagl1.1 zagl1.6
AZM4 − 91097	AZM4 − 91097	2	PZB01110.2 PZB01110.3
AZM4 − 116423	AZM4 − 116423	2	PZB01111.6 PZB01111.8
zea centroradialis	zen1	3	zen1.1 zen1.2 zen1.4
zea mays gigantea	ZmGI	3	PZB00049.2 PZB00049.4 PZB00049.7
BM660811	BM660811	3	PZB00092.4 PZB00229.3 PZB00229.4

Table A.2: Pairwise LD r^2 for the 9 SNPs detected in the single-SNP analysis.

	zagl1.1	PZD00073.5	PZD00006.1	PZD00022.3	PZD00073.8	te1.3	ba1.9	PZB00049.7	zagl1.6
PZD00073.5	.00004	–	–	–	–	–	–	–	–
PZD00006.1	.00435	.00660	–	–	–	–	–	–	–
PZD00022.3	.00458	.01258	.00202	–	–	–	–	–	–
PZD00073.8	.00021	.67708	.00295	.00221	–	–	–	–	–
te1.3	.00025	.00047	.00000	.00238	.00238	–	–	–	–
ba1.9	.00200	.00310	.00189	.00012	.00782	.00238	–	–	–
PZB00049.7	.00102	.00211	.00071	.00028	.00157	.00073	.00329	–	–
zagl1.6	.00083	.00004	.01189	.01830	.00013	.00147	.00410	.00105	–

B

APPENDIX B: Some extra figures

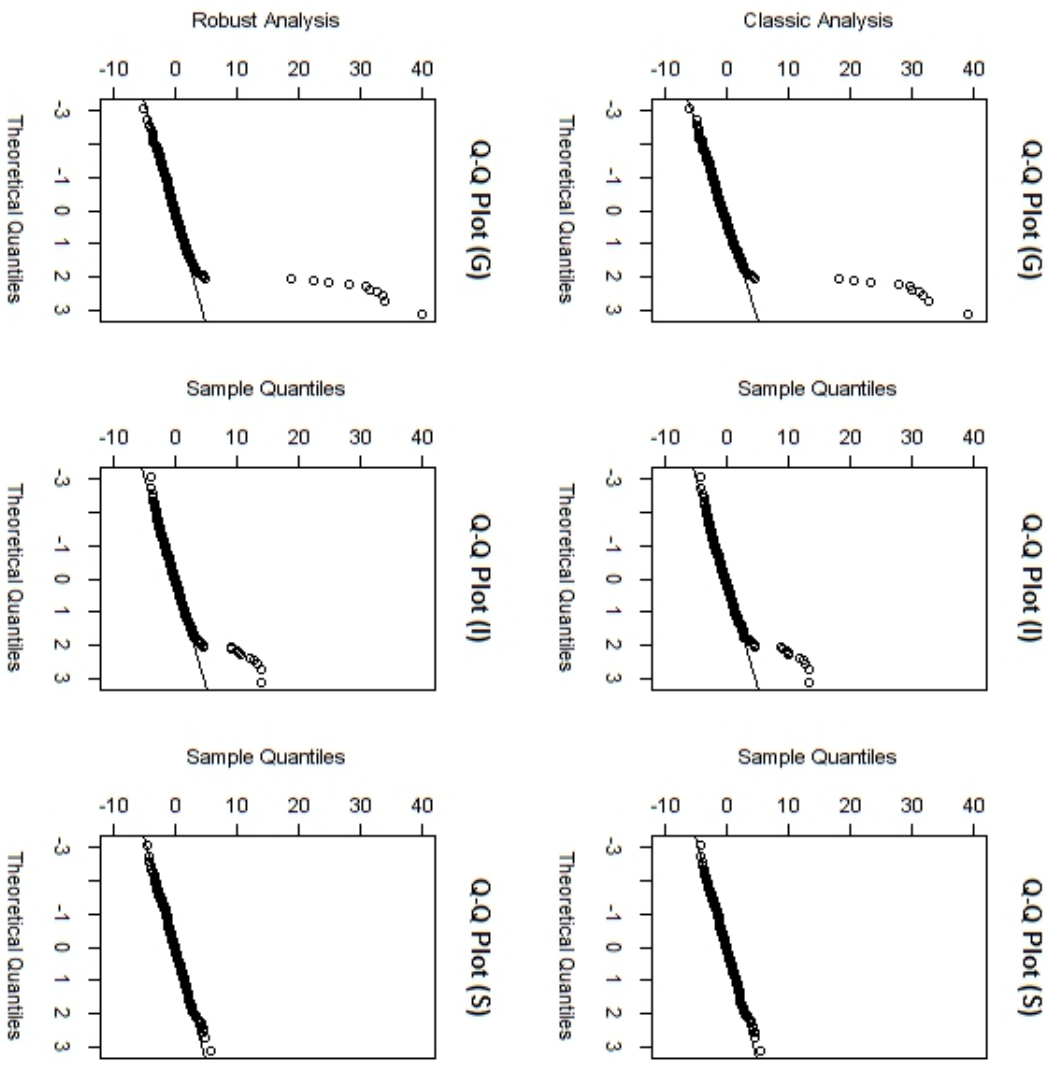


Figure B.1: Q-Q Plots for the 10 SNP model with 2% gross, intermediate and smooth contamination.

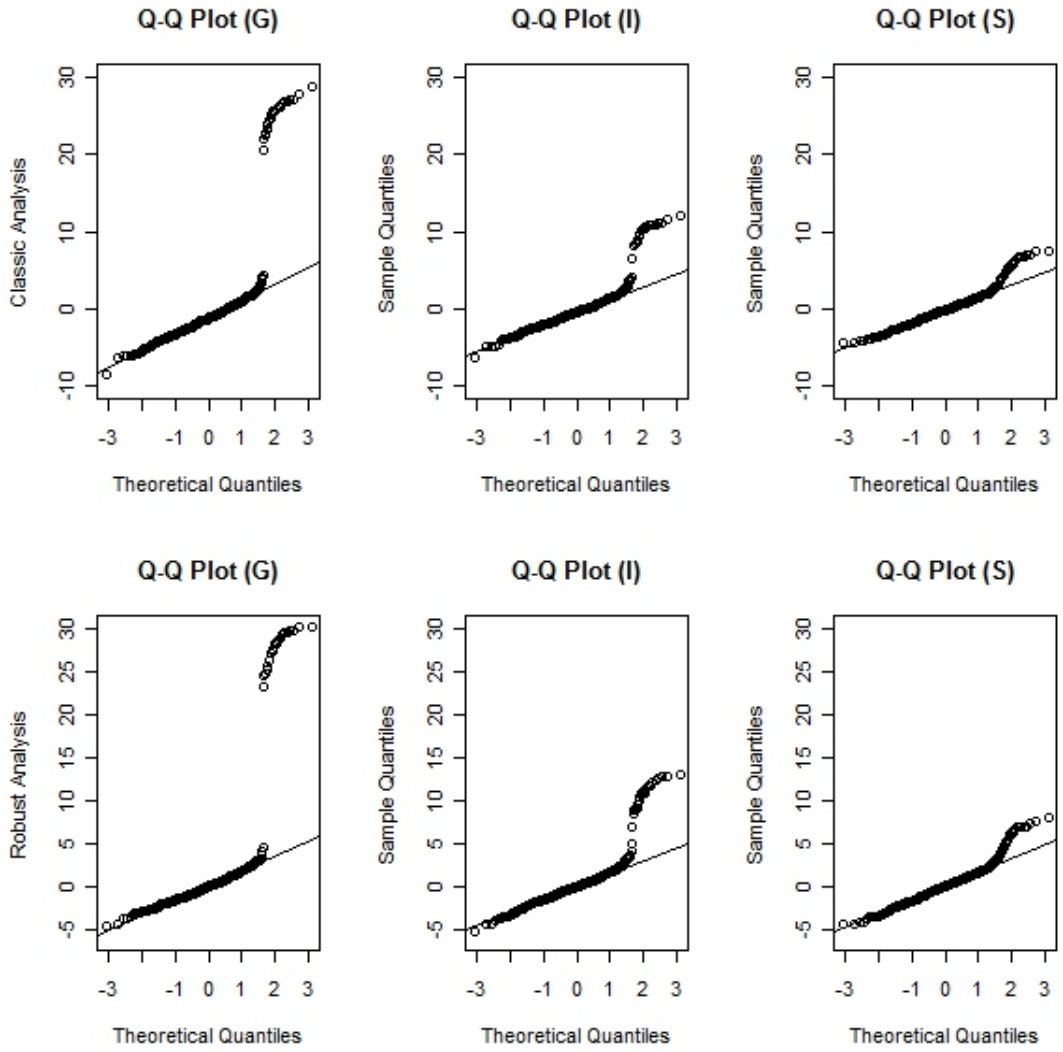


Figure B.2: Q-Q Plots for the 10 SNP model with 5% gross, intermediate and smooth contamination.

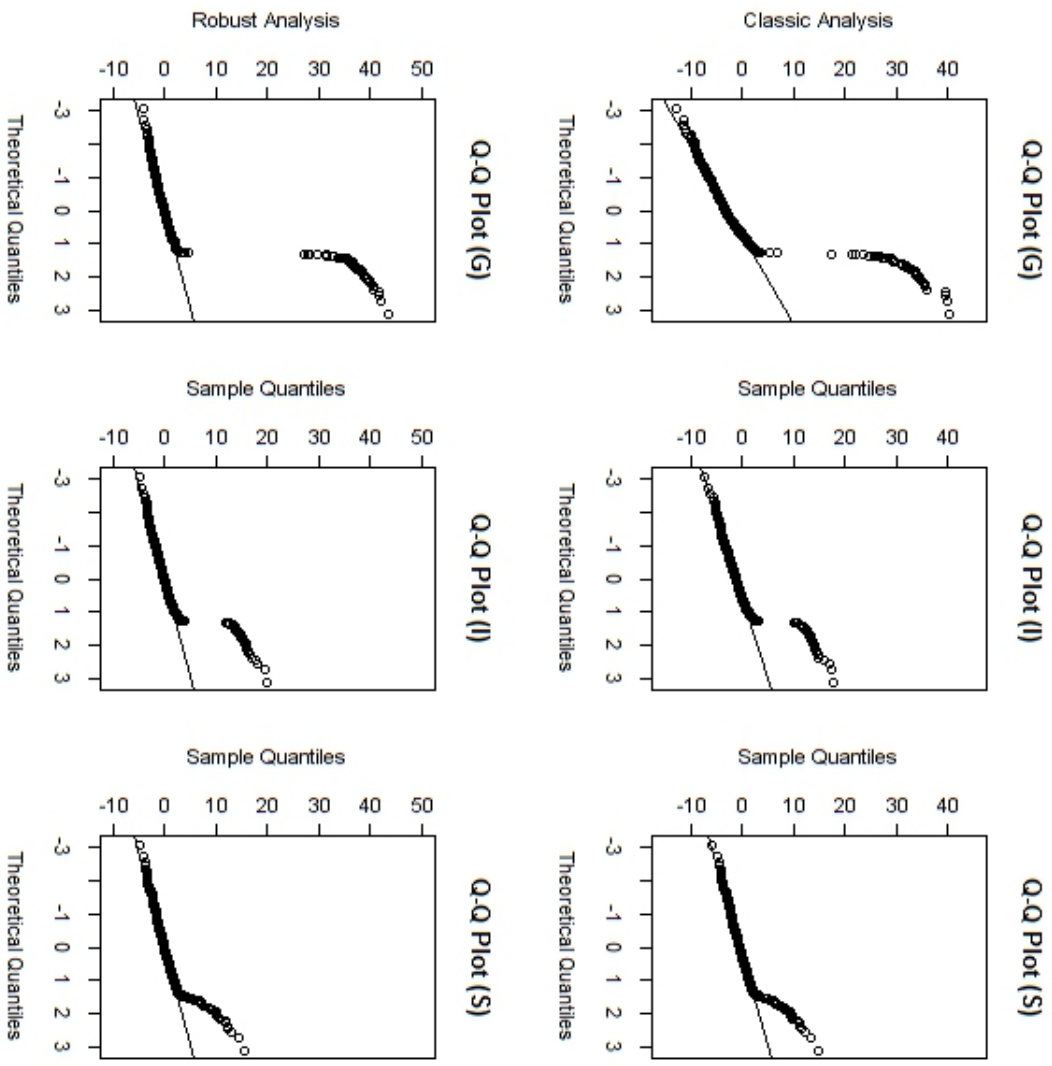


Figure B.3: Q-Q Plots for the 10 SNP model with 10% gross, intermediate and smooth contamination.

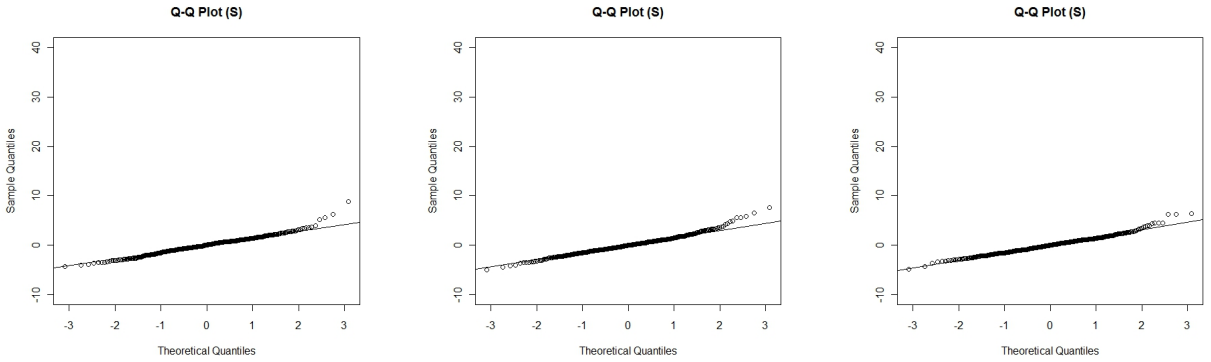


Figure B.4: Q-Q Normal plots of residuals from the 10 SNP model association analysis with 2% smooth contamination, with SF p-values of 0.00001, 0.00003 and 0.00045, respectively, corresponding to Pearson p-values of 0.134, 0.256 and 0.379.

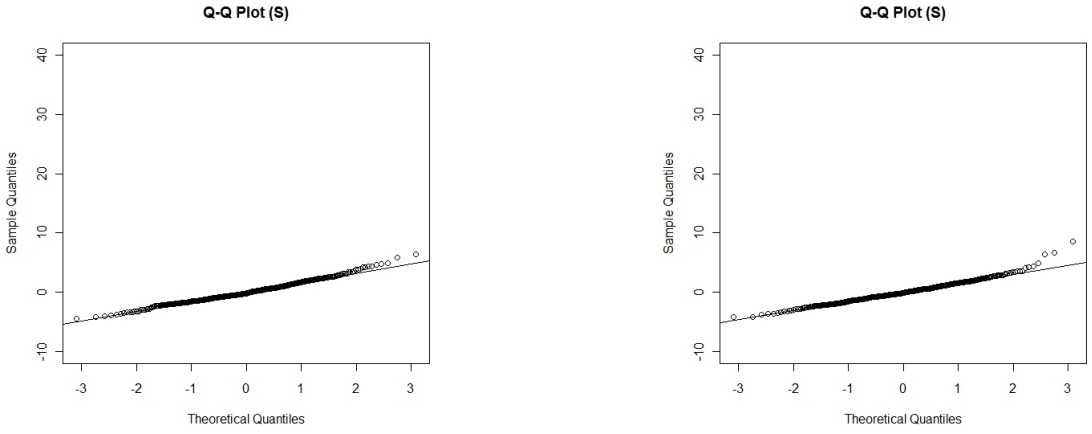


Figure B.5: Q-Q Normal plots of residuals from the 10 SNP model association analysis with 2% smooth contamination, with Pearson p-values of 0.0249 and 0.0352 corresponding to SF p-values of 0.000001 and 0.0009, respectively.

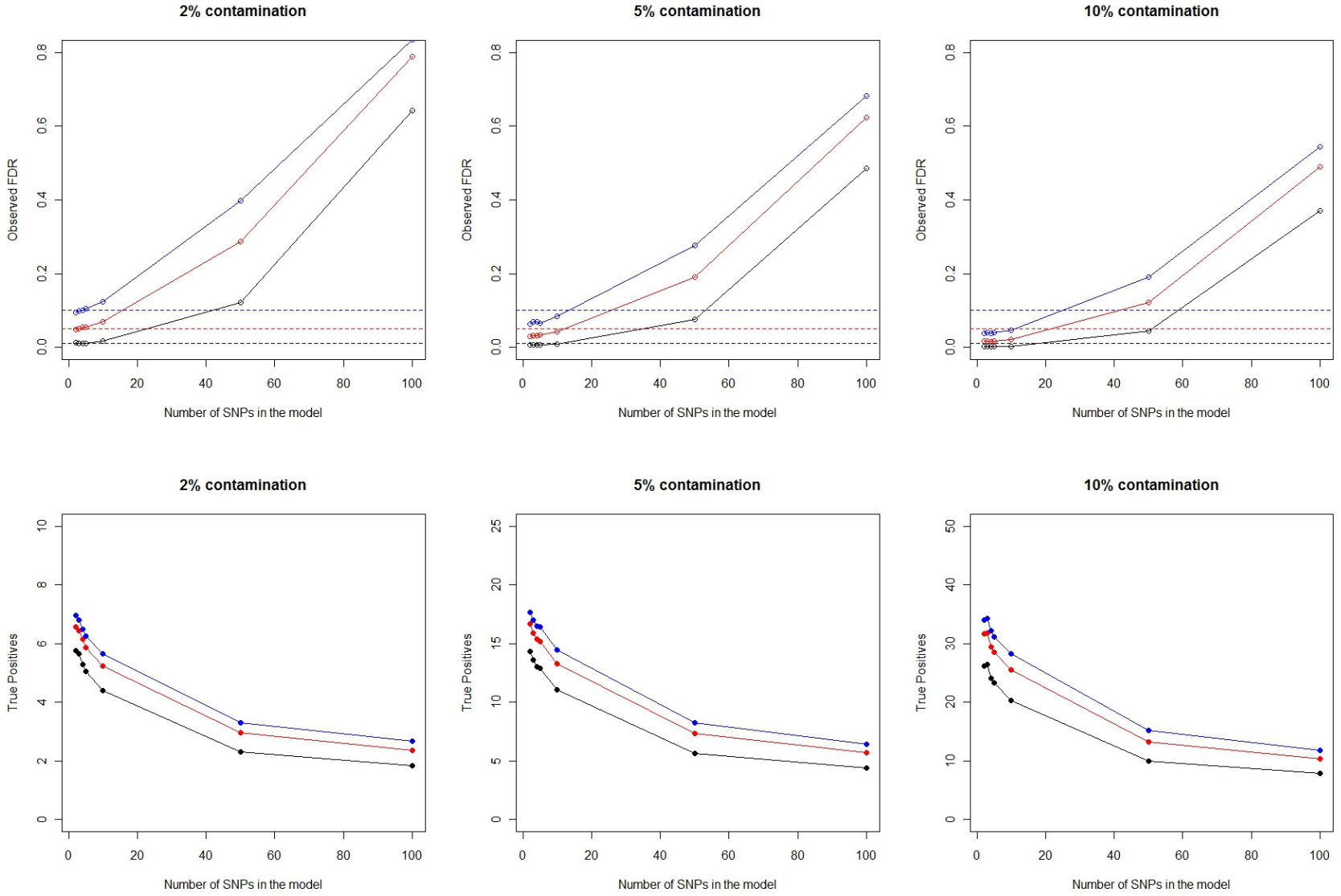


Figure B.6: Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.3$, a percentage of $2, 5$ and 10% of smooth data contamination and robust parameter of scale s. BH-1995 correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust standardized residuals (4.32). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1.

