Aus dem Institut für Medizinische Biometrie und Informatik
(Geschäftsführender Direktor: Prof. Dr. Meinhard Kieser)

Abteilung Medizinische Biometrie

Robust Generalised Linear Regression Models in Genetic Studies: Assessment of Standard Techniques and Their Generalisation to Incorporate Hampel's Function

Inauguraldissertation
zur Erlangung des Doctor scientiarum humanarum (Dr.sc.hum.)
an der
Medizinischen Fakultät Heidelberg
der
Ruprecht-Karls-Universität

vorgelegt von
Miriam Kesselmeier
aus
Köln
2016

Dekan: Prof. Dr. Wolfgang Herzog

Doktorvater: Prof. Dr. Justo Lorenzo Bermejo

On ne fait jamais attention à ce qui a été fait;
on ne voit que ce qui reste à faire.
(Marie Curie)

Contents

Mathematical notations

Notation	**Description**
Matrix operations	
X^{-1}	Inverse of matrix X
X^T	Transpose of matrix X
Numbers	
$\mathbb{R}_{>0}$	Positive real numbers
$\mathbb{R}^n$	n-dimensional space of real numbers (vector space)
$\mathbb{R}^{n\times m}$	$(n \times m)$-dimensional space of real numbers (matrix space)
Operators	
$\mathbb{1}_A$	Indicator function that is 1 if case A occurs and 0 otherwise
$x \in [a, b]$	x is an element of the interval $[a, b]$
$\|a\|$	Absolute value of a
$\lfloor a \rfloor$	Rounding down a to integer
a^{-b}	$\left(\frac{1}{a}\right)^b$
$f^{'}(x)$	1st partial derivative of the function f with respect to x, i.e. $\frac{\partial f(x)}{\partial x}$
$j := a$	j is set to a
$\exp(a)$	Natural exponential function of the real number a
$\max(x)$	Maximum of x
$\log(a)$	Natural logarithm of the positive real number a
$\log_b(a)$	Logarithm of the positive real number a to base b
$\mathrm{sign}(r)$	Sign of r
$\sin(r)$	Sine function with respect to r
$\sum_{i=1}^{n} f(i)$	Sum of all $f(i)$ with respect to all natural numbers i from 1 to n

(continued on next page)

(continued)

Notation	**Description**
Regression analysis	
$y \sim x_1 + x_2$	y is described by x_1 and x_2 (regression model)
$\hat{\beta}$	Estimate of β
Theory of probability	
$E[Y]$	Expectation of random variable Y
$P[Y = y]$	Probability of random variable Y having realisation y
$P[Y = y \| X]$	Conditional probability of random variable Y having realisation y given event X
$V[Y]$	Variance of random variable Y
$Y \sim \text{Bin}(m, p)$	Random variable Y is binomial distributed with m trials and success probability p
$Y \sim \mathcal{N}(\mu, \sigma^2)$	Random variable Y is normally distributed with expectation μ and standard deviation σ
$Y \sim \text{Poi}(\lambda)$	Random variable Y is Poisson distributed with parameter λ
Other	
$\exists$	It exists

Abbreviations

Abbreviation	Description
aCGH	Array-based comparative genomic hybridisation
AUC	Area under the receiver operating characteristic curve
BMI	Body mass index
cf.	Compare (confer)
CI	Confidence interval
CO_2	Carbon dioxide
CpG	Cytosine-phosphate-Guanine (dinucleotide)
DCM	Dilated cardiomyopathy
DNA	Deoxyribonucleic acid
e.g.	For example (exempli gratia)
FPR	False positive rate
GAW 18	18th Genetic Analysis Workshop
GRR	Genotype relative risk
GWAS	Genome-wide association study
HBV	Hepatitis B virus
HCC	Human hepatocellular carcinoma
HCV	Hepatitis C virus
HTN	Hypertension
IDI	Integrated discrimination improvement
i.e.	That is (id est)
1 kb	1 kilobase pair
LD	Linkage disequilibrium
LR	Logistic regression
MAD	Median absolute deviation
MAF	Minor allele frequency
MSE	Mean squared error

(continued on next page)

(continued)

Abbreviation	**Description**
NAFLD	Non-alcoholic fatty liver disease
OR	Odds ratio
Q1	1st quartile
Q3	3rd quartile
ROC	Receiver operating characteristic
RT-qPCR	Real-time quantitative polymerase chain reaction
SNP	Single nucleotide polymorphism

1 Introduction

Genetics comprise the area of research dealing with the principles of heritable traits and their inheritance. A major engagement of human genetics research is aimed at metabolic regulation of organisms and the development of heritable diseases. Hence, the transfer of such genetic findings into medical research and medical application is of prime importance. To gain and handle this knowledge, genetic studies must be analysed by the application of statistical methods. Due to the diversity of genetic data, adequate statistical methods must be found.

Section 1.1 introduces the design and standard approaches for the analysis of genetic studies. As data never are homogeneous, the term "outliers" is explained in section 1.2. The following sections introduce into robust statistics and familiarises with terms needed to discuss the robust generalised linear model framework. Section 1.3 gives a short overview about the essential terms as well as section 1.4 about the standard approach in generalised linear models, and the robust generalised linear model framework considered in this work will be explained in section 1.5. The chapter ends with the presentation of the objectives and the structure of this thesis (section 1.6) as well as of the work performed by myself (section 1.7).

1.1 Design and standard analysis of genetic studies

Statistical genetics share only a small part in the wide field of biostatistics. Their task is to analyse genetic data and, hence, to improve and to develop useful statistical techniques to answer specific questions dealing with relations between genetics and diseases. Genetic data often exhibit a high dimensionality, which is defined by a much higher number of variables than observations. In the framework of statistical genetics, data under investigation often describe genotypes, gene expression or DNA methylation. These genetic features will be introduced later in greater detail. To analyse these kinds of data, one uses for example linear or generalised linear models (see for example Kesselmeier et al. (accepted) for the analysis of DNA methylation data). These kinds of models are not only applied in statistical genetics but also for other purposes such as non-genetical clinical research to quantify the Hawthorne effect in hand hygiene compliance (Hagel et al., 2015) or biogeosciences to estimate effects of elevated carbon dioxide in the atmosphere (global change) on the exchange of climatic relevant trace gases (Sandoval-Soto et al., 2012). Among generalised linear models, one often applies logistic and Poisson regression. With logistic regression, one investigates for example the association of a binary disease phenotype or a pathway affiliation and some clinical or genetic measure (e.g. Kesselmeier et al. (in preparation); Ali et al. (in preparation)). If the response variable has countable values, one can use Poisson regression for analysis.

1.2 Outliers, outliers in genetic data and robust statistics

In high-dimensional data, few data points usually occur being distinct from the majority of the data. In other words,

> *"Many times values occur which are 'dubious' in the eyes of the analyst."*

as stated by Dixon (1950) on page 488. Citing the "intuitive" definition of Hawkins (1980) on page 1, an outlier is

> *"an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism".*

Their handling is a great challenge because they can significantly affect the analysis – e.g. significance levels, confidence intervals or hypothesis tests can be concerned (Sarkar et al., 2011; Wilcox, 1998; Muhlbauer et al., 2009). Standard methods used to estimate the parameters of regression models (e.g. iteratively re-weighted least squares) are of limited value due to their dependence on few outliers. This understanding contrasts the purpose of genetic risk models predicting a particular health outcome for the bulk of individuals and identifying persons with a deviating high-risk of disease. Therefore, the handling of outliers is a major scope of "robust statistics". Hampel et al. (1986) suggests on pages 6-7 two definitions for robustness; the more informal one is

> *"In a broad informal sense, robust statistic is a body of knowledge, partly formalized into 'theories of robustness', relating to deviations from idealized assumptions in statistics."*

and the more formal one is given by

> *"Robust statistics, as a collection of related theories, is the statistics of approximative parametric models."*

In particular, robust statistics aim at integrating outliers into the analyses but with consideration of their difference. Hence, outliers are down-weighted in robust statistics and one does not need to delete them a priori. In general, there are two different kinds of robustness (Heritier and Ronchetti, 1994). In statistical questions, it is desirable to control the amount of true but rejected null hypotheses (false negative rate). Outlier may hamper this goal. Hence, one aim is a non-increasing type I error rate in the presence of a small amount of outliers. This stability is called *robustness of validity.* The second aspect is *robustness of efficiency* concerning the amount of not rejected but false null hypotheses, i.e. the false positive rate. The related true negative rate is called power of a test and a non-decreasing power in the presence of outliers is desirable.

In the case of expected outliers, it is possible to robustify linear and generalised linear models in different ways. In this work, the method proposed by Cantoni and Ronchetti (2001) for robust logistic and Poisson regression will be investigated and extended. Consequently, this extension will be analysed and compared to its basis as well as standard methods. Note that the combination of logistic and Poisson regression models results in a hurdle model. This type of model at first estimates the probability of an event to occur. In case of a positive answer, the model estimates the positive value using a truncated Poisson model, i.e. a Poisson model restricted to positive counts.

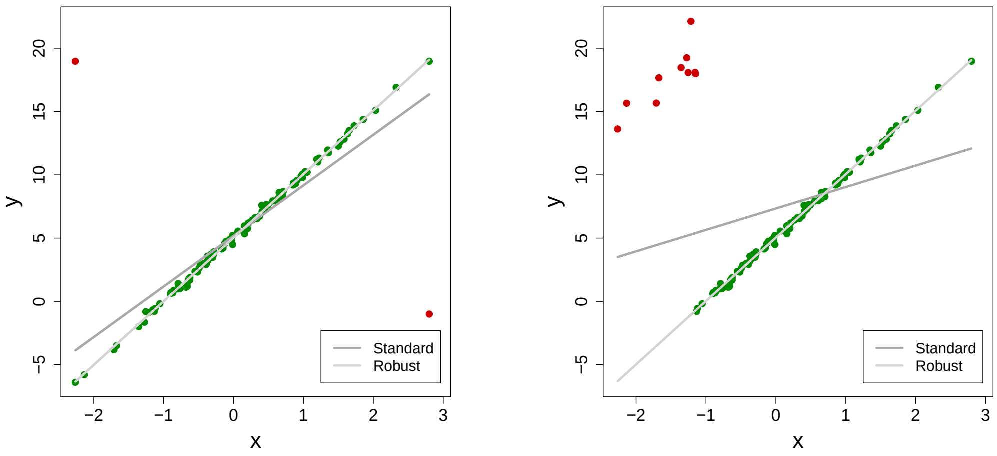
1.3 Illustration of and essential terms in robust statistics

It is essential to understand the term "robust". In parametric statistics, assumptions are used to specify the desired model such as normality, independence or linearity. If these assumptions are valid, parametric models describe the data well. But reality can only be approximated using such assumptions for several reasons. Measured data can be erroneous. In close accordance to this practical problem, the underlying theoretical model might not fit as well. Classical parametric methods rely on the central limit theorem but the normal approximation for large data sets is still an approximation (Hampel, 1968; Hampel et al., 1986).

Figure 1 gives an illustrative example to visualise the impact of the violation of these assumptions on standard estimates and their handling by robust regression. There, the influence of few and several outliers on standard and robust linear regression is demonstrated. These two simulated situations possibly occur in case of observations arbitrarily differing from the majority of the data. Figures 1(a) and (b) differ in the amount of mismatching observations, i.e. two (radial) outliers versus a group of ten (clustered) outliers. It is obvious that the robust approach is not noticeably affected by the distinct observations in both scenarios whereas the standard method is highly sensitive to this minority of data – even in the presence of only two outliers. This is in close accordance with Rousseeuw (1984). In case of the clustered outliers one must ask whether these two groups of data really represent one event or if there is a confounder so that it might be necessary to perform an adjusted analysis. In a group-adjusted analysis, the standard and robust estimates given as triples (intercept, slope, group) are $(5.034, 5.008, 20.234)$ and $(5.038, 5.009, 19.026)$, respectively, compared to the unadjusted estimates as tuples (intercept, slope) $(7.339, 1.694)$ and $(5.037, 5.012)$. There are two main observations:

- The standard and robust estimates are comparable in the adjusted analysis.
- The robust intercept and slope estimates of the adjusted and the unadjusted analysis are comparable but this is not the case for the standard estimates.

Summarising this small simulation, the robust estimates reliably approximate the simulated data even in the unadjusted analysis. This demonstrates the strong



(a) Two radial outliers, one in each outer region, and 100 inliers.

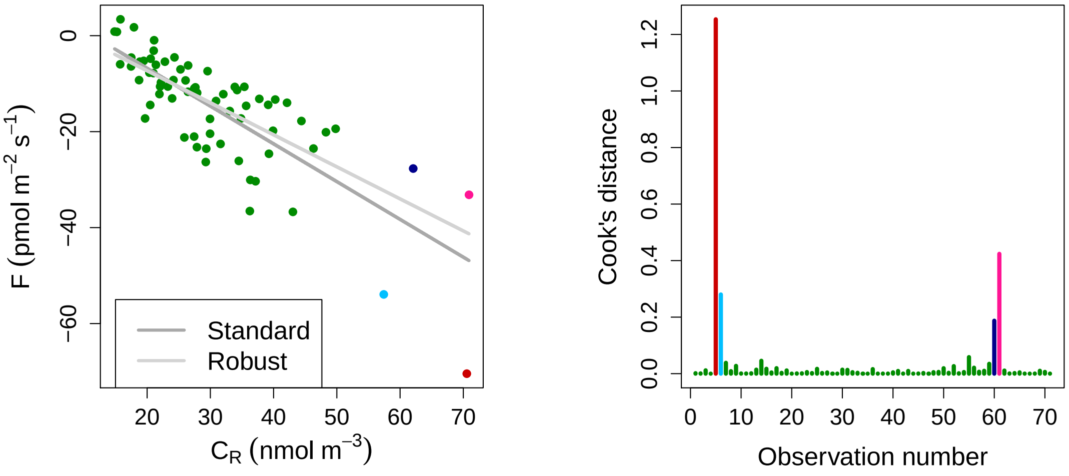
(b) Group of 10 clustered outliers in the left outer region in relation to 90 inliers.

Figure 1: Influence of outliers on standard and robust linear regression. The x value is normally distributed as $\mathcal{N}(0,1)$. The y values of the inliers (y_{in}) are $4+5x+\mathcal{N}(1,0.2^2)$. The y values for the two radial outliers are elements of $\{-1,\max(y_{in})\}$ and the ten clustered outliers are equal to $y_{in}+\mathcal{N}(20,1)$. Hence, the expected intercept is 5, the expected slope is 5 and the expected group difference is 20.

influence of outliers on standard linear regression estimates and the handling of the same by robust linear regression.

Using standard estimates, the outlier might influence the regression model in a way that the residual value of the outlier is reduced whereas the residuals of the non-outliers are enlarged. This can lead to outlier hiding (*masking effect*) or spurious identification of an observation as outlier (*swamping effect*). These results might cause problems in the naive approach to control the outlier influence by identification and removal based on classical techniques (Cantoni and Ronchetti, 2006; Jajo, 2005). This removing might additionally lead to sample selection bias (Heckman, 1979). Consequently, robust approaches aim at identifying the model best describing the structure of the majority of the data under consideration of outliers and mismatching substructures.

As demonstrated in figure 1, identification of outliers might be easier with a robust approach when relying on residuals because in this case the difference between estimated and observed values should be larger for the extreme observations as compared to standard techniques. An alternative to the use of residuals for outlier identification is to use the Cook's distances based on standard regression techniques. The Cook's distance of an observation is the difference between the estimated response variable with and without this observation. To give an example for Cook's distances, one can inspect parts of the real data set analysed in Sandoval-Soto et al. (2012) with linear regression. Figure 2(a) shows this real data of trace gas exchange between the atmosphere and European or common beech (*Fagus sylvatica*) grown under elevated CO_2 (carbon dioxide). Based on this distribution, assuming a linear relation and, hence, applying a linear model are reasonable. The Cook's distances for each observation of this data set are given in figure 2(b) indicating that some observations have a large impact on the standard regression result. Thus, it is not surprising that the robust estimates differ from the standard estimates – now accounting for the influential observations (figure 2(a)).



(a) Observations in relation to the standard and robust linear regression lines. F describes a trace gas exchange related to leaf area and time. C_R denotes atmospheric CO_2 concentration.

(b) Cook's distances for the standard linear regression model.

Figure 2: Existence proven by Cook's distance and impact of outliers in real data. Colour indicates the Cook's distance size of the observation, i.e. the Cook's distance is smaller than 0.1, in the interval $[0.1, 0.2]$, in the interval $[0.2, 0.4]$, in the interval $[0.4, 1.0]$ or larger than 1.

Applying regression methods, the identification of data points with high impact is desirable. If such data points exhibit departing values in the independent variable, they are called *high leverage points*. For example, all four observations in figure 2(a) with large impact on the standard estimate are points with high leverage. The need of identifying deviations from assumed structures leads to the definition of a *breakdown point* which indicates the percentage of the maximal admitted amount of outliers to still get reliable results. The maximal achievable breakdown point is 0.5 for location estimators treating observations at both estimator sides symmetrically; the median for example has such a maximal breakdown point whereas the mean has a breakdown point of 0 (Hampel, 1968; Hoaglin et al., 1983). An estimator with a breakdown point of 0.5 is called *globally robust*. Robust estimators with a contamination sensitivity larger than 0 are called *locally robust* (Ferretti et al., 1999). The *influence function* represents the standardized effect of a specific amount of outliers. It is possible to investigate the influence of a given amount of outliers with this function. The influence function is an indicator for the stability of an estimator and one can use it to answer questions about *gross error sensitivity* (i.e. supremum of the influence function over all possible values for its argument), *local shift sensitivity* (i.e. supremum of the difference in the influence functions between two possible data values divided by the distance of these two values) and the *rejection point* (i.e. minimal value defining the influence function to be zero for all values larger than this value) (Hampel et al., 1986). It should be clear, however, that the use of robust statistics does not allow to use a model that does not fit the data at all. The underlying model has always to be considered as reasonable and useful.

1.4 Linear and generalised linear models

Linear models are used to describe a linear relation between a q-dimensional explanatory variable $X = (1, x_1, \ldots, x_{q-1}) \in \mathbb{R}^{n \times q}$ and a continuous/discrete independently normally distributed response variable $Y \in \mathbb{R}^{n \times 1}$ with mean $\mu \in \mathbb{R}^{n \times 1}$ and constant variance σ^2, i.e. it is assumed that

$$Y = X\beta \tag{1.1}$$

with an unknown parameter vector $\beta \in \mathbb{R}^{q \times 1}$. Then, β can be determined by

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y \tag{1.2}$$

If the rank of $X^T X$ is not full, the inverse has to be replaced by a generalised inverse. Although the solution is no more unique, the variance and covariance are correct. To estimate the coefficient vector β, the inverse matrix of X^TX has to be calculated. Several approaches are possible such as Gaussian elimination, Cholesky decomposition or direct decomposition (McCullagh and Nelder, 1996).

Generalised linear models use a function of the response's mean instead of the mean itself (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1996). Hence, these models can be used on various data sources. They are characterized by

$$\eta = g(\mu) = X\beta \tag{1.3}$$

with the link function $g(\cdot)$. Due to the central limit theorem in large samples, the violation of the normal assumption might only lead to a modest reduced efficiency (McCullagh and Nelder, 1996).

Obviously, using the identical link, generalised linear models reduce themselves to classical linear models. For binomial and Poisson distributed random variables, the link function is defined as

$$g(\mu) = \begin{cases} \text{logit}(\mu) = \log\left(\frac{\mu}{1-\mu}\right) & \text{if } Y \text{ is binomially distributed} \\ \log(\mu) & \text{if } Y \text{ is Poisson distributed} \end{cases} \tag{1.4}$$

The identical link in linear regression is meaningful because all real numbers are reasonable. This is not the case for logistic (binomial) or Poisson regression. Poisson regression is used for count data which are non-negative and integer valued. Based on the binomial distribution, logistic regression is applied if the response is 0-1 coded and, hence, the regression response value has to be between 0 and 1. These two conditions are fulfilled by those link functions mentioned in equation (1.4). The coefficients of logistic regression can be interpreted as log odds ratios. The larger the mean of the Poisson distribution, the more the distribution tends towards the normal distribution (Ramsey and Schafer, 2002). Under these conditions, a linear regression is an alternative to Poisson regression.

To estimate the parameter vector, the needed maximum likelihood estimation equation equals

$$\sum_{i=1}^{n} W_i(y_i - \mu_i)\frac{d\eta_i}{d\mu_i}x_{ij} = 0$$

where the weight function W is defined as

$$W^{-1} = \left(\frac{d\eta}{d\mu}\right)^2 V \tag{1.5}$$

with variance function V (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1996). As an example, the explicit log-likelihood for logistic regression is given in chapter 2.3 on page 46. Then, the parameter vector β can be iteratively determined. Initial values have to be calculated based on the data itself. Let be $\hat{\eta}_0$ the current estimate of η, the linear predictor, and $\hat{\mu}_0$ the current estimator of μ, the fitted value derived from the link function. Then, the following procedure has to be repeated till convergence, i.e. the changes between two iterations are sufficiently small (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1996). First, calculate the adjusted dependent variable via

$$z_0 = \hat{\eta}_0 + (y - \hat{\mu}_0)\left(\frac{d\eta}{d\mu}\right)_0$$

evaluated at μ_0 and the quadratic weight via

$$W_0^{-1} = \left(\frac{d\eta}{d\mu}\right)_0^2 V_0$$

with variance function V_0 evaluated at $\hat{\mu}_0$ according to equation (1.5). Then, determine new parameter estimates $\hat{\beta}_1$ by regressing z_0 on X using weight W_0, i.e.

$$\hat{\beta}_1 = \left(X^T W_0 X\right)^{-1} X^T W_0 z_0$$

(weighted least squares) and derive new estimate $\hat{\eta}_1$ from

$$\hat{\eta}_1 = \hat{\beta}_1 X$$

Note that the method used to calculate z_0 follows the first order Taylor series of the link function. To begin the iteration, $\hat{\mu}_0 = y$ can be taken and $\hat{\eta}_0$ can then be deduced from $\hat{\mu}_0$ because they are related via the link function (Nelder and Wedderburn, 1972).

1.5 Robust approaches to generalised linear regression models

These classical estimators are sensitive to outliers. These are observations with large response deviations from the response's mean or observations departing from the majority of the data in the explanatory variable (leverage points). For example, Rousseeuw (1984) stated that the breakdown point of least square estimates is 0 compared to a highest possible breakdown point of 0.5 (Muhlbauer et al., 2009). Based on the central limit theorem, this is more important in small than in larger samples. To robustify generalised linear models, several methods have been developed and discussed. There were proposals

- to correct the score function (Nakamura, 1990),
- to use a bootstrap approach (Haukka, 1995),
- to generalise globally robust estimators to become locally and globally robust (Ferretti et al., 1999),
- to adjust the estimation function's scale (Adimari and Ventura, 2001),
- to rely on influence functions (Kordzakhia et al., 2001),
- to rely on flexible nonparametric extensions of the underlying model (Bednarski, 2002),
- to adapt the likelihood (Li and Hsiao, 2004),
- to average the mean squared error of predictions over the parameter space that defines the class of the unknown true model, i.e. a neighbourhood of the true model (Adewale and Xu, 2010),
- to apply a variance stabilising transformation to the response and to use then an M-estimator (Valdora and Yohai, 2014) or
- to combine the use of a bounded exponential score function and leverage-based weights (Lv et al., 2015).

The method investigated in this thesis was suggested by Cantoni and Ronchetti (2001) accounting for outliers in the response as well as for leverage points separately. They proposed to solve the estimation equation

$$0 = \sum_{i=1}^{n} \nu(y_i, \mu_i) w(x_i) \mu_i' - \alpha(\beta) \tag{1.6}$$

with weighting functions $\nu(\cdot)$ (outlier in the response) and $w(\cdot)$ (outlier in the explanatory variable). According to equation (1.3)

$$\mu_i = \mu_i(\beta) = g^{-1}\left(x_i^T \beta\right)$$

and μ_i' denotes the derivative of μ_i with respect to β. Finally, the α function ensures Fisher consistency, i.e. the estimate equals the true value when deduced from the whole population (Fisher, 1922), and is defined as

$$\alpha(\beta) = \frac{1}{n} \sum_{i=1}^{n} E[\nu(y_i, \mu_i)] w(x_i) \mu_i' \tag{1.7}$$

where the expectation is taken with respect to the conditional distribution of $y|x$. This estimator is based on a quasi-likelihood and is asymptotically normally distributed with asymptotic variance

$$\Omega = \left(E\left[\frac{\partial}{\partial\beta}\Psi(y,\mu)\right]\right)^{-1} E\left[\Psi(y,\mu)\Psi(y,\mu)^T\right] \left(E\left[\frac{\partial}{\partial\beta}\Psi(y,\mu)\right]\right)^{-1}$$

where Ψ denotes the score function

$$\Psi(y_i, \mu_i) = \nu(y_i, \mu_i) w(x_i) \mu_i' - \alpha(\beta)$$

The influence function of this M-estimator is defined as

$$IF(y;\Psi) = -\left(E\left[\frac{\partial}{\partial\beta}\Psi(y,\mu)\right]\right)^{-1} \Psi(y,\mu)$$

The objective of the weighting functions is to confine the influence of outlying observations to get a more reliable result describing the majority of the data. A usual choice for the weighting function w is

$$w(x_i) = \sqrt{1 - h_{ii}}$$

where h_{ii} is the i^{th} diagonal element of the hat matrix $H = X(X^TX)^{-1}X^T$. Due to its definition based on the hat matrix, one must consider that this weight does not have a high breakdown point (Cantoni and Ronchetti, 2001). A different and often used choice is related to the Mahalanobis distance d_i and is defined as

$$w(x_i) = \frac{1}{\sqrt{1 + 8 \max\left(0, \frac{d_i^2 - q}{\sqrt{2q}}\right)}}$$

This weighting function has a high breakdown point but its use is only reasonable in case of continuous explanatory variables (Heritier et al., 2009). The second weighting function ν can be defined via

$$\nu(y_i, \mu_i) = \frac{\psi(r_i)}{V^{1/2}(\mu_i)}$$

with an appropriate $\psi(\cdot)$. This weighting function, in particular ψ, aims for down-weighting outlier influence and assigning usual weights to inliers. Several choices for this function are possible. Huber (1964) proposed

$$\psi_{\text{Huber}}(r; c) = \begin{cases} r & \text{if } |r| < c \\ \operatorname{sign}(r)\, c & \text{otherwise} \end{cases} \tag{1.8}$$

with $c \in \mathbb{R}_{>0}$. This function does not weight extreme outliers down to zero but to a selected size. The asymptotic efficiency is 95% for c equal to 1.345 (Cantoni and Zedini, 2011). A different class of estimates are the re-descending M-estimators which set the weights of outliers with an impact larger than a pre-specified size to zero. Considering the tuning constant triplet (a, b, c) with $a,\ b,\ c\ \in \mathbb{R}_{>0}$ and $a < b < c$, the Hampel M-estimator

$$\psi_{\text{Hampel}}(r; a, b, c) = \operatorname{sign}(r) \cdot \begin{cases} |r| & \text{if } |r| < a \\ a & \text{if } a \le |r| < b \\ \frac{c - |r|}{c - b}\, a & \text{if } b \le |r| < c \\ 0 & \text{otherwise} \end{cases} \tag{1.9}$$

belongs to this class as well as Tukey's biweight M-estimator

$$\psi_{\text{Tukey}}(r; c) = \begin{cases} \left[\left(\frac{r}{c}\right)^2 - 1\right]^2 r & \text{if } |r| \le c \\ 0 & \text{otherwise} \end{cases} \tag{1.10}$$

or Andrews' sine wave M-estimator

$$\psi_{\text{Andrews}}(r;c) = \begin{cases} \sin\left(\frac{r}{c}\right) & \text{if } |r| < c\pi \\ 0 & \text{otherwise} \end{cases} \tag{1.11}$$

with $c \in \mathbb{R}_{>0}$ (Hampel et al., 1986; Beaton and Tukey, 1974; Andrews et al., 1972). For 95% efficiency when using the Hampel function the tuning constant triplet can be chosen as $0.902 \cdot (1.5, 3.5, 8)$ for a slope of the re-descending part equal to 1/3 or as $0.691 \cdot (2, 4, 8)$ for a slope of the re-descending part equal to 1/2 (as originally proposed) (Rousseeuw et al., 2012; Koller and Mächler, 2014; Koller and Stahel, 2011). To assure 95% efficiency for the Tukey function, the tuning constant has to be equal to 4.6851 (Alamgir et al., 2013). The Andrews function is often used with $c = 1$ (Alamgir et al., 2013). The shapes of these functions are shown in figure 3 with the tuning constants mentioned above to assure 95% efficiency for the Huber, Hampel and Tukey function. The rejection point for the use of the Huber function is infinity, $c\pi$ for Andrews and c for Hampel and Tukey function. Hence, re-descending functions always have a finite rejection point.