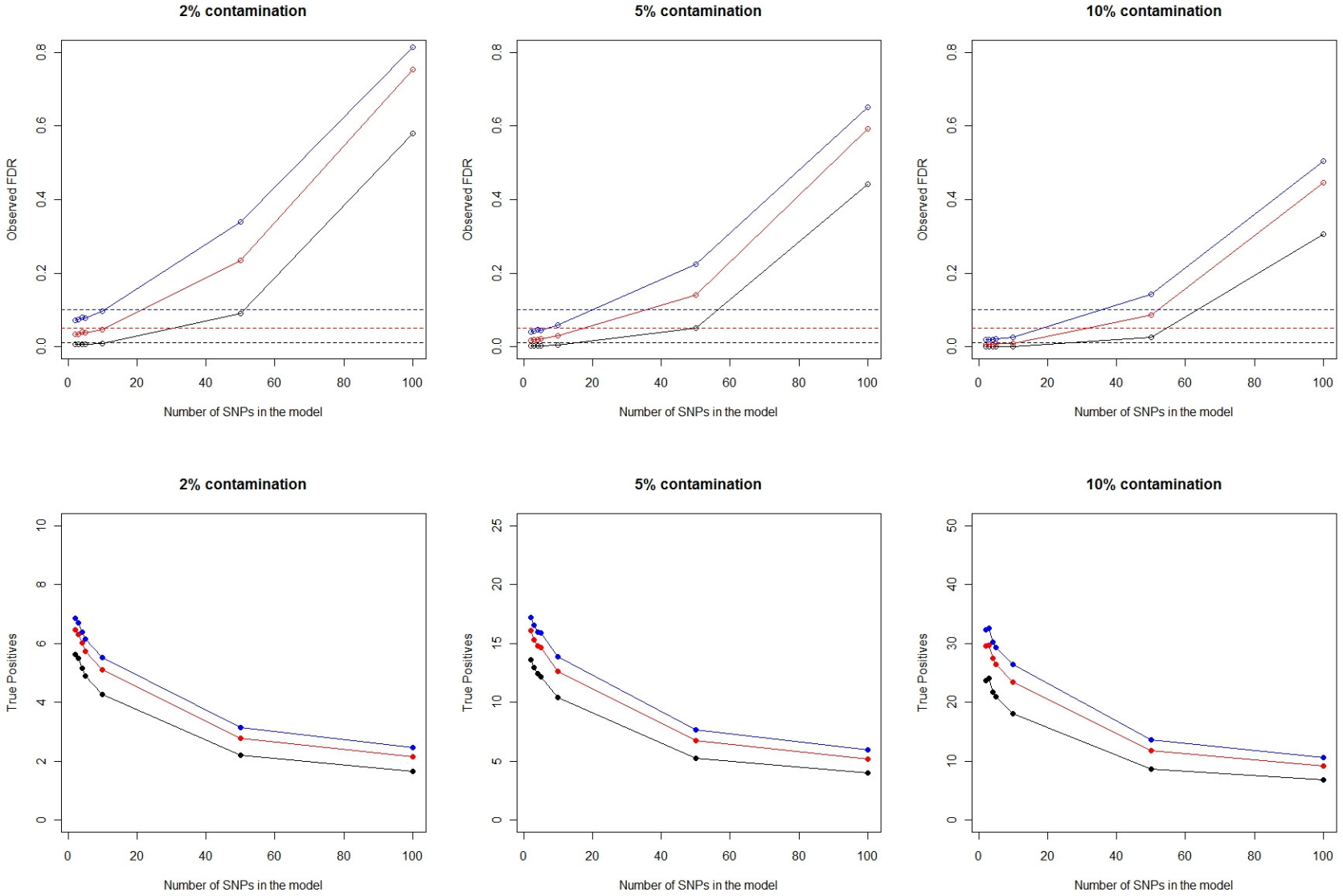


Figure B.7: Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.3$, a percentage of $2, 5$ and 10% of smooth data contamination and robust parameter of scale Q_n. BH-1995 correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust standardized residuals (4.32). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1.

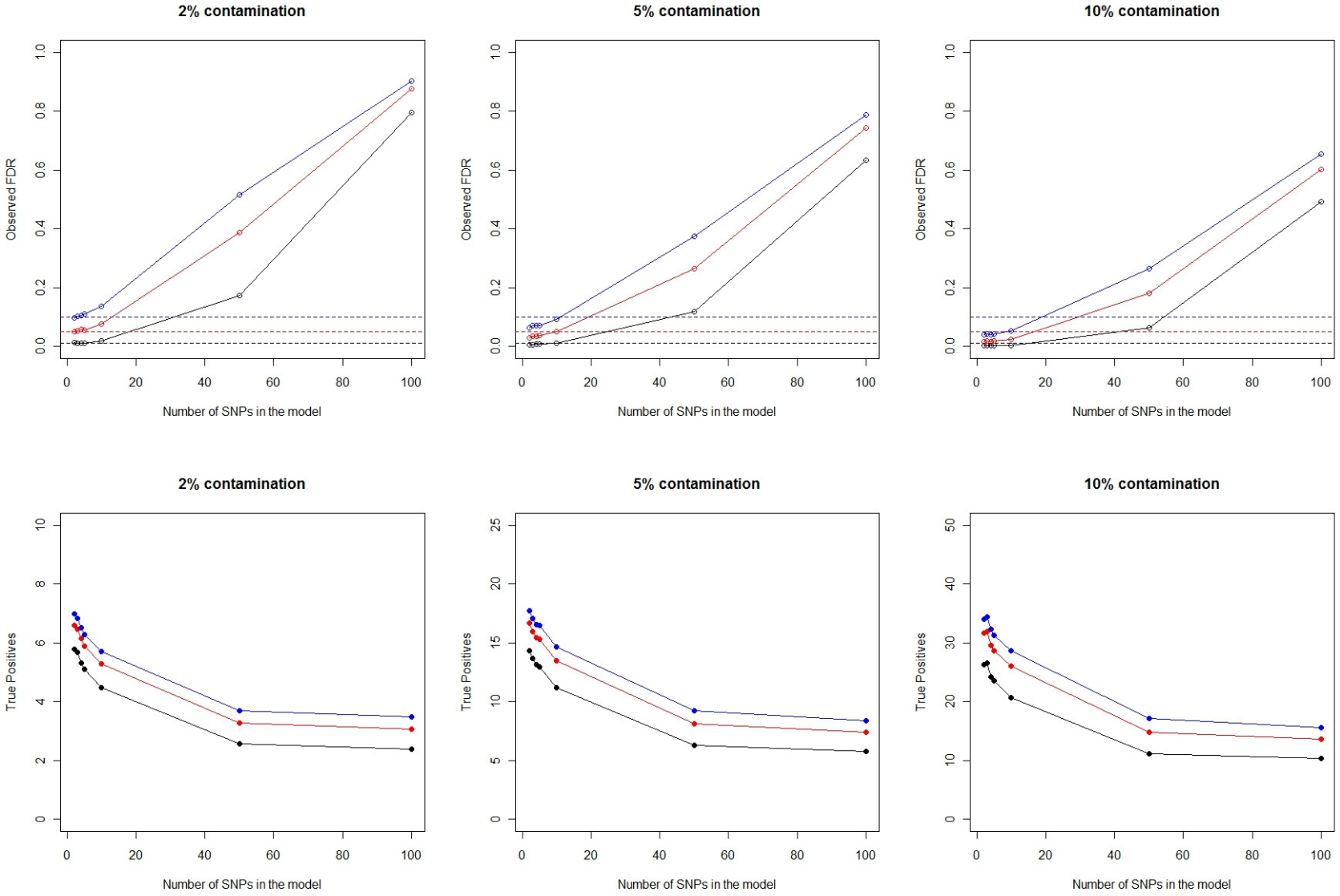


Figure B.8: Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.3$, a percentage of $2, 5$ and 10% of smooth data contamination and robust parameter of scale s. BH-1995 correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust studentized residuals (6.1). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1.

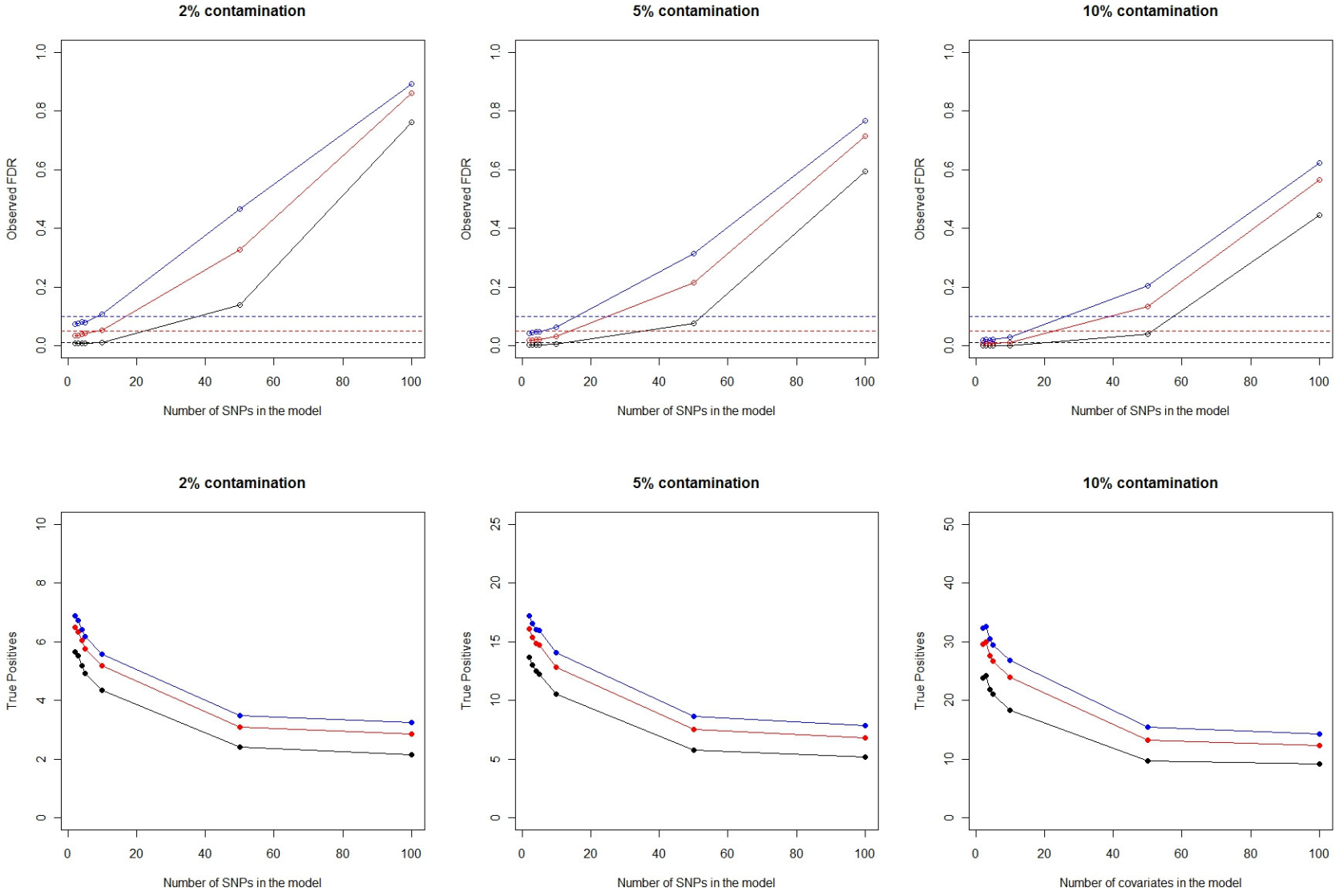


Figure B.9: Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.3$, a percentage of $2, 5$ and 10% of smooth data contamination and robust parameter of scale Q_n. BH-1995 correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust studentized residuals (6.1). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1.

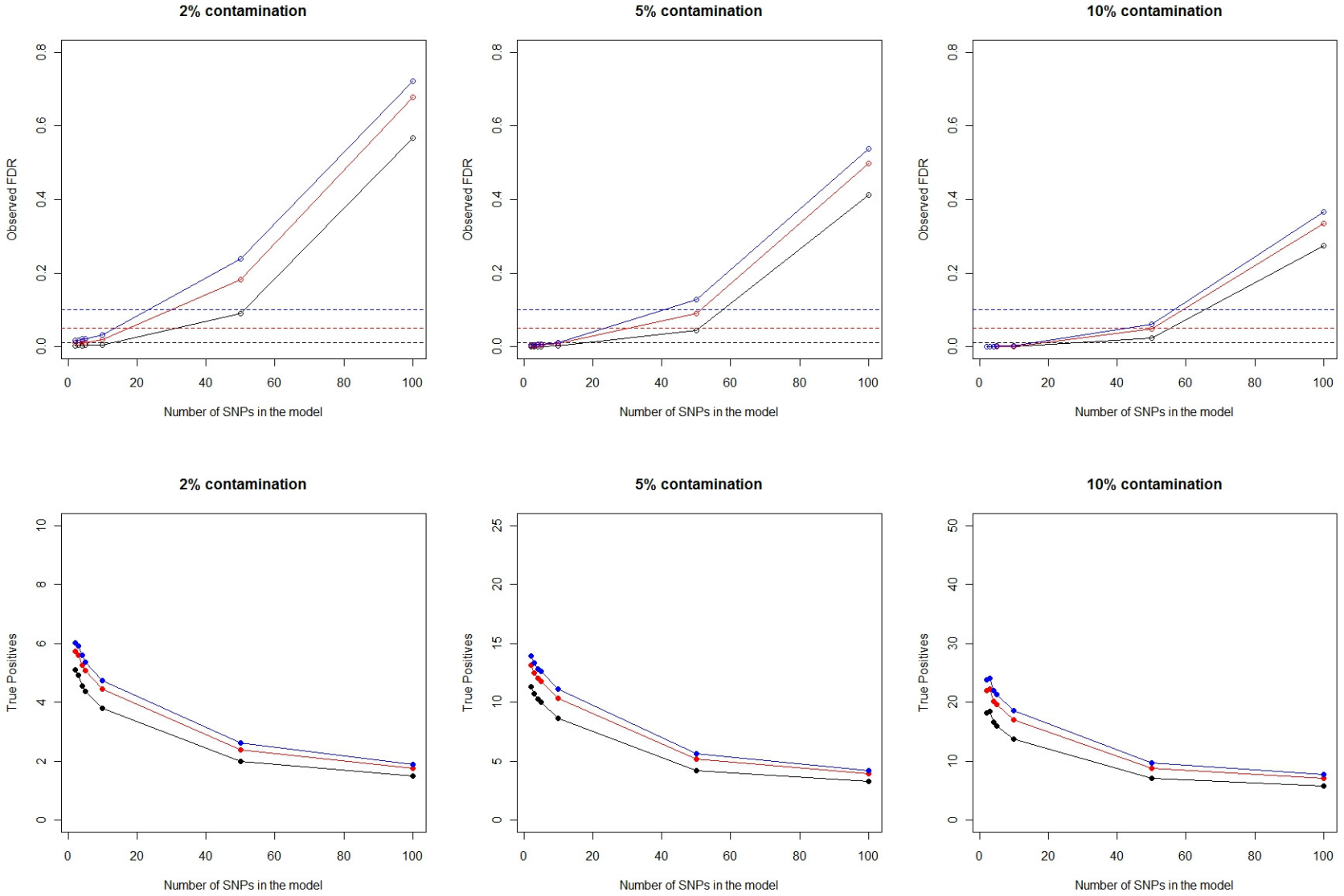


Figure B.10: Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.3$, a percentage of $2, 5$ and 10% of smooth data contamination and robust parameter of scale s. Bonferroni correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust standardized residuals (4.32). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1.

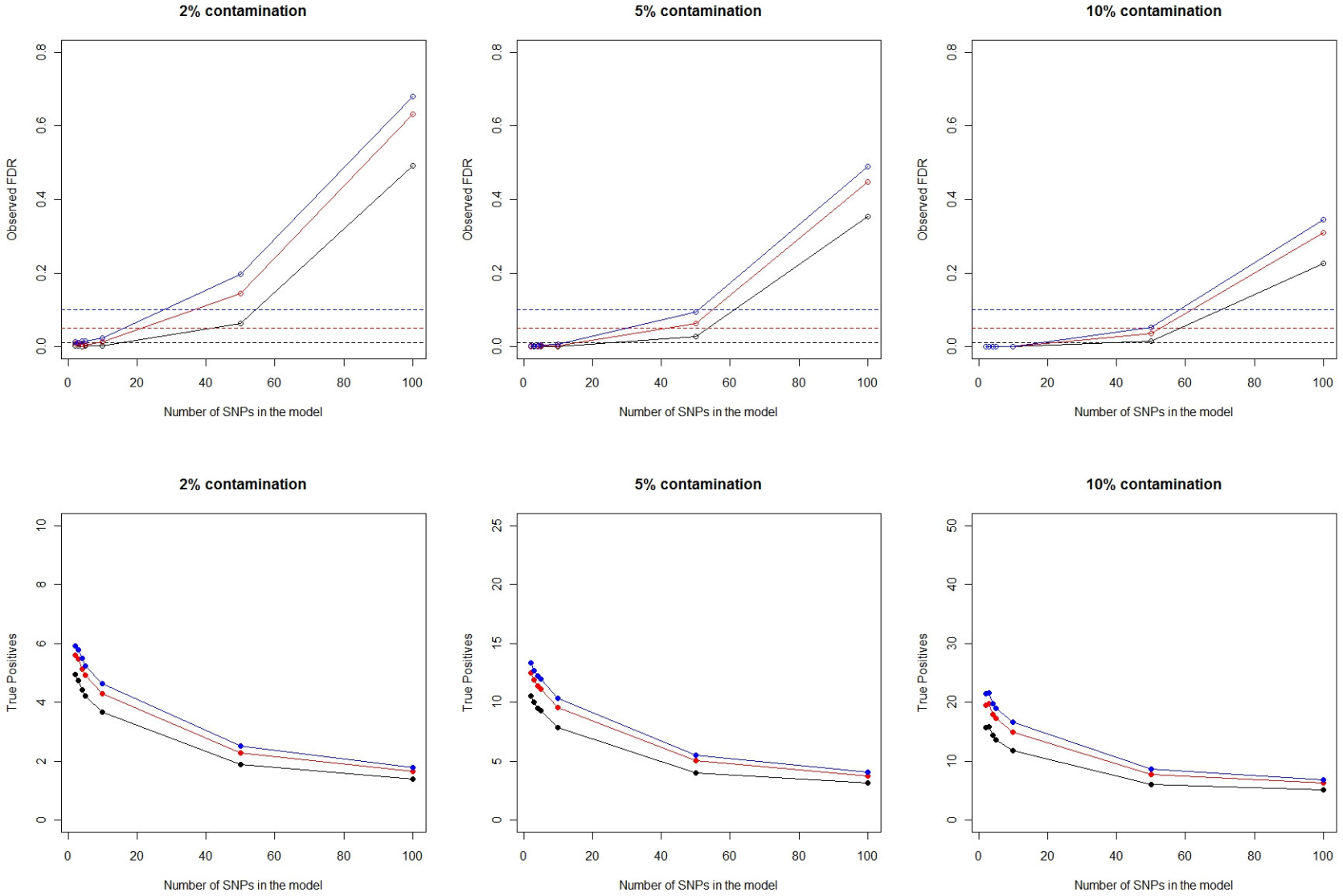


Figure B.11: Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.3$, a percentage of $2, 5$ and 10% of smooth data contamination and robust parameter of scale Q_n. Bonferroni correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust standardized residuals (4.32). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1.

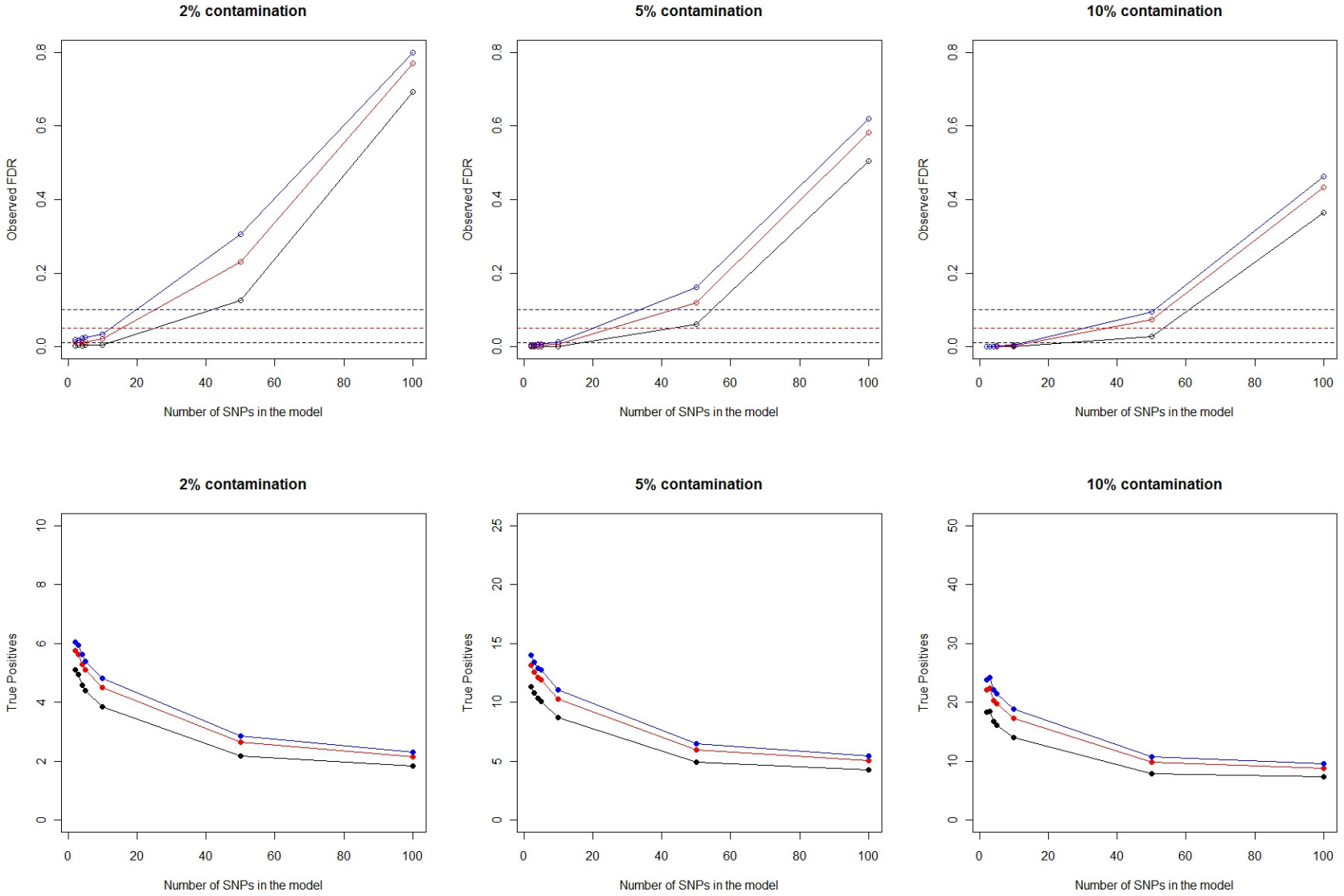


Figure B.12: Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.3$, a percentage of $2, 5$ and 10% of smooth data contamination and robust parameter of scale s. Bonferroni correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust studentized residuals (6.1). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1.

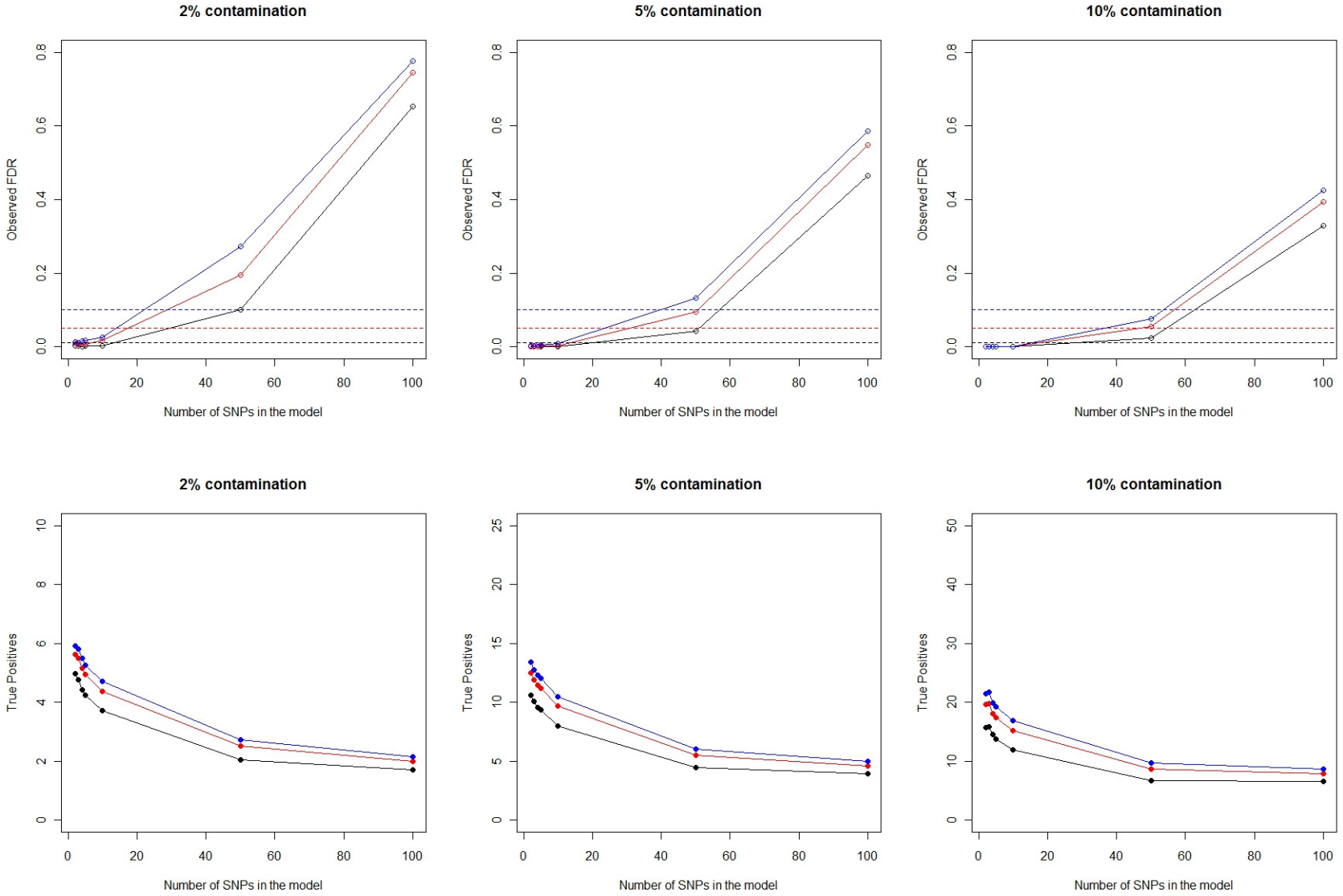


Figure B.13: Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.3$, a percentage of $2, 5$ and 10% of smooth data contamination and robust parameter of scale Q_n. Bonferroni correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust studentized residuals (6.1). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1.

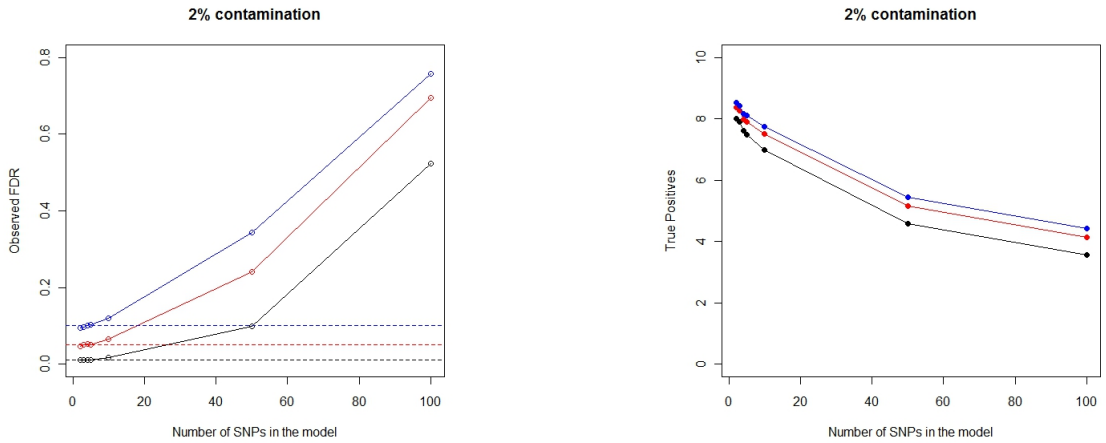


Figure B.14: Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.5$, 2% of smooth data contamination and robust parameter of scale s. BH-1995 correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust standardized residuals (4.32). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1.

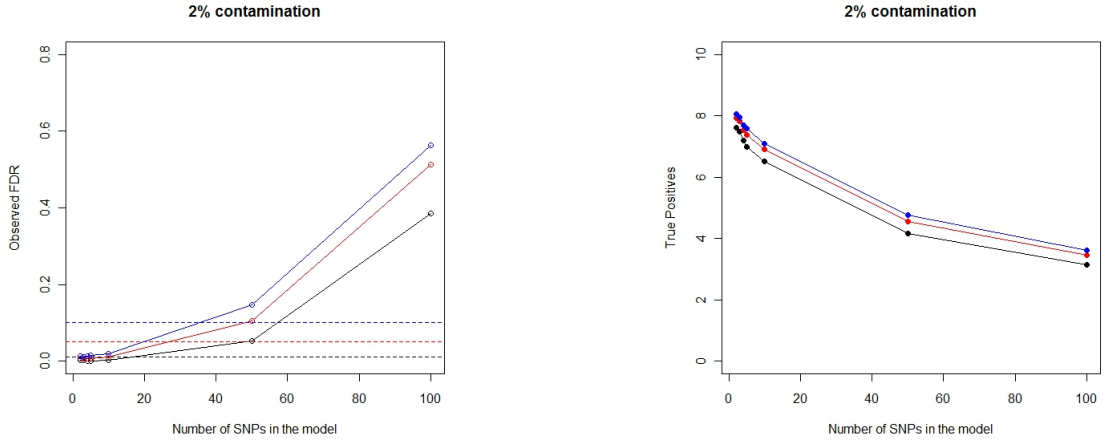


Figure B.15: Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.5$, 2% of smooth data contamination and robust parameter of scale s. Bonferroni correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust standardized residuals (4.32). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1.

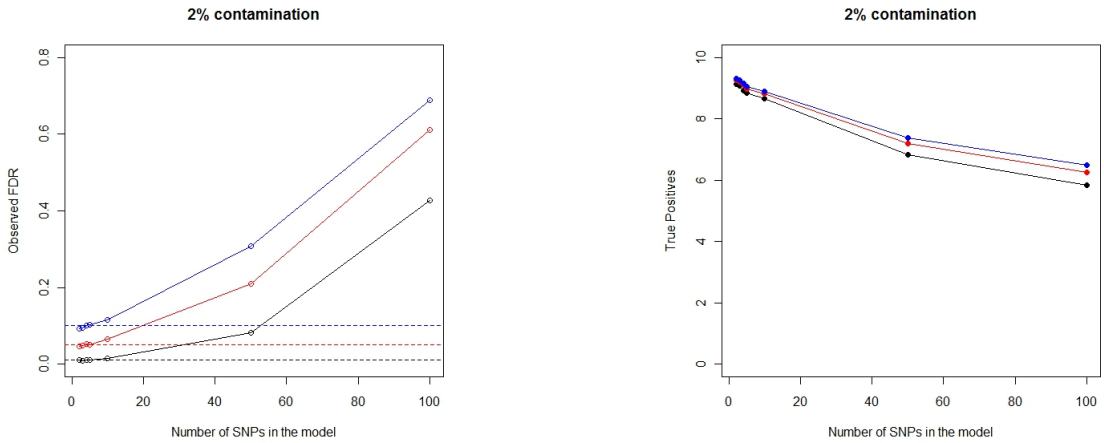


Figure B.16: Observed FDR and True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.75$, 2% of smooth data contamination and robust parameter of scale s. BH-1995 correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust standardized residuals (4.32). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1.

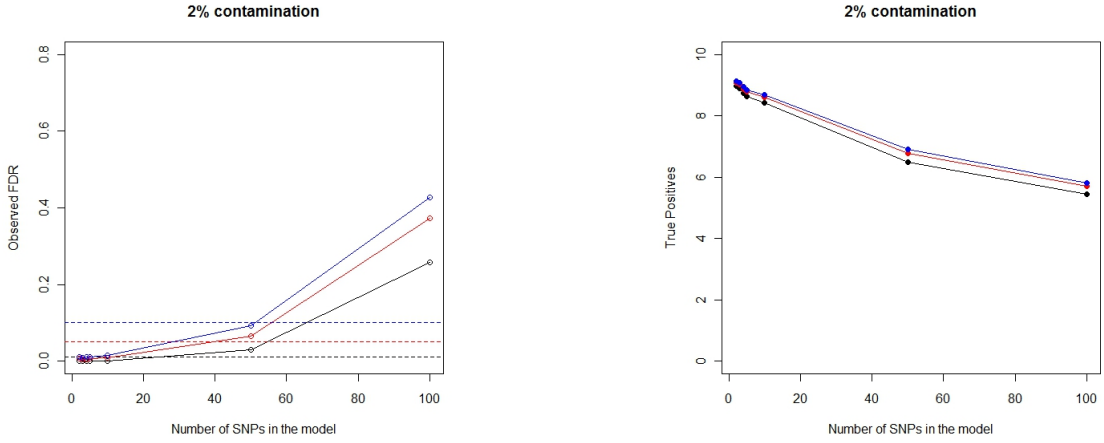


Figure B.17: Observed FDR True Positives for models with $2, 3, 4, 5, 10, 50$ and 100 SNPs, trait heritability $H^2 = 0.75$, 2% of smooth data contamination and robust parameter of scale s. Bonferroni correction used at the nominal levels of $1, 5$ and 10%. Outlier tests performed with the robust standardized residuals (4.32). Blue, red and black lines refer respectively, to the nominal FDR levels of $0.01, 0.05$ and 0.1.

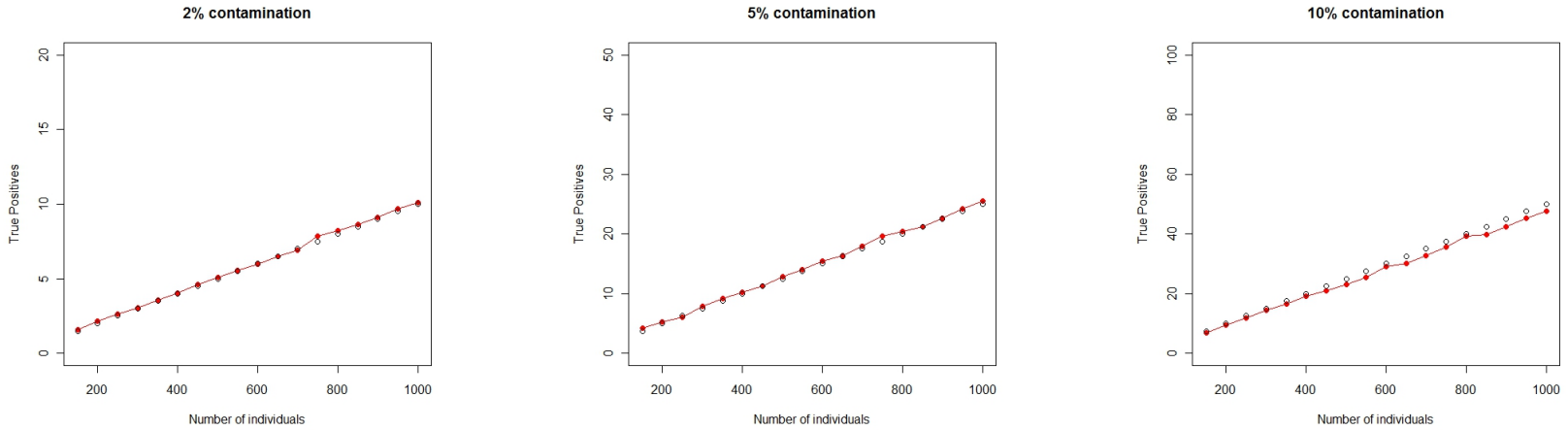


Figure B.18: Number of True Outliers observed for nominal level of $\alpha = 0.05$ in models with 10 SNPs. BH-2000 FDR correction was used.

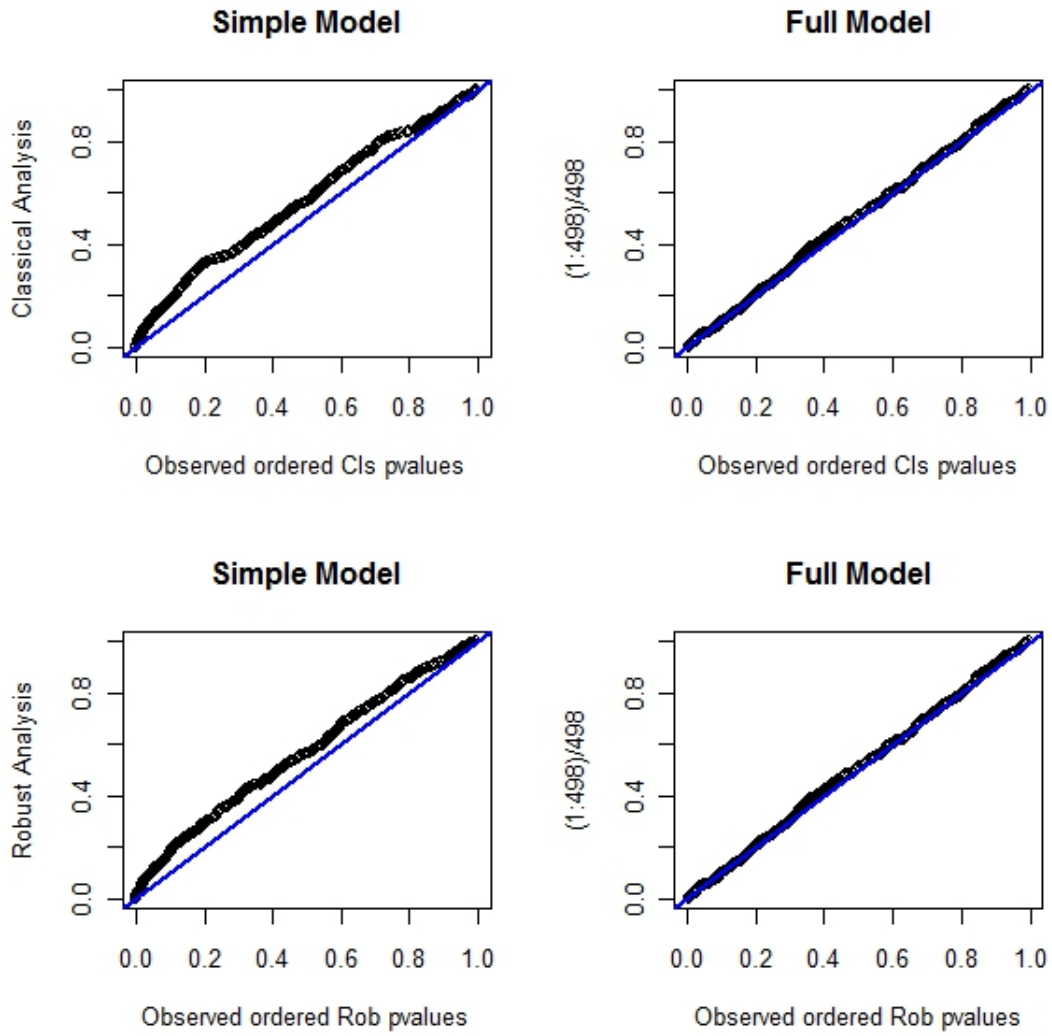


Figure B.19: Cumulative distribution of the 498 p-values for both simple and full models, in the classical and robust analysis.