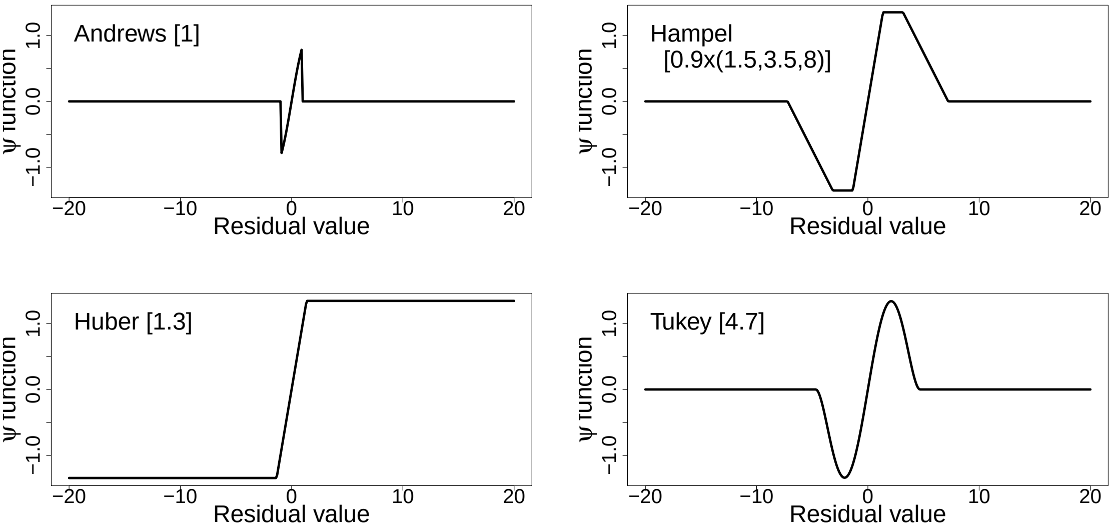


Figure 3: Shapes of different ψ functions. The tuning constants for this plot are given in brackets.

To use this robust regression framework, the R package `robustbase` is available (R Core Team, 2013; Rousseeuw et al., 2012). In this package, only the use of the Huber function was implemented at the time of method application.

1.6 Objectives and structure of this thesis

As discussed above, the scope of robust methods is to control the outlier influence by assigning each observation a weight using, e.g., the bounded Huber function. In literature, the use of re-descending weighting functions is advised which vanish outside a pre-specified region (Müller, 2004; Shevlyakov et al., 2008). Among them, the Hampel function performs well in many situations (Andrews et al., 1972; Alamgir et al., 2013). The re-descending Hampel function down-weights outliers more strongly than the bounded Huber function. Hence, the Hampel function can be useful if extreme outliers are expected. Its application has a possible impact on the type I error rate, bias, mean squared error (MSE) and statistical power for associated estimators. Under these circumstances, the aim of this thesis was threefold:

1. Comparison of standard and existing robust regression methods relying on the Huber function regarding different aims of analysis based on simulated and real data
2. Adaptation of the regression framework proposed by Cantoni and Ronchetti (2001) for logistic and Poisson models to the use of the Hampel function as an example for re-descending influence functions. As this proposed method has been already implemented in the R package `robustbase`, it takes advantage of the Huber function (R Core Team, 2013; Cantoni, 2004; Rousseeuw et al., 2012). To implement this procedure using the Hampel function into R, I derived the necessary quantities and adapted the existing programme.
3. Investigation of this extended method by comparing it to the use of the Huber function and to standard regression techniques regarding simulated and real data

These aims are reflected in the thesis structure as described as follows: Evaluation methods for the different regression approaches as well as the evaluation results and their discussion are described in the chapters 2, 4 and 5. There, the standard and the already existing robust regression approaches are compared first (sections 2.2, 4.1 and 5.1) with respect to consistency of model selection and prediction accuracy, influence of one single outlier and influence of genotyping errors on estimates (*Aim* 1). The required calculations and adjustments of the R code for the explicit use of the extended method will be presented in chapter 3 together with a plausibility check (*Aim* 2). This extended logistic regression approach is compared to the standard and the existing robust logistic regression approaches (sections 2.3, 4.2 and 5.2) with respect to several statistical properties (*Aim* 3). Chapter 5 also comprises a final conclusion and a perspective for further work (section 5.4). Special attention will be given to the hurdle model arising from the combination of logistic and (truncated) Poisson models. Chapter 6 summarizes the complete thesis in few words.

1.7 Collaborations and own work

As usual, real data applications need collaborations. These collaborations will be mentioned by article citation during the thesis when introducing the relevant data. In case of the use of real data for my methodological investigations, the collaborators only provided the data but where not involved in the research.

For the simulation described in section 2.3, the genotype relative risk (GRR) at the marker locus must be deduced from the GRR at the causal variant locus and, secondly, the randomly drawn allele frequency concerning the so-called null marker loci and the minor allele frequency at the marker have to be corrected for the prevalence in the population. The two corresponding R scripts were provided by my supervisor Prof. Dr. Justo Lorenzo Bermejo.

For the necessary calculations to extend the robust approach, I relied on the appendix A of Cantoni and Ronchetti (2001). To implement this extended approach into the statistical language R, I adapted the R scripts provided in the R package `robustbase` (Rousseeuw et al., 2012).

Main parts of this thesis are based on my methodological publications and I presented several parts on research conferences:

- Section 3.1.1 is based on the *Supplemental Note* of Kesselmeier and Lorenzo Bermejo (in preparation) [currently submitted to *Briefings in Bioinformatics*].
- Parts concerning figure 7 in section 3.3 are based on Kesselmeier and Lorenzo Bermejo (in preparation) [currently submitted to *Briefings in Bioinformatics*].

- Sections 2.1.1, 2.2.1, 4.1.1 and 5.1.1 were presented at the *Annual Meetings of the International Genetic Epidemiology Society (IGES) 2011* and *2012* (conference abstracts: Kesselmeier et al., Genetic Epidemiology 2012, 36:157–157; Kesselmeier et al., Genetic Epidemiology 2012, 36:768–769).
- Sections 2.1.2, 2.2.2, 4.1.2 and 5.1.2 are based on Kesselmeier et al. (2014) and were also presented within the *18th Genetic Analysis Workshop* (GAW 18).
- Sections 2.1.4, 2.2.3, 4.1.3 and 5.1.3 were presented at the 3*rd Joint Statistical Meeting of the DAGStat "Statistics under one Umbrella"* 2013.
- Sections 2.1.3, 2.1.5, 2.3, 4.2 and 5.2 are based on Kesselmeier and Lorenzo Bermejo (in preparation) [currently submitted to *Briefings in Bioinformatics*]. Parts were also presented at the *42nd European Mathematical Genetics Meeting* 2014 (conference abstract: Kesselmeier and Lorenzo Bermejo, Human Heredity 2013, 76:104–105) and at the *4th Joint Statistical Meeting of the DAGStat "Statistics under one Umbrella"* 2016.

2 Material and methods to compare standard and robust regression approaches

Simulated and real data are usually used to investigate performance of statistical methods. An advantage of simulated data is the knowledge about data properties, e.g. underlying distribution, relations between variables and effect sizes. But simulations stay artificial despite all invested efforts to create them in a realistic way. Thus, it is a common wish to observe the method performance in real data applications. In this chapter, the simulated and the real data are presented with background information at first (section 2.1). Then, the methods to compare different standard and robust regression approaches on these simulated and real data sets are developed (sections 2.2 and 2.3).

All calculations were done with R version 3.0.2 (R Core Team, 2013). The standard regression models were estimated using `lm` for linear models and `glm` for generalised linear models, such as logistic and Poisson regression. For the robust linear models, the function `rlm` from the package `MASS` was used (Venables and Ripley, 2002). For robust logistic and Poisson regression considering the Huber function, the existing function `glmrob` from the R package `robustbase` was applied (Rousseeuw et al., 2012).

2.1 Data sets

This section provides a description of the data sets. The first three data sets are real data sets dealing with DNA methylation and chromosomal instability in individuals suffering from human hepatocellular carcinoma (section 2.1.1), genotypes and phenotypes in individuals with hypertension (section 2.1.2) and genetic data from the Personal Genome Project (section 2.1.3). Then, the simulation of two data sets are presented in the sections 2.1.4 and 2.1.5.

2.1.1 DNA methylation and chromosomal instability in individuals suffering from human hepatocellular carcinoma

Genetic background

Gene expression is usually quantified by the amount of resulting products, for example proteins. Amongst others, expression of a gene depends on its accessibility. If DNA is densely packed around histones (alkaline proteins) a gene cannot be expressed. Epigenetical control, such as DNA methylation, influences gene expression as well. This process means that a methyl group binds to the DNA chain, typically occurring at CpG sites (dinucleotide **C**ytosine-**p**hosphate-**G**uanine). This event can cause gene silencing (Seyffert, 2003). Furthermore, gene expression can be influenced by copy number variations. These variations are either deletions or local duplications of chromosomal regions compared to a reference. In particular, a chromosomal region is called “normal” if there are two copies of this region. A situation with at least one copy less than in the reference is called “loss” and a situation is called “gain” if there is at least one more copy than in the reference. So it is clear that duplication can result in a higher expression compared to a deletion which usually leads to loss of expression. It is possible to measure copy number variation via array-based comparative genomic hybridisation (aCGH) (Stratton et al., 2009; van Wieringen et al., 2013). aCGH information can be used to define chromosomal instability which is related to tumour stage in several kinds of cancer (van Wieringen et al., 2013). To later define chromosomal instability via aCGH data, the term “centromer” is needed. The centromer is the part near the middle of a chromosome linking sister chromatids. It divides each chromosome into a short and a long arm (Laird and Lange, 2011). Since gene expression is related to both chromosomal instability and DNA methylation, it is of interest to explore the relationship between DNA methylation and chromosomal instability.

Epidemiological and medical background

This data set comprises information about persons suffering from human hepatocellular carcinoma (HCC) which is one of the most frequent malignancies worldwide – among men the 5th and among women the 7th most common. The incidence depends on the geographic region with incidence rates from more than 20 per 100,000 individuals in Sub-Saharan Africa and Eastern Asia to less than 5 per 100,000 individuals in North and South America (Mittal and El-Serag, 2013). An early diagnosis is essential for a successful treatment (de Lope et al., 2012). HCC develops mostly in a process lasting several years which is normally initiated by a chronic liver disease. The underlying aetiology is often unknown (El-Serag and Rudolph, 2007; Breuhahn, 2010). Most prominent risk factors are cirrhosis, the infections with the hepatitis B or C virus (HBV, HCV), non-alcoholic fatty liver disease (NAFLD) and high alcohol consumption whereas coffee intake seems to reduce the risk of developing liver cirrhosis and thus of HCC. Because only a small number of persons infected with HBV or HCV develops HCC, genetic factors might influence the progression to HCC as well (Mittal and El-Serag, 2013). Because cancer has its origin in DNA alteration, a better understanding of the relationship between chromosomal alterations and gene methylation may advance the identification of relevant steps in the development of HCC (van Wieringen et al., 2013).

Data

The SFB/TRR77 Consortium has generated a collection of patients with aCGH information, gene expression and DNA methylation (Neumann et al., 2012; Kesselmeier et al., in preparation). Information on DNA methylation for 600 selected genes as proposed by Hoshida et al. (2009) as well as aCGH data was available for 54 HCC samples. CpG sites in the selected genes were included in the analysis if they fulfilled the data quality criteria which were (i) a detection p-value below 0.01, (ii) methylation values between 0 and 1 and (iii) a positive median absolute deviation (MAD) which is defined as

$$\mathrm{MAD}(X) = \mathrm{median}\left(|X - \mathrm{median}(X)|\right)$$

with $X \in \mathbb{R}^n$ (Bortz and Schuster, 2010). A chromosome arm was defined as instable if it contained at least one region larger than 1000 kb with a gain or a loss. Genomic instability of one sample was defined as the number of instable chromosome arms divided by the total number of investigated arms.

2.1.2 Genotypes and real phenotypes in individuals with hypertension

Epidemiological and medical background

Data for the second real data application describes persons with and without hypertension. Hypertension is a common chronic disease characterised by elevated arterial blood pressure. High blood pressure is associated with an increased risk of stroke, heart attack and other serious diseases. Age, gender, tobacco smoking, alcohol consumption and high body mass index (BMI) constitute established risk factors for hypertension (Jonas et al., 1997). A genetic component has also been postulated. It has been shown that individuals with a family history of hypertension have on average a higher blood pressure than individuals without. For siblings of affected persons Yanek et al. (1998) found a 44% higher prevalence of hypertension than in the general reference population. In a Canadian study, standardised risk ratios of hypertension were reported to be higher for first-degree relatives than for spouses of probands with hypertension (Katzmarzyk et al., 2001). In genetic studies, a large number of polymorphisms has been associated with hypertension and validated in independent collectives; fourteen loci have been identified until 2010 and many genetic studies are currently in progress (Levy et al., 2009; Newton-Cheh et al., 2009; Wang et al., 2009; Ehret, 2010; Padmanabhan et al., 2012).

Data

The analysed data (real phenotypes) from the 18th Genetic Analysis Workshop (GAW 18) were derived from 142 unrelated individuals who participated in the San Antonio Family Heart or Family Diabetes/Gallbladder studies. Longitudinal information on hypertension, age, gender and current tobacco smoking was measured up to four times per individual. The present analyses relied on the first available measurement. Further information is provided in several articles (Mitchell et al., 1996; Duggirala et al., 1999; Hunt et al., 2005; Almasy et al., 2014).

2.1.3 Genetic data from the Personal Genome Project

Project background

There is a global network of Personal Genome Projects (Church, 2005) including projects in the United States of America at Harvard Medical School, in Canada at the University of Toronto and the Hospital for Sick Children in Toronto, in the United Kingdom at the University College London and in Austria at the CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences in Vienna (`http://www.personalgenomes.org/`). The aim of this project is to provide public genetic data comprising information about individual's genome, its health / environment and human traits to extend the knowledge about human health and to guide the evidence-based medicine. All participants voluntarily provide their personal data in a non-anonymous manner. Currently (2016/06/20), only the site at Harvard Medical School provides enough data for public use because the other projects are still under construction.

Data

The public genetic data was accessed at the Personal Genome Project website (`https://my.pgp-hms.org/public_genetic_data`) on 2015/06/11. The filter "genetic data - 23andMe (e.g., exome or genotyping data)" was applied and text files with genome data downloaded. The individual's age and body height information was manually extracted from the participant profiles (`https://my.pgp-hms.org/users`) on 2015/06/11. When several files were available per person, only the most recent one was included. Files with incomplete and inconsistent genotype data were excluded as well as files without the corresponding information on age and body height. The first 1000 genotypes were extracted from each file, variants measured in all individuals selected and individuals with missing genotypes subsequently excluded. Figure 4 shows the flow chart for data pre-processing.

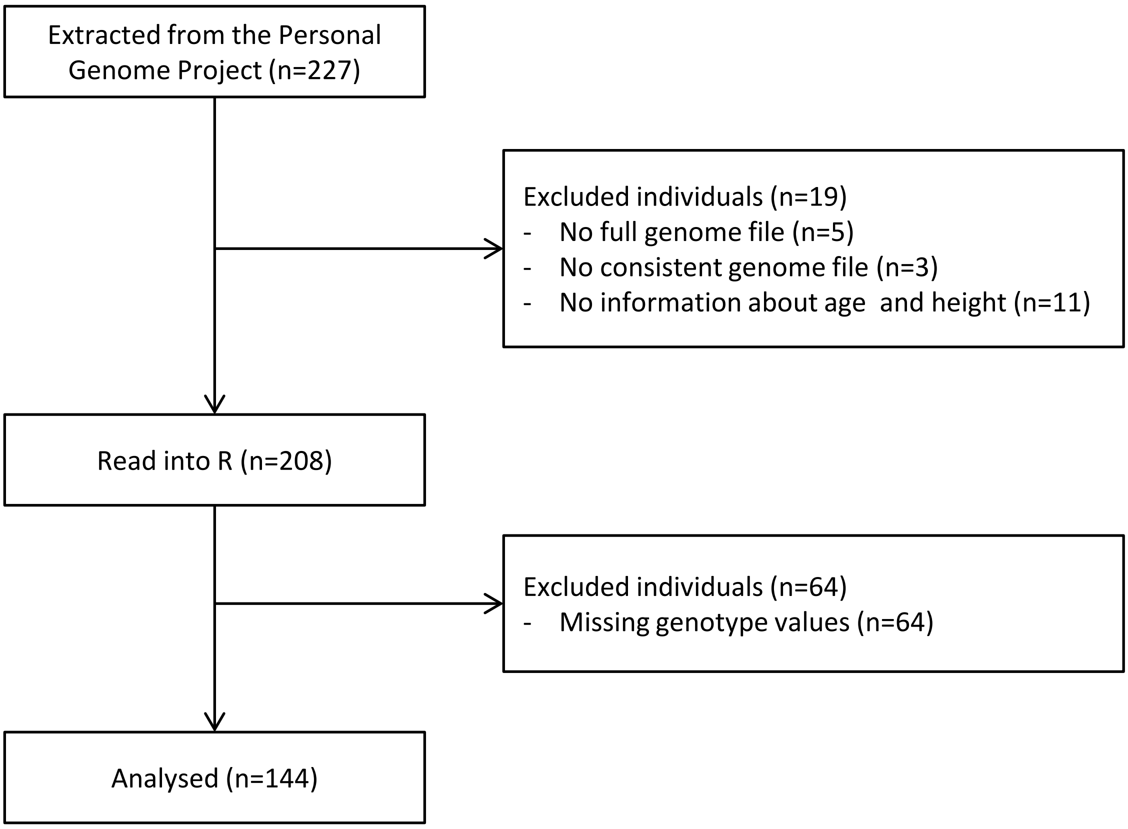

Extracted from the Personal Genome Project (n=227)
Excluded individuals (n=19)
- No full genome file (n=5)
- No consistent genome file (n=3)
- No information about age and height (n=11)
Read into R (n=208)
Excluded individuals (n=64)
- Missing genotype values (n=64)
Analysed (n=144)

Figure 4: Flow chart of data processing in the real data application.

2.1.4 Simulated data for increasing genotyping error rates

The simulation will start with given genotypes and age to deduce the case control status. The genotypes will afterwards be modified by introducing genotyping errors. The simulation strategy is given in figure 5.

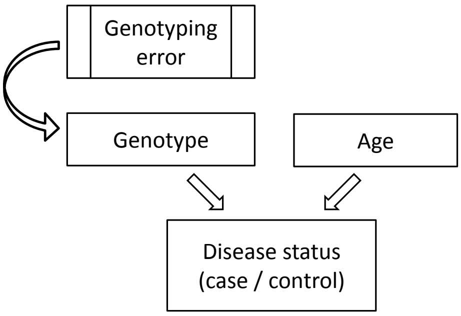


Figure 5: Simulation strategy. Straight arrows indicate the simulation direction. A bent arrow denotes the influence direction.

One must consider in advance some probabilities for data generation. Given are the causal allele frequency (p_C), the genotype relative risks for homozygous and heterozygous causal allele carriers ($\mathrm{GRR}_{\mathrm{C,hom}}$ and $\mathrm{GRR}_{\mathrm{C,het}}$) and the prevalence in the population (π). p_C can be transformed into the minor allele frequency (MAF). The MAF is defined as the frequency of the less common allele in a given population. Hence, this frequency equals the minimum of p_C and $1 - p_C$. The prevalence π in the population is

$$\kappa_0 \left[p_C^2 \ \mathrm{GRR}_{\mathrm{C,hom}} + 2p_C \left(1 - p_C\right) \mathrm{GRR}_{\mathrm{C,het}} + \left(1 - p_C\right)^2 \right]$$

where κ_0 is the prevalence for non causal allele carriers (π_{nc}). Knowing π, κ_0 can be calculated as

$$\kappa_0 = \frac{\pi}{p_C^2 \ \mathrm{GRR}_{\mathrm{C,hom}} + 2p_C \left(1 - p_C\right) \mathrm{GRR}_{\mathrm{C,het}} + \left(1 - p_C\right)^2}$$

The prevalence for causal allele carriers (π_c) can be calculated based on κ_0 and equals $\kappa_0 \cdot \mathrm{GRR}$.

For simulation, some of these values have to be defined in advance. The causal allele frequency is set to three different values: 0.005, 0.05 and 0.13. The age-specific prevalence is oriented at the NORDCAN project for breast cancer in the year 2010 (Engholm et al., 2010, 2012). The genotype relative risk was taken from results investigating the association of the *CHECK2* gene to breast cancer risk (CHEK2 Breast Cancer Case-Control Consortium, 2004). The age distribution is orientated at the age distribution in Sweden 2010. These population-specific characteristics are shown in table 3.

Table 3: Age, prevalence and genotype relative risk for simulation.

Age interval (numeric ID)	Age frequency	π(Age)	$\mathrm{GRR}_{\mathrm{Hom}}$(Age)
20-29 (0)	0.172	0.0001	7.91
30-39 (1)	0.023	0.0015	2.65
40-49 (2)	0.192	0.0084	2.80
50-59 (3)	0.212	0.0245	2.13
60-69 (4)	0.401	0.0478	1.95

For heterozygous allele carriers, the genotype relative risk ($\mathrm{GRR}_{\mathrm{C,het}}$) depends on the penetrance model and is defined as

$$\mathrm{GRR}_{\mathrm{C,het}} = \begin{cases} \frac{1}{2}\left(\mathrm{GRR}_{\mathrm{C,hom}}+1\right) & \text{for an additive penetrance model} \\ \mathrm{GRR}_{\mathrm{C,hom}} & \text{for a dominant penetrance model} \\ 1 & \text{for a recessive penetrance model} \end{cases} \tag{2.1}$$

Following these preparations, a dominant penetrance model was assumed. The simulation started with genotypes and age intervals. For the genotype, one drew two uniform distributed random numbers on the interval $[0, 1]$ (rn_{g_1}, rn_{g_2}). Genotype G was then defined via

$$G = \begin{cases} CC & \text{if } rn_{g_1} < p_C \text{ and } rn_{g_2} < p_C \\ cc & \text{if } rn_{g_1} \geq p_C \text{ and } rn_{g_2} \geq p_C \\ Cc & \text{otherwise} \end{cases}$$

The corresponding age interval A was estimated according to the defined age distribution using a uniform distributed random number on the interval $[0, 1]$ rn_a. This means that

$$A = \begin{cases} 0 & \text{if } rn_a < 0.172 \\ 1 & \text{if } 0.172 \leq rn_a < 0.195 \\ 2 & \text{if } 0.195 \leq rn_a < 0.387 \\ 3 & \text{if } 0.387 \leq rn_a < 0.599 \\ 4 & \text{otherwise} \end{cases}$$

Using these information and a uniform distributed random number $rn_{cc} \in [0, 1]$, the case control status (CaCo) follows as

$$\mathrm{CaCo} = \begin{cases} \text{case} & \text{if } (G \neq cc \text{ and } rn_{cc} < \pi_c) \text{ or } (G = cc \text{ and } rn_{cc} < \pi_{nc}) \\ \text{control} & \text{otherwise} \end{cases}$$

Two scenarios were investigated: (i) no genotyping errors occurred (model M_0) and (ii) different genotyping error rates were considered, namely 0.005, 0.010, 0.025 and 0.050 (model M_1). In case of genotyping errors, the fixed proportion of true genotypes were randomly assigned to one of the two other possible genotypes. According to these parameters, 1000 data sets with 1000 cases and 1000 controls were simulated.

2.1.5 Simulated data for varying population characteristics

A data set with information on age and genotype was generated for 3.5 million cases and 3.5 million controls. Genetic association studies were simulated by random sampling from this large data set. The age distribution of controls relied on data from the European Union (Office for Official Publications of the European Communities, 2006). The age of cases mirrored the incidence of colorectal cancer in European women (Ferlay et al., 2013). The two age distributions are shown in table 4.

Table 4: Age distribution, prevalence and age-dependent genotype relative risk (GRR) for the simulation study

Age interval [years]	Control frequency	Disease prevalence	Age-dependent GRR
≤ 35	0.14	0.0001	20.00
$36 - 40$	0.14	0.0007	15.00
$41 - 45$	0.13	0.0019	10.00
$46 - 50$	0.13	0.0040	5.00
$51 - 55$	0.12	0.0073	1.57
$56 - 60$	0.10	0.0122	1.00
$61 - 65$	0.09	0.0189	1.00
$66 - 70$	0.08	0.0273	1.00
> 70	0.07	0.0389	1.00

Null marker genotypes were simulated independently of case-control status. For associated markers, the age of the individual was first drawn according to case-control status. Then, causal variant genotypes were simulated assuming a given penetrance model. In more detail, let GRR_{hom} represent the relative risk for homozygous carriers of the causal variant. The GRR for carriers of only one copy of the causal variant (GRR_{het}) was

$$\text{GRR}_{\text{het}} = \begin{cases} \frac{1}{2}\left(\text{GRR}_{\text{hom}} + 1\right) & \text{for an additive penetrance model} \\ \text{GRR}_{\text{hom}} & \text{for a dominant penetrance model} \\ 1 & \text{for a recessive penetrance model} \end{cases}$$

Let C denote the high-risk allele and c the low-risk allele at the causal locus. Let M denote the high-risk allele and m the low-risk allele at the marker locus. Let p_C be the causal allele frequency and p_M the marker allele frequency. p_C and p_M were related via

$$p_M = \left[\frac{r^2(1-p_C)}{p_C D'^2} + 1\right]^{-1}$$

where r^2 represents the correlation and $D^{'}$ Lewinson's measure of the relative linkage disequilibrium between causal and marker loci. The expected distribution of genotypes (G) at the marker locus in controls (D=0) was

$$P[G = mm|D = 0] = (1 - p_M)^2$$

$$P[G = Mm|D = 0] = 2p_M(1 - p_M)$$

and

$$P[G = MM|D = 0] = p_M^2$$

The expected distribution of genotypes in cases (D=1) was

$$\mathrm{GRR}_{\mathrm{hom}} = \frac{P[D = 1|G = MM]}{P[D = 1|G = mm]}$$

and

$$\mathrm{GRR}_{\mathrm{het}} = \frac{P[D = 1|G = Mm]}{P[D = 1|G = mm]}$$

with $P[D = 1|G = mm] = \kappa_0$ representing the disease prevalence among low-risk allele homozygotes. Let κ denote the disease prevalence in the total population. Then,

$$P[G = mm|D = 1] = (1 - p_M)^2\ \frac{\kappa_0}{\kappa}$$

$$P[G = Mm|D = 1] = 2p_M(1 - p_M)\mathrm{GRR}_{\mathrm{het}}\ \frac{\kappa_0}{\kappa}$$

and

$$P[G = MM|D = 1] = p_M^2\mathrm{GRR}_{\mathrm{hom}}\ \frac{\kappa_0}{\kappa}$$

In summary, genotypes depended on genetic parameters (MAF, GRR, penetrance model for the causal allele, association (r^2 and $D^{'}$) between causal and marker loci) and also on study characteristics (sample size and genotyping error rate) that were specified considering different scenarios (Hemminki and Lorenzo Bermejo, 2007; Lorenzo Bermejo et al., 2011; Lewontin, 1964; Hill and Robertson, 1968). Ten null marker loci and one marker locus were simulated.

Under the reference scenario, the MAF was fixed to 0.05 for a dominant causal variant, no genotyping errors assumed and 400 studies simulated with 1000 cases and 1000 controls each. The GRR was set to 1.43 to reach a statistical power equal to 0.6. This reference scenario built the basis for sensitivity analyses where just one parameter was changed at once: the penetrance model fitted to the data (additive, recessive), the MAF (from 0.001 to 0.25) and r^2 (from 0.8 to 1.0) were modified. In addition to a constant GRR of 1.43 under the reference scenario, decreasing GRRs with increasing age as specified in table 4 were also considered. Age-dependent GRRs were consistent with the overall GRR of 1.43 assumed in the reference scenario. Genotyping errors were considered, too. For this purpose, a fixed proportion of true genotypes were randomly assigned to one of the two other possible genotypes. Genotyping arrays generally show error rates below 0.01, but genotyping errors seem to be more frequent for sequence data (Kennedy et al., 2003; Montgomery et al., 2005; Hong et al., 2012). In the present simulations, genotyping error rates varied from 0 to 0.05. Because the aim was to stay as realistic as possible and it was expected that some of these parameters introduced extreme observations with a possible impact on the different approaches, no further contamination was introduced.

Preliminary results motivated a closer investigation of rare and recessive variants. For rare variants, study and effect sizes were accommodated to reach around 0.6 statistical power using standard logistic regression. This led to the triplets (MAF, number of cases / controls, assumed GRR) equal to (0.001, 5000/5000, 2.53), (0.005, 1000/1000, 2.65) and (0.01, 1000/1000, 2.07). Genotyping error rates from 0 to 0.05 were considered. The remaining parameters were fixed to the same values as in the reference scenario. For recessive variants, the GRR was fixed to 6.32 in order to achieve 0.6 statistical power (MAF=0.05 and 1000 cases / 1000 controls). Additive, dominant and recessive models were fitted to recessively simulated data. Again, the genotyping error rate varied from 0 to 0.05. The remaining parameters were fixed to the same values as in the reference scenario.

2.2 Methods to compare standard versus existing robust regression methods

Sections 2.2.1 and 2.2.2 deal with two real data applications. The first one (section 2.2.1) is about the comparison of standard and robust linear as well as Poisson regression with respect to model selection consistency and prediction accuracy. In the second real data application (section 2.2.2), standard and robust logistic regression clarify their handling of one single outlier. In section 2.2.3, the influence of genotyping errors on standard and robust logistic regression estimates are investigated in simulated data.