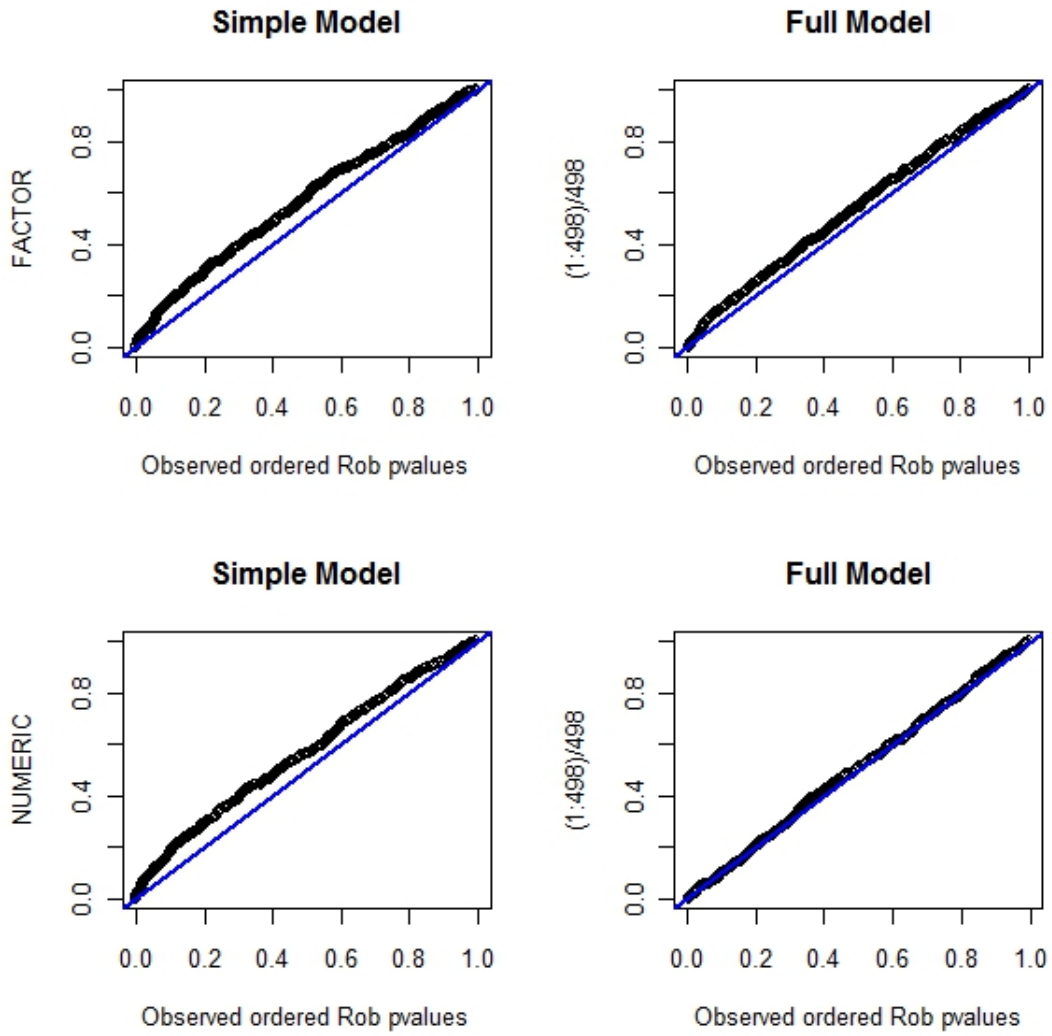


Figure B.20: Cumulative distribution of the 498 p-values for both simple and full models - robust methodology.

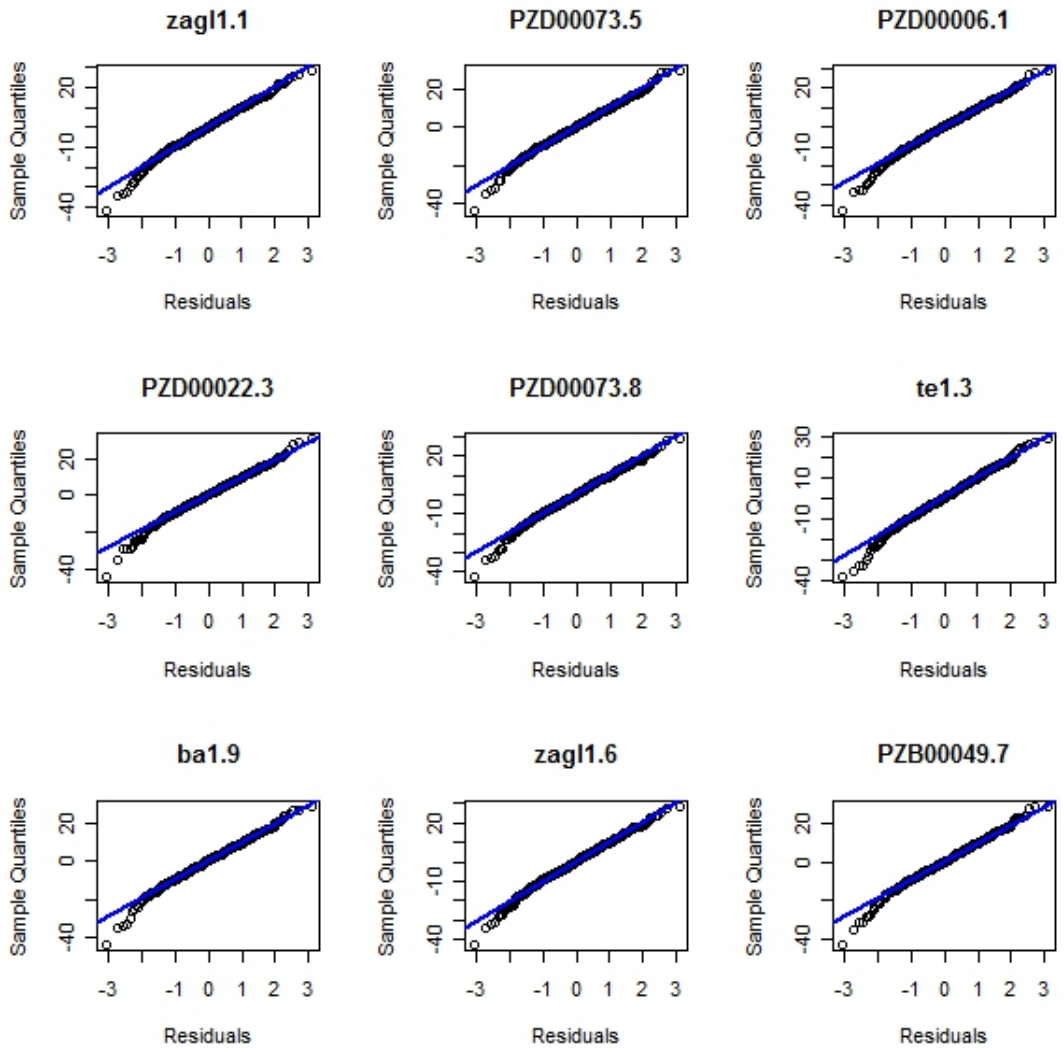


Figure B.21: QQ-normal residual plots from the classic single-SNP analysis.

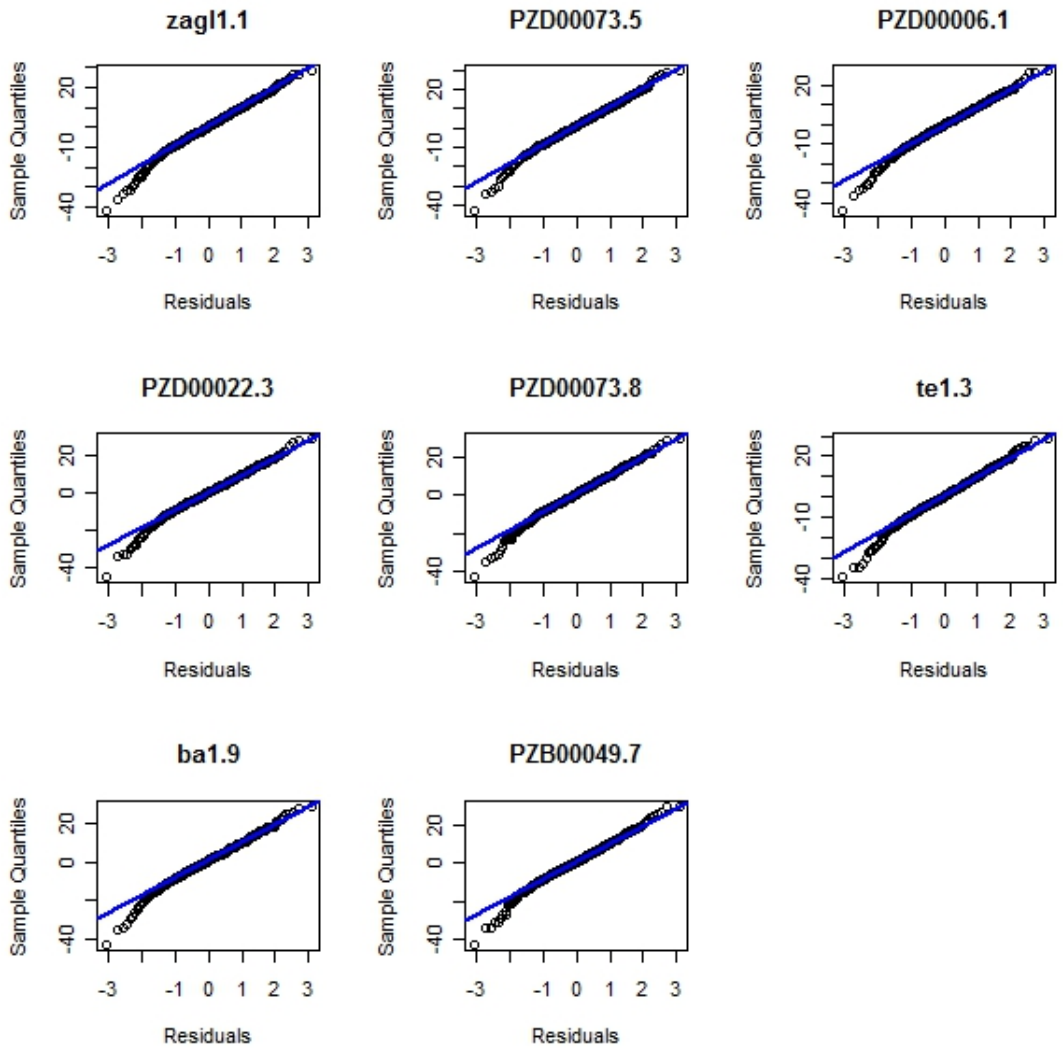


Figure B.22: QQ-normal residual plots from the robust single-SNP analysis.

C

APPENDIX C: R routines

C.1 R code for the phones data example on Section 4.6

```
library(MASS); library(GenABEL)

data(phones); attach(phones)

modeloR<-rlm(calls~year)
modelo<-lm(calls~year)
diagnost<-ls.diag(modelo)

p<- 2*pnorm(-abs(diagnost$std.res))

pR<- 2*pnorm(-abs(modeloR$residuals)/modeloR$s)

a<-qvaluebh95(p)$qvalue; aR<-qvaluebh95(pR)$qvalue

op <- par(mfrow=c(1, 2))
plot(year,calls,main="classical plot")
lines(year,200*a,type="o",col=4,pch=16)
abline(modelo,lwd=1,lty=2)
abline(20,0,col=3,lty=3,lwd=2)

plot(year,calls,main="robust plot")
lines(year,200*aR,type="o",col=4,pch=16)
abline(modeloR,lwd=1,lty=2)
abline(20,0,col=3,lty=3,lwd=2)
```

C.2 Description of the R `lmekin` routine

```
lmekin<-function (fixed, data = parent.frame(), random, varlist =
NULL, variance, sparse = c(20, 0.05), rescale = T, pdcheck = T,
subset, weight, na.action) {
    call <- match.call()
    m <- match.call(expand.dots = FALSE)
    temp <- c("", "data", "weights", "subset", "na.action")
    m <- m[match(temp, names(m), nomatch = 0)]
    if (missing(variance))
        theta <- NULL
    else theta <- variance
    reSt <- reStruct(random, REML = F, data = NULL)
    gform <- getGroupsFormula(reSt)
    if (is.null(gform)) {
        temp.fixed <- fixed
        gvars <- NULL
    }
    else {
        gvars <- all.vars(random)
        fvars <- all.vars(formula)
        gvars <- gvars[is.na(match(gvars, fvars))]
        temp.fixed <- paste(deparse(as.vector(fixed)), collapse = "")
        temp.fixed <- paste(temp.fixed, paste(gvars, collapse = "+"),
            sep = "+")
        temp.fixed <- as.formula(temp.fixed)
    }
    m$formula <- temp.fixed
    m[[1]] <- as.name("model.frame")
    m <- eval(m, sys.parent())
    Terms <- terms(fixed)
    X <- model.matrix(Terms, m)
    Y <- model.extract(m, "response")
    n <- length(Y)
    weights <- model.extract(m, "weights")
    offset <- attr(Terms, "offset")
    tt <- length(offset)
    offset <- if (tt == 0)
```

```
    rep(0, n)
else if (tt == 1)
    m[[offset]]
else {
    ff <- m[[offset[1]]]
    for (i in 2:tt) ff <- ff + m[[offset[i]]]
    ff
}
ncluster <- length(gvars)
if (ncluster == 0)
    stop("No grouping variables found")
groups <- getGroups(m, gform)
temp <- coxme.varcheck(ncluster, varlist, n, gvars, groups,
    sparse, rescale, pdcheck)
varlist <- temp$varlist
kindex <- temp$kindex
ntheta <- temp$ntheta
theta.names <- NULL
for (i in 1:ncluster) {
    if (ntheta[i] == 1)
        theta.names <- c(theta.names, gvars[i])
    else theta.names <- c(theta.names, paste(gvars[i], 1:ntheta[i],
        sep = ""))
}
if (length(theta) == 0)
    theta <- rep(0, sum(ntheta))
else if (length(theta) != sum(ntheta))
    stop("Wrong length for theta")
names(theta) <- theta.names
theta.names <- NULL
for (i in 1:ncluster) {
    if (ntheta[i] == 1)
        theta.names <- c(theta.names, gvars[i])
    else theta.names <- c(theta.names, paste(gvars[i], 1:ntheta[i],
        sep = ""))
}
if (length(theta) == 0)
```

```
    theta <- rep(0, sum(ntheta))
else if (length(theta) != sum(ntheta))
    stop("Wrong length for theta")
names(theta) <- theta.names
tindex <- which(theta == 0)
if (ncluster > 1)
    stop("function can have only 1 random effect")
varlist <- varlist[[1]]
kindex <- kindex[, 1]
if (max(kindex) != n)
    stop("The random effect must be 1 per subject")
ntheta <- ntheta[1]
kindex2 <- integer(n)
kindex2[kindex] <- 1:n
logfun <- function(itheta, X, Y, varlist, theta, tindex,
    center) {
    theta[tindex] <- exp(itheta)
    tkmat <- varlist[[1]]
    tkmat@blocks <- tkmat@blocks * theta[1]
    diag(tkmat) <- diag(tkmat + 1)
    if (length(varlist) > 1) {
        for (i in 2:length(varlist)) tkmat@blocks <- varlist[[i]]@blocks *
            theta[i] + tkmat@blocks
    }
    tkmat@blocks <- tkmat@blocks/tkmat@blocks[1]
    gk <- gchol(tkmat)
    newx <- solve(gk, X, full = FALSE)
    newy <- solve(gk, Y, full = FALSE)
    resid <- qr.resid(qr(newx), newy)
    n <- length(Y)
    loglik <- (n/2) * (log(mean(resid^2)) - center) + sum(log(diag(gk)))/2
    loglik
}
newX <- X[kindex2, ]
newY <- as.vector(Y[kindex2])
dimnames(newX) <- NULL
if (length(tindex) > 0) {
```

```
    center <- log(mean((Y - mean(Y))^2))
    nfit <- optim(par = rep(-1, length(tindex)), logfun,
        method = "L-BFGS-B", lower = log(1e-05), X = newX,
        Y = newY, varlist = varlist, theta = theta, tindex = tindex,
        center = center)
    iter <- nfit$counts
    theta[tindex] <- exp(nfit$par)
}
else iter <- 0
tkmat <- varlist[[1]]
tkmat@blocks <- tkmat@blocks * theta[1]
diag(tkmat) <- diag(tkmat + 1)
if (length(varlist) > 1) {
    for (i in 2:length(varlist)) tkmat@blocks <- varlist[[i]]@blocks *
        theta[i] + tkmat@blocks
}
gk <- gchol(tkmat)
xok <- as.matrix(solve(gk, newX, full = F))
yok <- solve(gk, newY, full = FALSE)
lfit <- lm(yok ~ 0 + xok)
names(lfit$coefficients) <- dimnames(X)[[2]]
ls <- summary(lfit)
resid.var <- mean(lfit$residuals^2)
theta <- c(theta * resid.var, resid.var)
names(theta) <- c(theta.names, "resid")
fitted <- c(X %*% lfit$coef)
residuals <- Y - fitted
frail <- residuals[kindex2]
names(frail) <- groups
fcoef <- lfit$coef
call$fixed <- fixed
call$random <- random
fit <- list(coefficients = list(fixed = fcoef, random = frail),
    theta = theta, variance = ls$cov.unscaled * ls$sigma^2, ctable =
    ls$coefficients, residuals = residuals, fitted.values = fitted,
    effects = lfit$effects, rank = lfit$rank, assign = lfit$assign,
    df.residual = lfit$df.residual - length(theta),
```

```
        loglik = (-n/2)*(log(mean(lfit$residuals^2)) + 1 + log(2 * pi)) -
            sum(log(diag(gk)))/2,
        iter = iter, n = n, call = call, method = "ML")
    na.action <- attr(m, "na.action")
    if (length(na.action))
        fit$na.action <- na.action
    oldClass(fit) <- c("lmekin")
    fit
}
```

C.3 Description of the R robust version of the R lmekin routine

```
lmekinR<-function(fixed, data = parent.frame(), random, varlist =
NULL,variance, sparse = c(20, 0.05), rescale = T, pdcheck = T,
subset, weight, na.action) {
    call <- match.call()
    m <- match.call(expand.dots = FALSE)
    temp <- c("", "data", "weights", "subset", "na.action")
    m <- m[match(temp, names(m), nomatch = 0)]
    if (missing(variance))
        theta <- NULL
    else theta <- variance
    reSt <- reStruct(random, REML = F, data = NULL)
    gform <- getGroupsFormula(reSt)
    if (is.null(gform)) {
        temp.fixed <- fixed
        gvars <- NULL
    }
    else {
        gvars <- all.vars(random)
        fvars <- all.vars(formula)
        gvars <- gvars[is.na(match(gvars, fvars))]
        temp.fixed <- paste(deparse(as.vector(fixed)), collapse = "")
        temp.fixed <- paste(temp.fixed, paste(gvars, collapse = "+"),
            sep = "+")
        temp.fixed <- as.formula(temp.fixed)
```

```
}
m$formula <- temp.fixed
m[[1]] <- as.name("model.frame")
m <- eval(m, sys.parent())
Terms <- terms(fixed)
X <- model.matrix(Terms, m)
Y <- model.extract(m, "response")
n <- length(Y)
weights <- model.extract(m, "weights")
offset <- attr(Terms, "offset")
tt <- length(offset)
offset <- if (tt == 0)
    rep(0, n)
else if (tt == 1)
    m[[offset]]
else {
    ff <- m[[offset[1]]]
    for (i in 2:tt) ff <- ff + m[[offset[i]]]
    ff
}
ncluster <- length(gvars)
if (ncluster == 0)
    stop("No grouping variables found")
groups <- getGroups(m, gform)
temp <- coxme.varcheck(ncluster, varlist, n, gvars, groups,
    sparse, rescale, pdcheck)
varlist <- temp$varlist
kindex <- temp$kindex
ntheta <- temp$ntheta
theta.names <- NULL
for (i in 1:ncluster) {
    if (ntheta[i] == 1)
        theta.names <- c(theta.names, gvars[i])
    else theta.names <- c(theta.names, paste(gvars[i], 1:ntheta[i],
        sep = ""))
}
if (length(theta) == 0)
```

```
    theta <- rep(0, sum(ntheta))
else if (length(theta) != sum(ntheta))
    stop("Wrong length for theta")
names(theta) <- theta.names
theta.names <- NULL
for (i in 1:ncluster) {
    if (ntheta[i] == 1)
        theta.names <- c(theta.names, gvars[i])
    else theta.names <- c(theta.names, paste(gvars[i], 1:ntheta[i],
        sep = ""))
}
if (length(theta) == 0)
    theta <- rep(0, sum(ntheta))
else if (length(theta) != sum(ntheta))
    stop("Wrong length for theta")
names(theta) <- theta.names
tindex <- which(theta == 0)
if (ncluster > 1)
    stop("function can have only 1 random effect")
varlist <- varlist[[1]]
kindex <- kindex[, 1]
if (max(kindex) != n)
    stop("The random effect must be 1 per subject")
ntheta <- ntheta[1]
kindex2 <- integer(n)
kindex2[kindex] <- 1:n
logfun <- function(itheta,X, Y, varlist, theta, tindex,center) {
  theta[tindex] <- exp(itheta[1])
    tkmat <- varlist[[1]]
    tkmat@blocks <- tkmat@blocks * theta[1]
    diag(tkmat) <- diag(tkmat + 1)
    if (length(varlist) > 1) {
        for (i in 2:length(varlist)) tkmat@blocks <- varlist[[i]]@blocks *
            theta[i] + tkmat@blocks
    }
    tkmat@blocks <- tkmat@blocks/tkmat@blocks[1]
    gk <- gchol(tkmat)
```

```
    newx <- solve(gk, X, full = FALSE)
    newy <- solve(gk, Y, full = FALSE)
    modR  <- rlm(newy~newx)
    n <- length(Y)
    loglik <- (n/2) * (log(modR$s^2) - center)+ sum(log(diag(gk)))/2
    loglik
}
newX <- X[kindex2, ]
newY <- as.vector(Y[kindex2])
dimnames(newX) <- NULL
if (length(tindex) > 0) {
    center <- log(mad(Y)^2)
    nfit <- optim(par = rep(-1, length(tindex)+1), logfun, X = newX,
        Y = newY, varlist = varlist, theta = theta, tindex = tindex,
        center = center)
    iter <- nfit$counts
    theta[tindex] <- exp(nfit$par[1])
}
else iter <- 0
tkmat <- varlist[[1]]
tkmat@blocks <- tkmat@blocks * theta[1]
diag(tkmat) <- diag(tkmat + 1)
if (length(varlist) > 1) {
    for (i in 2:length(varlist)) tkmat@blocks <- varlist[[i]]@blocks *
        theta[i] + tkmat@blocks
}
gk <- gchol(tkmat)
xok <- as.matrix(solve(gk, newX, full = F))
yok <- solve(gk, newY, full = FALSE)
lfit <- rlm(yok ~ 0 + xok)
names(lfit$coefficients) <- dimnames(X)[[2]]
ls <- summary(lfit)
resid.var <- lfit$s^2
theta <- c(theta * resid.var, resid.var)
names(theta) <- c(theta.names, "resid")
fitted <- c(X %*% lfit$coef)
residuals <- Y - fitted
```

```
    frail <- residuals[kindex2]
    names(frail) <- groups
    fcoef <- lfit$coef
    call$fixed <- fixed
    call$random <- random
    fit <- list(coefficients = list(fixed = fcoef, random = frail),
        theta = theta, variance = ls$cov.unscaled * lfit$s^2, s=lfit$s,
        X=newX,Weigth=lfit$w, ctable = ls$coefficients, residuals = residuals,
        fitted.values = fitted, effects = lfit$effects, rank = lfit$rank,
        assign = lfit$assign, df.residual = lfit$df.residual - length(theta),
        loglik = (-n/2)*(log(lfit$s^2) + 1 + log(2 * pi)) -
            sum(log(diag(gk)))/2,
        iter = iter, n = n, call = call, method = "ML")
    na.action <- attr(m, "na.action")
    if (length(na.action))
        fit$na.action <- na.action
    oldClass(fit) <- c("lmekin")
    fit
}
```

C.4 R routine for the example robust single-SNP analysis

```
#loading Weber's data, i.e, data on phenotype and 61 SNPs for 493
#plants
candidate<-read.table("WeberDataSet.txt")

#creating a vector with the SNPs identification
SNPsNames<-colnames(candidate);SNPsNames<-SNPsNames[-1]

#creating a vector for the phenotypic values - 493x1 matrix
Y<-as.matrix(candidate[,1]);colnames(Y)<-"FERL"

#creating a matrix with SNP data - 493x61 matrix
SNPs<-candidate[,2:62]; nSNPs<-dim(candidate)[2]-1

#loading Weber's 10 Principal Components - 493x10 matrix
```

```
PCs<-read.table("WeberPrincComps493Plants.txt");cpsnames<-names(PCs)

#loading Weber's Kinship matrix for the 493 plants - it's a 493x493
#matrix - this is not a semi-definite positive matrix
ship<-read.table("WeberKinship493Plants.txt");KShip<-as.matrix(ship)

#plant identification is taken out of KShip columns
plantid<-colnames(KShip)

#SNPs rows identification
rownames(SNPs)<-plantid

#loading the necessary R libraries
library(kinship); library(MASS);
library(numDeriv); library(nortest)

#load the lmekinR routine that is not loaded with package Kinship by
#running its code on the R environment

#creating a vector to store the association tests p-values
teste<-vector()

#creating some more vectors to store information on additive and
#dominance effects, phenotypic standard error and number of
#observations used in the regression analysis respectively; additive
#effects are calculated as
#abs(mean(Y.fitted_Hom_AA)-mean(Y.fitted_Hom_aa))/2; dominance
#effects are calculated as
#mean(Y.fitted_Het_Aa)-(mean(Y.fitted_Hom_AA)+mean(Y.fitted_Hom_aa))/2
a<-vector(); dom<-vector(); sd_pheno<-vector(); Nobs<-vector()

#initiating the cycle where all 61 p-values will be calculated in a
#single-SNP analysis
for(w in 1:dim(SNPs)[2]){
        dat<-cbind(Y,PCs,SNPs[,w])
        #taking out obs with non-observable SNP genotype:
        dados<-dat[!is.na(SNPs[,w]),]
```

```
plantnames<-rownames(dados)
#defining the new kinship matrix:
KS<-KShip[plantnames,plantnames]
#testing if KS is semi-definite positive; if it's not an
#adjustment is made so that it becomes semi-definite positive
eigenKS<-eigen(KS)$values
if(length(eigenKS[eigenKS<0])>0){
    lambda<-abs(eigenKS[length(eigenKS)])+1e-9
    KSpd<-(KS+lambda*diag(dim(KS)[1]))/(1+lambda)
    rownames(KSpd)<-plantnames
    colnames(KSpd)<-plantnames
    KS<-KSpd
}
#separating the data in phenotype and PCs+SNPs matrices:
dat<-dados[,-1]
Y1<-dados[,1]
colnames(dat)[dim(PCs)[2]+1]<-colnames(SNPs)[w]

#running the robust mixed model
modelo<-lmekinR(Y1~dat[,1]+dat[,2]+dat[,3]+dat[,4]+dat[,5]+dat[,6]
                  +dat[,7]+dat[,8]+dat[,9]+dat[,10]+dat[,11],
                  data=dat,random=~1|plantnames,varlist=list(KS))

#calculating the additive and dominance effects of SNP[,w],
#the number of observations used in the regression model and the
#phenotypic standard error
Nobs[w]<-modelo$n
Y_fitted<-modelo$fitted.values
sd_pheno[w]<-sqrt(var(Y1))
a[w]<-abs(mean(Y_fitted[dat[,11]==names(summary(dat[,11])[1])])
          -mean(Y_fitted[dat[,11]==names(summary(dat[,11])[3])]))/2
dom[w]<-mean(Y_fitted[dat[,11]==names(summary(dat[,11])[2])])-
    ( mean(Y_fitted[dat[,11]==names(summary(dat[,11])[1])])
     +mean(Y_fitted[dat[,11]==names(summary(dat[,11])[3])]) )/2

#performing a Wald type test on the effect of SNP on the
#trait; we need to have the inverse of the var/covar matrix,
```

```
        #the number of estimated parameters for SNP w and its values
        x<-modelo
        l<-dim(x$variance)[1]
        covarmat<-x$variance[12:l,12:l]
        invcovarmat<-solve(covarmat)
        estim<-x$ctable[12:l,1];estim<-as.matrix(estim)
        q<-length(levels(SNPs[,w]))-1

        #writing the test statistic and calculating the test p-value
        Twald<-t(estim)%*%invcovarmat%*%estim/q
        pv<-1-pf(Twald,q,x$n-l-q)
        #storing the test p-value
        teste[w]<-pv
}
```

C.5 R routine for the example robust multiple-SNP analysis

```
#we define the vector of the SNPs declared significant in the single
#SNP analysis and then subset the data set of all SNPs
sigsNames<-c("zagl1.1","PZD00073.5","PZD00006.1","PZD00022.3",
             "PZD00073.8","te1.3","ba1.9","PZB00049.7")
SNPs2<-subset(SNPs,select=sigsNames) sigsdat<-cbind(Y,PCs,SNPs2)

#we now remove from the data all observations that do not have an
#observable genotype at a SNP and construct the new kinship matrix
#for the new data set

for(i in 1:dim(SNPs2)[2]){
    sigsdat<-sigsdat[!is.na(sigsdat[,i+11]),]
}
plantnames<-rownames(sigsdat)
KS<-KShip[plantnames,plantnames]

#we test if the kinship matrix is or not positive definite
eigenKS<-eigen(KS)$values
        if(length(eigenKS[eigenKS<0])>0){
```

```
            lambda<-abs(eigenKS[length(eigenKS)])+1e-9
            KSpd<-(KS+lambda*diag(dim(KS)[1]))/(1+lambda)

            rownames(KSpd)<-plantnames
            colnames(KSpd)<-plantnames
            KS<-KSpd
        }

#we define the sets for the multiple regression analysis
Y2<-sigsdat[,1] PCs2<-sigsdat[,2:11]
SNPs2<-sigsdat[,12:dim(sigsdat)[2]]

sett<-cbind(PCs2,SNPs2) teste<-vector() for(w in 1:dim(SNPs2)[2]){
        modelo<-lmekinR(Y2~sett[,1]+sett[,2]+sett[,3]+sett[,4]+sett[,5]+sett[,6]
                +sett[,7]+sett[,8]+sett[,9]+sett[,10]+sett[,11]+sett[,12]
                +sett[,13]+sett[,14]+sett[,15]+sett[,16]+sett[,17]+sett[,18],
                data=sett,random=~1|plantnames,varlist=list(KS))

        #when w=1 we are taking care of the last SNP in the data
        #set, i.e, SNP k=dim(SNPs2)[2]-w+1
        k<-dim(SNPs2)[2]-w+1
        n<-modelo$n

        Kin<-KS
        p<-dim(modelo$ctable)[1]
        q<-length(levels(SNPs2[,k]))-1


        x<-modelo
        l<-dim(x$variance)[1]
        covarmat<-x$variance[(l-q+1):l,(l-q+1):l]
        invcovarmat<-solve(covarmat)
        estim<-x$ctable[(l-q+1):l,1];estim<-as.matrix(estim)
        Twald<-t(estim)%*%invcovarmat%*%estim/q

        pv<-1-pf(Twald,q,n-l-q) #pvalue de teste
```

```
        teste[k]<-pv

        #we now "rotate" the SNP matrix, putting the last SNP in
        #first place; in this way in the next iteration we will
        #calculate the p-value for the new last SNP in the SNP data
        #set - we have to proceed this way because we are dealing with
        #unbalanced data
        lixo<-SNPs2[,dim(SNPs2)[2]];lixo<-as.matrix(lixo)
        colnames(lixo)<-names(SNPs2)[dim(SNPs2)[2]]
        SNPs2<-SNPs2[,-dim(SNPs2)[2]]
        SNPs2<-cbind(lixo,SNPs2)
        sett<-cbind(PCs2,SNPs2)
}
```

C.6 R routine for the example classic single-SNP analysis

The initial part of the R code for this section is equal to that of Section C.4. In this program we used the Satterthwaite approximation for the df of the F test distribution since this was the approach taken by Weber in her analysis.

```
#we define extra matrices to keep the values of the regression
#residuals and fitted values:
residuals<-matrix(NA,493,61)
fitted<-matrix(NA,493,61)

#we also want to keep track of the plants that were used in each
#regression:
plants<-matrix(NA,493,61)

#load the lmekinC routine that is not loaded with package Kinship by
#running its code on the R environment

#initiating the cycle where all 61 p-values will be calculated in a
#single-SNP analysis
for(w in 1:dim(SNPs)[2]){
        ... the same as in the robust counterpart ...
```

```
#running the classic mixed model
modelo<-lmekinC(Y1~dat[,1]+dat[,2]+dat[,3]+dat[,4]+dat[,5]
                  +dat[,6]+dat[,7]+dat[,8]+dat[,9]+dat[,10]+dat[,11],
                  data=dat,random=~1|plantnames,varlist=list(KS))

#keeping the regression residuals, fitted values and plants
#identification used in the regression:
residuals[1:modelo$n,w]<-modelo$residuals
fitted[1:modelo$n,w]<-modelo$fitted
plants[1:modelo$n,w]<-plantnames

#calculating the additive and dominance effects of SNP[,w],
#the number of observations used in the regression model and the
#phenotypic standard error:
... the same as in the robust counterpart ...

#we identify the number of observations used in the
#regressio, which equals the total number of the model estimated random
#parameters:
n<-modelo$n

#we change the name of the kinship matrix and define
#variables for the number of fixed parameters estimated in the model, p, an
#the number of parameters estimated for SNPw, q
Kin<-KS
p<-dim(modelo$ctable)[1]
q<-p-(dim(PCs)[2]+1)

#defining vectors to keep the estimates of the fixed and random
#parameters, respectively, b and u
u<-modelo$coefficients$random
b<-modelo$coefficients$fixed
b<-as.matrix(b)

#identifying the specification matrix relative to the
#model's fixed component:
```

```
X<-modelo$X

#computing the matrices needed to perform the hypothesis
#test on the fixed parameters coefficients
I<-diag(1,modelo$n,modelo$n)  #identity matrix nxn
G<-modelo$theta[1]*Kin        #matrix nxn
R<-modelo$theta[2]*I          #matrix nxn where modelo$theta[2]==sigma^2
V<-G+R                        #matriz de nxn

#we define the C matrix as a function of (thet,sig)
Cfunc<-function(thet,sig){
    G<-thet*Kin
    R<-sig*I
    V<-G+R
    C<-ginv(t(X)%*%solve(V)%*%X)
    #since sig is the usual and not the centered estimate of sigma,
    #we have still to center it; if we could get it from the
    #regression model it would only be necessary to consider
    #C=modelo$variance, but that is not the case here
    C1<-(modelo$n/(modelo$n-modelo$rank))*C
    return(C1)
}

#we evaluate C for the estimates (modelo$theta[1],modelo$theta[2])
#of (theta,sigma^2) obtained in the regression:
C<-Cfunc(modelo$theta[1],modelo$theta[2])

#we now need to determine the matrix of var/covar of the
#estimated parameters of (theta,sigma), A = - H where H is
#the Hessian matrix corresponding to the second derivatives
#of the likelihood function;
#following Walsh (pp789) Aij=[S^(-1)]ij para i=1,2
#where Sij=1/2tr(V^-1 Vi V^-1 Vj) and where V2=I and V1=Kin
S11<-0.5*sum(diag( solve(V) %*% Kin %*% solve(V) %*% Kin))
S22<-0.5*sum(diag( solve(V) %*% I %*% solve(V) %*% I))
S12<-0.5*sum(diag( solve(V) %*% Kin %*% solve(V) %*% I))
S21<-S12
```

```
S<-matrix(0,2,2);S[1,1]<-S11;S[2,2]<-S22;S[1,2]<-S12;S[2,1]<-S21
A<-matrix(0,2,2)
A[1,1]<-solve(S)[1,1]
A[2,2]<-solve(S)[2,2]
A[1,2]<-solve(S)[1,2]
A[2,1]<-A[1,2]

#we will now perform test t or F depending on the number of
#parameters estimated for the SNP
if(q==2){
    L<-matrix(0,q,p)
    L[q-1,p-1]<-1
    L[q,p]<-1
    #computing the test statistic
    F<-t(b)%*%t(L)%*%solve(L%*%C%*%t(L))%*%L%*%b/q
    #the df of the denominator have to be estimated;
    #we will use the WelchSatterthwaite approximation
    dec<-eigen(L%*%C%*%t(L)) #spectral decomposition of LCL'=PDP'
    d<-dec$values
    D<-diag(d)
    P<-t(dec$vectors)
    PL<-P%*%L                #matrix 2xn

    v<-vector()
    for(j in 1:2){
        l<-as.matrix(PL[j,])
        fun<-function(x){thet<-x[1];sig<-x[2];t(l)%*%Cfunc(thet,sig)%*%l}
        parms<-c(modelo$theta[1],modelo$theta[2])
        g<-grad( fun, parms );g<-as.matrix(g)
        v[j]<-2*D[j,j]^2/(t(g)%*%A%*%g)
    }
    bool<-vector();
    for(i in 1:2){if(v[i]>2){bool[i]<-1}else{bool[i]<-0}}
    vv<-vector();for(i in 1:2){vv[i]<-v[i]/(v[i]-2)*bool[i]}
    E<-sum(vv)
    if(E>q){
        dfF<-2*E/(E-q)
```

```
            pv<-1-pf(F,q,dfF)
        }else{pv<-0}
    }
    else{
        if(q==1){
            L<-vector();
            for(i in 1:p){L[i]<-0}
            L[p]<-1
            L<-as.matrix(L)
            T<-t(L)%*%b/sqrt(t(L)%*%C%*%L)

            #the df of the t distribution have to be estimated;
            #we will use the WelchSatterthwaite approximation
            fun<-function(x){thet<-x[1];sig<-x[2];t(L)%*%Cfunc(thet,sig)%*%L}
            parms<-c(modelo$theta[1],modelo$theta[2])
            g<-grad( fun, parms );g<-as.matrix(g)
            dfF<-2*(t(L)%*%C%*%L)^2/(t(g)%*%A%*%g)
            pv<-1-pf(T^2,q,dfF)
        }
    }
    teste[w]<-pv
}
```