2.2.1 Consistency of model selection and prediction accuracy in real data

With the data set described in section 2.1.1, the relationship between DNA methylation and chromosomal instability was investigated using standard and robust linear as well as Poisson regression with the Huber function relying on the regression models

$$\text{Chromosomal instability} \sim \sum_{i=1}^{n} \text{CpG}(i)$$

for linear regression and

$$\#\text{ instable arms} \sim \sum_{i=1}^{n} \text{CpG}(i) + \text{offset}(\#\text{ investigated arms})$$

for Poisson regression with the ith included CpG site CpG(i) and # denoting “number of”. n denotes the number of CpG sites that were included into the model. The model was forwardly selected. The decision on whether to include an additional variable into the model was made on a deviance criterion. If the deviance was not significantly reduced the model without the additional variable was the final model. In order to compare the four regression methods, 54 leave-one-out cross-validations were used. This means that all but one sample was used as a training data set to select the best model and the remaining sample was used to validate the model. This was repeated till every sample was once used as validation sample. Leave-**one**-out cross-validation was chosen to keep the original data distribution as unaffected as possible. The regression models were then compared based on two criteria: goodness-of-fit and reproducibility (consistency). Reproducibility was defined as the number of times a specific model was selected and goodness-of-fit as the difference between observed and predicted instability. The latter was examined with the Wilcoxon signed rank test to search for differences between the methods. Standard linear regression was used as reference because it is a common and often used technique. The differences between the methods were quantified by the median as well as 5th and 95th quantiles. The relationship between instability and methylation at CpG sites identified as relevant for the regression model as well as gene expression were investigated with Spearman's correlation coefficient with 95% confidence intervals. An overview about model selection and validation is given in figure 6.

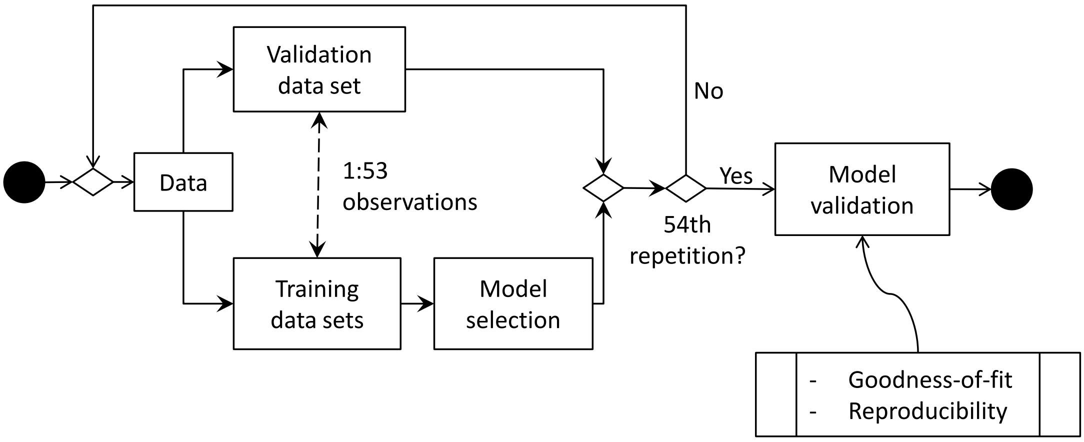


Figure 6: Model selection and validation via leave-one-out cross validation.

2.2.2 Influence of one single outlier in real data

The relationship between inherited genetic polymorphisms and a binary response variable (with/without hypertension) can be investigated using logistic regression models that simultaneously consider the effects of multiple risk factors. Here, data from GAW 18 (see section 2.1.2) was used to explore the possible benefit of robust parameter estimates in logistic regression models for the genetic prediction of hypertension risk. The original data was filtered according to the following criteria: (1) at least one measurement with complete information on hypertension and age, (2) monomorphisms were excluded and each polymorphism had to be represented by at least two individuals, (3) individuals with more than 5% missing genotypes were excluded and (4) variants with missing data in any individual were removed.

The relationship between hypertension and age, gender and current tobacco smoking was first investigated by χ^2 tests. Covariates significantly associated at the 0.05 confidence level entered the intercept-only model to build the baseline model. Subsequently, standard logistic regression (iteratively re-weighted least squares) was used to identify possible hypertension-associated SNPs (single nucleotide polymorphisms) with minimal deviance taking into account associated covariates. The goodness-of-fit criterion deviance D is defined as

$$D(y;\mu) = 2\,l(y;y) - 2\,l(\mu;y)$$

with maximal achievable log-likelihood $l(y;y)$ in an exact fit and the usual log-likelihood $l(\mu;y)$ of the observation y and the mean μ. Minimizing the deviance is equivalent to maximizing the log-likelihood as $l(y;y)$ is independent of the parameters (McCullagh and Nelder, 1996).

Genotypes were coded according to an additive penetrance model, i.e. 0, 1 and 2 indicating the number of causal alleles. Outliers according to standard logistic regression were identified based on the Cook's distance in the baseline model. The Cook's distance for observation i is defined as

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{q\,\mathrm{MSE}}$$

where $\hat{y}_j$ denotes the full regression model prediction for observation j, $\hat{y}_{j(i)}$ represents the regression model prediction for observation j estimated omitting observation i and MSE indicates the mean squared error of the regression model with q explanatory variables and n observations. Thus, the Cook's distance quantifies the impact of observation i on the regression model.

To investigate the possible benefit of robust parameter estimates in logistic regression, model coefficients were also estimated using the approach proposed by Cantoni and Ronchetti (2001) using the Huber function. Variable selection under robust logistic regression relied on the minimal quasi-deviance as described by Cantoni and Ronchetti (2001), which is a robust test statistic for model selection. The quasi-deviance between two nested models is defined as

$$\Lambda_{\mathrm{QM}} = 2\left[\sum_{i=1}^{n} Q_M(y_i, \hat{\mu}_i) - \sum_{i=1}^{n} Q_M(y_i, \dot{\mu}_i)\right]$$

with

$$Q_M(y_i, \mu_i) = \int_{\tilde{s}}^{\mu_i} v(y_i, t) w(x_i)\, dt - \frac{1}{n}\sum_{j=1}^{n} \int_{\tilde{t}}^{\mu_j} E[v(y_j, t) w(x_j)]\, dt$$

with $\tilde{s}$ such that $v(y_i, \tilde{s}) = 0$ and $\tilde{t}$ such that $E[v(y_i, \tilde{t})] = 0$ and the estimated linear predictor $\hat{\mu}$ is associated to the estimate $\hat{\beta}$ of β and $\dot{\mu}$ is associated to $\dot{\beta}$ which is the estimate of $(\beta_{(1)}, 0)$. Linkage disequilibrium was not accounted for during variant selection.

The comparison of the performance of standard and robust logistic regression was based on different statistics. First, standard and robust estimates of age effects were used to exemplify the potential influence of departing observations. Due to a different handling of outliers, it was expected that different age-genotype models were selected under standard and robust logistic regression. Therefore, the areas under the receiver operating characteristic curves (AUC) were subsequently compared in order to investigate the discriminative performance of the selected models. Comparisons were conducted for the complete data set and after exclusion of potential outliers.

In addition, concordance, sensitivity, specificity, clinical net benefit and AUCs were estimated for age-genotype models using a leave-one-out cross-validation approach (Vickers and Elkin, 2006). Concordance was defined as the proportion of correctly estimated hypertension statuses using several cut-off values for the predicted affection probability. The clinical net benefit (NB) was defined by

$$\begin{aligned}\mathrm{NB}(c) &= \frac{\text{True positive counts}}{\text{Sample size}} - \frac{c}{1-c}\,\frac{\text{False positive counts}}{\text{Sample size}} \\ &= \text{Sensitivity}\,(\%\ \text{Hypertensive}) - \frac{c}{1-c}\,(1 - \text{Specificity})\,(\%\ \text{Normotensive})\end{aligned}$$

where c is the chosen threshold for allocating an individual to the cases based on the logistic regression probability estimate. Note that the net benefit depends on the hypertension prevalence in the study population. The standard and robust logistic regression models were also compared based on the integrated discrimination improvement (IDI) estimated by cross-validation which is defined as

$$\mathrm{IDI} = \left(\frac{1}{n_A} \sum_{i=1}^{n_A} \hat{p}_{\mathrm{rob},i} - \frac{1}{n_N} \sum_{j=1}^{n_N} \hat{p}_{\mathrm{rob},j} \right) - \left(\frac{1}{n_A} \sum_{i=1}^{n_A} \hat{p}_{\mathrm{stand},i} - \frac{1}{n_N} \sum_{j=1}^{n_N} \hat{p}_{\mathrm{stand},j} \right)$$

where $\hat{p}_{\mathrm{rob},i}$, $\hat{p}_{\mathrm{rob},j}$, $\hat{p}_{\mathrm{stand},i}$ and $\hat{p}_{\mathrm{stand},j}$ denote the probability estimates from the robust and standard logistic regression models for cases and controls as well as n_A and n_N the number of cases and controls (Pencina et al., 2008). The IDI represents the difference in the discrimination slopes of the two compared models. A positive IDI indicates that the robust model discriminates better between hypertensive and normotensive individuals than the standard model.

2.2.3 Influence of genotyping errors on estimates in simulated data

For both scenarios described in section 2.1.4, standard and robust logistic regression with the Huber function was used to estimate genotype odds ratios (ORs). A value of 1.345 was introduced as tuning constant for the Huber function. To compare standard and robust logistic regression, the relative differences in the genotype ORs calculated based on standard and robust logistic regression were compared. Let ΔOR denote this relative difference. Based on preliminary results, a narrow genotyping error rate grid varying from 0 to 0.05 was additionally evaluated regarding $|\Delta\mathrm{OR}_{stand}| - |\Delta\mathrm{OR}_{rob}|$ for a causal allele frequency of 0.13.

2.3 Methods for statistical properties evaluation applying the Hampel function

Introduction

Logistic regression is an established technique used in genetic case-control association studies to investigate the relationship between genetic markers and a disease of interest simultaneously considering possible confounders. The large sample sizes required to identify novel low-penetrance susceptibility variants often result in some study individuals with genotypes and phenotypes departing from the majority of the population (outliers). It is well known that outliers strongly influence standard maximum likelihood estimators. For example, few patients diagnosed unusually early in life, and also healthy controls of advanced age, may outweigh the bulk of 'average individuals' in the calculation of standard probability values, point estimates and confidence intervals (Sarkar et al., 2011; Wilcox, 1998; Muhlbauer et al., 2009). Outlier identification can be extremely challenging due to the high-dimensionality of genetic data, which is often accompanied by reciprocal masking of outlier effects. Even if outliers can be flagged, outlier definition is always arbitrary and their handling often controversial. Robust statistics aim to estimate population parameters relying on the majority of the study population. Therefore, they constitute a valuable alternative to the state of the art outlier identification and subsequent arbitrary removal.

Standard and robust logistic regression

Logistic regression is a generalisation of the linear regression model. In logistic regression, the conditional mean of the response variable is linked to a linear combination of explanatory variables (linear predictor), usually via the logit or probit link functions. The model investigated in the present study is

$$\text{logit}(E[Y]) = \text{logit}(\mu) = X\beta + \varepsilon$$

where the n-dimensional vector Y represents the case-control status as response variable of n individuals, the $n \times 3$-dimensional matrix $X = (X_1, X_2, X_3)$ $(X_i \in \mathbb{R}^n)$ includes the intercept (X_1) and the individual genotype (X_2) as well as age (X_3) as explanatory variables, $\beta \in \mathbb{R}^3$ is a coefficient vector and ε is an error term. In standard logistic regression, β is estimated by maximizing the log-likelihood

$$L(\beta|Y, X) = Y^T X\beta - \sum_{i=1}^{n} \log\left(1 + \exp\left(X_{i\cdot}\beta\right)\right)$$

with $X_{i\cdot}$ denoting the ith row of X. Maximum likelihood estimators of β are found by solving the equation

$$\frac{\partial}{\partial\beta} L(\beta|Y, X) = Y^T X - \sum_{i=1}^{n} \frac{\exp\left(X_{i\cdot}\beta\right)}{1 + \exp\left(X_{i\cdot}\beta\right)} X_{i\cdot}$$

Cantoni and Ronchetti's robust estimator relies on the wider class of M-estimators of β

$$\sum_{i=1}^{n}\left[\psi(r_i;\beta,\phi,c)w(X_{i\cdot})\frac{1}{\sqrt{\phi\nu_i}}\mu_i^{'}-\alpha(\beta)\right]=0$$

where r_i represents Pearson residuals

$$r_i=\frac{Y_i-\mu_i}{\sqrt{\phi\nu_i}}$$

In particular, for logistic regression

$$\mu_i=n\pi_i$$

and

$$\phi\nu_i=n\pi_i(1-\pi_i)$$

with the number of individuals n and the disease probability π_i. Other components of the M-estimator equation are

$$\mu_i^{'}=\frac{\partial\mu_i}{\partial\beta}=\frac{\partial}{\partial\beta}\left[1+\exp\left(-X_{i\cdot}\beta\right)\right]^{-1}=\frac{\exp\left(-X_{i\cdot}\beta\right)}{\left[1+\exp\left(-X_{i\cdot}\beta\right)\right]^{2}}X_{i\cdot}$$

and

$$\alpha(\beta)=\frac{1}{n}\sum_{i=1}^{n}E\left[\psi(r_i;\beta,\phi,c)\right]w(X_{i\cdot})\frac{1}{\sqrt{\phi\nu_i}}\mu_i^{'}$$

which is a constant guaranteeing Fisher consistency of the estimator (Cantoni and Ronchetti, 2001). In the particular case of maximum likelihood estimators,

$$\psi(r_i;\beta,\phi,c)=r_i$$

and

$$w(X_{i\cdot})=1$$

for all observations. Different influence functions $\psi(r_i;\beta,\phi,c)$ and weight functions $w(X_{i\cdot})$ can be used for robust parameter estimation. Here, the weight function

$$w(X_{i\cdot})=\sqrt{1-h_{ii}}$$

with h_{ii} the ith diagonal element of the matrix $H=X(X^TX)^{-1}X^T$ was considered although it does not have a high breakdown point. An often used choice with a high breakdown point is related to the Mahalanobis distance but its use is only reasonable in case of continuous explanatory variables (Heritier et al., 2009). Different bounded influence functions can be used to constrain outlier influence. In this study the bounded Huber and the re-descending Hampel function were investigated (Huber, 1964; Hampel et al., 1986). Tuning constants can be selected to ensure 95% asymptotic efficiency when used in the Gaussian family with identity link (i.e. the linear model) in the absence of outliers. Obviously, they do not necessarily yield the 95% asymptotic efficiency in the logistic regression framework. According to this, a c value equal to 1.345 for the Huber function and the two (a, b, c) vector values $(1.5, 3.5, 8) \times 0.9$ and $(2, 4, 8) \times 0.7$ corresponding to slopes of the re-descending part of the Hampel function equal to $\frac{1}{3}$ and $\frac{1}{2}$, respectively, were chosen (Rousseeuw et al., 2012; Koller and Stahel, 2011; Koller and Mächler, 2014).

Computer simulations

Extensive simulations were conducted to examine the impact of differences between standard and robust GRR estimation on the type I error rate, bias, variance, mean squared error (MSE) and statistical power (see the data set described in section 2.1.5). The case-control status was regressed on individual genotype and age using standard and robust logistic regression using the above described influence functions and tuning constants. Then, standard and robust GRR estimates were compared with respect to type I error rate, bias, variance, mean squared error (MSE) and statistical power. The type I error rate was derived as the false positive rate at a 0.05 significance level across null marker loci (Majumdar et al., 2013). The bias of the GRR estimator was calculated as the difference between the mean estimated GRR in simulated studies and the true GRR used for simulation. The MSE was calculated as the sum of the squared bias and the variance of GRR estimates. Statistical power was estimated as the true positive rate at a 0.05 significance level.

Application to real data

In the data set described in section 2.1.3, genotypes were recoded assuming a recessive penetrance model (homozygous carriers of the minor variant versus others) and the body hight dichotomised (1 larger than the median, 0 otherwise). Then, the dichotomised body height was regressed on individual genotype and age for each genotype. Standard and robust logistic regression was used with the Huber and the Hampel functions and the tuning constants described above. Possible influential observations were identified based on Cook's distances and methods were compared with respect to p-values and estimated GRRs.

R code for simulations and analysis of the real data is provided in appendix D on page 153.

3 Incorporation of the Hampel function into robust generalised linear models

As mentioned before, one can use the R script provided in the `robustbase` package for robust generalised linear models with the Huber function only to account for outliers in the response variable (R Core Team, 2013; Rousseeuw et al., 2012). Thus, to use different weighting functions, one must adapt the Fisher consistency correction and the asymptotic variance according to Cantoni and Ronchetti (2001). The only part to be evaluated for the Fisher consistency correction is the expectation $E[\nu(y_i, \mu_i)]$ which can be rewritten as

$$\mathrm{E}[\nu(y_i, \mu_i)] = \frac{\mathrm{E}[\psi(r_i)]}{V^{1/2}(\mu_i)}$$

with the Pearson residuals

$$r_i = \frac{Y_i - \mu_i}{V^{1/2}(\mu_i)}$$

(cf. appendix A of Cantoni and Ronchetti (2001)). Hence, one has to derive

$$\mathrm{E}[\psi(r_i)] \tag{3.1}$$

For estimation of the asymptotic variance, one must determine two weighting function ψ-dependent expressions:

$$\mathrm{E}\left[\psi^2\left(r_i\right)\right] \tag{3.2}$$

and

$$\mathrm{E}\left[\psi\left(r_i\right)\frac{Y_i - \mu_i}{V(\mu_i)}\right] \tag{3.3}$$

(cf. appendix B of Cantoni and Ronchetti (2001)).

Due to the structure of the Hampel function (cf. equation (1.9)), one must distinguish different cases for the Pearson residual value r: $r \in (-\infty, -c]$ (case D_1), $r \in (-c, -b]$ (case C_1), $r \in (-b, -a]$ (case B_1), $r \in (-a, a)$ (case A), $r \in [a, b)$ (case B_2), $r \in [b, c)$ (case C_2) or $r \in [c, \infty)$ (case D_2). In the case of case A, it is

$$-a \leq \frac{j - \mu_i}{V^{1/2}(\mu_i)} \leq a$$

for a realisation j of a random variable. This is equivalent to

$$\mu_i - aV^{1/2}(\mu_i) \leq j \leq \mu_i + aV^{1/2}(\mu_i)$$

Accordingly, limits to allocate a realisation j of a random variable to a specific case are

$$j_{z_1} := \lfloor \mu_i - zV^{1/2}(\mu_i) \rfloor \quad \text{and} \quad j_{z_2} := \lfloor \mu_i + zV^{1/2}(\mu_i) \rfloor$$

with $z \in \{a, b, c\}$. For example, the condition $j_{a_1} + 1 \leq j \leq j_{a_2}$ for realisation j of a random variable limits this random variable to case A.

In this chapter, the required calculations for the Fisher consistency correction and asymptotic variance to use the Hampel function are given for binomial and Poisson distributed random variables Y_i, $i = 1, \ldots, n$ (section 3.1). The implementation in R for the Hampel function follows in section 3.2. To finish the development, a plausibility check is given in section 3.3 for a first check of the performance of the implementation.

3.1 Calculation of Fisher consistency correction and asymptotic variance

This section is about the calculations of the Fisher consistency correction and the asymptotic variance. Subsection 3.1.1 provides these calculations for a binomial distributed random variable with the results in theorems 1-3 on pages 53, 56 and 59. Subsection 3.1.2 provides the same for a Poisson distributed random variable with the results in theorems 4-6 on pages 61, 63 and 67.

3.1.1 Binomial distributed random variable

Let $Y_i \sim \text{Bin}(m_i, p_i)$, $\tilde{Y}_i \sim \text{Bin}(m_i{-}1, p_i)$ and $\tilde{\tilde{Y}}_i \sim \text{Bin}(m_i - 2, p_i)$ be three binomial distributed random variables related by

$$j \ \text{P}[Y_i = j] = \mu_i \ \text{P}[\tilde{Y}_i = j] \tag{3.4}$$

and

$$j(j-1) \ \text{P}[Y_i = j] = m_i(m_i-1)p_i^2 \ \text{P}[\tilde{\tilde{Y}}_i = j] \tag{3.5}$$

Note that

$$j^2 - 2\mu_i j = j(j-1) + j(1-2\mu_i) \tag{3.6}$$

Then, it holds for expectation (3.1) to correct for Fisher consistency:

Theorem 1 (Binomial distributed random variable: Fisher consistency correction).

$$\begin{aligned}
&\text{E}\left[\psi_{Hampel}\left(\frac{Y_i-\mu_i}{V^{1/2}(\mu_i)}\right)\right] \\
&\quad= \frac{\mu_i}{V^{1/2}(\mu_i)}\left(\text{P}\Big[j_{a_1} \le \tilde{Y}_i \le j_{a_2}-1\Big] - \text{P}[j_{a_1}+1 \le Y_i \le j_{a_2}]\right) \\
&\qquad+ a\left(\text{P}[j_{a_2}+1 \le Y_i \le j_{b_2}] - \text{P}[j_{b_1}+1 \le Y_i \le j_{a_1}]\right) \\
&\qquad+ \frac{a}{c-b}\Bigg\{\left(c+\frac{\mu_i}{V^{1/2}(\mu_i)}\right)\text{P}[j_{b_2}+1 \le Y_i \le j_{c_2}] \\
&\qquad- \left(c-\frac{\mu_i}{V^{1/2}(\mu_i)}\right)\text{P}[j_{c_1}+1 \le Y_i \le j_{b_1}] \\
&\qquad- \frac{\mu_i}{V^{1/2}(\mu_i)}\left(\text{P}\Big[j_{b_2} \le \tilde{Y}_i \le j_{c_2}-1\Big] + \text{P}\Big[j_{c_1} \le \tilde{Y}_i \le j_{b_1}-1\Big]\right)\Bigg\}
\end{aligned}$$

Proof. Expectation (3.1) for the Hampel function equals

$$\sum_{j=0}^{m_i} \left\{ \frac{j-\mu_i}{V^{1/2}(\mu_i)} \mathbb{1}_A + a\,(\mathbb{1}_{B_2} - \mathbb{1}_{B_1}) + \frac{c - \left|\frac{j-\mu_i}{V^{1/2}(\mu_i)}\right|}{c-b}\, a\,(\mathbb{1}_{C_2} - \mathbb{1}_{C_1}) \right\} \mathrm{P}[Y_i = j] \quad (3.7)$$

This expression will be calculated in three steps. For case A, one splits the sum into two sums, i.e.

$$V^{-1/2}(\mu_i) \left(\sum_{j=0}^{m_i} j\,\mathrm{P}[Y_i = j]\ \mathbb{1}_A - \mu_i \sum_{j=0}^{m_i} \mathrm{P}[Y_i = j]\ \mathbb{1}_A \right) \quad (3.8)$$

Allocating realisation j of Y_i to the probability $\mathrm{P}[Y_i = j]$ in the first sum and the direct derivation of the second sum of the probabilities leads to

$$\frac{\mu_i}{V^{1/2}(\mu_i)} \left(\sum_{j=1}^{m_i} \mathrm{P}\left[\tilde{Y}_i = j\right] \mathbb{1}_A - \mathrm{P}[j_{a_1} + 1 \le Y_i \le j_{a_2}] \right)$$

because the second sum of equation (3.8) equals the probability of Y_i having a realisation according to case A. The remaining sum is equal to the probability of $\tilde{Y}_i$ having a realisation between j_{a_1} and $j_{a_2}-1$, i.e. $\mathrm{P}\left[j_{a_1} \le \tilde{Y}_i \le j_{a_2} - 1\right]$. Combination of these results leads to

$$\frac{\mu_i}{V^{1/2}(\mu_i)} \left(\mathrm{P}\left[j_{a_1} \le \tilde{Y}_i \le j_{a_2} - 1\right] - \mathrm{P}[j_{a_1} + 1 \le Y_i \le j_{a_2}] \right) \quad (3.9)$$

as result for case A. One can directly determine the term for cases B_1 and B_2 as the probability of the random variable Y_i to have a realisation according to these two cases. This equals

$$a \cdot (\mathrm{P}[j_{a_2} + 1 \le Y_i \le j_{b_2}] - \mathrm{P}[j_{b_1} + 1 \le Y_i \le j_{a_1}])$$

The last remaining term of equation (3.7) is the expression for the cases C_1 and C_2. Due to the absolute value of the residuals, one must distinguish between case C_1 and C_2 in more detail. First, one splits the sum into one residual-independent sum and in one residual-dependent sum which results in

$$\frac{a\,c}{c-b} \sum_{j=0}^{m_i} \mathrm{P}[Y_i = j]\ (\mathbb{1}_{C_2} - \mathbb{1}_{C_1}) - \frac{a}{c-b} \sum_{j=0}^{m_i} \frac{j-\mu_i}{V^{1/2}(\mu_i)}\ \mathrm{P}[Y_i = j]\ (\mathbb{1}_{C_2} + \mathbb{1}_{C_1})$$

One directly calculates the residual-independent expression and derives the dependent part similarly to equation (3.9). Hence, this leads to

$$\begin{aligned}
&\frac{a\,c}{c-b}\left(\mathrm{P}[j_{b_2}+1 \le Y_i \le j_{c_2}] - \mathrm{P}[j_{c_1}+1 \le Y_i \le j_{b_1}]\right) \\
&\qquad - \frac{a}{c-b}\,\frac{\mu_i}{V^{1/2}(\mu_i)}\Big(\mathrm{P}\Big[j_{b_2} \le \tilde{Y}_i \le j_{c_2}-1\Big] + \mathrm{P}\Big[j_{c_1} \le \tilde{Y}_i \le j_{b_1}-1\Big] \\
&\qquad - \mathrm{P}[j_{b_2}+1 \le Y_i \le j_{c_2}] - \mathrm{P}[j_{c_1}+1 \le Y_i \le j_{b_1}]\Big)
\end{aligned}$$

Then, a slight simplification of the result of these two sums by combining the prefactors of identical probabilities leads to

$$\begin{aligned}
&\frac{a}{c-b}\left\{\left(c+\frac{\mu_i}{V^{1/2}(\mu_i)}\right)\mathrm{P}\left[j_{b_2}+1 \le Y_i \le j_{c_2}\right]\right. \\
&\qquad - \left(c-\frac{\mu_i}{V^{1/2}(\mu_i)}\right)\mathrm{P}\left[j_{c_1}+1 \le Y_i \le j_{b_1}\right] \\
&\qquad \left. - \frac{\mu_i}{V^{1/2}(\mu_i)}\left(\mathrm{P}\left[j_{b_2} \le \tilde{Y}_i \le j_{c_2}-1\right] + \mathrm{P}\left[j_{c_1} \le \tilde{Y}_i \le j_{b_1}-1\right]\right)\right\}
\end{aligned}$$

Combination of the partial results results in the statement. □

For the first expectation for the asymptotic variance (expectation (3.2)), it is:

Theorem 2 (Binomial distributed random variable: Asymptotic variance I)**.**

$$
\begin{aligned}
&\mathrm{E}\left[\psi^2_{Hampel}\left(\frac{Y_i-\mu_i}{V^{1/2}(\mu_i)}\right)\right]\\
&\quad= \frac{\mu_i^2}{V(\mu_i)}\,\mathrm{P}[j_{a_1}+1\le Y_i\le j_{a_2}] + \frac{m_i\,(m_i-1)\,p_i^2}{V(\mu_i)}\,\mathrm{P}\Big[j_{a_1}-1\le \tilde{\tilde{Y}}_i\le j_{a_2}-2\Big]\\
&\quad+\frac{(1-2\mu_i)\mu_i}{V(\mu_i)}\,\mathrm{P}\Big[j_{a_1}\le \tilde{Y}_i\le j_{a_2}-1\Big]\\
&\quad+a^2\left(\mathrm{P}[j_{a_2}+1\le Y_i\le j_{b_2}]+\mathrm{P}[j_{b_1}+1\le Y_i\le j_{a_1}]\right)\\
&\quad+\frac{a^2}{(c-b)^2}\left\{\left[c+\frac{\mu_i}{V^{1/2}(\mu_i)}\right]^2\mathrm{P}[j_{b_2}+1\le Y_i\le j_{c_2}]\right.\\
&\quad+\left[c-\frac{\mu_i}{V^{1/2}(\mu_i)}\right]^2\mathrm{P}[j_{c_1}+1\le Y_i\le j_{b_1}]\\
&\quad+\frac{m_i(m_i-1)\,p_i^2}{V(\mu_i)}\left(\mathrm{P}\Big[j_{b_2}-1\le \tilde{\tilde{Y}}_i\le j_{c_2}-2\Big]+\mathrm{P}\Big[j_{c_1}-1\le \tilde{\tilde{Y}}_i\le j_{b_1}-2\Big]\right)\\
&\quad-\left(\frac{2c\mu_i}{V^{1/2}(\mu_i)}-\frac{(1-2\mu_i)\mu_i}{V(\mu_i)}\right)\mathrm{P}\Big[j_{b_2}\le \tilde{Y}_i\le j_{c_2}-1\Big]\\
&\quad\left.+\left(\frac{2c\mu_i}{V^{1/2}(\mu_i)}+\frac{(1-2\mu_i)\mu_i}{V(\mu_i)}\right)\mathrm{P}\Big[j_{c_1}\le \tilde{Y}_i\le j_{b_1}-1\Big]\right\}
\end{aligned}
$$

Proof. The expectation (3.2) for the Hampel function equals

$$
\begin{aligned}
&\sum_{j=0}^{m_i}\left(\frac{j-\mu_i}{V^{1/2}(\mu_i)}\right)^2\mathrm{P}[Y_i=j]\;\mathbb{1}_A + a^2\sum_{j=0}^{m_i}\mathrm{P}[Y_i=j]\;(\mathbb{1}_{B_2}+\mathbb{1}_{B_1})\\
&\qquad+\frac{a^2}{(c-b)^2}\sum_{j=0}^{m_i}\left(c-\frac{|j-\mu_i|}{V^{1/2}(\mu_i)}\right)^2\mathrm{P}[Y_i=j]\;(\mathbb{1}_{C_2}+\mathbb{1}_{C_1})
\end{aligned}
\tag{3.10}
$$

To calculate the first term, one expands the Pearson residuals and split the sum then in a j-dependent and a j-independent part. Hence, calculation of the probability of Y_i having realisation j according to case A yields the second part equalling

$$
V^{-1}(\mu_i)\sum_{j=0}^{m_i}\left(j^2-2\mu_i j\right)\mathrm{P}[Y_i=j]\;\mathbb{1}_A+\frac{\mu_i^2}{V(\mu_i)}\,\mathrm{P}[j_{a_1}+1\le Y_i\le j_{a_2}]
\tag{3.11}
$$

Relying on the trick mentioned in equation (3.6), the sum of the first part of equation (3.11) equals

$$V^{-1}(\mu_i) \sum_{j=0}^{m_i} \left[j\,(j-1) + j\,(1-2\mu_i)\right] \mathrm{P}[Y_i = j]\ \mathbb{1}_A$$

which can be calculated to

$$V^{-1}(\mu_i) \left\{ \sum_{j=0}^{m_i} j\,(j-1)\,\mathrm{P}[Y_i = j]\ \mathbb{1}_A + (1-2\mu_i) \sum_{j=0}^{m_i} j\ \mathrm{P}[Y_i = j]\ \mathbb{1}_A \right\}$$

The first summand describes the probability of a random variable $\tilde{\tilde{Y}}_i$ with $m_i - 2$ trials and probability p_i to take the value j. The summand of the second sum corresponds again to the probability of a random variable $\tilde{Y}_i$ with $m_i - 1$ trials and probability p_i to take the value j. Considering that $m_i p_i$ equals μ_i, the sum of equation (3.11) can be derived as

$$m_i\,(m_i-1)\,p_i^2\ \mathrm{P}[j_{a_1} - 1 \le \tilde{\tilde{Y}}_i \le j_{a_2} - 2] + \mu_i(1-2\mu_i)\,\mathrm{P}[j_{a_1} \le \tilde{Y}_i \le j_{a_2} - 1] \quad (3.12)$$

Combination leads to the final result for the first term of equation (3.10) and this is

$$\begin{aligned} &\frac{\mu_i^2}{V(\mu_i)}\,\mathrm{P}[j_{a_1} + 1 \le Y_i \le j_{a_2}] + \frac{m_i\,(m_i-1)\,p_i^2}{V(\mu_i)}\,\mathrm{P}\Big[j_{a_1} - 1 \le \tilde{\tilde{Y}}_i \le j_{a_2} - 2\Big] \\ &\quad + \frac{(1-2\mu_i)\mu_i}{V(\mu_i)}\,\mathrm{P}\Big[j_{a_1} \le \tilde{Y}_i \le j_{a_2} - 1\Big] \end{aligned} \quad (3.13)$$

Direct derivation of the second term of equation (3.10) leads to

$$a^2\,(\mathrm{P}[j_{a_2} + 1 \le Y_i \le j_{b_2}] + \mathrm{P}[j_{b_1} + 1 \le Y_i \le j_{a_1}])$$

For the last term of equation (3.10), one splits this term into two sums, one for each case because of the absolute value of the residuals. So, this results in the reformulation

$$\begin{aligned} &\frac{a^2}{(c-b)^2} \left\{ \sum_{j=0}^{m_i} \left(c - \frac{j-\mu_i}{V^{1/2}(\mu_i)}\right)^2 \mathrm{P}[Y_i = j]\ \mathbb{1}_{C_2} \right. \\ &\qquad \left. + \sum_{j=0}^{m_i} \left(c + \frac{j-\mu_i}{V^{1/2}(\mu_i)}\right)^2 \mathrm{P}[Y_i = j]\ \mathbb{1}_{C_1} \right\} \end{aligned} \quad (3.14)$$

Only differing regarding their sign of the expression in the squared brackets, one gains the second expression directly from the first expression. After extending the squared expression of the first term, one splits the sum into three sums, i.e.

$$\frac{a^2}{(c-b)^2} \sum_{j=0}^{m_i} \left(\left[c + \frac{\mu_i}{V^{1/2}(\mu_i)} \right]^2 - \frac{2\,c}{V^{1/2}(\mu_i)}\, j + \frac{j\,(j-2\mu_i)}{V(\mu_i)} \right) \mathrm{P}[Y_i = j]\ \mathbb{1}_{C_2}$$

Based on equations (3.8) and (3.12), this expression equals

$$\begin{aligned} \frac{a^2}{(c-b)^2} \Bigg\{ & \left[c + \frac{\mu_i}{V^{1/2}(\mu_i)} \right]^2 \mathrm{P}[j_{b_2} + 1 \le Y_i \le j_{c_2}] - \frac{2c\mu_i}{V^{1/2}(\mu_i)}\, \mathrm{P}\Big[j_{b_2} \le \tilde{Y}_i \le j_{c_2} - 1 \Big] \\ & + \frac{m_i(m_i-1)\,p_i^2}{V(\mu_i)}\, \mathrm{P}\Big[j_{b_2} - 1 \le \tilde{\tilde{Y}}_i \le j_{c_2} - 2 \Big] \\ & + \frac{(1-2\mu_i)\mu_i}{V(\mu_i)}\, \mathrm{P}\Big[j_{b_2} \le \tilde{Y}_i \le j_{c_2} - 1 \Big] \Bigg\} \end{aligned}$$

In the end, one combines the partial results and simplifies somewhat the final result. This leads to

$$\begin{aligned} \frac{a^2}{(c-b)^2} \Bigg\{ & \left[c + \frac{\mu_i}{V^{1/2}(\mu_i)} \right]^2 \mathrm{P}[j_{b_2} + 1 \le Y_i \le j_{c_2}] \\ & - \left(\frac{2c\mu_i}{V^{1/2}(\mu_i)} - \frac{(1-2\mu_i)\mu_i}{V(\mu_i)} \right) \mathrm{P}\Big[j_{b_2} \le \tilde{Y}_i \le j_{c_2} - 1 \Big] \\ & + \frac{m_i(m_i-1)\,p_i^2}{V(\mu_i)}\ \mathrm{P}\Big[j_{b_2} - 1 \le \tilde{\tilde{Y}}_i \le j_{c_2} - 2 \Big] \Bigg\} \end{aligned}$$

Then, the second part of the last term of equation (3.14) equals

$$\begin{aligned} \frac{a^2}{(c-b)^2} \Bigg\{ & \left[c - \frac{\mu_i}{V^{1/2}(\mu_i)} \right]^2 \mathrm{P}[j_{c_1} + 1 \le Y_i \le j_{b_1}] \\ & + \left(\frac{2c\mu_i}{V^{1/2}(\mu_i)} + \frac{(1-2\mu_i)\mu_i}{V(\mu_i)} \right) \mathrm{P}\Big[j_{c_1} \le \tilde{Y}_i \le j_{b_1} - 1 \Big] \\ & + \frac{m_i(m_i-1)\,p_i^2}{V(\mu_i)}\ \mathrm{P}\Big[j_{c_1} - 1 \le \tilde{\tilde{Y}}_i \le j_{b_1} - 2 \Big] \Bigg\} \end{aligned}$$

Combination of the partial results results in the statement. □

Finally, for the second expectation for the asymptotic variance (expectation (3.3)), it holds:

Theorem 3 (Binomial distributed random variable: Asymptotic variance II)**.**

$$\begin{aligned}
&\mathrm{E}\left[\psi_{Hampel}\left(\frac{Y_i-\mu_i}{V^{1/2}(\mu_i)}\right)\frac{Y_i-\mu_i}{V(\mu_i)}\right]\\
&\quad=\frac{\mu_i^2}{V^{3/2}(\mu_i)}\,\mathrm{P}[j_{a_1}+1\le Y_i\le j_{a_2}]+\frac{m_i\,(m_i-1)\,p_i^2}{V^{3/2}(\mu_i)}\,\mathrm{P}\Big[j_{a_1}-1\le\tilde{\tilde{Y}}_i\le j_{a_2}-2\Big]\\
&\quad+\frac{(1-2\mu_i)\mu_i}{V^{3/2}(\mu_i)}\,\mathrm{P}\Big[j_{a_1}\le\tilde{Y}_i\le j_{a_2}-1\Big]\\
&\quad+\frac{a\,\mu_i}{V(\mu_i)}\Big(\mathrm{P}\Big[j_{a_2}\le\tilde{Y}_i\le j_{b_2}-1\Big]-\mathrm{P}[j_{a_2}+1\le Y_i\le j_{b_2}]\\
&\quad-\mathrm{P}\Big[j_{b_1}\le\tilde{Y}_i\le j_{a_1}-1\Big]+\mathrm{P}[j_{b_1}+1\le Y_i\le j_{a_1}]\Big)\\
&\quad+\frac{a\,c\,\mu_i}{(c-b)\,V(\mu_i)}\Big(\mathrm{P}\Big[j_{b_2}\le\tilde{Y}_i\le j_{c_2}-1\Big]-\mathrm{P}[j_{b_2}+1\le Y_i\le j_{c_2}]\\
&\quad-\mathrm{P}\Big[j_{c_1}\le\tilde{Y}_i\le j_{b_1}-1\Big]+\mathrm{P}[j_{c_1}+1\le Y_i\le j_{b_1}]\Big)\\
&\quad-\frac{a}{(c-b)\,V^{3/2}(\mu_i)}\Big\{\mu_i^2\,(\mathrm{P}[j_{b_2}+1\le Y_i\le j_{c_2}]+\mathrm{P}[j_{c_1}+1\le Y_i\le j_{b_1}])\\
&\quad+m_i\,(m_i-1)\,p_i^2\Big(\mathrm{P}\Big[j_{b_2}-1\le\tilde{\tilde{Y}}_i\le j_{c_2}-2\Big]+\mathrm{P}\Big[j_{c_1}-1\le\tilde{\tilde{Y}}_i\le j_{b_1}-2\Big]\Big)\\
&\quad+(1-2\mu_i)\mu_i\Big(\mathrm{P}\Big[j_{b_2}\le\tilde{Y}_i\le j_{c_2}-1\Big]+\mathrm{P}\Big[j_{c_1}\le\tilde{Y}_i\le j_{b_1}-1\Big]\Big)\Big\}
\end{aligned}$$

Proof. Expectation (3.3) for the Hampel function equals

$$\begin{aligned}
&\sum_{j=0}^{m_i}\frac{(j-\mu_i)^2}{V^{3/2}(\mu_i)}\,\mathrm{P}\left[Y_i=j\right]\mathbb{1}_A+\frac{a}{V(\mu_i)}\sum_{j=0}^{m_i}(j-\mu_i)\,\mathrm{P}\left[Y_i=j\right](\mathbb{1}_{B_2}-\mathbb{1}_{B_1})\\
&\quad+\frac{a}{c-b}\sum_{j=0}^{m_i}\left(c-\left|\frac{j-\mu_i}{V^{1/2}(\mu_i)}\right|\right)\frac{j-\mu_i}{V(\mu_i)}\,\mathrm{P}\left[Y_i=j\right](\mathbb{1}_{C_2}-\mathbb{1}_{C_1})
\end{aligned}\tag{3.15}$$

One directly deduces the first two terms from equations (3.13) and (3.9). One also derives the third term from these two equations after reformulation of this term, this implies splitting the sum into two sums. Hence, splitting the third term leads to

$$\begin{aligned}
&\frac{a\,c}{(c-b)\,V(\mu_i)} \sum_{j=0}^{m_i} (j-\mu_i)\,\mathrm{P}\left[Y_i=j\right]\left(\mathbb{1}_{C_2}-\mathbb{1}_{C_1}\right) \\
&\qquad - \frac{a}{(c-b)\,V^{3/2}(\mu_i)} \sum_{j=0}^{m_i} (j-\mu_i)^2\,\mathrm{P}\left[Y_i=j\right]\left(\mathbb{1}_{C_2}+\mathbb{1}_{C_1}\right)
\end{aligned}$$

Application of equations (3.9) and (3.13) results in

$$\begin{aligned}
&\frac{a\,c\,\mu_i}{(c-b)\,V(\mu_i)} \Big(\mathrm{P}\Big[j_{b_2} \le \tilde{Y}_i \le j_{c_2}-1\Big] - \mathrm{P}[j_{b_2}+1 \le Y_i \le j_{c_2}] \\
&\quad -\mathrm{P}\Big[j_{c_1} \le \tilde{Y}_i \le j_{b_1}-1\Big] + \mathrm{P}[j_{c_1}+1 \le Y_i \le j_{b_1}]\Big) \\
&\quad - \frac{a}{(c-b)\,V^{3/2}(\mu_i)} \Big\{\mu_i^2\left(\mathrm{P}[j_{b_2}+1 \le Y_i \le j_{c_2}] + \mathrm{P}[j_{c_1}+1 \le Y_i \le j_{b_1}]\right) \\
&\quad + m_i\,(m_i-1)\,p_i^2\,\Big(\mathrm{P}\Big[j_{b_2}-1 \le \tilde{\tilde{Y}}_i \le j_{c_2}-2\Big] + \mathrm{P}\Big[j_{c_1}-1 \le \tilde{\tilde{Y}}_i \le j_{b_1}-2\Big]\Big) \\
&\quad +(1-2\mu_i)\mu_i\,\Big(\mathrm{P}\Big[j_{b_2} \le \tilde{Y}_i \le j_{c_2}-1\Big] + \mathrm{P}\Big[j_{c_1} \le \tilde{Y}_i \le j_{b_1}-1\Big]\Big)\Big\}
\end{aligned}$$

Combination of the partial results results in the statement. □

3.1.2 Poisson distributed random variable

For the Poisson distribution, one performs the same calculations as for the binomial distribution. One derives the sums in a different manner caused by the different definition of the probability function. So, assume $Y_i \sim \mathrm{Poi}(\mu_i)$ with $\mathrm{E}[Y_i] = \mathrm{V}[Y_i] = \mu_i$. Although expectation and variance are equal, they will not be cancelled during the calculations. Note for a realisation j of Y_i that

$$j\ \mathrm{P}\left[Y_i=j\right] = \mu_i \cdot \mathrm{P}\left[Y_i=j-1\right] \tag{3.16}$$

and

$$j(j-1)\ \mathrm{P}\left[Y_i=j\right] = \mu_i^2 \cdot \mathrm{P}\left[Y_i=j-2\right] \tag{3.17}$$

Then, it holds for expectation (3.1) to correct for Fisher consistency:

Theorem 4 (Poisson distributed random variable: Fisher consistency correction)**.**

$$\begin{aligned}
&\mathrm{E}\left[\psi_{\mathit{Hampel}}\left(\frac{Y_i-\mu_i}{V^{1/2}(\mu_i)}\right)\right] \\
&= \frac{\mu_i}{V^{1/2}(\mu_i)}\left\{\mathrm{P}[Y_i=j_{a_1}]-\mathrm{P}[Y_i=j_{a_2}]-\frac{a}{c-b}\left(\mathrm{P}[Y_i=j_{b_2}]-\mathrm{P}[Y_i=j_{c_2}]\right.\right. \\
&\quad \left.\left.+\mathrm{P}[Y_i=j_{c_1}]-\mathrm{P}[Y_i=j_{b_1}]\right)\right\} \\
&\quad + a\left\{\mathrm{P}[j_{a_2}+1\le Y_i\le j_{b_2}]-\mathrm{P}[j_{b_1}+1\le Y_i\le j_{a_1}]\right. \\
&\quad \left.+\frac{c}{c-b}\left(\mathrm{P}[j_{b_2}+1\le Y_i\le j_{c_2}]-\mathrm{P}[j_{c_1}+1\le Y_i\le j_{b_1}]\right)\right\}
\end{aligned}$$

Proof. Expectation (3.1) for the Hampel function equals

$$\begin{aligned}
&\sum_{j=j_{a_1}+1}^{j_{a_2}}\frac{j-\mu_i}{V^{1/2}(\mu_i)}\mathrm{P}[Y_i=j]+a\left(\sum_{j=j_{a_2}+1}^{j_{b_2}}\mathrm{P}[Y_i=j]-\sum_{j=j_{b_1}+1}^{j_{a_1}}\mathrm{P}[Y_i=j]\right) \\
&\quad+\frac{a}{c-b}\left\{\sum_{j=j_{b_2}+1}^{j_{c_2}}\left(c-\frac{j-\mu_i}{V^{1/2}(\mu_i)}\right)\mathrm{P}[Y_i=j]\right. \qquad (3.18) \\
&\quad\left.-\sum_{j=j_{c_1}+1}^{j_{b_1}}\left(c+\frac{j-\mu_i}{V^{1/2}(\mu_i)}\right)\mathrm{P}[Y_i=j]\right\}
\end{aligned}$$

Splitting the first sum results in

$$V^{-1/2}(\mu_i)\left(\sum_{j=j_{a_1}+1}^{j_{a_2}} j\,\mathrm{P}[Y_i=j]-\sum_{j=j_{a_1}+1}^{j_{a_2}}\mu_i\,\mathrm{P}[Y_i=j]\right)$$

Settling the probability with the sum index according to equation (3.16) compensated by adjusting the sum indices leads to

$$V^{-1/2}(\mu_i)\left(\mu_i\sum_{j=j_1}^{j_2-1}\mathrm{P}[Y_i=j]-\mu_i\sum_{j=j_1+1}^{j_2}\mathrm{P}[Y_i=j]\right)$$

Allocation of the two sums with respect to their indices results in

$$\frac{\mu_i}{V^{1/2}(\mu_i)} \left(\mathrm{P}[Y_i = j_{a_1}] - \mathrm{P}[Y_i = j_{a_2}]\right) \tag{3.19}$$

Direct derivation of the second part of equation (3.18) leads to

$$a \left(\mathrm{P}[j_{a_2} + 1 \leq Y_i \leq j_{b_2}] - \mathrm{P}[j_{b_1} + 1 \leq Y_i \leq j_{a_1}]\right)$$

Excluding $\frac{a}{c-b}$ from the third part of equation (3.18) and splitting its first sum into two sums lead to a sum of probabilities of Y_i having a realisation in the interval $[j_{b_2} + 1,\, j_{c_2}]$ and a second sum similar to the first sum of equation (3.18). Hence, the first sum of the third part equals

$$\frac{a}{c-b} \left\{ c\,\mathrm{P}[j_{b_2} + 1 \leq Y_i \leq j_{c_2}] - \frac{\mu_i}{V^{1/2}(\mu_i)} \left(\mathrm{P}[Y_i = j_{b_2}] - \mathrm{P}[Y_i = j_{c_2}]\right) \right\}$$

With respect to different signs of the two sums of the third part, it follows for its second sum by adjusting for these signs

$$-\frac{a}{c-b} \left\{ c\,\mathrm{P}[j_{c_1} + 1 \leq Y_i \leq j_{b_1}] + \frac{\mu_i}{V^{1/2}(\mu_i)} \left(\mathrm{P}[Y_i = j_{c_1}] - \mathrm{P}[Y_i = j_{b_1}]\right) \right\}$$

So, the calculation of equation (3.18) results in

$$\begin{aligned}
&\frac{\mu_i}{V^{1/2}(\mu_i)} \left(\mathrm{P}[Y_i = j_{a_1}] - \mathrm{P}[Y_i = j_{a_2}]\right) \\
&\quad + a \left(\mathrm{P}[j_{a_2} + 1 \leq Y_i \leq j_{b_2}] - \mathrm{P}[j_{b_1} + 1 \leq Y_i \leq j_{a_1}]\right) \\
&\quad + \frac{a}{c-b} \left\{ c\,\mathrm{P}[j_{b_2} + 1 \leq Y_i \leq j_{c_2}] - \frac{\mu_i}{V^{1/2}(\mu_i)} \left(\mathrm{P}[Y_i = j_{b_2}] - \mathrm{P}[Y_i = j_{c_2}]\right) \right\} \\
&\quad - \frac{a}{c-b} \left\{ c\,\mathrm{P}[j_{c_1} + 1 \leq Y_i \leq j_{b_1}] + \frac{\mu_i}{V^{1/2}(\mu_i)} \left(\mathrm{P}[Y_i = j_{c_1}] - \mathrm{P}[Y_i = j_{b_1}]\right) \right\}
\end{aligned}$$

Combination of the partial results leads to the statement. □

For the first expectation for the asymptotic variance (expectation (3.2)), it is

Theorem 5 (Poisson distributed random variable: Asymptotic variance I)**.**

$$\begin{aligned}
&\mathrm{E}\left[\psi^2_{\mathit{Hampel}}\left(\frac{Y_i-\mu_i}{\mu_i}\right)\right] \\
&= \frac{\mu_i^2}{V(\mu_i)}\Big\{\mathrm{P}[Y_i=j_{a_1}-1]-\mathrm{P}[Y_i=j_{a_2}-1]-\mathrm{P}[Y_i=j_{a_1}]+\mathrm{P}[Y_i=j_{a_2}] \\
&\quad+\frac{a^2}{(c-b)^2}\left(\mathrm{P}[Y_i=j_{b_2}-1]-\mathrm{P}[Y_i=j_{c_2}-1]-\mathrm{P}[Y_i=j_{b_2}]+\mathrm{P}[Y_i=j_{c_2}]\right. \\
&\quad\left.+\mathrm{P}[Y_i=j_{c_1}-1]-\mathrm{P}[Y_i=j_{b_1}-1]-\mathrm{P}[Y_i=j_{c_1}]+\mathrm{P}[Y_i=j_{b_1}]\right)\Big\} \\
&\quad+\frac{\mu_i}{V(\mu_i)}\Big[\mathrm{P}[j_{a_1}\le Y_i\le j_{a_2}-1] \\
&\quad+\frac{a^2}{(c-b)^2}\left(\mathrm{P}[j_{a_1}\le Y_i\le j_{a_2}-1]+\mathrm{P}[j_{c_1}\le Y_i\le j_{b_1}-1]\right)\Big] \\
&\quad+a^2\Big[\mathrm{P}[j_{a_2}+1\le Y_i\le j_{b_2}]+\mathrm{P}[j_{b_1}+1\le Y_i\le j_{a_1}] \\
&\quad+\frac{c^2}{(c-b)^2}\left(\mathrm{P}[j_{b_2}+1\le Y_i\le j_{c_2}]+\mathrm{P}[j_{c_1}+1\le Y_i\le j_{b_1}]\right)\Big] \\
&\quad-\frac{2c\mu_i}{V^{1/2}(\mu_i)}\frac{a^2}{(c-b)^2}\left(\mathrm{P}[Y_i=j_{b_2}]-\mathrm{P}[Y_i=j_{c_2}]-\mathrm{P}[Y_i=j_{c_1}]+\mathrm{P}[Y_i=j_{b_1}]\right)
\end{aligned}$$

Proof. Expectation (3.2) for the Hampel function equals

$$\begin{aligned}
&\sum_{j=j_{a_1}+1}^{j_{a_2}}\frac{(j-\mu_i)^2}{V(\mu_i)}\mathrm{P}[Y_i=j]+a^2\left(\sum_{j=j_{a_2}+1}^{j_{b_2}}\mathrm{P}[Y_i=j]+\sum_{j=j_{b_1}+1}^{j_{a_1}}\mathrm{P}[Y_i=j]\right) \\
&\quad+\frac{a^2}{(c-b)^2}\left\{\sum_{j=j_{b_2}+1}^{j_{c_2}}\left(c-\frac{j-\mu_i}{V^{1/2}(\mu_i)}\right)^2\mathrm{P}[Y_i=j]\right. \qquad (3.20)\\
&\quad\left.+\sum_{j=j_{c_1}+1}^{j_{b_1}}\left(c+\frac{j-\mu_i}{V^{1/2}(\mu_i)}\right)^2\mathrm{P}[Y_i=j]\right\}
\end{aligned}$$

The calculations for this expression will be performed separately. One splits the first sum as usual into three sums which leads to

$$V^{-1}(\mu_i)\left\{\sum_{j=j_{a_1}+1}^{j_{a_2}} j(j-1)\,\mathrm{P}[Y_i=j]+(1-2\mu_i)\sum_{j=j_{a_1}+1}^{j_{a_2}} j\,\mathrm{P}[Y_i=j]\right.$$
$$\left.+\mu_i^2\sum_{j=j_{a_1}+1}^{j_{a_2}}\mathrm{P}[Y_i=j]\right\}$$

Using equations (3.16) and (3.17), it follows

$$V^{-1}(\mu_i)\left\{\mu_i^2\,\mathrm{P}[j_{a_1}-1\le Y_i\le j_{a_2}-2]+(1-2\mu_i)\mu_i\,\mathrm{P}[j_{a_1}\le Y_i\le j_{a_2}-1]\right.$$
$$\left.+\mu_i^2\,\mathrm{P}[j_{a_1}+1\le Y_i\le j_{a_2}]\right\}$$

and equals rewritten

$$\frac{\mu_i^2}{V(\mu_i)}\left\{\mathrm{P}[j_{a_1}-1\le Y_i\le j_{a_2}-2]-2\,\mathrm{P}[j_{a_1}\le Y_i\le j_{a_2}-1]+\mathrm{P}[j_{a_1}+1\le Y_i\le j_{a_2}]\right\}$$
$$+\frac{\mu_i}{V(\mu_i)}\,\mathrm{P}[j_{a_1}\le Y_i\le j_{a_2}-1] \tag{3.21}$$

One allocates probabilities with a coefficient of μ_i^2 to each other so that there will not be anymore a probability for a realisation within an interval but equalling a specific value because

$$\mathrm{P}[j_{a_1}-1\le Y_i\le j_{a_2}-2]-2\ \mathrm{P}[j_{a_1}\le Y_i\le j_{a_2}-1]+\mathrm{P}[j_{a_1}+1\le Y_i\le j_{a_2}]$$

equals

$$(\mathrm{P}[j_{a_1}-1\le Y_i\le j_{a_2}-2]-\mathrm{P}[j_{a_1}\le Y_i\le j_{a_2}-1])$$
$$+(\mathrm{P}[j_{a_1}+1\le Y_i\le j_{a_2}]-\mathrm{P}[j_{a_1}\le Y_i\le j_{a_2}-1])$$

Writing the probabilities as finite sums leads to

$$\left(\sum_{j=j_{a_1}-1}^{j_{a_2}-2}\mathrm{P}[Y_i=j]-\sum_{j=j_{a_1}}^{j_{a_2}-1}\mathrm{P}[Y_i=j]\right)+\left(\sum_{j=j_{a_1}+1}^{j_{a_2}}\mathrm{P}[Y_i=j]-\sum_{j=j_{a_1}}^{j_{a_2}-1}\mathrm{P}[Y_i=j]\right)$$

Allocation results in

$$(\mathrm{P}[Y_i = j_{a_1} - 1] - \mathrm{P}[Y_i = j_{a_2} - 1]) + (\mathrm{P}[Y_i = j_{a_2}] - \mathrm{P}[Y_i = j_{a_1}]) \tag{3.22}$$

Hence, one simplifies expression (3.21) to

$$\begin{aligned} &\frac{\mu_i^2}{V(\mu_i)} \{\mathrm{P}[Y_i = j_{a_1} - 1] - \mathrm{P}[Y_i = j_{a_2} - 1] - \mathrm{P}[Y_i = j_{a_1}] + \mathrm{P}[Y_i = j_{a_2}]\} \\ &\quad + \frac{\mu_i}{V(\mu_i)} \mathrm{P}[j_{a_1} \le Y_i \le j_{a_2} - 1] \end{aligned} \tag{3.23}$$

The sums between the round brackets of equation (3.20) directly equal

$$a^2 \left(\mathrm{P}[j_{a_2} + 1 \le Y_i \le j_{b_2}] + \mathrm{P}[j_{b_1} + 1 \le Y_i \le j_{a_1}]\right)$$

The sums in the curly brackets of equation (3.20) only differ in the sign. So, one directly deduces this second sum from this first sum. This first sum will be calculated now. One splits the sum into three sums leading to

$$\begin{aligned} \frac{a^2}{(c-b)^2} \Bigg\{ & c^2 \sum_{j=j_{b_2}+1}^{j_{c_2}} \mathrm{P}[Y_i = j] - 2c \sum_{j=j_{b_2}+1}^{j_{c_2}} \frac{j - \mu_i}{V^{1/2}(\mu_i)} \mathrm{P}[Y_i = j] \\ & + \sum_{j=j_{b_2}+1}^{j_{c_2}} \frac{(j-\mu_i)^2}{V(\mu_i)} \mathrm{P}[Y_i = j] \Bigg\} \end{aligned}$$

One either directly calculates or deduces these sums from previous calculations. The first sum equals

$$c^2 \, \mathrm{P}[j_{b_2} + 1 \le Y_i \le j_{c_2}]$$

Using the results of the first part of equation (3.18) for the second sum, it follows

$$-2c \frac{\mu_i}{V^{1/2}(\mu_i)} (\mathrm{P}[Y_i = j_{b_2}] - \mathrm{P}[Y_i = j_{c_2}])$$

Deduction of the third sum from the expression of the first line of equation (3.20) leads to

$$\begin{aligned} &\frac{\mu_i^2}{V(\mu_i)} (\mathrm{P}[Y_i = j_{b_2} - 1] - \mathrm{P}[Y_i = j_{c_2} - 1] - \mathrm{P}[Y_i = j_{b_2}] + \mathrm{P}[Y_i = j_{c_2}]) \\ &\quad + \frac{\mu_i}{V(\mu_i)} \mathrm{P}[j_{b_2} \le Y_i \le j_{c_2} - 1] \end{aligned}$$