C.7 R routine for the example classic multiple-SNP analysis

```
#we define the vector of the SNPs declared significant in the single
#SNP analysis and then subset the data set of all SNPs
sigsNames<-c("zagl1.1","PZD00073.5","PZD00006.1","PZD00022.3",
             "PZD00073.8","te1.3","ba1.9","zagl1.6","PZB00049.7")
SNPs2<-subset(SNPs,select=sigsNames)
sigsdat<-cbind(Y,PCs,SNPs2)

#we now remove from the data all observations that do not have an
#observable genotype at a SNP and construct the new kinship matrix
#for the new data set
```

```
for(i in 1:dim(SNPs2)[2]){
    sigsdat<-sigsdat[!is.na(sigsdat[,i+11]),]
}
plantnames<-rownames(sigsdat)
KS<-KShip[plantnames,plantnames]

#we test if the kinship matrix is or not positive definite
eigenKS<-eigen(KS)$values
        if(length(eigenKS[eigenKS<0])>0){
            lambda<-abs(eigenKS[length(eigenKS)])+1e-9
            KSpd<-(KS+lambda*diag(dim(KS)[1]))/(1+lambda)
            rownames(KSpd)<-plantnames
            colnames(KSpd)<-plantnames
            KS<-KSpd
        }

#we define the sets for the multiple regression analysis
Y2<-sigsdat[,1] PCs2<-sigsdat[,2:11]
SNPs2<-sigsdat[,12:dim(sigsdat)[2]] sett<-cbind(PCs2,SNPs2)
teste<-vector()

for(w in 1:dim(SNPs2)[2]){
        modelo<-lmekinC(Y2~sett[,1]+sett[,2]+sett[,3]+sett[,4]+sett[,5]+sett[,6]
                +sett[,7]+sett[,8]+sett[,9]+sett[,10]+sett[,11]+sett[,12]
                +sett[,13]+sett[,14]+sett[,15]+sett[,16]+sett[,17]+sett[,18]
                +sett[,19],data=sett,random=~1|plantnames,varlist=list(KS))

        #when w=1 we are taking care of the last SNP in the data
        #set, i.e, SNP k=dim(SNPs2)[2]-w+1
        k<-dim(SNPs2)[2]-w+1
        n<-modelo$n         #num obs used in the regression == num params
                            #estimated in the model
        plantsused<-names(modelo$coefficients$random)   #plants used
                                                        #in the analysis
        Kin<-KS
        p<-dim(modelo$ctable)[1]                #num fixed params estimated
        q<-length(levels(SNPs2[,k]))-1          #num params estimated for SNPw
```

```
#getting the estimated coefficients and matrix X
u<-modelo$coefficients$random           #random params estimates
b<-modelo$coefficients$fixed            #fixed params estimates
b<-as.matrix(b)
X<-modelo$X                pecification matrix of the fixed component
I<-diag(1,modelo$n,modelo$n)        #nxn identity matrix
G<-modelo$theta[1]*Kin              #nxn matrix
R<-modelo$theta[2]*I                #nxn matrix
modelo$theta[2]==sigma^2
V<-G+R                              #nxn matrix

Cfunc<-function(thet,sig){
    G<-thet*Kin
    R<-sig*I
    V<-G+R
    C<-ginv(t(X)%*%solve(V)%*%X)
    return(C)
}

#C function of (theta,sigma^2); we call sig to sigma^2
C<-Cfunc(modelo$theta[1],modelo$theta[2])

#we must calculate the variance/covariance matrix of the estimated theta
#and sigma parameters, A = - H where H is the hessian matrix
#corresponding to the second derivatives of the likelihood function
#also, A12=A21=0, i.e, cov(theta,sigma)=cov(sigma,theta)=0;
#taking delta=(delta1,delta2)=(sigma,theta)
#according to Walsh (pags789) Aii=[S^(-1)]ii para i=1,2
#where Sii=1/2tr(V^-1 V1 V^-1 V1)
#and where V1=I if deltai=sigma and V1=Kin;

S11<-0.5*sum(diag( solve(V) %*% Kin %*% solve(V) %*% Kin))
S22<-0.5*sum(diag( solve(V) %*% I %*% solve(V) %*% I))
S12<-0.5*sum(diag( solve(V) %*% Kin %*% solve(V) %*% I))
S21<-S12
S<-matrix(0,2,2);S[1,1]<-S11;S[2,2]<-S22;S[1,2]<-S12;S[2,1]<-S21
```

```
A<-matrix(0,2,2)
A[1,1]<-solve(S)[1,1]
A[2,2]<-solve(S)[2,2]
A[1,2]<-solve(S)[1,2]
A[2,1]<-A[1,2]

#choosing between the t and the F test
if(q==2){
    L<-matrix(0,q,p)
    L[q-1,p-1]<-1
    L[q,p]<-1
    F<-t(b)%*%t(L)%*%solve(L%*%C%*%t(L))%*%L%*%b/q

    #estimating the denominator degrees of freedom of the F distribution
    #(WelchSatterthwaite equation);
    #we begin with the spectral decomposition of LCL'=PDP'
    dec<-eigen(L%*%C%*%t(L))
    d<-dec$values
    D<-diag(d)
    P<-t(dec$vectors)
    PL<-P%*%L       #2 x p matrix

    v<-vector()
    for(j in 1:2){
        l<-as.matrix(PL[j,])
        #the next function is a real function of 2 variables, thet and sig
        fun<-function(x){thet<-x[1];sig<-x[2];t(l)%*%Cfunc(thet,sig)%*%l}
        parms<-c(modelo$theta[1],modelo$theta[2])
        #calculating the gradient of fun in (theta,sigma)
        g<-grad( fun, parms );g<-as.matrix(g)
        v[j]<-2*D[j,j]^2/(t(g)%*%A%*%g)
    }
    bool<-vector();
    for(i in 1:2){if(v[i]>2){bool[i]<-1}else{bool[i]<-0}}
    vv<-vector();for(i in 1:2){vv[i]<-v[i]/(v[i]-2)*bool[i]}
    E<-sum(vv)
    if(E>q){
```

```
            dfF<-2*E/(E-q)
            pv<-1-pf(F,q,dfF)
        }else{pv<-0}
    }
    else{
        if(q==1){
            L<-vector();
            for(i in 1:p){L[i]<-0}
            L[p]<-1
            L<-as.matrix(L)
            T<-t(L)%*%b/sqrt(t(L)%*%C%*%L)

            fun<-function(x){thet<-x[1];sig<-x[2];t(L)%*%Cfunc(thet,sig)%*%L}
            parms<-c(modelo$theta[1],modelo$theta[2])
            g<-grad( fun, parms );g<-as.matrix(g)
            dfF<-2*(t(L)%*%C%*%L)^2/(t(g)%*%A%*%g)
            pv<-1-pf(T^2,q,dfF)
        }
    }

    teste[k]<-pv

    #we now "rotate" the SNP matrix, putting the last SNP in
    #first place; in this way in the next iteration we will
    #calculate the p-value for the new last SNP in the SNP data
    #set - we have to proceed this way because we are dealing with
    #unbalanced data
    lixo<-SNPs2[,dim(SNPs2)[2]];lixo<-as.matrix(lixo)
    colnames(lixo)<-names(SNPs2)[dim(SNPs2)[2]]
    SNPs2<-SNPs2[,-dim(SNPs2)[2]]
    SNPs2<-cbind(lixo,SNPs2)
    sett<-cbind(PCs2,SNPs2)
}
```

References

[1] Abecassis, G.R., Gosh, D. and Nichols, T.E. (2005). Linkage disequilibrium: ancient history drives the new genetics. *Human Heredity* **59**, 118–124.

[2] Abdurakhmonov, I.Y. and Abdukarimov, A. (2008). Application of association mapping to understanding the genetic diversity of plant germplasm resources. *International Journal of Plant Genomics*. ID 574927, 1–18.

[3] Alnosaier W.S. (2007). *Kenward-Roger approximate F test for fixed effects in mixed linear models*. Phd Thesis.

[4] Anderson-Sprecher, R.A. (1994). Model comparison and R^2. *The American Statistician* **48**, 113–117.

[5] Anderson, M.J. and Legendre, P. (1999). An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *Journal of Statistical Computation and Simulation* **62**, 271–303.

[6] Arshadi, N., Chang, B. and Kustra, R. (2009). Predictive modeling in case-control single-nucleotide polymorphism studies in the presence of population stratification: a case study using Genetic Analysis Workshop 16 Problem 1 dataset. *BMC Proceedings* **3(7)**, S60.

[7] Atkinson, A.C. (1985). *Plots, transformations and regression. An introduction to graphical methods of diagnostic regression analysis*. New York: Oxford University Press.

[8] Bacanu S.A., Devlin B. and Roeder K. (2002). Association studies for quantitative traits in structured populations. *Genetic Epidemiology* **22**, 78–93.

[9] Bacanu S.A., Devlin B. and Roeder K. (2000). The power of genomic control. *American Journal of Human Genetics* **66**, 1933–1944.

[10] Balding D.J., Bishop M. and Cannings C. (2003). *Handbook of statistical genetics*. Chichester, West Sussex, England. John Wiley and Sons.

[11] Balding D.J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics* **7**, 781–791.

[12] Barnett, V. and Lewis, T. (1978). *Outliers in statistical data*. New York: Wiley.

[13] Beasley, T.M. and Erickson, S. (2009). Rank-based inverse normal transformations are increasingly used, but are they merited? *Behavior Genetics* **39(5)**, 580–595. doi:10.1007/s10519-009-9281-0.

[14] Belmonte, M. and Yurgelun, T.D. (2001). Permutation testing made practical for functional magnetic resonance image analysis. *IEEE Transactions on Medical Imaging* **20(3)**, 243-248.

[15] Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression diagnostics: identifying influential data and sources of collinearity*. New York: John Wiley.

[16] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* **57(1)**, 289–300.

[17] Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics* **25(1)**, 60–83.

[18] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* **29(4)**, 1165–1188.

[19] Benjamini, Y., Krieger, A.M. and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, **93(3)**, 491–507.

[20] Benjamini, Y. and Heller, R. (2007). False discovery rate for spatial signals. *Journal of the American Statistical Association* **102**, 1272–1281.

[21] Benjamini, Y. and Gavrilov, Y. (2009). A simple forward selection procedure based on false discovery rate control. *Annals of Applied Statistics* **3**, 179–198.

[22] Benjamini, Y., Heller, R. and Yekutieli, Y. (2009). Selective inference in complex research. *Philosophical Transactions of the Royal Society A* **367**, 4255–4271.

[23] Benjamini, Y. (2010). Simultaneous and selective inference: current successes and future challenges. *Biometrical Journal* **52(6)**, 708–721.

[24] Bianco, A., Boente, G. and Rienzo, J. (2000). Some results for robust GM-based estimators in heteroscedastic regression models. *Journal of Statistical Planning and Inference* **89(1-2)**, 215–242.

[25] Blanchard, G. and Roquain, E. (2009). Adaptive FDR control under independence and dependence. *Journal of Machine Learning Research* **10**, 2837–2871.

[26] Boente, G., Pires, A.M. and Rodrigues, I.M. (2002). Influence functions and outlier detection under the common principal components model: a robust approach. *Biometrika* **89(4)**, 861–875.

[27] Bonferroni, C.E. (1935). Il calcolo delle assicurazioni su gruppi di teste. *In Studi in Onore del Professore Salvatore Ortu Carboni*. Rome: Italy, 13–60.

[28] Bonferroni, C.E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze* **8**, 3–62.

[29] Box, G.E.P. (1953) Non-normality and tests on variances. *Biometrika*, **40**, 318–335.

[30] Buckler, E.S. and Stevens, N.M. (2006). Maize origins, domestication, and selection. *Darwin's Harvest*. Columbia University Press, NY.

[31] Bureau A., Dupuis J., Falks K. *et al.* (2005). Identifying SNPs predictive of phenotype using random forests. *Genetic Epidemiology* **28(2)**, 171–82.

[32] Byng, M.C., Whittaker, J.C., Cuthbert, A.P. *et al.* (2003). SNP subset selection for genetic association studies. *Annals of Human Genetics* **67**, 543–556.

[33] Campbell, M. (1976). Explanation of Mendel's results. *Centarurus* **20(2)**, 159–174.

[34] Cardon, L.R. and Palmer, L.J. (2003). Population stratification and spurious allelic association. *Lancet* **361**, 598–604.

[35] Carlson, C.S., Eberle, M.A., Rieder, M.J. *et al.* (2004). Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *American Journal of Human Genetics* **74**, 106–120.

[36] Carroll, R.J. and Ruppert, D. (1982). Robust estimation in heteroscedatic linear models. *Annals of Statistics* **10(2)**, 429–441.

[37] Chapman, J. and Whittaker, J. (2008). Analysis of multiple SNPs in candidate gene or region. *Genetic Epidemiololgy* **32(6)**, 560–566.

[38] Chatterjee, S. and Hadi, A.S. (1986). Influential observations, high leverage points, and outliers in linear regression. *Statistical Science* **1(3)**, 379–393.

[39] Chen, B.E., Sakoda, L., Hsing, A.W. and Rosenberg, P.S. (2006). Resampling-based multiple hypothesis testing procedures for genetic case-control association studies. *Genetic Epidemiology* **30**, 495–507.

[40] Cheverud, J.M. (2001). A simple correction for multiple comparisons in interval mapping genome scans. *Heredity* **87**, 52-58.

[41] Colhoun, H.M., McKeigue, P.M. and Davey, S.G. (2003). Problems of reporting genetic associations with complex outcomes. *Lancet* **361**, 865-872.

[42] Conover, W.J. and Iman, R.L. (1981) Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician* **35**, 121–129.

[43] Cook, R.D. (1977). Detection of influential observations in linear regression. *Technometrics* (American Statistical Association) **19(1)**, 15–18. doi:10.2307/1268249.

[44] Cook, R.D. and Weisberg, S. (1982). *Residuals and influence in regression.* Chapman and Hall, New York.

[45] Copt, S. and Feser, V. (2006). High-breakdown inference for mixed linear models. *Journal of the American Statistical Association* **101**, 292–300.

[46] Copt, S. and Heritier, S. (2007). Robust alternatives to the F-test in mixed linear models based on MM-estimates. *Biometrics*, **63**, 1045–1052.

[47] Cousin, E., Genin, E., Ricard, S., Chansac, C., Zompo, M. and Deleuze, J.F. (2003). Association studies in candidate genes: strategies to select SNPs to be tested. *Human Heredity* **56**, 151–159.

[48] Croux, C. and Rousseeuw, P.J. (1992). Time-efficient algorithms for two highly robust estimators of scale. *Computational Statistics* **1**, ed. Dodge and Whittaker, Physica-Verlag Heidelberg, 411–428.

[49] Croux, C. and Dehon, C. (2003). Estimators of the multiple correlation coefficient: local robustness and confidence intervals. *Statistical Papers* **44(3)**, 315–334.

[50] Daszykowski, M., Kaczmarek, K., Heyden, Y.V. and Walczak, B. (2007). Robust statistics in data analysis - a review, basic concepts. *Chemometrics and Intelligent Laboratory Systems* **85**, 203–219.

[51] de Haan, J.R., Bauerschmidt, S., van Schaik, R.C. *et al.* (2009). Robust ANOVA for microarray data. *Chemometrics and Intelligent Laboratory Systems* **98**, 38–44.

[52] Dell'Aquila, R. and Ronchetti, E. (2006). *Robust statistics and econometrics with economic and financial applications*. New York, Wiley.

[53] Devlin, B. and Roeder, K. (1997). The heritability of IQ. *Nature* **388**, 468–471.

[54] Devlin, B. and Roeder, K. (1999). Genomic control for Association Studies. *Biometrics* **55**, 997-1004.

[55] Devlin, B. and Roeder, K. (2001). Genomic control a new approach to genetic-based association studies. *Theoretical Population Biology* **60**, 155–166.

[56] Doebley, J. (2004). The genetics of maize evolution. *Annual Reviews on Genetics* **38**, 37-59.

[57] Donoho, D.L. and Huber, P.J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehman* (P.J. Bickel, K.A. Doksum and J.L. Hodges, Jr., eds.) 157–184, Wadsworth, Belmont, Calif.

[58] Dudoit, S., Shaffer, J.P. and Boldrick, J.C. (2003). Multiple hypotheses testing in microarray experiments. *Statistical Science* **18(1)**, 71–103.

[59] Dudoit, S. and Van der Laan, M.J. (2008). *Multiple Testing Procedures with Applications to Genomics*. Springer.

[60] Dunn, O.J. (1961). Multiple Comparisons Among Means. *Journal of the American Statistical Association* **56**, 52–64.

[61] Edwards, L.J., Muller, K.E., Wolfinger, R.D. *et al.* (2008). An R^2 statistic for fixed effects in the linear mixed model. *Statistics in Medicine* **27**, 6137–6157.

[62] Edwin, O. (2008). Controlling false discoveries in genetic studies. *American Journal of Medical Genetics* **147B**, 637–644.

[63] Eyre-Walker, A., Gaut, R.L., Hilton, H. *et al.* (1998). Investigation of the bottleneck leading to the domestication of maize. *Proceedings of the National Academy of Sciences USA* **95**, 4441-4446.

[64] Eskin, E., Halperin, E. and Karp, R.M. (2003). Large scale reconstruction of haplotypes from genotype data. RECOMB '03 *Proceedings of the 7th International Annual Conference on Research in Computational Molecular Biology*, 104–113.

[65] Emig, T.H. (1980). A Comparison of tests for Hardy-Weinberg equilibrium. *Biometrics* **36**, 627–642.

[66] Falconer, D.S. and Mackay, T.F.C. (1996). *Introduction to quantitative genetics* (fourth edition). Longman Group Limited, London.

[67] Falush, D., Stephens, M. and Pritchard, J.K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587.

[68] Fairbanks, D.J. and Rytting, B. (2001). Mendelian controversies: A botanical and historical review. *American Journal of Botany* **88(5)**, 737–752.

[69] Farcomeni, A. (2008). A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research* **17**, 347–388.

[70] Farrall, L.A. (1975). Controversy and conflict in science: A case study - The English biometric school and Mendel's laws. *Social Studies of Science* **5**, 269–301.

[71] Finner, H. and Strassburger, K. (2002). The partioning principle: a powerful tool in multiple decision theory. *Annals of Statistics* **30**, 1194-1213.

[72] Fisher, R. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Philosophical Transactions of the Royal Society of Edinburgh* **52**, 399–433.

[73] Fisher, R.A. (1930). *The genetical theory of natural selection.* Clarendon Press.

[74] Fisher, R. (1936). Has Mendel's work been rediscovered? *Annals of Science* **1**, 115–137.

[75] Foulkes, A.S. (2009). *Applied statistical genetics with R.* Springer.

[76] Franklin, A., Edwards, A.W.F., Fairbanks, D.J. *et al.* (2008). *Ending the Mendel-Fisher controversy.* University of Pittsburgh Press.

[77] Freedman, M., Reich, D., Penney, K. *et al.* (2004). Assessing the impact of population stratification on genetic association studies. *Nature Genetics* **36(4)**, 388–393.

[78] Froggatt, P. and Levin, N.C. (1971). The law of ancestral herefity and the Mendelian-ancestrian controversy in England, $1886-1906$. *Journal of Medical Genetics* **8**, 1–36.

[79] Galinat, W.C. (2001). A reconstrucion of a possible role of crucial observations leading to a rapid domestic transformation of wild teosint into the first maize. *Economic Botany* **55**, 570–574.

[80] Gao, X., Starmer, J. and Martin, E. (2008). A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic Epidemiology* **32**, 361–369.

[81] Garcia, S.A.F., Thornsberry, J.M. and Buckler, E.S. IV (2003). Structure of linkage disequilibrium in plants. *Annual Review of Plant Biology* **54**, 357–374.

[82] Garcia, L.V. (2004). Escaping the Bonferroni iron claw in ecological studies. *In Opinion*. IRNAS-CSIC.

[83] Ge, Y., Dudoit, S. and Speed, T.P. (2003). Resampling-based multiple testing for microarray data analysis. *TEST* **12(1)**, 1–77. doi: 10.1007/BF02595811.

[84] Gelman, A. and Pardoe, I. (2006). Bayesian measures of explained variance and poooling in multilevel (hierarchical) models. *Technometrics* **48**, 241–251.

[85] Genovese, C.R., and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society B* **64**, 419–437.

[86] Genovese, C.R and Wasserman, L. (2004). A stochastic process approach to false discovery control. *Annals of Statistics* **32**, 1035–1061.

[87] Genovese, C.R and Wasserman, L. (2006). Exceedance control of the false discovery proportion. *Journal of the American Statistical Association* **101**, 1408-1417.

[88] Gerhard, D. and Hothorn, L.A. (2010). Rank transformation in Haseman-Elston regression using scores for location-scale alternatives. *Human Heredity* **69**, 143–151. doi: 10.1159/000267994.