Combination of these results leads to the value of the forth sum of equation (3.20) and this is

$$\begin{aligned}\frac{a^2}{(c-b)^2}\Bigg\{ & c^2\,\mathrm{P}[j_{b_2}+1 \le Y_i \le j_{c_2}] - 2c\frac{\mu_i}{V^{1/2}(\mu_i)}\left(\mathrm{P}[Y_i = j_{b_2}] - \mathrm{P}[Y_i = j_{c_2}]\right)\\ & + \frac{\mu_i^2}{V(\mu_i)}\left(\mathrm{P}[Y_i = j_{b_2}-1] - \mathrm{P}[Y_i = j_{c_2}-1] - \mathrm{P}[Y_i = j_{b_2}] + \mathrm{P}[Y_i = j_{c_2}]\right)\\ & + \frac{\mu_i}{V(\mu_i)}\,\mathrm{P}[j_{b_2} \le Y_i \le j_{c_2}-1]\Bigg\}\end{aligned}$$

One deduces the last sum of equation (3.20) from the previously calculated sum. Then, this sum equals

$$\begin{aligned}\frac{a^2}{(c-b)^2}\Bigg\{ & c^2\,\mathrm{P}[j_{c_1}+1 \le Y_i \le j_{b_1}] + 2c\frac{\mu_i}{V^{1/2}(\mu_i)}\left(\mathrm{P}[Y_i = j_{c_1}] - \mathrm{P}[Y_i = j_{b_1}]\right)\\ & + \frac{\mu_i^2}{V(\mu_i)}\left(\mathrm{P}[Y_i = j_{c_1}-1] - \mathrm{P}[Y_i = j_{b_1}-1] - \mathrm{P}[Y_i = j_{c_1}] + \mathrm{P}[Y_i = j_{b_1}]\right)\\ & + \frac{\mu_i}{V(\mu_i)}\,\mathrm{P}[j_{c_1} \le Y_i \le j_{b_1}-1]\Bigg\}\end{aligned}$$

Summing up the partial results and simplification lead to the statement. □

For the second expectation for the asymptotic variance (expectation (3.3)), it holds:

Theorem 6 (Poisson distributed random variable: Asymptotic variance II)**.**

$$
\begin{aligned}
\mathrm{E}&\left[\psi_{\mathit{Hampel}}\left(\frac{Y_i-\mu_i}{V^{1/2}(\mu_i)}\right)\frac{Y_i-\mu_i}{V(\mu_i)}\right]\\
&=\frac{\mu_i^2}{V^{3/2}(\mu_i)}\Big\{\mathrm{P}[Y_i=j_{a_1}-1]-\mathrm{P}[Y_i=j_{a_2}-1]-\mathrm{P}[Y_i=j_{a_1}]+\mathrm{P}[Y_i=j_{a_2}]\\
&\quad-\frac{a}{c-b}\left(\mathrm{P}[Y_i=j_{b_2}-1]-\mathrm{P}[Y_i=j_{c_2}-1]-\mathrm{P}[Y_i=j_{b_2}]+\mathrm{P}[Y_i=j_{c_2}]\right.\\
&\quad\left.+\mathrm{P}[Y_i=j_{c_1}-1]-\mathrm{P}[Y_i=j_{b_1}-1]-\mathrm{P}[Y_i=j_{c_1}]+\mathrm{P}[Y_i=j_{b_1}]\right)\Big\}\\
&\quad+\frac{\mu_i}{V^{3/2}(\mu_i)}\Big\{\mathrm{P}[j_{a_1}\le Y_i\le j_{a_2}-1]\\
&\quad-\frac{a}{c-b}\left(\mathrm{P}[j_{b_2}\le Y_i\le j_{c_2}-1]+\mathrm{P}[j_{c_1}\le Y_i\le j_{b_1}-1]\right)\Big\}\\
&\quad+\frac{a\mu_i}{V(\mu_i)}\Big\{\mathrm{P}[Y_i=j_{a_2}]-\mathrm{P}[Y_i=j_{b_2}]-\mathrm{P}[Y_i=j_{b_1}]+\mathrm{P}[Y_i=j_{a_1}]\\
&\quad+\frac{c}{c-b}\left(\mathrm{P}[Y_i=j_{b_2}]-\mathrm{P}[Y_i=j_{c_2}]-\mathrm{P}[Y_i=j_{c_1}]+\mathrm{P}[Y_i=j_{b_1}]\right)\Big\}
\end{aligned}
$$

Proof. Expectation (3.3) for the Hampel function equals

$$
\begin{aligned}
&\sum_{j=j_{a_1}+1}^{j_{a_2}}\frac{(j-\mu_i)^2}{V^{3/2}(\mu_i)}\,\mathrm{P}[Y_i=j]\\
&\quad+\frac{a}{V(\mu_i)}\left(\sum_{j=j_{a_2}+1}^{j_{b_2}}(j-\mu_i)\,\mathrm{P}[Y_i=j]-\sum_{j=j_{b_1}+1}^{j_{a_1}}(j-\mu_i)\,\mathrm{P}[Y_i=j]\right)\\
&\quad+\frac{a}{(c-b)V(\mu_i)}\left(\sum_{j=j_{b_2}+1}^{j_{c_2}}\left(c-\frac{j-\mu_i}{V^{1/2}(\mu_i)}\right)(j-\mu_i)\,\mathrm{P}[Y_i=j]\right.\\
&\qquad\left.-\sum_{j=j_{c_1}+1}^{j_{b_1}}\left(c+\frac{j-\mu_i}{V^{1/2}(\mu_i)}\right)(j-\mu_i)\,\mathrm{P}[Y_i=j]\right)
\end{aligned}
\tag{3.24}
$$

The first sum equals (cf. equation (3.23))

$$
\begin{aligned}
&\frac{\mu_i^2}{V^{3/2}(\mu_i)}\left\{\mathrm{P}[Y_i=j_{a_1}-1]-\mathrm{P}[Y_i=j_{a_2}-1]-\mathrm{P}[Y_i=j_{a_1}]+\mathrm{P}[Y_i=j_{a_2}]\right\}\\
&\quad+\frac{\mu_i}{V^{3/2}(\mu_i)}\,\mathrm{P}[j_{a_1}\le Y_i\le j_{a_2}-1]
\end{aligned}
$$

and one rewrites the next two sums deduced from equation (3.19) as

$$\frac{a\mu_i}{V(\mu_i)}\left(\mathrm{P}[Y_i = j_{a_2}] - \mathrm{P}[Y_i = j_{b_2}] - \mathrm{P}[Y_i = j_{b_1}] + \mathrm{P}[Y_i = j_{a_1}]\right)$$

Reordering the last two sums of equation (3.24) leads to

$$\frac{a}{(c-b)V(\mu_i)}\left\{c\left(\sum_{j=j_{b_2}+1}^{j_{c_2}}(j-\mu_i)\,\mathrm{P}[Y_i=j] - \sum_{j=j_{c_1}+1}^{j_{b_1}}(j-\mu_i)\,\mathrm{P}[Y_i=j]\right)\right.$$
$$\left. -V^{-1/2}(\mu_i)\left(\sum_{j=j_{b_2}+1}^{j_{c_2}}(j-\mu_i)^2\,\mathrm{P}[Y_i=j] + \sum_{j=j_{c_1}+1}^{j_{b_1}}(j-\mu_i)^2\,\mathrm{P}[Y_i=j]\right)\right\}$$

Using the results of the first expression in equations (3.18) and (3.20) now, the desired expression is equal to

$$\frac{a}{(c-b)V(\mu_i)}\left\{c\mu_i\left(\mathrm{P}[Y_i = j_{b_2}] - \mathrm{P}[Y_i = j_{c_2}] - \mathrm{P}[Y_i = j_{c_1}] + \mathrm{P}[Y_i = j_{b_1}]\right)\right.$$
$$-\left[\frac{\mu_i^2}{V^{1/2}(\mu_i)}\left(\mathrm{P}[Y_i = j_{b_2}-1] - \mathrm{P}[Y_i = j_{c_2}-1] - \mathrm{P}[Y_i = j_{b_2}]\right.\right.$$
$$\left. + \mathrm{P}[Y_i = j_{c_2}] + \mathrm{P}[Y_i = j_{c_1}-1] - \mathrm{P}[Y_i = j_{b_1}-1] - \mathrm{P}[Y_i = j_{c_1}] + \mathrm{P}[Y_i = j_{b_1}]\right)$$
$$\left.\left. + \frac{\mu_i}{V^{1/2}(\mu_i)}\left(\mathrm{P}[j_{b_2} \le Y_i \le j_{c_2}-1] + \mathrm{P}[j_{c_1} \le Y_i \le j_{b_1}-1]\right)\right]\right\}$$

Combination of the partial results and simplification lead to the statement. □

3.2 Implementation in R

After the theoretical preparation of the expressions required for implementing the Hampel function, the topic of this section is the practical implementation in R exactly orientated on the `glmrobMqle.R` file of the package `robustbase` version 0.9-8 (Date: 14/06/2013). First, general adjustments are given in section 3.2.1 followed by the sections concerning the changes needed for binomial and Poisson distributed random variables (sections 3.2.2 and 3.2.3). One has to place all code sections exactly where they can be found in the original file for the Huber function. Afterwards, one must install the package as a whole under a modified name. For further instructions see supplemental chapter D on page 153.

3.2.1 General adjustments

In all calculated expectations, there are probabilities to derive. These probabilities always depend on the limits given by the definition of the Hampel function. These values are the same for both distributions. Hence, one can define them earlier. The values for the Hampel function are given in R language by

```
Ha <- floor(mu*ni-tcc[1]*sni*sV); Ka <- floor(mu*ni+tcc[1]*sni*sV);
Hb <- floor(mu*ni-tcc[2]*sni*sV); Kb <- floor(mu*ni+tcc[2]*sni*sV);
Hc <- floor(mu*ni-tcc[3]*sni*sV); Kc <- floor(mu*ni+tcc[3]*sni*sV);
```

The aim of Cantoni and Ronchetti (2001) was to solve the estimation equation (1.6) which was

$$0 = \sum_{i=1}^{n} [\nu(y_i, \mu_i) w(x_i) \mu_i' - \alpha(\beta)]$$

In comparing the sum of this equation and the α function by substituting the α function in the estimation equation, i.e.

$$0 = \sum_{i=1}^{n} \left[\nu(y_i, \mu_i) w(x_i) \mu_i' - \frac{1}{n} \sum_{j=1}^{n} E[\nu(y_j, \mu_j)] w(x_j) \mu_j' \right]$$

one observes that most of their arguments are the same. The second summand is independent of i and hence the right side of this equation equals

$$\sum_{i=1}^{n} \nu(y_i, \mu_i) w(x_i) \mu_i' - n \frac{1}{n} \sum_{j=1}^{n} E[\nu(y_j, \mu_j)] w(x_j) \mu_j'$$

which is the same as

$$\sum_{i=1}^{n} \left(\nu(y_i, \mu_i) - E[\nu(y_i, \mu_i)]\right) w(x_i) \mu_i'$$

$\nu(y_i, \mu_i)$ is equal to the ψ function divided by the square root of the variance of μ_i and so this expression equals

$$\sum_{i=1}^{n} \left(\psi(y_i, \mu_i) - E[\psi(y_i, \mu_i)]\right) \frac{w(x_i) \mu_i'}{V^{1/2}(\mu_i)}$$

Hence, the only part to be adjusted for the Hampel function is the expression $\psi(y_i, \mu_i) - E[\psi(y_i, \mu_i)]$. The expectation is different for binomial and Poisson distribution and will be implemented later on in the adequate sections. But nevertheless, the calculations can be prepared by defining the difference using for the expectation an `expression` that will be called up in the programme. So, call this difference cpsi and write it in R as

```
cpsi <-
  ifelse((0<=abs(residPS)) & (abs(residPS)<tcc[1]), residPS,
    ifelse((tcc[1]<=abs(residPS)) & (abs(residPS)<tcc[2]),
        sign(residPS)*tcc[1],
      ifelse((tcc[2]<=abs(residPS)) & (abs(residPS)<tcc[3]),
        sign(residPS)*tcc[1]*(tcc[3]-abs(residPS))/(tcc[3]-tcc[2]),
        0))) - eval(Epsi)
```

where `Epsi` indicates the expectation of $\mathrm{E}\left[\psi_{\text{Hampel}}\left(r_i\right)\right]$ with Pearson residual r_i.

The last general step will be the adjustment of the residual weights w_r. Therefore, consider that these weights are given by $\psi_{\text{Hampel}}(r)/r$ (Fox and Weisberg, 2011). Thus,

$$w_r = \begin{cases} 1 & \text{if } |r| < a \\ \frac{a}{|r|} & \text{if } a \leq |r| < b \\ \frac{(c-|r|)\, a}{(c-b)\, |r|} & \text{if } b \leq |r| < c \\ 0 & \text{otherwise} \end{cases}$$

Hence, the modified R code is

```
w.r <- ifelse(abs(residPS)<tcc[1], 1,
  ifelse((tcc[1]<=abs(residPS)) & (abs(residPS)<tcc[2]),
      tcc[1]/abs(residPS),
    ifelse((tcc[2]<=abs(residPS)) & (abs(residPS)<tcc[3]),
        (tcc[3]-abs(residPS))/(tcc[3]-tcc[2])*tcc[1]/abs(residPS),
      0)))
```

3.2.2 Adjustments for a binomial distributed random variable

The binomial distribution-specific adjustments are the three needed expectations. First, define the probabilities needed for the expectations as

```
EpsiBin.init <- expression({
  # P[Y_i <= j_z2]
  pKa <- pbinom(Ka, ni, mu); pKb <- pbinom(Kb, ni, mu);
  pKc <- pbinom(Kc, ni, mu)
  # P[Y_i <= j_z1]
  pHa <- pbinom(Ha, ni, mu); pHb <- pbinom(Hb, ni, mu);
  pHc <- pbinom(Hc, ni, mu)
  # P[Y_i~ <= j_z2-1]
  pKam1 <- pbinom(Ka-1, pmax.int(0, ni-1), mu)
  pKbm1 <- pbinom(Kb-1, pmax.int(0, ni-1), mu)
  pKcm1 <- pbinom(Kc-1, pmax.int(0, ni-1), mu)
  # P[Y_i~ <= j_z1-1]
  pHam1 <- pbinom(Ha-1, pmax.int(0, ni-1), mu)
  pHbm1 <- pbinom(Hb-1, pmax.int(0, ni-1), mu)
  pHcm1 <- pbinom(Hc-1, pmax.int(0, ni-1), mu)
  # P[Y_i~~ <= j_z2-2]
  pKam2 <- pbinom(Ka-2, pmax.int(0, ni-2), mu)
  pKbm2 <- pbinom(Kb-2, pmax.int(0, ni-2), mu)
  pKcm2 <- pbinom(Kc-2, pmax.int(0, ni-2), mu)
  # P[Y_i~~ <= j_z1-2]
  pHam2 <- pbinom(Ha-2, pmax.int(0, ni-2), mu)
  pHbm2 <- pbinom(Hb-2, pmax.int(0, ni-2), mu)
  pHcm2 <- pbinom(Hc-2, pmax.int(0, ni-2), mu)
})
```

Write the expectation of the Hampel function with respect to the Pearson residuals for the Fisher consistency correction as

```
EpsiBin <- expression({
  tcc[1]*(pKb-pKa-pHa+pHb) +
  tcc[1]*tcc[3]/(tcc[3]-tcc[2])*(pKc-pKb-pHb+pHc) +
  (pKam1-pHam1-pKa+pHa)*mu*sni/sV -
  tcc[1]/(tcc[3]-tcc[2])*((pKcm1-pKbm1-pKc+pKb)*mu*sni/sV +
    (pHbm1-pHcm1-pHb+pHc)*mu*sni/sV)
})
```

The first expectation for the asymptotic variance is the expectation of the squared Hampel function with respect to the Pearson residuals. This is equal to

```
Epsi2Bin <- expression({
  tcc[1]*tcc[1]*(pKb-pKa+pHa-pHb) +
  tcc[1]*tcc[1]*tcc[3]*tcc[3]/(tcc[3]-tcc[2])/(tcc[3]-tcc[2])*
    (pKc-pKb+pHb-pHc) +
  mu*mu*ni/Vmu*(pKa-pHa) +
  tcc[1]*tcc[1]/(tcc[3]-tcc[2])/(tcc[3]-tcc[2])*mu*mu*ni/Vmu*
    (pKc-pKb+pHb-pHc) +
  mu/Vmu*(ni-1)*mu*(pKam2-pHam2) +
  tcc[1]*tcc[1]/(tcc[3]-tcc[2])/(tcc[3]-tcc[2])*mu/Vmu*(ni-1)*
    mu*((pKcm2-pKbm2)+(pHbm2-pHcm2)) +
  mu/Vmu*(1-2*mu*ni)*(pKam1-pHam1) +
  tcc[1]*tcc[1]/(tcc[3]-tcc[2])/(tcc[3]-tcc[2])*mu/Vmu*
    (1-2*mu*ni)*((pKcm1-pKbm1)+(pHbm1-pHcm1)) -
  tcc[1]*tcc[1]/(tcc[3]-tcc[2])/(tcc[3]-tcc[2])*2*tcc[3]*
    ((pKcm1-pKbm1-pKc+pKb)*mu*sni/sV-(pHbm1-pHcm1-pHb+pHc)*
    mu*sni/sV)
})
```

The second expectation for the asymptotic variance is the expectation of the Hampel function with respect to the Pearson residuals multiplied by the Pearson residuals divided by the square root of the variance of μ_i. This equals

```
EpsiSBin <- expression({
  Q2V + ifelse(ni==0, 0, Q1V/sni/sV)
})
```

where

```
Q1V <- mu*mu*ni/Vmu*(pKa-pHa) -
  tcc[1]/(tcc[3]-tcc[2])*mu*mu*ni/Vmu*(pKc-pKb+pHb-pHc) +
  mu*mu*(ni-1)/Vmu*(pKam2-pHam2) -
  tcc[1]/(tcc[3]-tcc[2])*mu*mu*(ni-1)/Vmu*
    ((pKcm2-pKbm2)+(pHbm2-pHcm2)) +
  mu/Vmu*(1-2*mu*ni)*(pKam1-pHam1) -
  tcc[1]/(tcc[3]-tcc[2])*mu/Vmu*(1-2*mu*ni)*
    ((pKcm1-pKbm1)+(pHbm1-pHcm1))
```

and

```
Q2V <- tcc[1]*((pKbm1-pKam1-pKb+pKa)*mu/Vmu -
    (pHam1-pHbm1-pHa+pHb)*mu/Vmu) +
  tcc[1]*tcc[3]/(tcc[3]-tcc[2])*
    ((pKcm1-pKbm1-pKc+pKb)*mu/Vmu-(pHbm1-pHcm1-pHb+pHc)*mu/Vmu)
```

One must include the calculations of Q1V and Q2V in EpsiBin.init.

3.2.3 Adjustments for a Poisson distributed random variable

The required probabilities are

```
EpsiPois.init <- expression({
  # P[Y_i = j_z1]
  dpHa <- dpois(Ha, mu); dpHb <- dpois(Hb, mu);
  dpHc <- dpois(Hc, mu)
  # P[Y_i = j_z2]
  dpKa <- dpois(Ka, mu); dpKb <- dpois(Kb, mu);
  dpKc <- dpois(Kc, mu)
  # P[Y_i = j_z1-1]
  dpHa1 <- dpois(Ha-1, mu); dpHb1 <- dpois(Hb-1, mu);
  dpHc1 <- dpois(Hc-1, mu)
  # P[Y_i = j_z2-1]
  dpKa1 <- dpois(Ka-1, mu); dpKb1 <- dpois(Kb-1, mu);
  dpKc1 <- dpois(Kc-1, mu)
  # P[Y_i <= j_z1-1]
  pHam1 <- ppois(Ha-1, mu); pHbm1 <- ppois(Hb-1, mu);
  pHcm1 <- ppois(Hc-1, mu)
  # P[Y_i <= j_z2-1]
  pKam1 <- ppois(Ka-1, mu); pKbm1 <- ppois(Kb-1, mu);
  pKcm1 <- ppois(Kc-1, mu)
  # P[Y_i <= j_z1]
  pHa <- pHam1 + dpHa; pHb <- pHbm1 + dpHb;
  pHc <- pHcm1 + dpHc
  # P[Y_i <= j_z2]
  pKa <- pKam1 + dpKa; pKb <- pKbm1 + dpKb;
  pKc <- pKcm1 + dpKc
})
```

Implement the expectation needed for Fisher consistency as

```
EpsiPois <- expression({
  mu/sV*(dpHa-dpKa-tcc[1]/(tcc[3]-tcc[2])*(dpKb-dpKc+dpHc-dpHb))+
    tcc[1]*(pKb-pKa-pHa+pHb + tcc[3]/(tcc[3]-tcc[2])*
      (pKc-pKb-pHb+pHc))
})
```

The two expectations for the asymptotic variance are

```
Epsi2Pois <- expression({
  mu*(dpHa1-dpKa1-dpHa+dpKa +
    tcc[1]/(tcc[3]-tcc[2])*tcc[1]/(tcc[3]-tcc[2])*
      (dpKb1-dpKc1-dpKb+dpKc+dpHc1-dpHb1-dpHc+dpHb)) +
  (pKam1-pHam1 +
    tcc[1]/(tcc[3]-tcc[2])*tcc[1]/(tcc[3]-tcc[2])*
      (pKam1-pHam1+pHbm1-pHcm1)) -
  tcc[1]*tcc[1]*(pKb-pKa+pHa-pHb +
    tcc[3]/(tcc[3]-tcc[2])*tcc[3]/(tcc[3]-tcc[2])*
    (pKc-pKb+pHb-pHc)) -
  2*tcc[3]*mu/sV*tcc[1]/(tcc[3]-tcc[2])*tcc[1]/
    (tcc[3]-tcc[2])*(dpKb-dpKc-dpHc+cpHb)
})
```

and

```
EpsiSPois <- expression({
  mu/sV*(dpHa1-dpKa1-dpHa+dpKa - tcc[1]/(tcc[3]-tcc[2])*
    (dpKb1-dpKc1-dpKb+dpKc+dpHc1-dpHb1-dpHc+dpHb)) +
  1/sV*(pKam1-pHam1 -
    tcc[1]/(tcc[3]-tcc[2])*(pKcm1-pKbm1+pHbm1-pHcm1)) +
  tcc[1]*(dpKa-dpKb-dpHb+dpHa +
    tcc[3]/(tcc[3]-tcc[2])*(dpKb-dpKc-dpHc+dpHb))
})
```

3.3 Plausibility check

To verify the developed approach and its implementation in R, one can check its outlier control and compare it to the use of the Huber function as well as to standard logistic regression. In the first step, logistic regression was evaluated. There, by checking for plausibility, the lack of robustness of standard genotype relative risk (GRR) estimators against single cases and controls in relatively large association studies were also illustrated and the bounded influence of outliers on robust GRR estimates depicted. Let the logistic regression model

$$\text{case/control} \sim \exists\text{variant} + \text{age}$$

describe the relation between the case-control status and two explaining variables (existence of a genetic variant and age). In order to examine the influence of single outliers on standard and robust estimators of the GRR, a 1000 case / 1000 control study investigating a rare variant was simulated with a minor allele frequency (MAF) of 0.0075 and a dominant GRR equal to 1.84. In a dominant genetic model, the genotypes are coded 0 or 1 indicating whether the genotype comprises at least one causal allele. The two genetic parameters were chosen in consistency with the moderate-penetrance breast cancer susceptibility variant CHEK2*1100delC (Meijers-Heijboer et al., 2002). Hypothetical outliers were mimicked by single cases and controls aged between 0 and 125 years who carried the high-risk variant. Standard and robust logistic regression models with Huber and Hampel influence functions were fitted to the baseline data set with 2000 individuals and to the extended data sets with 2001 individuals. The corresponding standard and robust odds ratios (ORs) were used as GRR estimates. In this context, the curves of the influenced standard and robust estimates are of special interest. Standard estimates can be influenced without any bound. By comparison, bounded and re-descending estimators limit the outlier influence to a pre-specified amount. In the latter case, the influence is cancelled if the outlier strength exceeds the defined limits. The estimate curves have to represent these characteristics and figure 7 shows results from this small simulation exercise.

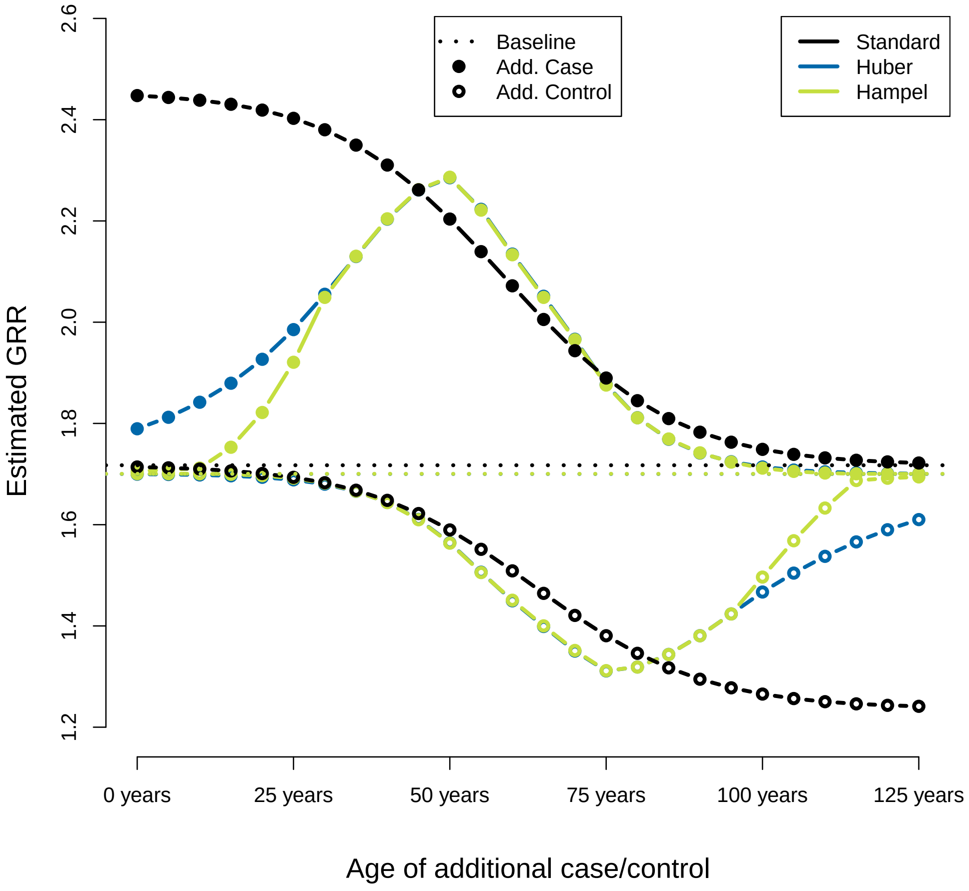


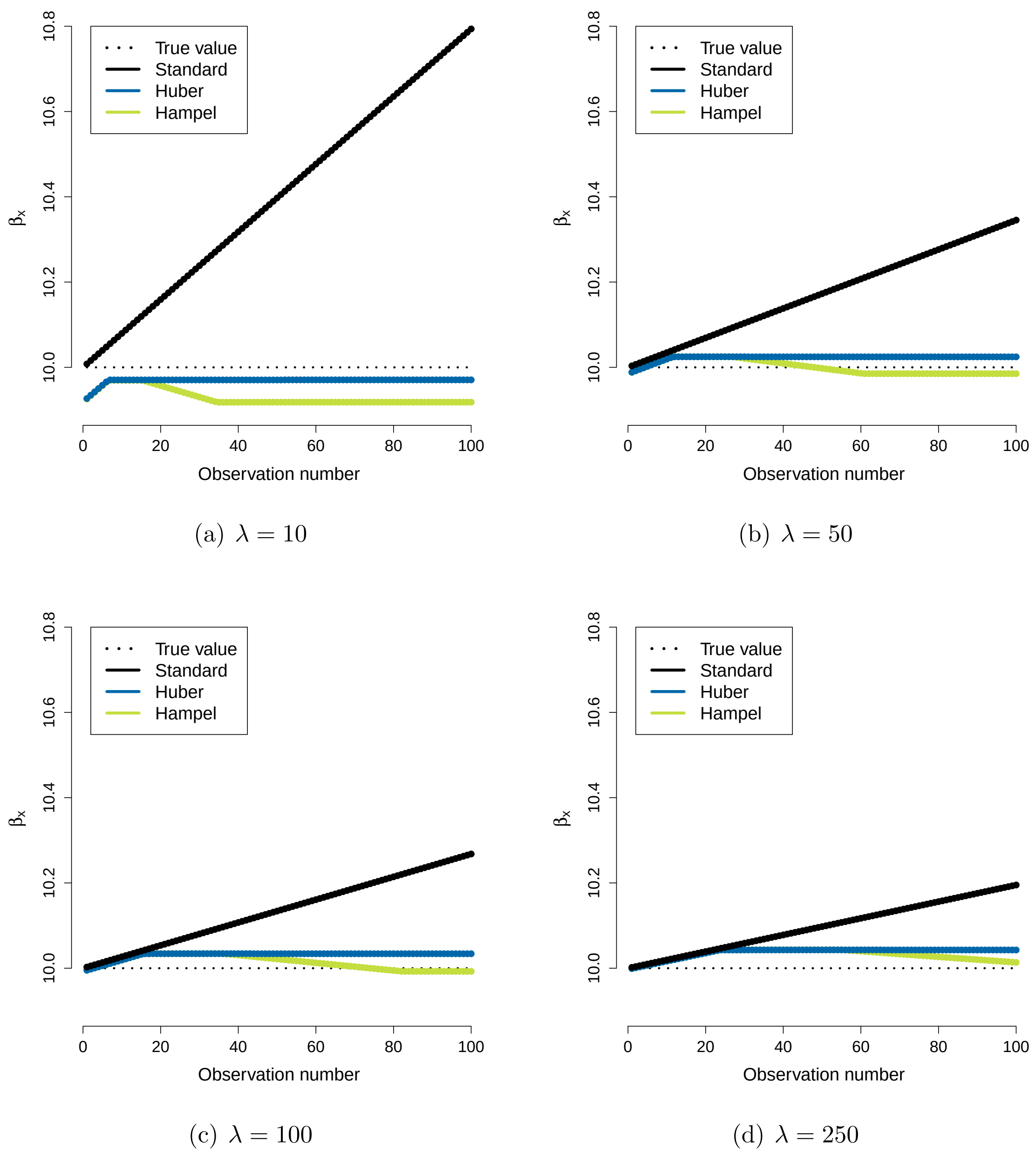
Figure 7: Influence of outliers on standard and robust estimates of the genotype relative risk (GRR). The influence was examined by including single cases and controls that carried the high-risk variant to the baseline data set with 1000 cases and 1000 controls, a dominant GRR of about 2 and a minor allele frequency (MAF) of 0.0075. The fitted logistic regression model was disease status (case/control) explained by genotype and age. The tuning constants were for the Huber function 1.345 and for the Hampel function $(1.5, 3.5, 8) \times 0.9$.

For example, one single case diagnosed at age 25 years who carried the high-risk variant increased the estimated standard GRR from approximately 1.70 (1000 cases and 1000 controls) to 2.40 (study with 2001 individuals). The influence of the same single case was less accentuated on robust GRR estimates. The GRR increased to around 1.9 (2.0) when the Hampel (Huber) influence function was used. Note that the influence of single outliers on standard GRR estimates was unbounded. By contrast, the larger the departure of the outlier from the bulk of the study population, the smaller its influence on robust GRR estimates. Summarising, additional observations aged between 45 and 80 years were better handled by standard than by robust logistic regression. Outside this range, robust regression is less influenced by departing observations. The difference between standard and robust logistic regression heavily increases with increasing outlier strength. Furthermore, the Hampel function controls the outlier influence even better than the Huber function. Thus, the incorporation of the Hampel function resulted in a clear improvement for extreme outliers in both additional cases and controls.

For consideration of Poisson regression, 1000 Poisson distributed random variables were simulated for several Poisson parameter λ to create the response variable y, $\lambda \in \{10, 50, 100, 250\}$. The independent variable x was calculated as

$$x = \frac{\log(y) - \beta_{\text{Int}}}{\beta_{\text{x}}}$$

with regression coefficients $\beta_{\text{Int}} = 0$ and $\beta_{\text{x}} = 10$. Hence, the true regression coefficient for x equals 10 and the intercept is 0. Hypothetical outliers were mimicked by single observations defined by the tuples $(\max(x), \max(y) + i)$ with $i \in \{1, 2, \ldots, 100\}$. Standard and robust logistic regression models with Huber and Hampel influence functions were fitted to the baseline data set with 1000 observations and to the extended data sets with 1001 observations. Results are shown in figure 8. Again, robust methods controlled the influence of departing observations whereas the standard estimates increased without any bounds. Once again, the use of the Hampel function reflected the characteristic of a re-descending weighting function. Figure 8 demonstrates that, in case of the existence of a highly influential outlier, both robust estimates were more reliable because the outlier influence was bounded. In absence of outlying observations the accuracy of both robust Poisson regression methods depended on the underlying distribution although the response was perfectly Poisson distributed and the explanatory variable was exactly log-linear related with the response. This underestimation of the slope by the robust methods decreased with increasing Poisson parameter λ. The robust approaches overestimated the intercept for λ equal to 10 and 50. The standard Poisson regression estimated both coefficients correctly. Table 5 summarises the estimated coefficients (intercept and slope) for standard and robust Poisson regression regarding the case without outliers.



(a) $\lambda = 10$

(b) $\lambda = 50$

(c) $\lambda = 100$

(d) $\lambda = 250$

Figure 8: Investigation of the influence of one single outlier on standard and robust Poisson regression estimates. The response variable y was simulated as Poisson distributed considering different Poisson parameters λ. The independent variable x was calculated according to the regression model with an intercept equal to 0 and a slope equal to 10. Outliers were created as data points $(\max(x), \max(y) + i)$ with $i \in \{1, 2, \dots, 100\}$. The observation number indicates the distance between the additional and the maximal value of the independent variable in the outlier-free scenario, i.e. the value of i. The tuning constants were for the Huber function 1.345 and for the Hampel function $(1.5, 3.5, 8) \times 0.9$.

Table 5: Estimated regression coefficients for intercept and slope (β_{Int}, β_{x}) in absence of additional observations. The true values are 0 for the intercept and 10 for the slope.

λ	β_{Int}			β_{x}		
	Standard	Huber	Hampel	Standard	Huber	Hampel
10	0.00	0.03	0.03	10.00	9.92	9.92
50	0.00	0.01	0.01	10.00	9.99	9.98
100	0.00	0.00	0.00	10.00	9.99	9.99
250	0.00	0.00	0.00	10.00	10.00	10.00

In view of figures 7 and 8, it seems that the use of the Hampel function resulted in a robuster method as compared to the use of the Huber function. To explore this assumption, the simulated data of the illustrative example with ten clustered outliers in chapter 1 was used again with the response rounded to non-negative integer (see figure 9 for the modified data). This time, the amount of outliers were increased in each iteration by one additional outlier from the outlying group of observations. Thus, the influence of an increasing amount of outliers could be demonstrated. Figure 10 shows that both robust methods are considerably less influenced than the standard method. Additionally, the Hampel function inhibited almost any noticeable reaction on contamination whereas the use of the Huber function to weight deviating observations still led to a clear decrease of the regression coefficient. This is in close accordance with Arora and Biegler (2001) who showed that re-descending estimators are very robust as compared to non re-descending (robust as well as standard) estimators.

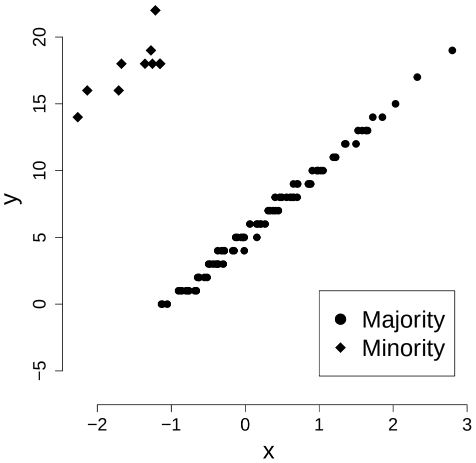


Figure 9: Simulated data with ten clustered outliers.

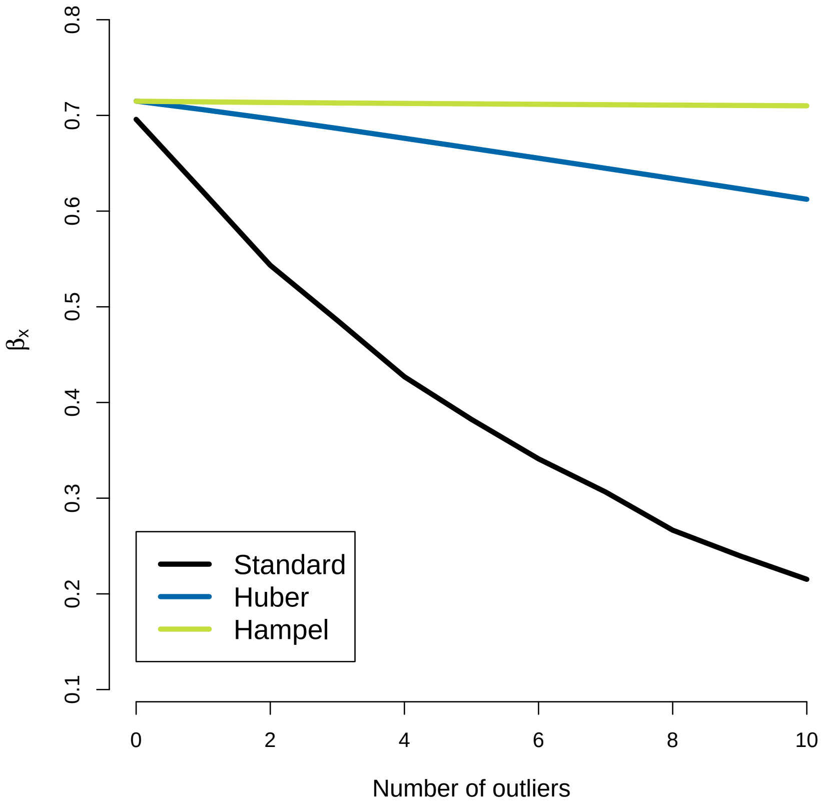


Figure 10: Investigation of the influence of different amounts of outliers on standard and robust Poisson regression estimate β_x by inclusion of 0 to 10 outliers. The tuning constants were for the Huber function 1.345 and for the Hampel function $(1.5, 3.5, 8) \times 0.9$.

4 Results of method application

After deducing the theoretical background and explaining the practical application to analyse simulated and real data in chapter 2, the results are presented in this chapter starting with the results of the comparison between standard and existing robust logistic regression approaches (section 4.1) followed by the comparison between standard and robust logistic regression applying both the Huber and the Hampel function (section 4.2).

4.1 Standard versus existing robust regression methods

Ordering follows the method section. The results of consistency of model selection and prediction accuracy are given first (section 4.1.1). Then, the results of the influence of one single outlier (section 4.1.2) and of genotyping errors on estimates (section 4.1.3) are provided.

4.1.1 Consistency of model selection and prediction accuracy in real data

The distribution of the chromosomal instability is shown in figure 11. Very similar results were found for different sizes, e.g. 500 kb (data not shown).

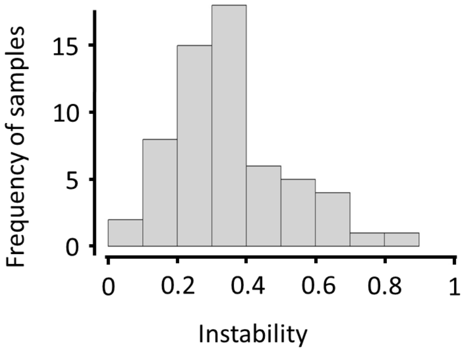


Figure 11: Genomic instability distribution. Instability is given as the ratio between the number of instable chromosome arms and the total number of investigated arms.

Using the relation between instability and DNA methylation in (robust) linear and Poisson regression models (relying on the Huber function), goodness-of-fit was analysed based on the residuals. The comparison of residuals' magnitudes of all four regression types is shown in table 6. No significant difference between the standard linear regression (reference) and the other three approaches could be identified (all p-values > 0.05). For standard linear regression, a median residual value of 5.2 was found. Robust linear regression and standard Poisson regression were very similar compared to the reference (median residual values of 5.1 and 4.9). The highest accuracy was reached by robust Poisson regression (median residual value of 3.4) but it also showed the highest residual variation.

Table 6: Goodness-of-fit regarding the residuals. The p-value results from the residual two-sided Wilcoxon signed rank test.

Residuals	Linear regression	Robust linear regression	Poisson regression	Robust Poisson regression
Median	5.2	5.1	4.9	3.4
5^{th} & 95^{th} quantile	0.6, 12.5	0.5, 13.0	1.0, 10.9	0.3, 15.7
p-value	reference	0.09	0.86	0.27

The second aspect was reproducibility. Figure 12 shows that standard and robust linear regression selected identical models. In particular, they selected the methylation of the same single gene in all iterations; this gene was *GNS* (Glucosamine (N-acetyl)-6-sulfatase). Standard and robust Poisson regression models built several different models including the methylation of two or three genes; *GNS* was always one of them.

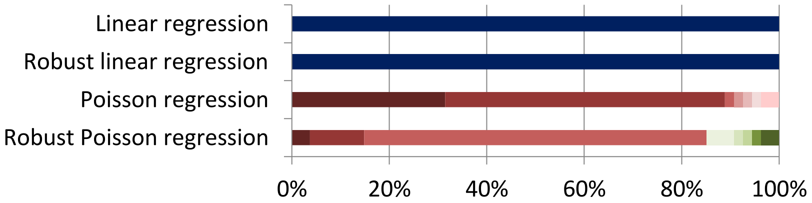
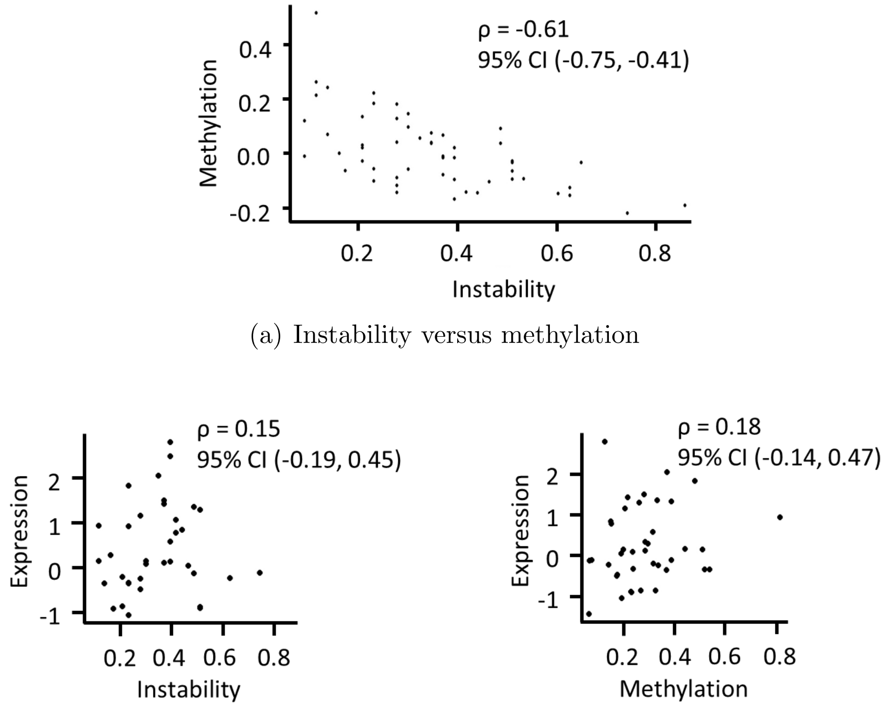


Figure 12: Reproducibility (frequency of selected models). Each colour represents a selection of a different model.

The monotone relationship between chromosomal instability and *GNS* methylation shown in figure 13 (a) resulted in a Spearman's correlation coefficient of −0.61. However, a relationship between *GNS* expression and chromosomal instability or *GNS* expression and *GNS* methylation was not found for this gene (see figures 13 (b) and (c)). The correlation coefficients for these investigated relations were 0.15 and 0.18, respectively.



(a) Instability versus methylation

(b) Instability versus expression

(c) Methylation versus expression

Figure 13: Comparative scatterplots for genomic instability, DNA methylation and gene expression of *GNS*. ρ indicates Spearman's correlation coefficient given in combination with its 95% confidence interval (CI).

4.1.2 Influence of one single outlier in real data

χ^2 tests revealed no influence of gender ($p = 0.95$) and tobacco smoking ($p = 1.00$) on hypertension risk. Hence, only age was included in the logistic regression models as covariate. Filter criteria resulted in 130 individuals (43 cases and 87 controls) with complete genotype and phenotype information. The age of the individuals ranged between 20 and 95 years with a median age of 52 years. The total number of measured SNPs on chromosome 3 in the investigated GAW 18 data set was 35,045.

A plot of Cook's distances under the age-only standard logistic regression model (figure 14) revealed several observations that departed from the majority of the sample. Considering a threshold of 0.05 for the Cook's distance, four observations could be defined as outliers. Information on disease status and age of deviating individuals is shown in table 7. Individuals 62, 58 and 24 were older than 80 years and normotensive. On the other hand, individual number 60 was affected by the condition early in life (38 years old).

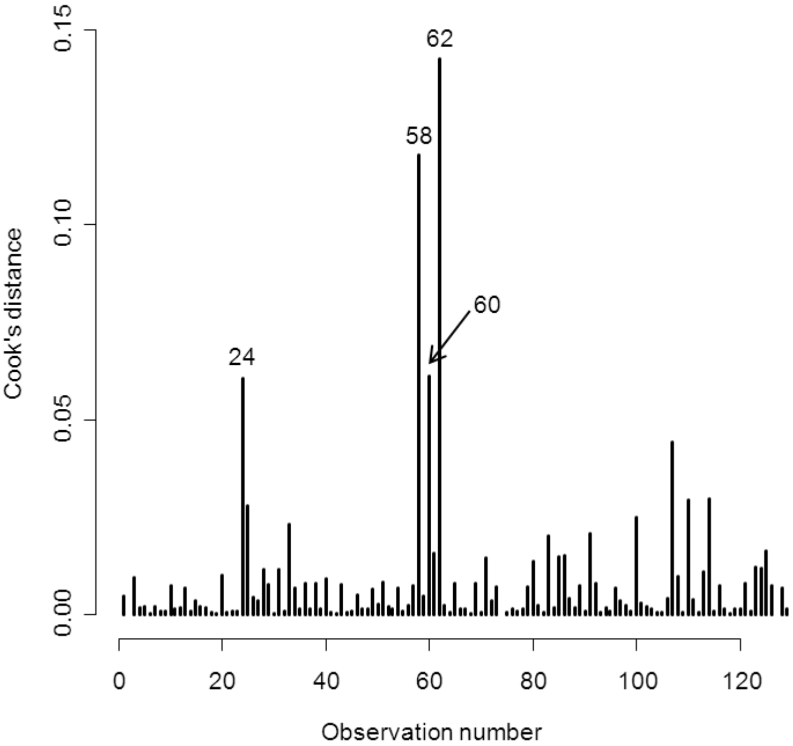


Figure 14: Cook's distances from the age-only standard logistic regression model. The four most prominent observations are indicated by their observation number.

Table 7: Estimated odds ratios (OR) per year of age. Odds ratios were estimated based on standard and robust logistic regression (LR) models for the complete set of individuals and after exclusion of the four most remarkable outliers. HTN: Hypertension

Excluded individuals	HTN	Age	OR – Age (% Change)			
			Standard LR		Robust LR	
None			1.085	(reference)	1.084	(reference)
62	0	90.23	1.095	(+11.2)	1.091	(+7.8)
58	0	87.66	1.094	(+10.0)	1.091	(+7.9)
60	1	38.44	1.091	(+ 6.5)	1.089	(+5.1)
24	0	80.27	1.091	(+ 6.6)	1.091	(+7.6)

The influence of these outliers on standard and robust parameter estimates of age effects is also shown in table 7. For example, the exclusion of individual 62 resulted in an 11.2% increase of the excess risk of hypertension per year according to standard logistic regression compared to a 7.8% increase for robust logistic regression. The odds of hypertension by age interval are shown in table 8.

Table 8: Overall odds of hypertension per age interval. Age intervals were defined by the age quartiles in controls.

Age interval	(Number of cases : controls)	Odds
< 39.0	(1 : 22)	0.05
$[39.0, 46.0)$	(2 : 20)	0.10
$[46.0, 56.2)$	(9 : 23)	0.39
≥ 56.2	(31 : 22)	1.41

Standard logistic regression identified SNP rs3934103 located in the *ULK4* gene as the variant that most improved the model fit. Robust logistic regression identified SNP rs11918360 in *RP11-408H1.3* as the variant with the strongest association signal. Under both standard and robust regression, model selection clearly favoured the two identified SNPs as represented in figure 15. The pairwise r^2 between SNP rs3934103 and SNP rs11918360 was 0.003.

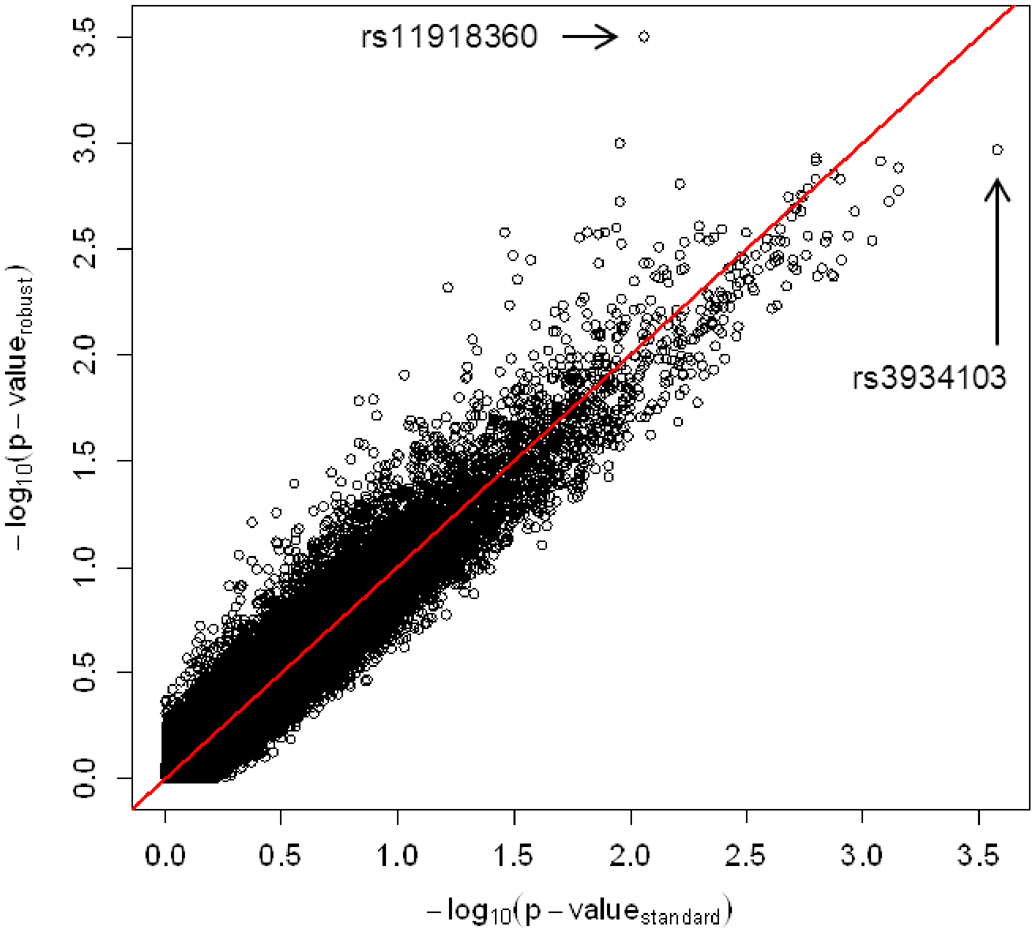


Figure 15: Quantile-quantile plots from the age-genotype standard and robust logistic regression models. The two selected SNPs are indicated by their reference SNP ID number.

The influence of the four outliers on the AUCs from the standard and robust logistic regression models is shown in table 9. Robust and standard AUCs for the age-only models were identical. For the age-genotype models, the AUCs were slightly smaller and also slightly less outlier-dependent for robust than for standard logistic regression.

Table 9: Area under the receiver operating characteristic curve (AUC) for standard (upper table) and robust logistic regression (lower table). AUCs were calculated for the complete set of individuals and after exclusion of the four most remarkable outliers. The relative contributions of the variables age and SNP (rs3934103 and rs11918360) are also shown.

Standard

Excluded individuals	AUC – Age (% Change)		AUC – Age + SNP (% Change)	
None	0.811	(reference)	0.852	(reference)
62	0.820	(+1.1)	0.861	(+1.1)
58	0.820	(+1.1)	0.861	(+1.1)
60	0.825	(+1.7)	0.859	(+0.9)
24	0.819	(+1.0)	0.859	(+0.9)

Robust

Excluded individuals	AUC – Age (% Change)		AUC – Age + SNP (% Change)	
None	0.811	(reference)	0.843	(reference)
62	0.820	(+1.1)	0.852	(+1.0)
58	0.820	(+1.1)	0.853	(+1.2)
60	0.825	(+1.7)	0.851	(+0.9)
24	0.819	(+1.0)	0.844	(+0.0)

Table 10 summarizes the results from the leave-one-out cross-validation. The concordance was better for the robust logistic regression model at every cut-off probability compared to standard logistic regression. Both models allocated best at probability 0.5 and almost identically at probability 0.3 (the investigated population included 43 cases and 87 controls, i.e. 33% hypertension prevalence). At a probability of 0.3, sensitivities were identical and the specificity was slightly higher under robust regression. Standard and robust estimates showed similar discriminative performances supported by an IDI of -0.07 at every cut-off probability. AUCs were also almost identical. The clinical net benefit was slightly larger for the robust logistic regression model in the probability range between 0.2 and 0.6.

Table 10: Overview: Concordance, sensitivity, specificity, clinical net benefit and overall AUCs. These characteristics rely on standard (upper table) and robust logistic regression models (lower table) estimated based on leave-one-out cross validation.

Standard	Probability cut-off	Concordance N (%)	Sensi-tivity	Speci-ficity	Net benefit
	0.0	43 (33.1)	1.00	0.00	0.33
	0.1	79 (60.8)	0.95	0.44	0.27
	0.2	90 (69.2)	0.86	0.61	0.22
	0.3	98 (75.4)	0.81	0.72	0.19
	0.4	98 (75.4)	0.70	0.78	0.13
	0.5	101 (77.7)	0.60	0.86	0.11
	0.6	97 (74.6)	0.40	0.92	0.05
	0.7	99 (76.2)	0.35	0.97	0.06
	0.8	93 (71.5)	0.19	0.98	0.00
	0.9	91 (70.0)	0.12	0.99	−0.03
	1.0	87 (66.9)	0.00	1.00	–
	AUC		0.835		

Robust	Probability cut-off	Concordance N (%)	Sensi-tivity	Speci-ficity	Net benefit
	0.0	43 (33.1)	1.00	0.00	0.33
	0.1	82 (63.1)	0.88	0.51	0.26
	0.2	97 (74.6)	0.86	0.69	0.23
	0.3	99 (76.2)	0.81	0.74	0.19
	0.4	102 (78.5)	0.72	0.82	0.16
	0.5	107 (82.3)	0.67	0.90	0.15
	0.6	102 (78.5)	0.51	0.92	0.09
	0.7	100 (76.9)	0.42	0.94	0.05
	0.8	97 (74.6)	0.30	0.97	0.01
	0.9	93 (71.5)	0.19	0.98	−0.08
	1.0	87 (66.9)	0.00	1.00	–
	AUC		0.830		

4.1.3 Influence of genotyping errors on estimates in simulated data

Figure 16 represents the investigation of the genotyping error influence on odds ratios (OR). The OR dependence on genotyping error rate was large for rare variants (causal allele frequency of 0.005) and much smaller for common variants (causal allele frequency of 0.13). Differences existed between the model M_1 (model with genotyping errors) and the error free model M_0 as well as in the effect of genotyping errors on standard and robust estimates. For a 0.05 causal allele frequency and 0.005 genotyping error rate, the effect of genotyping errors was smaller for standard than for robust logistic regression. This contrasted the results considering larger genotyping error rates where the effect of mis-genotyping was smaller for robust estimates. The results of standard and robust logistic regression were practically identical for rare and common variants.

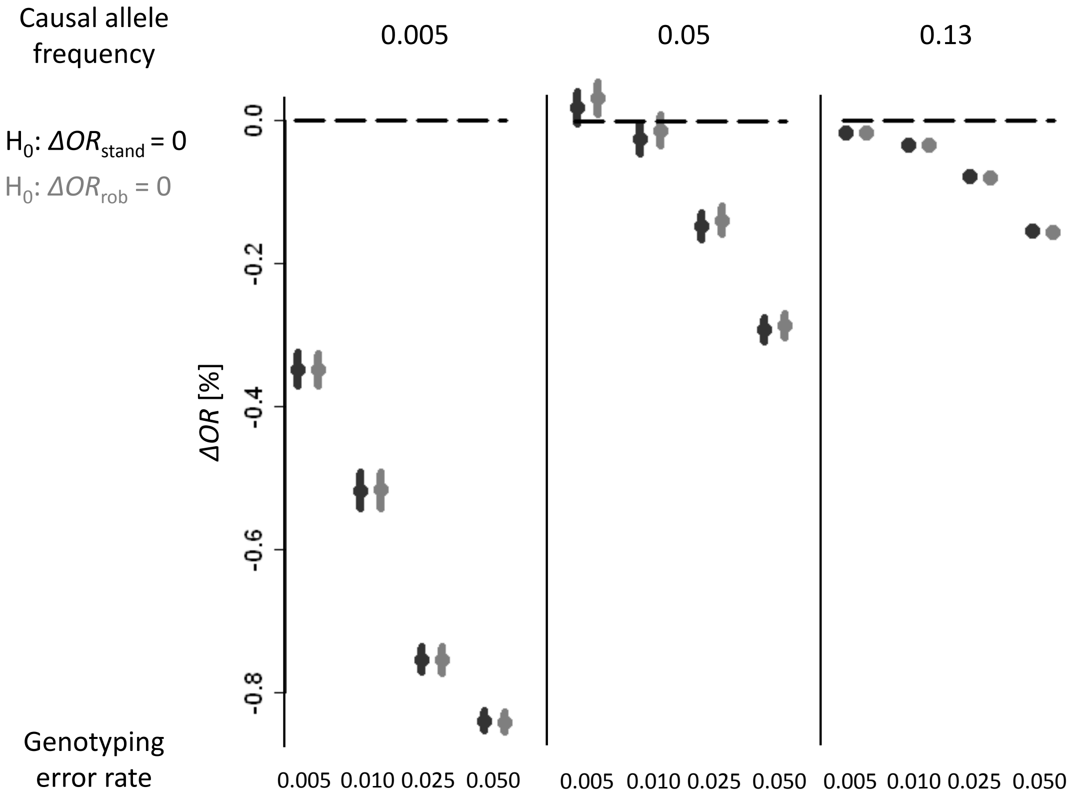


Figure 16: Influence of different genotyping error rates on odds ratios for different causal allele frequencies estimated by standard (dark grey) and robust logistic regression (light grey). Median estimates are indicated by points and their 95% confidence intervals by vertical bars.