[89] Giltinan, D.M. and Carroll, R.J. (1986). Some new estimation methods for weighted regression when there are possible outliers. *Technometrics* **28(3)**, 219–230.

[90] Gomes, I., Collins, A., Lonjou, C., Thomas, N.S., Wilkinson, J., Watson, M. and Morton, N. (1999). Hardy-Weinberg quality control. *Annals of Human Genetics* **63**, 535-538.

[91] Gonick, L. and Wheelis, M. (1991). *The cartoon guide to genetics*. Harper Perennial, New York, NY, updated edition.

[92] González, J.R., Armengol, L., Solé, X. *et al.* (2008). SNPassoc: an R package to perform whole genome association studies. *Bioinformatics Application Note* **23(5)**, 644-645.

[93] Gopalakrishnan, S. and Qin, Z.S. (2006). TagSNP Selection Based on Pairwise LD Criteria and Power Analysis in Association Studies. *Pacific Symposium on Biocomputing* **11**, 511–522.

[94] Graser, H.U., Smith, S.P. and Tier, B. (1987). A derivative-free approach for estimating variance components in animal models by restricted maximum likelihood. *Journal of Animal Science* **64**, 1362–1370.

[95] Gravilov, J., Benjamini, Y. and Sarkar, S.K. (2009). An adaptive step-down procedure with proven FDR control under independence. *The Annals of Statistics* **37(2)**, 619–629.

[96] Greene, W. (1997). *Econometric Analysis* (third ed.). Prentice Hall.

[97] Gudbjartsson, D.F., Holm, H., Indridason, O.S. *et al.* (2010). Association of variants at UMOD with chronic kidney disease and kidney stonesrole of age and comorbid diseases. *PLoS Genetics* **6(7)**, e1001039.

[98] Guo, W. (2009). A note on adaptive Bonferroni and Holm procedures under dependence. *Biometrika* **96(4)**, 1012–1018. doi: 10.1093/biomet/asp048.

[99] Györffy, B., Kocsis, I. and Vásárhelyi, B. (2004). Biallelic genotype distributions in papers published in Gut between 1998 and 2003: altered conclusions after recalculating the Hardy-Weinberg equilibrium. *Gut* **53**, 614-616.

[100] Haldane, J.B.S. (1932). *The causes of evolution.* Longman, Green and Co. Princeton University Press.

[101] Halperin, E., Kimmel, G. and Shamir, R. (2005). Tag SNP selection in genotype data for maximizing SNP prediction accuracy. *Bioinformatics* **21(1)**, i195-i203.

[102] Hampel, F.R. (1968). *Contributions to the theory of robust estimation.* PhD Thesis, University of California, Berkeley.

[103] Hampel, F.R. (1971). A general qualitative definition of robustness. *Annals of Mathematical Statistics* **42**, 1887–1896.

[104] Hampel, F.R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association* **69**, 383–393.

[105] Hampel, F.R., Rochetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). *Robust statistics - the approach based on influence functions*. John Wiley and Sons.

[106] Harvey, A.C. (1976). Estimating regression models with multiplicative heteroscedasticity. *Econometrica* **44**, 460–465.

[107] Harville, D.A.(1974). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* **72**, 320–338.

[108] Hawkins, D.M. (1980). *Identification of outliers*. Chapman and Hall, London.

[109] Hayter, A.J. and Hsu, J.C. (1994). On the relationship between stepwise decision procedures and confidence sets. *Journal of the American Statistical Association* **89**, 128–136.

[110] He, J. and Zelikovsky, A. (2006). Tag SNP selection based on multivariate linear regression. *Lecture Notes in Computer Science* **3992**, 750–757. doi: 10.1007/11758525_101.

[111] Hekimoglu, S. and Erenoglu, R.C. (2009). New Method for outlier diagnostics in linear regression. *Journal of Surveying Engineering* **135(3)**, 85–89.

[112] Heritier, S., Cantoni, E., Copt, S. and Victoria-Feser, M.P. (2009). *Robust methods in biostatistics*. Wiley Series in Probability and Statistics.

[113] Hettmansperger, T.P. and McKean, J.W. (2001). *Robust non-parametric statistical methods*. Second Edition. CRC Press.

[114] Hicks, B.M., Krueger, R.F., Iacono, W.G., McGue, M. and Patrick, C.J. (2004). Family transmission and heritability of externalizing disorders: a twin-family study. *Archives of General Psychiatry* **61**, 922-928. doi:10.1001/ archpsyc.61.9.922.

[115] Hill, W.G. and Robertson, A. (1968). Linkage desiquilibrium in finite populations. *Theoretical and Applied Genetics (Der Züchter)* **38**, 226–231.

[116] Hoaglin, D.C. and Welsch, R.E. (1978). The hat matrix in regression and ANOVA. *The American Statistician* **32**, 17–22.

[117] Hochberg, Y. and Tamhane, A.C. (1987). *Multiple comparisons procedures*. Wiley.

[118] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800-802.

[119] Hochberg, Y. and Benjamini, Y. (1990). More powerful procedures for multiple significance testing. *Statistics in Medicine* **9(7)**, 811 - 818.

[120] Holland, J.B. and Cervantes-Martinez, C.T. (2003). Estimating and interpreting heritability for plant breeding. *Plant Breeding Reviews* **22**, 9–112.

[121] Holland, B. and Cheung, S.H. (2002). Familywise robustness criteria for multiple comparisons procedures. *Journal of the Royal Statistical Society B* **94**, 63–77.

[122] Holland, B.S. and Copenhaver, M.D. (1987). An improved sequentially rejective Bonferroni test procedure. *Biometrics* **43(2)**, 417–423

[123] Holm, S. (1979). A simple sequential rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65–70.

[124] Hommel, G. (1983). Tests of the overall hypothesis for arbitrary dependence structures. *Biometrical Journal* **25**, 423–430.

[125] Hommel, G. (1988). A stagewise rejective multiple test procedure on a modified Bonferroni test. *Biometrika* **75**, 383–386.

[126] Hosgood, H.D., Menashe, I., Shen, M. *et al.* (2008). Pathway-based evaluation of 380 candidate genes and lung cancer susceptibility suggests the importance of the cell cycle pathway. *Carcinogenesis* **29(10)**, 1938–1943.

[127] Hosking, L., Lumsden, S., Lewis, K. *et al.* (2004). Detection of genotyping errors by Hardy-Weinberg equilibrium testing. *European Journal of Human Genetics* **12**, 395-399.

[128] Howie, B.N., Carlson, C.S., Rieder, M.J. and Nickerson, D.A. (2006). Efficient selection of tagging single-nucleotide polymorphisms in multiple populations. *Human Genetics* **120**, 58–68.

[129] Huang, Y., Hsu, J.C. (2007). Hochberg's step-up method: cutting corners off Holm's step-down method. *Biometrika* **94**, 965-975.

[130] Huber, P.J. (1964) Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, **35**, 73–101.

[131] Huber, P.J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* **I**, 221–33.

[132] Huber, P.J. (1972). Robust statistics: a review. *The Annals of Mathematical Statistics* **43**, 1041–1067.

[133] Huber, P.J. (1973). Robust regression: Asymptotics, conjectures, and Monte Carlo. *Annals of Statistics* **1**, 799-821.

[134] Huber, P.J. and Ronchetti, E.M.(1981). *Robust statistics*. John Wiley & Sons, New York.

[135] Hubert, M. and Rousseeuw, P.J. (1997). Robust regression with both continuous and binary regressors. *Journal of Statistical Planning and Inference* **57**,153-163.

[136] Inouye, M., Silander, K., Hamalainen, E. *et al.* (2010). An immune response network associated with blood lipid levels. *PLoS Genetics* **6(9)**, e1001113. doi:10.1371/journal.pgen.1001113.

[137] Jajo, N.K. and Hussain, S.S. (1989). Robust estimators in linear regression model. *Journal of Management and Economic* **10**, 1-15.

[138] Jajo, N.K. (1993). A comparative study of two robust alternatives to least squares estimatores. *Journal of Engineering and Technology* **10**, 23-28.

[139] Jajo, N.K. (2005). A Review of robust regression and diagnostic procedures in linear regression. *Acta Mathematicae Applicatae Sinica, English Series* **21(2)**, 209-224.

[140] Jannot, A.S., Essioux, L., Reese, M.G. *et al.* (2003). Improved use of SNP information to detect the role of genes. *Genetic Epidemiology* **25**, 158–167.

[141] Jiang, L., Liu, J., Sun, D. *et al.* (2010). Genome wide association studies for milk production traits in chinese holstein population. *PLoS ONE* **5(10)**, e13661. doi:10.1371/journal.pone.0013661.

[142] Kardia, S.L.R., Greene, M.T., Boerwinkle, E. *et al.* (2008). Investigating the complex genetic architecture of ankle-brachial index, a measure of peripheral arterial disease, in non-Hispanic whites. *BMC Medical Genomics* **1(16)**, 1–11.

[143] Kempthorne, O. and Tandon, O.B. (1953). The estimation of heritability by regression of offspring on parent. *Biometrics* **9(1)**, 90–100.

[144] Kenward, M.G. and Roger, J.H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* **53(3)**, 983-997.

[145] Kimmel, G., Jordan, M.I., Halperin, E., Sharmir, R. and Karp, R.M. (2007). A randomization test for controlling population stratification in whole-genome association studies. *American Journal of Human Genetics* **81(5)**, 895–905.

[146] Kramer, M. (2005). R^2 Statistics for mixed models. Presented at the *17th Annual Kansas State University Conference on Applied Statistics in Agriculture*, Manhattan, Kansas.

[147] Kvalseth, T.O. (1985). A cautionary note about R^2. *The American Statistician* **39**, 279–285.

[148] Kwee, L.C., Liu, D., Lin, X. *et al.* (2008). A powerful and flexible multilocus association test for quantitative traits. *American Journal of Human Genetics* **82**, 386–397.

[149] LaMotte, L.R. (2007). A direct derivation of the REML likelihood function. *Statistical Papers* **48**, 321–337. Springer-Verlag.

[150] Langsrud, O. (2003). ANOVA for unbalanced data: use type II instead of type III sums of squares. *Statistics and Computing* **13**, 163–167.

[151] Lee, J.Y., Park, A.K., Lee, K.M. *et al.* (2009). Candidate gene approach evaluates association between innate immunity genes and breast cancer risk in Korean women. *Carcinogenesis* **30(9)**, 1528–1531.

[152] Legendre, P. and Legendre, L. (1998). *Numerical ecology*. 2nd English edition. Elsevier Science BV, Amsterdam. [Chapter 11].

[153] Lehmann, E.L., Romano, J.P. and Shaffer, J.P. (2005). On optimality of stepdown and steup multiple test procedures. *The Annals of Statistics* **33(3)**, 1084-1108.

[154] Lewin, B. (2006). *Essencial genes*. Pearson Prentice Hall.

[155] Li, J. and Ji, L. (2005). Adjusting multiple testing in multilocus analysis using the eigenvalues of a correction matrix. *Heredity* **95**, 221-227.

[156] Li, M. and Li, C. (2008). Assessing departure from Hardy-Weinberg equilibrium in the presence of disease association. *Genetic Epidemiology* **32(7)**, 589–599.

[157] Li, M., Reilly, C. and Hanson, T. (2008). A semiparametric test to detect associations between quantitative traits and candidate genes in structured populations. *Bioinformatics* **24**, 2356–2362.

[158] Li, M., Wang, K., Grant, S.F.A. *et al.* (2009). ATOM: a powerful gene-based association test by combining optimally weighted markers. *Bioinformatics* **25**, 497–503.

[159] Little, R.J.A. and Rubin, D.B. (2002). *Statistical analysis with missing data*. Wiley, New York.

[160] Liu, B.H. (1997). *Statistical genomics: linkage, mapping, and QTL analysis*. CRC Press.

[161] Liu, K. and Muse, S.V. (2005). PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics* **21**, 2128–2129.

[162] Lynch, M. and Walsh, B. (1997). *Genetics and analysis of quantitative traits*. Sinauer Associates, Sunderland, MA.

[163] Lopuhaä, H.P. and Rousseeuw, P.J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics* **19**, 229–248.

[164] Lourenço, V.M., Pires, A.M. and Kirst, M. (2011). Robust linear regression methods in association studies. *Bioinformatics*. **27(6)**, 815–821. doi: 10.1093/bioinformatics/btr006.

[165] Maechler, M. (2006). *Basic robust statistics: the robustbase package*. Reference Manual.

[166] Magee, L. (1990). R^2 measures based on Wald and likelihood ration joint signidicance tests. *The American Statistician* **44**, 250–253.

[167] Mallows, C.L. (1973). Some comments on C_p. *Technometrics* **15**, 661–675.

[168] Malo, N., Libiger, O. and Schork, N.J. (2008). Accommodating linkage disequilibrium in genetic association analysis via Ridge regression. *American Journal of Human Genetics* **82**, 375–385.

[169] Mander, A.P. (2001). Haplotype analysis in population-based association studies. *The Stata Journal* **1(1)**, 58–75.

[170] Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective sudies of disease. *Journal of the National Cancer Institute* **22**, 719–748.

[171] Maronna, R.A., Bustos, O.H. and Yohai, V.J. (1979). Bias and efficiency-robustness of general M-estimators for regression with random carriers. In *Smoothing Techniques for Curve Estimation* (T. Gasser and M. Rosenblatt, eds.), 91–116. Lecture Notes in Mathematics **757**, Springer, Berlin.

[172] Maronna, R.A. and Yohai, V.J. (2000). Robust regression with both continuous and categorical predictors. *Journal of Statistical Planning and Inference* **89**, 197–214.

[173] Maronna, R.A., Martin, R.D. and Yohai, V.J. (2006). *Robust statistics, theory and methods*. Chichester: Wiley.

[174] Martínez, S., Wheeler, N.C., Ersoz, E. *et al.* (2007). Association genetics in *Pinus taeda* L.I. wood property traits. *Genetics* **175**, 399–409.

[175] McKean, J.W., Sheather, S.J. and Hettmansperger, T.P. (1993). The use and interpretation of residuals based on robust estimation. *Journal of the American Statistical Association* **88(424)**, 1254–1263.

[176] McKean, J.W., Terpstra, J.T. and Kloke, J.D. (2009). Computational rank-based statistics. *Wiley Interdisciplinary Reviews: Computational Statistics* **1**, 132–140.

[177] Mendel, G. (1866). Experiments in plant hybridization (translation). *Classic papers in genetics*. Edited by J.A. Peters 1959. Englewood Cliffs, NJ: Prentice Hall.

[178] Meyer, K. (1983). Maximum likelihood procedures for estimating genetic parameters for later lactations of dairy cattle. *Journal of Dairy Science* **66**, 1988–1997.

[179] Meyer, K. (1985). Maximum likelihood estimation of variance components for a multivariate mixed model with equal design matrices. *Biometrics* **41**, 153-165.

[180] Myers, R.H., Montgomery, D.C. and Vining, G.G. (2001). *Generalized linear models: with applications in engineering and the sciences*. New York: Wiley.

[181] Moe, M., Lien, S., Aasmundstad, T. *et al.* (2009) Association between SNPs within candidate genes and compounds related to boar taint and reproduction. *BMC Genetics* **10**, 32.

[182] Morkûnienë, A. and Kuèinskas, V. (2005). Validation and selection of an informative set of sinlge-nucleotide polymorphisms in the TGFB3 gene. *Acta Medica Lituanica* **12(2)**, 1–7.

[183] Nakagawa, S. (2004). A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behavioral Ecology* **15**, 1044–1045

[184] Netter, J., Kutner, M.H., Nachtsheim, C.J. and Wasserman, W. (1996). *Applied linear statistical models*. McGraw-Hill, 4th edition.

[185] Nicolas, P., Sun, F. and Li, L.M. (2006). A model-based approach to selection of tag SNPs. *BMC Bioinformatics* **7(303)**, 1–14. doi:10.1186/1471-2105-7-303.

[186] Nielsen, D.M., Ehm, M.G. and Weir, B.S. (1998). Detecting markerdisease association by testing for Hardy-Weinberg disequilibrium at a marker locus. *American Journal of Human Genetics* **63**, 15311540.

[187] Nyholt, D.R. (2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage desiquibilbrium with each other. *American Journal of Human Genetics* **74**, 765-769

[188] Nyholt, D.R. (2005). Evaluation of Nyholt's procedure for multiple testing correction - author's reply. *Human Heredity* **60**, 61-62

[189] Novistki, C.E. (2004a). On Fisher's criticism of Mendel's results with the garden pea. *Genetics* **166**, 1133-1136.

[190] Novistki, C.E. (2004b). Revision of Fisher's analysis of Mendel's garden pea experiments. *Genetics* **166**, 1139-1140.

[191] Oh, J.H. and Gao, J. (2009). A kernel-based approach for detecting outliers of high-dimensional biological data. *BMC Bioinformatics* **10(4)**, S7.

[192] Oord, E.J.C.G. (2005). Controlling false discoveries in candidate gene studies. *Molecular Psychiatry* **10**, 230–231.

[193] Park, R.E. (1966). Estimation with heteroscedastic error terms. *Econometrica* **34**, 888.

[194] Paschou, P., Drineas, P., Lewis, J. *et al.* (2008). Tracing Sub-Structure in the European American Population with PCA-Informative Markers. *PLoS Genetics* **4(7)**, e1000114.

[195] Padilla, M.A. and Algina, J. (2004). Type I Error Rates for a One-Factor Within Subjects Design with Missing Values. *Journal of Modern Applied Statistical Methods* **3(2)**, 406–416.

[196] Pereira, I.G., Oliveira, H.N. and Rosa, G.J.M. (2007). Estudo de simulação de modelos lineares mistos com distribuição normal contaminada no melhoramento genético animal. *Revista Brasileira de Zootecnia* **36(5)**, 1304–1315.

[197] Pilgrim, I. (1984). The too-good-to-be-true paradox and Gregor Mendel. *Journal of Heredity* **75**, 501-502.

[198] Pilpel, A. (2007). Statistics is not enough: revisiting Ronald A. Fisher's critique (1936) of Mendel's experimental results (1866). *Studies in History and Philosophy of Biological and Biomedical Sciences* **38**, 618-626.

[199] Pinheiro, J.C. and Bates, D.M. (2000). *Mixed effects models in S and S-Plus*. Springer Verlag, New York.

[200] Pires, A.M. and Branco, J.A. (2007). *Introdução aos métodos estatísticos robustos*. Edições SPE.

[201] Pires, A.M. and Branco, J.A. (2010). A statistical model to explain the Mendel-Fisher controversy. *Statistical Science* **25(4)**, 545–565.

[202] Pires, A.M. and Rodrigues, I.M. (2007). Multiple linear regression with some correlated errors: classical and robust methods. *Statistics in Medicine* **26**, 2901-2918.

[203] Pollard, K.S. and Van Der Laan, M.J. (2003). Resampling based multiple testing with asymptotic strong control of type I error. *U.C. Berkeley Division of Biostatistics Working Paper Series*. University of California, Berkeley. Paper 121.

[204] Pollard, K.S., Birkner, M.D., Van Der Laan, M.J. and Dudoit, S. (2005). Test statistics null distributions in multiple testing: simulation studies and applications to genomics. *Journal de la Société Française de Statistique* **146(1-2)**, 77–116.

[205] Price, A.L., Patterson, N.J., Plenge, R.M. *et al.* (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904–909.

[206] Pritchard, J.K., Stephens, M. and Donelly, P. (2000a). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959.

[207] Pritchard, J.K., Stephens, M. (2000b). Association mapping in structured populations. *American Journal of Human Genetics* **67**, 170–181.

[208] Pritchard, J.K. and Przeworski, M. (2001). Linkage disequilibrium in humans: models and data (Review Article). *American Journal of Human Genetics* **69**, 1–14.

[209] Rafalski, J.A. and Morgante, M. (2003). Corn and humans: recombination and linkage desiquilibrium in two genomes of similar size. *Trends in Genetics* **20(2)**, 103–111.

[210] Reich, D.E., Cargill, M., Bolk, S. *et al.* (2001). Linkage desiquilibrium in the human genome. *Nature* **411**, 199–204.