As seen before for the causal allele frequency of 0.05, there is a value of the genotyping error rate where the benefit of standard logistic regression changes to a benefit of robust logistic regression. Figure 17 shows the result of the genotyping error rate screen to identify the point of benefit change between standard and robust logistic regression. With current genotyping platforms, genotyping error rates around 0.005 are plausible (Kennedy et al., 2003; Montgomery et al., 2005; Hong et al., 2012). It can bee seen that considering a genotyping error rate of 0.005 the benefit of robust logistic regression is small though about 0.25‰. In general, it can be observed that the advantage of robust logistic regression increases with increasing genotyping error rates. For almost every genotyping error rate there is an advantage for robust logistic regression.

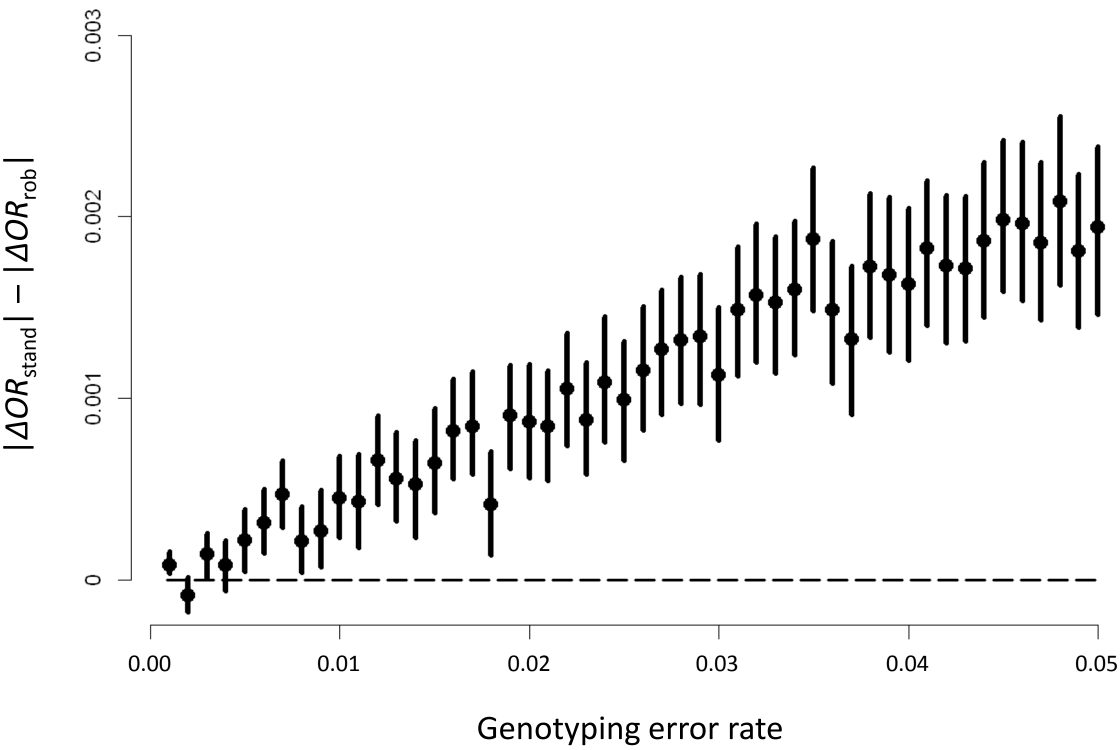


Figure 17: Differences between the effects of genotyping errors on standard and robust estimates for several genotyping error rates. Median estimates are indicated by points and their 95% confidence intervals by vertical bars.

4.2 Statistical properties of robust logistic regression applying the Hampel function

Computer simulations

Type I error rates did not exceed the nominal 0.05 level in the simulated null scenarios (table S1 on page 124). For example, the type I error rate was 0.044 (95% CI: $0.038 - 0.050$) when a standard recessive penetrance model was fitted to null data. The corresponding type I error rate for robust logistic regression with the Huber function was 0.046 ($0.040 - 0.052$). Robust logistic regression using the Hampel function resulted in type I error rates equal to 0.045 ($0.039 - 0.051$) for tuning constants $(1.5, 3.5, 8) \cdot 0.9$, and equal to 0.044 ($0.038 - 0.050$) for tuning constants $(2, 4, 8) \cdot 0.7$.

Simulation results revealed appreciable differences between standard and robust GRR estimates. In the reference scenario, the median of the standard GRR estimates was 1.44, slightly higher than robust counterparts, which were around 1.43 (figure 18). Standard GRR estimates were also higher than robust estimates when a recessive penetrance model was fitted to data generated under the reference scenario (median standard (robust) GRR estimate: 1.40 (1.32), figure S1 on page 144) as well as for rare variants (MAF= 0.001: median standard (robust) GRR estimate = 1.50 (1.38); MAF= 0.005: 1.41 (1.39), figure S2 on page 145). Age-dependent GRRs constituted an exception with higher estimates for robust than standard methods (median standard (robust) GRR estimate 2.28 (2.42), figure S4 on page 147). Boxplots of the estimated genotype relative risk (GRR) for all scenarios are given in the figures S1-S9 on pages 144-152.

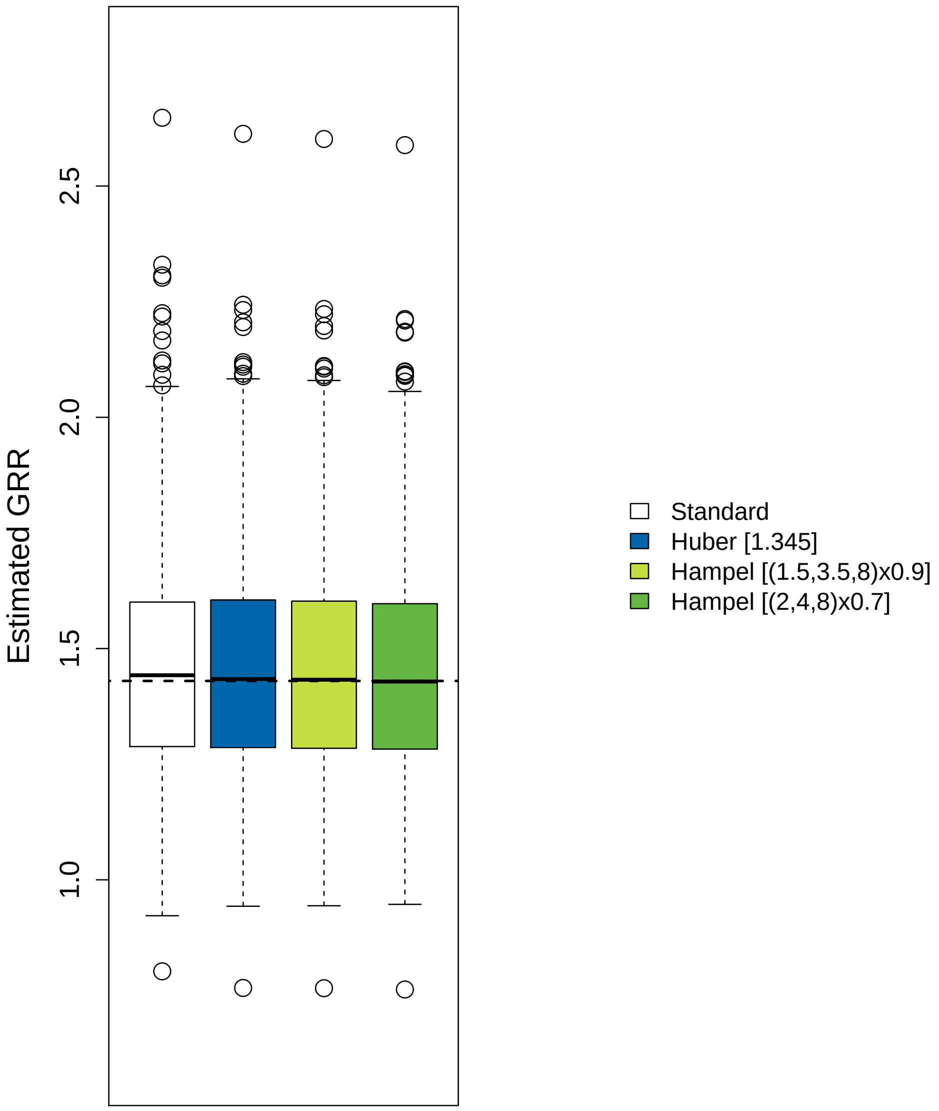


Figure 18: Boxplots of the estimated genotype relative risk (GRR) under the reference scenario. Settings for the reference scenario were minor allele frequency (MAF) = 0.05, simulated dominant GRR = 1.43 age-independent, $D^{'} = 1$, $r^2 = 1$, no genotyping errors, 400 simulated studies with 1000 cases and 1000 controls and a fitted dominant penetrance model. Tuning constants for the robust logistic regression models are given in brackets in the legend. The dotted line indicates the simulated true effect.

Tables S2-S5 on pages 125-131 presents the bias, variance, MSE and statistical power of standard and robust GRR estimates for all simulation scenarios. Under the reference simulation scenario, the bias amounted +0.010 for standard compared to +0.002 to +0.006 for robust GRR estimates. The GRR overestimation by standard logistic regression translated into a larger statistical power than for robust logistic regression – in spite to the fact that the variance was higher for standard than for robust GRR estimates. In practically all simulated scenarios, the statistical power was higher for standard than for robust logistic regression. For age-dependent GRRs and in the presence of genotyping errors, biases and variances were higher for robust compared to standard GRR estimates. By contrast, rare variants and recessive fitted models showed markedly smaller biases and variances for robust than for standard GRR estimates. For example, for a variant with MAF equal to 0.001, the variance was 23.5 for the standard compared to 1.0 for robust GRR estimates and for a fitted recessive penetrance model 19.3 compared to 1.0 to 1.2. Figure 19 shows the MSE of standard and robust GRR estimates according to the penetrance model fitted to the data and the MAF of the associated variant. The large bias, variance and MSE differences motivated a closer comparison of standard and robust methods for rare and recessive variants to exclude the possibility of spurious observations due to a lack of statistical power. Tables S6-S13 on pages 133-140 present results consistent with a statistical power of approximately 0.6 for standard logistic regression in the absence of genotyping errors for rare variants and recessively simulated data. The left panel of figure 20 represents standard and robust GRR estimates for rare variants with MAFs equal to 0.001 and 0.005, with corresponding median standard (robust) estimates of 2.6 (2.5). The two right panels of figure 20 depict standard and robust GRR estimates for a recessive variant with corresponding median standard (robust) estimates equal to 7.0 (5.5). In general, standard GRR estimates showed higher biases and higher variances resulting in a higher MSE than their robust counterparts – consistent with the previous results for rare variants and for a fitted recessive penetrance model. Tables S6-S13 on pages 133-140 show complete results for increasing genotyping error rates and for different penetrance models fitted to recessively simulated data.

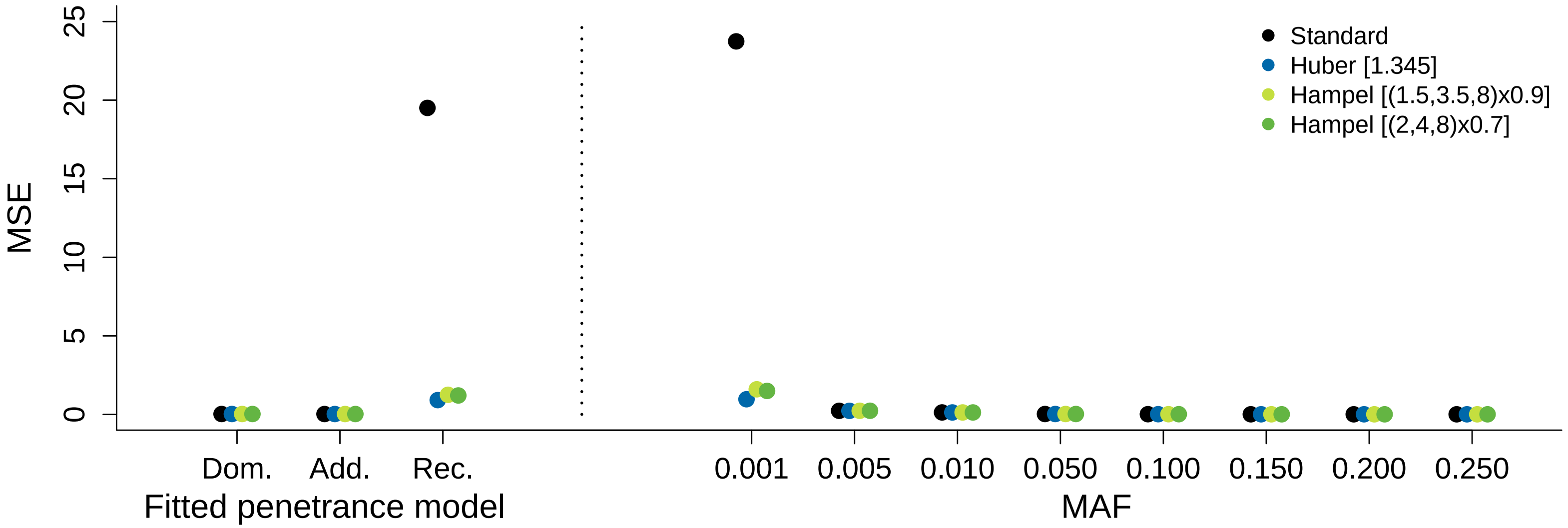


Figure 19: Mean squared error (MSE) of the estimated genotype relative risk (GRR) under different fitted penetrance models for minor allele frequency (MAF) = 0.05 (left) and different MAFs for a fitted dominant penetrance model (right). The assumed parameters were: simulated dominant GRR = 1.43 age-independent, $D' = 1$, $r^2 = 1$, no genotyping errors and 400 simulated studies with 1000 cases and 1000 controls. The tuning constants for the robust logistic regression methods are given in brackets in the legend.

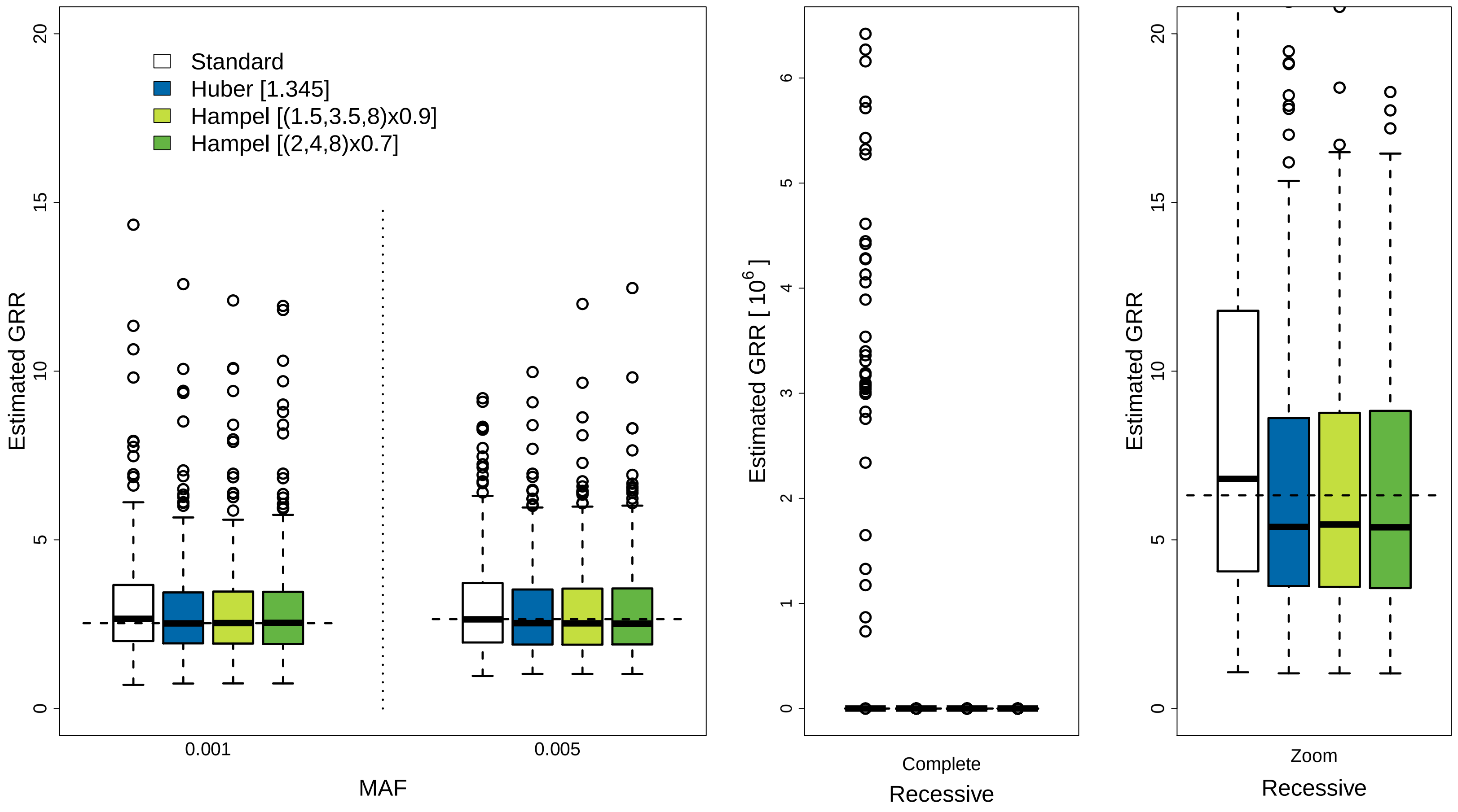


Figure 20: Boxplots of estimated genotype relative risks (GRR) for selected scenarios. The left panel shows the estimated GRR for a genetic variant with minor allele frequency (MAF) equal to 0.001 (left; simulated dominant GRR = 2.53 age-independent, 5000 cases, 5000 controls) and 0.005 (right; simulated dominant GRR = 2.65 age-independent, 1000 cases, 1000 controls) with $D' = 1$, $r^2 = 1$, 400 simulated studies with a fitted dominant penetrance model. The middle and right panels display the estimated GRRs of the recessive simulated data with a fitted recessive penetrance model (MAF = 0.05, simulated recessive GRR = 6.32 age-independent, $D' = 1$, $r^2 = 1$, 400 simulated studies with 1000 cases / 1000 controls). The middle panel shows the complete domain. The display is limited to 0-20 in the right panel. The robust logistic regression tuning constants are given in brackets in the legend.

Real data application

The investigated real data set included 245 genetic variants and 144 individuals with a median body height of 175 cm (Q1-Q3: 170 − 181 cm) and a median age of 45 years (Q1-Q3: 37-54 years). Genetic variant and individual identifiers are given in the tables S14 and S15 on pages 141 and 142.

The smallest p-value ($p = 0.004$) was reached for variant rs7519458 (MAF = 0.50). The estimated GRR for this variant was 1.2 according to both standard and robust methods (no outlying observation: left panel of figure 21). Figure 22 represents p-values and estimated GRRs for the 245 investigated variants. Differences between standard and robust results were apparent regarding p-values (left panels) and in particular estimated GRRs (right panels, please note the different scale of the y-axis for standard GRR estimates). All robust GRR estimates were below 12. By contrast, 22 variants resulted in standard GRRs over 12. Interestingly, robust methods did not converge for 21 out of these 22 variants. Variant rs2500262 (MAF = 0.34) was the only exception (standard results: p-value = 0.014, GRR = 13.7; robust results: p-values = 0.019 to 0.021, GRRs = 11.0 to 12.0). One strongly influential observation was identified for this variant (Cooks distance about 0.4; right panel of figure 21).

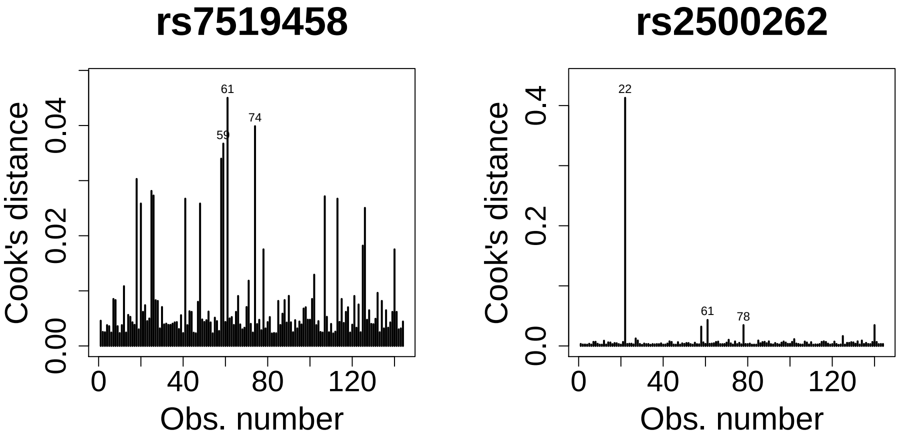


Figure 21: Cooks distances for the single nucleotide polymorphisms (SNPs) rs7519458 and rs2500262 in the real data application. Note the different axis scaling.

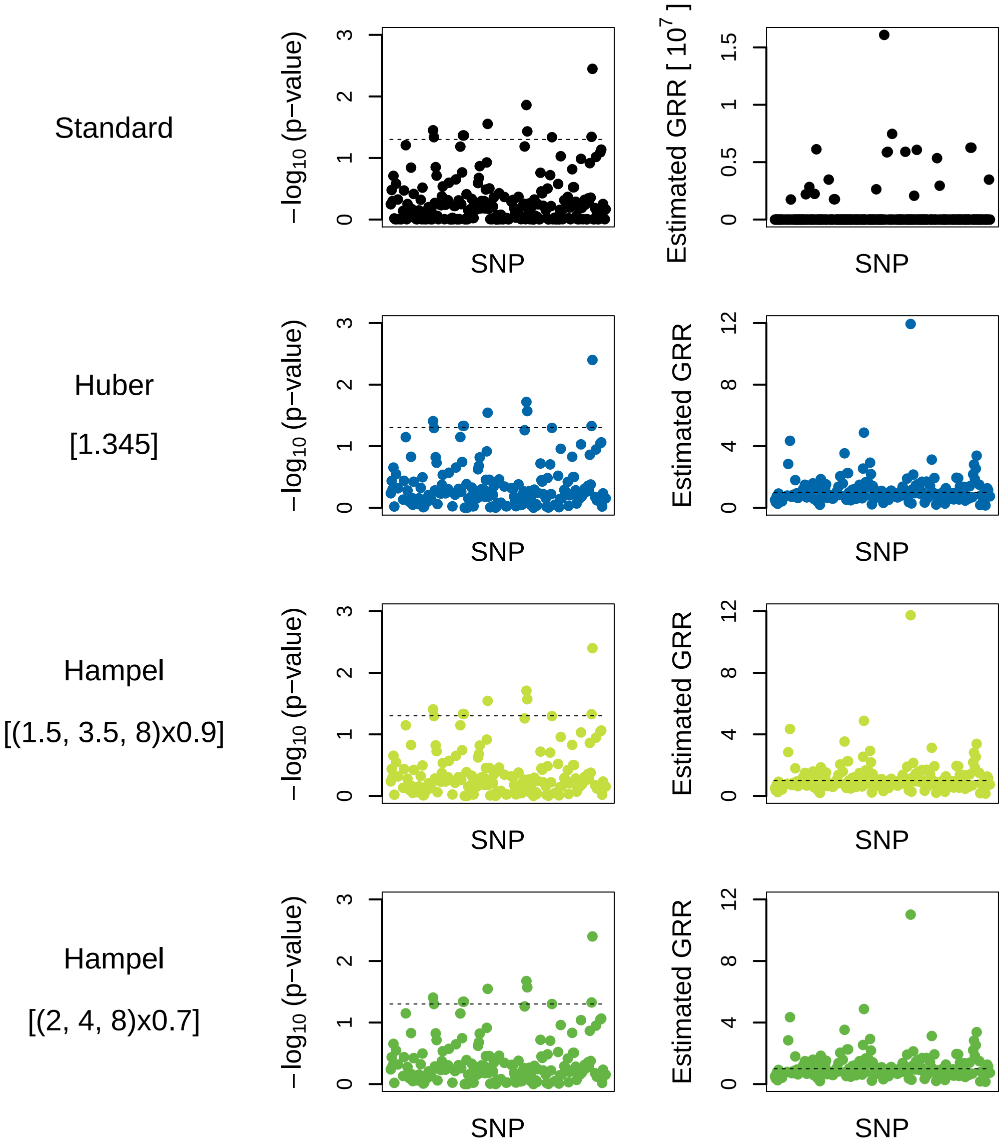


Figure 22: Manhattan plots ($-\log_{10}$(p-value)) and estimated genotype relative risks (GRR) of the real data application. Robust tuning constants are given in brackets behind the influence function on the left side of the plots. The horizontal lines indicate $-\log_{10}(0.05)$ in the Manhattan plots (left column) and exp(0) for the estimated GRRs (right column). Note the different axis scaling for the estimated GRRs.

5 Discussion

The presented results delivered insight that remains to be discussed. Section 5.1 provides the discussion of the comparison of standard and existing robust regression approaches and section 5.2 the comparison of standard and robust logistic regression applying both the Huber and the Hampel function. Since the R package `robustbase` was updated in the meantime, the improvement is presented in section 5.3. A final conclusion and a perspective complete this chapter in section 5.4. The final conclusion also includes an overall summary of the important results of this thesis and overall strengths and limitations.

5.1 Standard versus existing robust regression methods

The structure is the same as in the previous chapters. Section 5.1.1 treats model selection consistency and prediction accuracy, section 5.1.2 deals with the influence of one single outlier on estimated odds ratios and section 5.1.3 is about the influence of genotyping errors on estimates.

5.1.1 Consistency of model selection and prediction accuracy in real data

Genomic instability was defined by using available methylation data and aCGH information. Relying on instability and its relationship to methylation, model selection and prediction accuracy was analysed and compared within the already existing standard and robust logistic regression frameworks. The models built by standard and by robust linear regression coincided in the included single gene and this gene was incorporated in every selected Poisson model. Prediction accuracy was best supported by robust Poisson regression models. The median difference between observed and predicted counts of instability was about 10%.

It is known that the Poisson distribution tends to the normal distribution with increasing mean (Ramsey and Schafer, 2002). Under these conditions, a linear regression is an alternative to Poisson regression. In this real data application, however, linear and Poisson regression built different models to predict the instability. Interestingly, the variable of the linear models was always included in the Poisson models but these models were extended by one or two additional variables. Hence, their is some consistency between the methods in model selection. Based on the selected prediction models, differences in prediction accuracy between linear and Poisson regression could not be identified.

The aim of robust statistics is to handle outliers so that the result is valid for the majority of the data. This handling is without a priori exclusion of departing observations but with assignment of different weights depending on their location in relation to the bulk of data. This weighting led to different regression models in case of Poisson regression.

Real data applications are normally characterised by a set of independent variables. The goal is to find a subset that describes the current data and also predicts future observations well. For predictions, overfitting must be avoided as it leads to a small mean squared error (MSE) in the current data set but to large MSEs in future data sets (James et al., 2015). Hence, model selection is required. There are several different approaches. In this analysis, a subset of variables was searched for the prediction model. When applying the procedure that is called "best subset selection", all possible models are fitted to the data and the best model is chosen. In high-dimensional data, computational costs are important but the issue "overfitting" is essential. Both are concerns in view of best subset selection. Hence, forward stepwise selection was applied, although this model selection technique does not guarantee to find the best model. However, this algorithm is a computationally efficient alternative and can be applied for high-dimensional data (James et al., 2015).

For the assessment of the differences in method performance and model selection, the resampling method leave-one-out cross-validation was applied. In general, the investigated data set is divided for cross-validation randomly into a training and into a validation data set. Based on the training data set, the regression model is built. This model is then used to predict the response based on the validation data set to assess the model fit on an independent data set. If the training and the validation data sets are of comparable size, the model performance measure can vary depending on the split of the observations because the distributions in the training and in the validation data sets do not need to correspond to each other. Additionally, the distribution used for model fitting can differ from the distribution of the complete data set resulting in an inappropriate model. Furthermore, the sample size in the training data set is drastically reduced (James et al., 2015). To overcome these limitations, leave-one-out cross-validation was applied. There, the validation data set only comprises one observation and all other observations belong to the training data set. This splitting is repeated until every observation once built the validation data set. This results in less bias and avoids randomness in the data set selection (James et al., 2015).

There was a linear relationship between chromosomal instability and methylation of *GNS* (Glucosamine (N-acetyl)-6-sulfatase) leading to the high consistency in model selection. Surprisingly, the chromosomal instability decreased with increasing methylation of *GNS* and there was no correlation with the corresponding gene expression. A literature search did not help to identify a biological/medical background. This gene occurs in every cell. The deficiency of *GNS* causes Sanfilippo Syndrome Type D (Mucopolysaccharidosis type IIID). Progressive neurodegeneration is the clinical feature of this disorder (Elçioglu et al., 2009).