[211] Reiner, A., Yekutieli, D. and Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* **19**, 368–375.

[212] Renaud, O. and Victoria-Feser, M.P. (2010). A robust coefficient of determination for regression. *Journal of Statistical Planning and Inference* **140(7)**, 1852–1862 .

[213] Rice, W.R. (1989). Analyzing tables of statistical tests. *Evolution*, **43**, 223–225.

[214] Rice, K. and Spiegelhalter, D. (2006). A simple diagnostic plot connecting robust estimation, outlier detection, and false discovery rates. *Journal of Applied Statistics* **33(10)**, 1131–1147.

[215] Ritland, K. (2000). Marker-inferred relatedness as a tool for detecting heritability in nature. *Molecular Ecology* **9**, 1195–1204.

[216] Robinson, T.R. (2010). *Genetics for dummies*. Wiley Publishing Inc., 2nd Edition.

[217] Rohatgi, V.K. and Saleh, A.K. (2001). *An Introduction to Probability and Statistics*. Wiley Series in Probability and Statistics, 2nd edition.

[218] Rom, D.M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika* **77**, 663-665.

[219] Ronchetti, E. (1982). Robust alternatives to the F-test for the linear model (STMA V24 1026). *In Probability and Statistical Inference*. W. Grossmann, G. C. Pflug, and W. Wertz (eds), 329–342. Dordrecht: D. Reidel Publishing Co.

[220] Ronchetti, E. (1987) Robust C(α)-type tests for linear models. *The Indian Journal of Statistics A* **49**, 1–16.

[221] Ronchetti, E. (1997). Robustness aspects of model choice. *Statistica Sinica* **7**, 327–338.

[222] Ronchetti, E. (2006). The historical development of robust statistics. *7th International Conference on Teaching Statistics (ISI)*. Salvador, Bahia (Brazil).

[223] Rosa, G.J.M., Padovani, C.R. and Gianola, D. (2003). Robust linear mixed models with normal/independent distributions and Bayesian MCMC implementation. *Biometrical Journal* **45(5)**, 573–590.

[224] Rousseeuw, P.J. (1984). Least median of squares regression. *Journal of the American Statistical Association* **79**, 871-880.

[225] Rousseeuw, P.J. and Yohai, V.J. (1984). Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series Analysis*, Franke J, Hardle W, Martin RD (eds). Lecture Notes in Statistics, vol. 26. Springer: New York, 256–272.

[226] Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust regression and outlier detection.* New York: Wiley.

[227] Rousseeuw, P.J. and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association* **88(424)**, 1273–1283.

[228] Ruger, B. (1978). Das maximale Signifikanzniveau des Tests: Lehne ab, wenn k unter n gegebenen Tests zur Ablehnung fuhren. *Metrika* **25**, 171-178.

[229] Russel, P.J. (1992). *Genetics.* HarperCollins, New York.

[230] Rutemiller, H.C. and Bowers, D.A. (1968). Estimation in a heteroscedastic regression model. *Journal of the American Statistical Association* **63**, 552–557.

[231] Ryan, T.P. (2008). *Modern regression methods.* John Wiley and Sons Inc.

[232] Salibian-Barrera, M. and Zamar, R.H. (2002). Bootstrapping robust estimates of regression. *The Annals of Statistics* **30**, 556–582.

[233] Salibian-Barrera, M. and Zamar, R.H. (2004). Uniform asymptotics for robust location estimates when the scale is unknown. *The Annals of Statistics* **32(4)**, 1434–1447.

[234] Salibian-Barrera, M. (2005). Estimating the p-values of robust tests for the linear model. *Journal of Statistical Planning and Inference* **128(1)**, 241–257.

[235] Salyakina, D., Seaman, S.R., Browning, B.L. *et al.* (2005). Evaluation of Nyholt's procedure for multiple testing correction. *Human Heredity* **60**, 19-25.

[236] Sarkar, S.K. and Chang, C.K. (1997). The Simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association* **92**, 1601–1608.

[237] Sarkar, S.K. (1998). Some probability inequalities for ordered MTP_2 random variables: a proof of the Simes conjecture. *Annals of Statistics* **26**, 494–504.

[238] Sarkar, S.K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Annals of Statistics* **30**, 239257.

[239] Sarkar, S.K. (2004). FDR-controlling stepwise procedures and their false negatives rates. *Journal of Statistical Planning and Inference* **125(1-2)**, 119–137.

[240] Sarkar, S.K. (2006). False discovery and false non discovery rates in single-step multiple testing procedures. *The Annals of Statistics* **34(1)**, 394–415.

[241] Sarkar, S.K. and Guo, W. (2009). On a generalized false discovery rate. *Annals of Statistics* **37**, 337–363.

[242] Schaalje, B.G., McBride J.B. and Fellingham G.W. (2001). Approximation to distribution of test statistics in complex mixed linear models using SAS Proc Mixed. *Proceedings of SUGI conference*. Paper 262-26.

[243] Schweder, T. and Spjotvoll, E. (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika* **69**, 493–502.

[244] Schrader, R.M. and Hettmansperger, T.P. (1980) Robust analysis of variance based upon a likelihood ratio criterion. *Biometrika*, **67**, 93–101.

[245] Schwartzbaum, J.A., Xiao, Y., Liu, Y. *et al.* (2010). Inherited variation in immune genes and pathways and glioblastoma risk. *Carcinogenesis* **31(10)**, 1770–1777. doi:10.1093/carcin/bgq152.

[246] Searl, S.R. (1979). Notes on variance components estimation. A detailed account of maximum likelihood and kindred methodology. *Technical Report* BU-673-M. Biometrics Unit, Cornell University, Ithaca, New York.

[247] Searle, S.R., Casella, G. and McCulloch, C.E. (1992). *Variance components*. John Wiley and Sons, Inc.

[248] Sebastiani, P., Gussoni, E., Kohane, I.S. and Ramoni, M.F. (2003). Statistical challenges in functional genomics. *Statistical Science* **18**, 33-70.

[249] Sebastiani, P., Timofeev, N., Dworkis, D.A. *et al.* (2009). Genome-wide association studies and the genetic dissection of complex traits. *American Journal of Hematology*, 504–515.

[250] Seeger, P. (1968). A note on a method for the analysis of significances en mass. *Technometrics* **10**, 586–593.

[251] Shaw, R. and Olds, T.M. (1993). ANOVA for unbalanced data: an overview. *Ecology* **74(6)**, 1638-1645.

[252] Shaffer, J.P. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association* **81**, 826–831.

[253] Sidák, Z. (1968). On multivariate normal probabilities of rectangles: their dependence on correlations. *Annals of Mathematical Statistics* **39**, 1425–1434.

[254] Sidák, Z. (1971). On probabilities of rectangles in multivariate normal Student distributions: their dependence on correlations. *Annals of Mathematical Statistics* **41**, 169–175

[255] Simes, J.R. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 75–754.

[256] Smith, J.A., Turner, S.T., Sun, Y.V. *et al.* (2009). Complexity in the genetic architecture of leukoaraiosis in hypertensive sibships from the GENOA Study. *BMC Medical Genomics* **2(16)**, 1-12.

[257] Smyth, G.K., Huele, A.F. and Verbyla, A.P. (2001). Exact and approximate REML for heteroscedastic regression. *Statistical Modelling* **1**, 161–175.

[258] Smyth, G.K. (2002). An efficient algorithm for REML in heteroscedastic regression. *Journal of Graphical and Computational Statistics* **11**, 836–847.

[259] Snjiders, T.A.B. and Bosker, R.J. (1994). Modeled variance in two-level models. *Sociological Methods and Research* **22**, 342–363.

[260] Soriç, B. (1989). Statistical discoveries and effect size estimation. *Journal of the American Statistical Association* **84**, 608–610.

[261] Souverein, O.W., Zwinderman, A.H. and Tanck, M.W.T. (2006). Multiple imputation of missing genotype data for unrelated individuals. *Annals of Human Genetics* **70**, 372-381.

[262] Spike, J., Piepho, H.P. and Meyer, U. (2004). Approximating the degrees of freedom for contrasts of genotypes laid out as subplots in an alpha-design in a split-plot experiment. *Plant Breeding* **123**, 193–197.

[263] Spike, J., Piepho, H.P. and Hu, X. (2005). Analysis of unbalanced data by mixed linear models using the Mixed procedure of the SAS system. *Journal of Agronomy and Crop Science* **191**, 47–54. Blackwell Verlag, Berlin.

[264] Stefansson G., Kim, W. and Hsu, J.C. (1988). On confidence sets in multiple comparisons. *In Statistical Decision Theory and Related Topics IV*, Vol.2, Ed. Gupta, S.S. and Berger, J.O., 89–104. New York Springer Verlag.

[265] Stephens, M. and Balding, D.J. (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews: Genetics* **10(10)**, 681–690.

[266] Storey, J.D. and Tibshirani, R. (2001). *Estimating false discovery rates under dependence, with applications to DNA microarrays*. Department of Statistics, Stanford University.

[267] Storey, J.D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society B* **64**, 479-498.

[268] Storey, J.D. and Tibshirani, R. (2003). Statistical significance for genome-wide studies. *In Proceedings of the National Academy of Sciences* **100**, 9440-9445.

[269] Storey, J.D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *Annals of Statistics* **31**, 2013-2035.

[270] Storey, J.D, Taylor, J.E, Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society B* **66**, 187-205.

[271] Stram, D.O. (2004). Tag SNP selection for association studies. *Genetic Epidemiology* **27**, 365–374.

[272] Sun, G. (2006). R^2 statistics with applications to association mapping. *A report.* B.S., China University of Geosciences, Beijing.

[273] Tabangin, M.E., Woo, J.G., Liu C. *et al.* (2007). Comparison of false-discovery rate for genome-wide and fine mapping regions. *BMC Proceedings* **I**, (Suppl I): SI48.

[274] Takeuchi, F., Yanai, K., Morii, T. *et al.* . (2005) Linkage disequilibrium grouping of single nucleotide polymorphisms (SNP) reflecting haplotype phylogeny for efficient selection of tag SNP. *Genetics*, **170**, 291–304.

[275] Tan, W.Y. (1982). Sampling distribution and robustness of t, F and variance-ratio of two samples and ANOVA models with respect to departure from normality. *Communications in Statistics - Theory and Methods* **11**, 2485–2511.

[276] Tan, S., Scherag, A., Janssen, O.E. *et al.* (2010). Large effects on body mass index and insulin resistance of fat mass and obesity associated gene (FTO) variants in patients with polycystic ovary syndrome (PCOS). *BMC Medical Genetics* **11(12)**.

[277] Tang, S., Tsui, K.W. (2007). Distributional properties for the generalized p-value for the Behren-Fisher problem. *Statistics and Probability Letters* **77**, 1–8.

[278] Tao, W.J. and Boulding, E.G. (2003). Association between single nucleotide polymorphims in candidate gene and growth rate in the Artic Charr (*Salvelinus alpinus L.*). *Heredity* **91**, 60-69.

[279] Taplin, R.H. (1999). Robust F-tests for linear models. *The Canadian Journal of Statistics* **27(2)**, 361–371.

[280] Tenaillon, M.I., Sawkins, M.C., Long, A.D. *et al.* (2001). Patterns of DNA sequence polymorphisms along chromosome 1 of maize (*Zea mays ssp. mays L.*). *Proceedings of the National Academy of Sciences USA*, **98**, 9161–9166.

[281] Thomas, S.C. (2005). The estimation of genetic relationships using molecular markers and their efficiency in estimating heritability in natural populations. *Philosophical Transactions of the Royal Society B* **360**, 1457–1467.

[282] Troendle, J.F. (2000). Stepwise normal theory multiple test procedures controlling the false discovery rate. *Journal of Statistical Planning and Inference* **84**, 139–158.

[283] Tsui, K.W., Weerahandi, S. (1989). Generalized p-values in significance testing of hypothesis in the presence of nuisance parameters. *Journal of the American Statistical Associaton* **84(406)**, 602–607.

[284] Tukey, J.W. (1960) A survey of sampling from contaminated distributions. *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow and H. B. Mann, eds.) 448–485. Stanford Univ. Press.

[285] Tusher, V.G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences USA* **98**, 5116–5121.

[286] Tzeng, J.Y., Byerley, W., Devlin, B. *et al.* (2003). Outlier detection and false discovery rates for whole-genome DNA matching. *Journal of the American Statistical Association* **98(461)**, 236–246.

[287] Van Valen, L.M. (1987). Mendel was no fraud. *Nature*, **325**, 395. doi:10.1038/325395d0.

[288] Vissecher, P.M., Medland, S.E., Ferreira, M.A.R. *et al.* (2006). Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genetics* **2(3)**, e41, 316–325.

[289] Vonesh, E.F. and Chinchilli, V.M. (1997). *Linear and nonlinear models for the analysis of repeated measurements*. Marcel Dekker: New York.

[290] Waqar-Ul-Haq, Malik, M.F., Rashid, M. *et al.* (2008). Evaluation and estimation of heritability and genetic advancement for yield related attributes in wheat lines. *Pakistan Journal of Botany* **40(4)**, 1699–1702.

[291] Waite, T.A. and Campbell, L.G. (2006). Controlling the false discovery rate and increasing statistical power in ecological studies. *Ecoscience* **13(4)**, 439–442.

[292] Walsh, B. (2004). Bonferroni corrections and false discovery rates. *Lecture Notes for EEB 581*.

[293] Wang, T. and Elston, R. (2006) Improved power by use of a weighted score test for linkage disequilibrium mapping. *American Journal of Human Genetics*, **80** 353–360.

[294] Wang, K. and Abbot, D. (2008) A principal components regression approach to multilocus genetic association studies. *Genetic Epidemiology* **32**, 108–118.

[295] Wang, K. and Huang, J. (2002). A score-statistic approach for the mapping of quantitative-trait loci with sibships of arbitrary size. *American Journal of Human Genetics* **70**, 412–424.

[296] Weber, A.L., Clark, R.M., Laura, V. *et al.* (2007). Major regulatory genes in maize contribute to standing variation in Teosinte (*Zea mays* ssp. *parviglumis*). *Genetics* **177**, 2349–2359.

[297] Weber, A.L., Briggs, W.H., Rucker, J. *et al.* (2008). The genetic architecture of complex traits in Teosinte (*Zea mays* ssp. *parviglumis*): new evidence from association mapping. *Genetics* **180**, 1221–1232.

[298] Weiling, F. (1986). What about R.A. Fisher's statement of the "too good" data of J.G. Mendel's *Pisum* paper? *Journal of Heredity* **77**, 281–283.

[299] Weiling, F. (1989). Which points are incorrect in R.A. Fisher's statistical conclusion: Mendel's experimental data agree too closely with his expectations? *Angewandte Botanik* **63**, 129–143.

[300] Weir, B.S. (1996). *Genetic data analysis II*. Sinauer Associates, Sunderland, Massachusetts.

[301] Westfall, P.H. and Young, S.S. (1993). *Resampling-based multiple testing: examples and methods for p-value adjustment*. Wiley, New York.

[302] Westfall, P.H., Zaykin, D.V. and Young, S.S. (2001). Multiple tests for genetic effects in association studies. *In Biostatistical Methods (S. Looney, ed.)*, 143–168. Humana, Totowa, NJ.

[303] White, S. and Doebley, J. (1998). Of genes and genomes and the origin of maize. *Trends in Genetics* **14(8)**, 327–332.

[304] Wigginton, J.E., Cutler, D. and Abecasis, G. (2005). A note on exact tests of Hardy-Weinberg equilibrium. *American Journal of Human Genetics* **76**, 887–893.

[305] Wilcox, R.R. (2010). *Fundamentals of modern statistical methods*. Springer, New York. 2nd Edition.

[306] Wittke-Thompson, J.K., Pluzhnikov, A., and Cox, N.J. (2005). Rational inferences about departures from Hardy-Weinberg equilibrium. *American Journal of Human Genetics* **76**, 967-986.

[307] Wright, S. (1931). Evolution in Mendelian populations. *Genetics* **16**, 97–159.

[308] Wright, S.P. (1992). Adjusted p-values for simultaneous inference. *Biometrics* **48**, 1005–1012.

[309] Wright, S.I., Bi, I.V., Schroeder, S.G. *et al.* (2005). The effects of artificial selection on the maize genome. *Science* **398**, 1310–1314.

[310] Wu, R., Ma, C.X. and Casella, G. (2007). *Statistical genetics of quantitative traits: linkage, maps and QTL*. New York: Springer.

[311] Xiang, Z., Zou, F. and Wang, W. (2008). Fastanova: an efficient algorithm for genome-wide association study. *KDD*, 821–819.

[312] Xiang, Z., Zou, F. and Wang, W. (2009). Efficient algorithms for genome-wide association study. *ACM Transactions on Knowledge Discovery from Data* **3(4)**, 1–28. doi: 10.1145/1631162.1631167.

[313] Xu, J., Turner, A., Little, J., Bleecker, E.R. and Meyers, D.A. (2002) Positive results in association studies are associated with departure from Hardy-Weinberg equilibrium: hint for genotyping error? *Human Genetics* **111**, 573-574.

[314] Xu, R.H. (2003). Measuring explained variation in linear mixed effects models. *Satistics in Medicine* **22**, 3527–3541.

[315] Xu, Z., Kaplan, N.L. and Taylor, J.A. (2007). TAGster: efficient selection of LD tag SNPs in single or multiple populations. *Bioinformatics Applications Note* **23**, 3254–3255.

[316] Xu, C., Liu, Y., Wang, P. *et al.* (2010). Integrative analysis of DNA copy number and gene expression in metastatic oral squamous cell carcinoma identifies genes associated with poor survival. *Molecular Cancer* **9(143)**, 1–12.

[317] Yarmohammadi, M. and Mahmoudvand, R. (2010). The effect of outliers on robust and resistant coefficient of determination. *International Journal of Academic Research* **2(3)**, 133–138.

[318] Yekutieli, D. and Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical and Planning Inference* **82**, 171-196.

[319] Yu, J., Pressoir, G., Briggs, W.H. *et al.* (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* **38**, 203–208.

[320] Yuan, M. and Wahba, G. (2004). Doubly penalized likelihood estimator in heteroscedastic regression. *Statistics and Probability Letters* **69(1)**, 11–20.

[321] Zhang, J. and Li, G. (1998). The breakdown properties of location M-estimators. *The Annals of Statistics* **26(3)**, 1170–1189.

[322] Zhang, K., Deng, M., Chen, T. *et al.* (2002). A dynamic programming algorithm for haplo-type block partitioning. *Proceedings of the National Academy of Sciences USA* **99**, 7335–7339.

[323] Zhang, K., Zhu, X. and Zhao, H. (2003). On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. *Genetic Epidemiology* **24**, 44–56.

[324] Zhang, F., Wang, Y. and Deng, H.W. (2008). Comparison of population-based association study methods correcting for population stratification. *PLoS ONE* **3** e3392. doi: 10.1371/journal.pone.0003392

[325] Zhao, K., Aranzana, M.J., Kim, S. *et al.* (2007). An Arabidopsis example of association mapping in structured samples. *PLoS Genetetics* **3**, e4. doi:10.1371/journal.pgen.0030004

[326] Zhao, W., Canaran, P., Jurkuta, R. *et al.* (2006). Panzea: a database and resource for molecular and functional diversity in the maize genome. *Nucleic Acids Research* **34**, D752–D757.

[327] Zengh, B.Y. (2000). Summarizing the goodness of fit of generalized linear models for longitudinal data (2000). *Statistics in Medicine* **19**, 1265–1275.

[328] Zheng, G., Freidlin, B. and Gastwirth, J.L. (2006). Robust genomic control for association studies. *American Journal of Human Genetics* **78**, 350–356.

[329] Zimmerman, D.W. (1994). A note on the influence of outliers on parametric and non-parametric tests. *Journal of General Psychology* **121(4)**, 391–401.

[330] Zou, F., Yandell, B.S. and Fine, J.P. (2003). Rank-based statistical methodologies for quantitative trait locus mapping. *Genetics* **165**, 1599–1605.