Instability is related to DNA changes. It is known that cancer can originate from such changes. During cancer progression, these changes proliferate (van Wieringen et al., 2013). Because cancer development is a process of DNA aberrations, it seems reasonable that genomic instability is relatively independent of the size of the investigated region. This assumption was supported by the findings on the region size being practically identical for 500 kb and 1000 kb.

This investigation has strengths but also some limitations. As this was a real data application, the underlying truth was unknown. Hence, the methods could be investigated for differences and similarities in their results but a decision on the more appropriate method was not possible. The investigation was limited by a small sample size. Hence, only 600 selected genes were investigated for an association between chromosomal instability and DNA methylation. Furthermore, the chromosomal instability had to be deduced from the aCGH data. This required a definition of a threshold to define a chromosome arm as instable. However, it was observed that the chromosomal instability measure was relatively independent from this threshold. Overall, two different regression models and their robustifications could be compared to each other on a real data set with a genetically plausible question.

In summary and based on this real data application, the choice of the methods might depend on the aim of the analysis. There were differences between linear and Poisson regression with respect of the selected models but these models did not notably differ in their prediction accuracy. Hence, besides modelling accuracy, computation time should also be considered. The use of robust regression can be more expensive than standard approaches. Furthermore, the handling of numerical instabilities of robust Poisson regression is challenging.

5.1.2 Influence of one single outlier in real data

The influence of one single outlier on standard and robust logistic regression relying on the Huber function were compared investigating the relationship between hypertension and the explaining variables genotype and age. Present results confirmed that single individuals (1/130 = 0.8% of the observations) with a departing risk of hypertension may substantially affect the overall risk estimates in the baseline model causing up to 11.2% change in the estimated excess risk of hypertension per year according to standard logistic regression in the present exercise.

To investigate the influence of outliers on standard logistic regression estimates and to compare it to the handling by the robustification, one must identify the observations which influence standard estimates. Relying on residuals for outlier identification is one possibility. But there, the goal can be hampered by masking and swamping effects when residuals are used within standard logistic regression. Consequently, Cook's distances were used for this purpose. These distances directly quantify the magnitude of impact of each single observation on the estimated response variable when applying standard logistic regression. Furthermore, they simultaneously consider both observations in the independent and in the dependent variables. This identification of outliers was found to be relatively straightforward using the routine diagnostic plot for Cook's distances.

Once identified, outliers must be managed and this management is extremely challenging. Robust statistics aim at generating estimates that hold for the majority of the population using complete data sets. The unequal weighting of outliers by standard and robust regression resulted in prediction models containing different genetic variants.

When building a prediction model, there must be a measure to assess model performance (accuracy). Established techniques are the related measures concordance, sensitivity and specificity as well as AUCs. The AUC can be seen as a weighted average across all threshold values (Pencina et al., 2008). Pencina et al. (2008) proposed another weighted average measure: the integrated discrimination improvement (IDI). With respect to the mean predicted probability, this measure also quantifies an increase in sensitivity and specificity by relying on the group difference (e.g., cases versus controls). A disadvantage for clinical application of these measures is that they do not address clinical implications. To address clinical consequences, Vickers and Elkin (2006) proposed a decision curve analysis. In terminology of the analysis in this thesis, they assume that the logistic regression probability estimate is informative on balancing the impact of false negative and false positive predictions. Then, the clinical net benefit is calculated across a grid of threshold values to allocate an individual to the cases to get the decision curve considering the clinical consequences. As AUC and IDI average across the range of threshold values, it is not surprising that the cross-validation AUC and the IDI did not differ between standard and robust approaches in this analysis, whereas the non-averaging measures showed differences between the regression approaches.

The standard logistic regression model selected one variant in the *ULK4* gene. It has previously been shown that variants in this gene are associated with hypertension (Levy et al., 2009; Ho et al., 2011). Among others, four variants (rs2272007, rs3774372, rs1716975, rs1052501) mentioned in the two publications were also genotyped in the GAW 18 collective and found to be in linkage disequilibrium (r^2 values 0.83, 0.73, 0.83 and 0.83) with the associated SNP rs3934103. In contrast, a literature search for the variant in the *RP11-408H1.3* gene that was selected by robust logistic regression did not reveal an association with hypertension or blood pressure – neither for the variant nor for the gene.

In view of these results, one must consider strengths and limitations of this investigation. Common as well as not generally applied model performance measures were used to provide a sound basis for the method comparison. Nevertheless, it is important to keep in mind that the results are limited by moderate sample size, genetic effect sizes and proportion of outliers.

In summary, although robust estimates of age effects and AUCs for age-genotype models were less sensitive to outliers than standard estimates, cross-validation AUCs based on standard and robust logistic regression as well as IDIs were almost identical. The other investigated performance characteristics (concordance, sensitivity, specificity and clinical net benefit) were better for robust logistic regression around the probability that reflects the case-control-ratio. These preliminary findings indicated some advantage of robust statistics in the context of genetic association studies.

5.1.3 Influence of genotyping errors on estimates in simulated data

In simulated case-control data describing the relation between a disease status and a genotype as well as age, the influence strength of genotyping errors on standard and robust logistic regression estimates was investigated. The OR dependence on the genotyping error rate was large for rare variants and decreased with increasing causal allele frequency. Standard and robust estimates were similar for causal allele frequencies of 0.005 and 0.13 in this simulation. However, for a causal allele frequency of 0.05 the differences of the effects of genotyping errors on estimated ORs of standard and robust logistic regression increased with increasing genotyping error rate. In the latter case, robust estimates took advantage of the increasing rate.

Depending on the investigated data (e.g. its distribution, outlier characteristics), Alamgir et al. (2013) showed that there are situations where the use of the Huber function is advantageous compared to standard techniques. This is in close accordance with the simulation results – both for similar and for different strong impact of genotyping error rates on standard and robust estimates. The simulation results confirmed that the dependence of estimated ORs on genotyping errors decreased with the allele frequency (Powers et al., 2011).

This analysis has strengths but also limitations. Two parameters playing an important role in genetic association analyses, namely minor allele frequency and genotyping error rate, were varied over a wide range of values. But there are further population and genetic parameters that differ between, e.g., individuals, diseases and lifestyles (e.g. linkage disequilibrium, genotype relative risk, genetic penetrance model) and which also might have an impact on the analysis. Furthermore, only the OR change due to the genotyping errors were investigated but not statistical power, type I error rate or mean squared error.

In summary, this small simulation study indicated a possible benefit of robust logistic regression in genetic studies. However, the importance of genotyping accuracy in genetic association studies is accentuated.

5.2 Statistical properties of robust logistic regression applying the Hampel function

The aim of the present study was to investigate the benefits and limitations of robust logistic regression using the Huber and Hampel functions to down-weight outliers. After adapting the R package `robustbase` to accommodate the Hampel function, computer simulations, complemented with the analysis of a real data set, were conducted to assess the type I error rate, statistical power and MSE of standard and robust GRR estimates according to study characteristics as well as properties of the investigated markers. Both standard and robust methods controlled the type I error rates. Standard logistic regression consistently showed the highest statistical power, which was often attributable to an increased GRR overestimation in comparison to robust estimates. For rare and recessive variants, robust GRR estimates presented markedly lower biases and variances than standard GRR estimates. These results suggest that, after identification of novel susceptibility variants, robust regression may represent an interesting alternative to standard maximum likelihood estimation when the focus lies on accurate risk prediction.

As proof of concept, the simulation results confirmed that power depended on the fitted penetrance model, MAF, correlation between causal allele and marker locus, genotyping accuracy and sample size (Chen et al., 2011; Hein et al., 2008). In more detail, Hong and Park (2012) reported that the required sample size to achieve a power of 0.80 is smaller under a dominant penetrance model compared to other genetic models. Whether the indirect approach to detect a causal variant is successful depends heavily on the linkage between this variant and its marker (LD as well as correlation) (Howey and Cordell, 2014; Kraft et al., 2005). Power for this detection increases with increasing causal allele frequency and increasing linkage between marker and causal allele locus (Hein et al., 2008; Lin and Schaid, 2009). Even small genotyping error rates can seriously reduce power which can be further amplified if the MAF decreases (Powers et al., 2011).

Results from computer simulations revealed a power-variance paradox in standard versus robust GRR estimation. Often, the smaller the variance of a parameter estimate the larger the statistical power to reject the null hypothesis given the alternative hypothesis. This is, however, only the case for unbiased or equally biased parameter estimators. In general, and in particular for rare and recessive variants, it was found that larger variances usually came along with larger positive biases resulting in a higher statistical power for standard than for robust logistic regression. This situation was especially evident for the simulated recessive variant with a MAF equal to 0.05 and a true GRR equal to 6.32. The biases were $+1.2$ for standard compared to -0.1 (12 times lower) for Huber GRR estimates. Variances were 15 for standard compared with 0.5 (30 times lower) for Huber GRR estimates. In contrast, the statistical power was 0.6 for standard versus only 0.51 when the Huber function was used to constrain outlier influence in a study with 1000 cases and 1000 controls.

In agreement with computer simulations, the analysis of real data confirmed that standard logistic regression can be strongly influenced by single or few outliers, which may inflate estimated genetic effects. For example, the GRR estimated by standard logistic regression for SNP rs2500262 was 13.7 and, thus, 1.7 to 2.7 larger than the corresponding robust estimates – likely due to the influence of one outlier (Cooks distance about 0.4). This strong impact of few or even one outlier on standard logistic regression and their handling by robust approaches are in close accordance with observations from real data applications – Hosseinian and Morgenthaler (2011) as well as the investigation of the influence of a single outlier in this thesis. Large differences between estimated standard and robust GRRs as well as non-convergence of robust procedures may be indicative of the presence of departing observations. A practical recommendation of this study is to thoroughly inspect diagnostic plots when this happens.

Present results are relevant to genome-wide association studies (GWASs) where the "winner's curse" is a major issue (Göring et al., 2001; Hirschhorn et al., 2002; Zöllner and Pritchard, 2007). GWAS results can be strongly affected by ascertainment bias leading to effect overestimation. Since sample size calculation for adequately powered replication studies relies on possibly biased initial findings, the necessary sample size can be underestimated causing replication failure. In a simulation study, Zöllner and Pritchard (2007) observed that genetic effects were overestimated by about 20% in the absence of correction for ascertainment bias. Investigations revealed that ascertainment bias is particularly large when the power is small and ascertainment bias disappears when the statistical power approaches one (Xiao and Boehnke, 2009; Garner, 2007). Several methods have been proposed to deal with the winner's curse in linkage analysis, some of which could be extended to association studies. Göring et al. (2001) concluded that large, population-based samples of persons recruited independently of their phenotype would alleviate this issue. However, it is not clear if this is also true for association analysis (Zöllner and Pritchard, 2007). Based on a maximum likelihood which explicitly considers genome-wide scans, Zöllner and Pritchard (2007) proposed an ascertainment bias correction that tends to underestimate the true effect addressing the winner's curse. Two different conditional likelihood approaches have been proposed for point and interval estimators in GWAS (Zhong and Prentice, 2008; Ghosh et al., 2008). In this context, it is of special interest that the MSE of robust GRR estimates was smaller than of standard estimates in the simulations. So, robust logistic regression might be beneficial, especially for rare and recessive susceptibility variants as well as for variants with low penetrances narrowing down the winner's curse. This might translate into an increased replication rate of initial findings.

The bias-variance trade-off is another important aspect to consider in close relation to the MSE. Estimators are constructed in a way to describe the target variable best. One accuracy measure is the MSE which is the sum of the squared bias and the variance. The bias indicates how closely the estimator determines the target variable on average. A small variance accompanies an estimator that is stable against sampling variations (Friedman, 1997). Hence, it is desirable to have both a small bias and a small variance causing a small MSE. But in most situations, a bias decrease often results in an increased variance (Friedman, 1997; Geman et al., 1992). Hence, it is not clear whether an unbiased estimator is really the major aim because this does not guarantee minimisation of the estimation error (Kohavi and Wolpert, 1996). The variance decreases with increasing sample size so that the bias is the major component of the MSE for common genetic variants (Friedman, 1997). If the sample size is relatively small, a balance between small bias and small variance has to be found when building estimators to get a minimal (or small enough) MSE. In the simulation study, robust logistic regression controlled better the bias-variance trade-off than standard logistic regression for rare and recessive variants.

Building a phenotype prediction model relying on GWAS results is often used to identify persons at a high risk of a given disease. There are many limitations and pitfalls when building such a prediction model. Most limitations relate to availability of data and background knowledge, e.g. data sets with many genotyped markers possibly in LD with causal variants, data sets with cryptic relationships and differences in stratification between discovery/validation and target population as well as environmental factors resulting in stochastic events (Burga et al., 2011). A special issue are rare variants whose contributions might not be tagged by genotyped SNPs (Yang et al., 2010; Visscher et al., 2012). However, this has changed with advances in whole-genome sequencing. Once detected, rare variants can be included in prediction models in the same way as common variants and in sum their contribution might be relevant (Wray et al., 2013). The effects of rare and common variants on a phenotype can only be estimated with an error. This plays a more important role if effect sizes are small because large sample sizes are needed for sufficient accuracy. In this context, robust logistic regression might be relevant. It was found that robust GRR estimates were more accurate than standard counterparts in this simulation study.

While robust logistic regression might be beneficial regarding prediction based on rare and recessive variants, with respect to the winner's curse and the bias-variance trade-off, the advantage over standard GRR estimates depends on study characteristics as well as on the properties of associated variants. This conclusion is in close accordance with Çetin and Erar (2006) which considered variable selection in robust linear regression. Within this context it is of interest that the method performance can be influenced by sample size as well as by the outlier distribution and proportion, as reported by Wen et al. (2013) in their investigation of outlier impact on net-benefit regression models in cost-effectiveness analysis. Alamgir et al. (2013) and Muthukrishnan and Radha (2010) also reported that the comparative performance of the Hampel and the Huber function depend on investigated data and outlier characteristics. As the simulations and the illustrative example in section 3.3 showed, robust approaches might be even more useful in rare variant settings. In the illustrative example, a clear advantage was observed for robust logistic regression and especially for the use of the Hampel function when extreme outliers were present. This is in agreement with literature on the use of re-descending weighting functions (Müller, 2004; Shevlyakov et al., 2008). Several re-descending influence functions are available, such as the three-part Hampel, the biweight Tukey and the sine-wave Andrews function (Hampel et al., 1986; Beaton and Tukey, 1974; Andrews et al., 1972). Among them, the Hampel function seems to perform well in most situations (Alamgir et al., 2013; Andrews et al., 1972).

Another issue are the computational costs. Standard logistic regression needed on average about 9.5 ns for one model estimation as it is applied in the reference scenario of the simulation study (source: function `microbenchmark` of the R package `microbenchmark` (Mersmann, 2014)). The robust logistic regression approaches needed several times longer (Huber: about 3 times, Hampel: about 6 times).

A literature search in PubMed (`http://www.ncbi.nlm.nih.gov/pubmed`) on body height and rs7519458 as well as the corresponding gene symbols (LINC01346, LOC105376672) did not reveal any findings. There were no results in "Genopedia" of the HuGE Navigator (Yu et al., 2008), neither. Both were accessed at 2016/06/21.

The present study has strengths but also limitations. It was made an effort to simulate realistic data. Age and disease prevalence were based on real demographic data. To verify the bias and variance advantage of robust logistic regression for rare and recessive variants, the effect and sample sizes were varied. One limitation was the use of just one weight function and two bounded influence functions. Different combination could be investigated in future studies. But with respect to the similar results for the bounded Huber and the re-descending Hampel function in case of probably no extreme outlier, the main task is the decision whether extreme outliers are expected in the data. Furthermore, tuning constants were used that assure 95% asymptotic efficiency in linear models. The application of tuning constants that assure this efficiency for logistic regression models might be worth to consider. Here, it was focused on logistic regression but the generalisation of current results to other analytical approaches in statistical genetics, for example collapsing methods, is straightforward.

In conclusion and based on these analyses, the potential advantage of robust GRR estimates depends on the study aim – identification or characterisation of genetic effects. To achieve a large power, standard logistic regression is the best choice. For sufficiently large sample sizes, the use of robust logistic regression is recommended with regard to small bias, variance and MSE alleviating effect overestimation – especially when analysing rare variants and assuming a recessive penetrance model. Robust GRR estimation is computationally demanding, in particular for the Hampel function. On the other hand, the Hampel function may minimise biases when strongly departing outliers are present. An added value of the present study rests on demonstrating the use of an alternative influence function in the logistic regression framework proposed by Cantoni and Ronchetti to narrow down the winner's curse of rare and recessive susceptibility variants.

5.3 Recent updates of the R package `robustbase`

Meanwhile, the R package `robustbase` has been updated with version 0.92-6 (date: 2016/05/28) by 2016/06/19. Additional influence functions have been added to the function `glmrob` which was applied for both robust Poisson and robust logistic regression models as dealt within this thesis.

Additionally, the unweighted and weighted Bianco-Yohai estimators were available (Croux and Haesbroeck, 2003). The ψ-function of the unweighted Bianco-Yohai estimator (Bianco and Yohai, 1996) is defined as

$$\psi_{Bianco-Yohai}(r) = \begin{cases} 1 - \frac{r}{c} & \text{if } r \leq c \\ 0 & \text{otherwise} \end{cases}$$

with $c > 0$. This estimator is consistent and asymptotically normal. To get the weighted Bianco-Yohai estimator, an additional weighting function based on a robust distance measure is integrated into the algorithm. According to Croux and Haesbroeck (2003), an established choice for this weight function W is

$$W(r) = \begin{cases} 1 & \text{if } r^2 \leq \chi^2_{p,0.975} \\ 0 & \text{otherwise} \end{cases}$$

with the number of independent variables p. The "M Estimator based on Transformation" was currently (2016/06/19) only available for Poisson regression (Valdora and Yohai, 2014).

Among the newly implemented influence functions for logistic regression, only the unweighted Bianco-Yohai estimator would be applicable for logistic regression investigating the association between a binary outcome and a non-continuous independent variable (e.g. a genotype). The reason was that the weighted Bianco-Yohai estimator was calculated based on the Mahalanobis distance in this package. This distance is defined for individual i $(i = 1, \ldots, n)$ as

$$\sqrt{(x_i - \mu)\, S^{-1}\, (x_i - \mu)^T}$$

where $x_i \in \mathbb{R}^p$ is a row of the $(n \times p)$-dimensional matrix X, p denotes the number of measured variables, $\mu \in \mathbb{R}^p$ describes the mean values across the individuals and S is the covariance matrix (De Maesschalck et al., 2000). The application of this distance is only reasonable for continuous variables (Heritier et al., 2009).

For a short illustration of the unweighted Bianco-Yohai estimator of the `robustbase` package, the analysis of simulated data underlying figure 7 on page 76 was extended by the application of this estimator (figure 23). There, the disease status was regressed on age and the genotype of a rare variant in a dominant penetrance model. Hence, the second independent variable is binary. The Bianco-Yohai estimator was compared to the application of the Huber and the Hampel function as well as to standard logistic regression. In this small simulation study, the results of the robust logistic regression applying the unweighted Bianco-Yohai estimator were relatively similar to the results of standard logistic regression but they clearly differed for extreme outliers from the results when applying the Huber or the Hampel function. Both the Hampel and the Huber function better controlled the outlier impact. The similarity between the unweighted Bianco-Yohai estimator and the standard logistic regression might result from the characteristic that the ψ-function of the Bianco-Yohai estimator is bounded but still returns large values for outliers (Croux and Haesbroeck, 2003). Hence, only the influence of very extreme outliers are bounded by this estimator. This observation is also in accordance with the report by Hauser and Booth (2011) who compared the unweighted Bianco-Yohai estimator to the maximum likelihood estimator using a logistic regression analysis to predict bankruptcy based on five financial ratios (Altman, 1968). They concluded that the robust estimator could improve prediction and classification and produced at worst similar results as achieved by the maximum likelihood estimator. Consequently, they suggested to use the Bianco-Yohai estimator in logistic regression analysis as robustness check. Based on the small simulation exercise, the Huber and the Hampel function should be favoured if outliers are expected in the data.

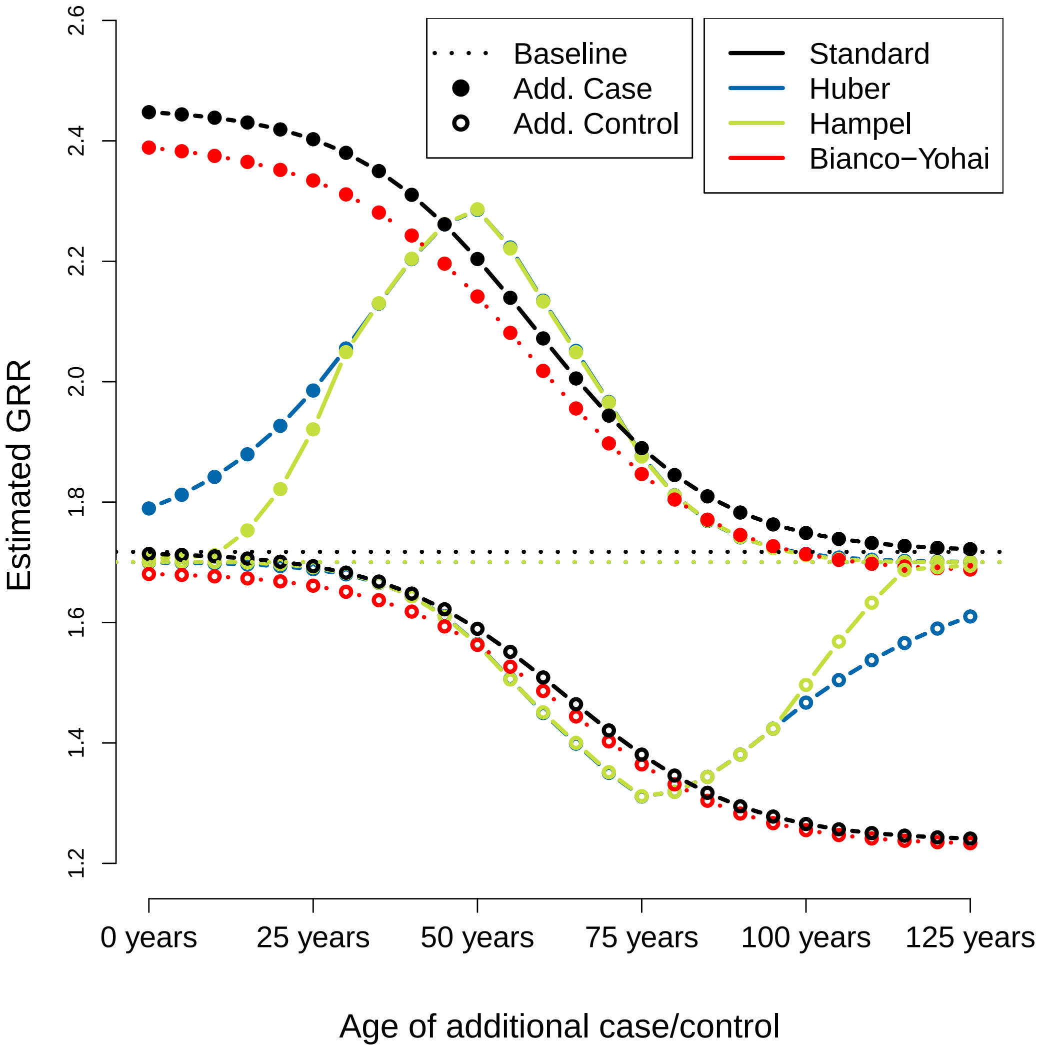


Figure 23: Influence of outliers on standard and robust estimates of the genotype relative risk (GRR). The influence was examined by including single cases and controls that carried the high-risk variant to the baseline data set with 1000 cases and 1000 controls, a dominant GRR of about 2 and a minor allele frequency (MAF) of 0.0075. The fitted logistic regression model was disease status (case/control) explained by genotype and age.

5.4 Final conclusion and perspective

Overall summary

The aim of this thesis was threefold. First, the comparison of standard and existing robust regression methods should provide theoretical and practical insight into the capabilities of these methods. For this purpose, different analysis aims were pursued in simulated and in real data. The second aim was to adapt the already existing framework for robust logistic and Poisson regression proposed by Cantoni and Ronchetti (2001) to practically apply it with the Hampel function for outlier weighting in addition to the Huber function. The resulting algorithms were successfully checked on plausibility. Then, the extended approach should be compared to standard and existing robust regression applying the Huber function. For logistic regression, this was done on simulated data and in a real data application. In brief, the main observations were:

- Model selection is influenced by the different observation weighting in standard and robust regression methods.
- Already one single outlier can have a large impact on estimates, especially on standard estimates.
- The statistical power of standard logistic regression is larger than the power of robust logistic regression.
- Estimates of robust logistic regression were less biased and had smaller variances causing smaller MSEs. This especially applied to rare variants and to a fitted recessive penetrance model.

These results demonstrated that robust generalised linear models can be advantageous as compared to standard generalised linear models but it always depended on the analysis' aim (e.g., identification or characterisation) and the underlying data structure (e.g., MAF or penetrance model). Çetin and Erar (2006) and Wen et al. (2013) stated this in similar circumstances.

Overall strengths and limitations

Overall, this work has several strengths but also some limitations. The capabilities of the several different robust regression models were investigated and related to the performance of the corresponding standard regression methods leading to a complex knowledge about advantages and disadvantages of robust regression. Furthermore, the effect of different influence functions on the robust estimation process could be compared within one theoretical framework because a second function for outlier weighting (namely the Hampel function) was implemented into an already existing robust logistic regression framework applying the Huber function. Additionally, the different regression models were inspected with respect to their reaction on different influence sources (only one outlier, genotyping errors, population and genetic characteristics) as well as applied with different analysis aims (identification, characterisation and prediction). Due to time constraints, robust Poisson regression applying the Hampel function has not been compared to standard and robust Poisson regression applying the Huber function, yet.

Perspective

For the Hampel function in robust Poisson regression, a first plausibility check already suggested an adequate functionality of this influence function in Poisson regression. As a future step, it is of interest to investigate the Hampel function in the context of Poisson regression in more detail on simulated data with respect to mean squared error, statistical power and type I error rate as well as in real data applications. Poisson regression is used to test for association of a countable response variable with explanatory variables, e.g. the relationship between the number of variants within one gene and the left ventricle ejection fraction in patients with dilated cardiomyopathy (DCM), the chromosomal instability depending on DNA methylation or length of hospital stay (in days) in relation to the disease severity and patient's age. To decide on simulation scenarios, real data has to be examined to get realistic explanatory values and Poisson parameter (λ) for the count variable depending on the explanatory value.

Subsequent to realisation of the performance of robust logistic and robust Poisson regression, the combination of these two regression types leads to hurdle models which are an example for two-part models. These models can be used for zero-inflated data, e.g. methylation data (Mullahy, 1986; Zeileis et al., 2007). Depicting the idea of such a model by an example leads to a decision making process. In the first step, the decision is taken whether to do something or not. Then in case of a positive decision, it is decided on how often. Logistic regression is used in the first step, i.e. to check whether an event has a zero or a positive outcome. If the value is positive, the truncated Poisson model estimates the positive count value (Duan et al., 1983; Min and Agresti, 2002; Cantoni and Zedini, 2011). Cantoni and Zedini (2011) proposed a robust version of the hurdle model based on the results of Cantoni and Ronchetti (2001). This method was again explicitly applied by using the Huber function to weight outliers in the response. Regarding the results of this thesis, it would be interesting to investigate this concept relying on the Hampel function. Therefore, the truncated Poisson part remains to be derived in close analogy to Cantoni and Zedini (2011). For the calculation see appendix A on page 119. To compare this method with existing methods, one must implement it in a functional language, for example in R (R Core Team, 2013). Thereafter, this method should again be tested on simulated data and in a real data application.

Besides binomial and Poisson distributions, the investigation of further distributions (e.g. Gamma with parameters ν and λ_i) would be of great interest. For a calculation example see Appendix A of Cantoni and Ronchetti (2006). Gamma distribution can be used in situations with a positive outcome where the data is highly skewed and the outcome belongs to the exponential family. The expectation μ_i is given by $\frac{\nu}{\lambda_i}$ and the variance σ^2 is equal to $\frac{1}{\nu}$. Cantoni and Ronchetti (2006) used the Gamma distribution in generalised linear models with the logarithmic link function to model the cost of staying in a hospital considering different explanatory variables (e.g. length of stay, insurance, sex or age).

Compared to the Hampel and the Huber function in some circumstances, Çetin and Erar (2006) already showed some advantage for Andrews' M-estimator in variable selection for linear regression models and Wen et al. (2013) for Tukey's M-estimator in net-benefit regression models. Furthermore, Alamgir et al. (2013) stated that the Hampel function has the disadvantage not to be differentiable and a smooth differentiable influence function might be desired. Thus, the theoretical development and practical implementation for additional weighting functions such as Andrews' sine wave or Tukey's biweight function would be desirable although the benefit is unclear due to the observed small differences between the use of the Hampel and the Huber function. The small difference is in accordance with Alamgir et al. (2013). But one should take this effort to probably increase statistical power without decreasing the prediction accuracy. The incorporation of the Tukey and the Andrews function into the framework proposed by Cantoni and Ronchetti (2001) will be challenging due to the weighting function structures. Calculations for the Tukey function are extensive due to several squared expressions. But feasibility increases by using polynomial long division to get a similar device as in equation (3.6). The Andrews function probably provides a different kind of time consuming task caused by the sine function. This function can be written as infinite sum but this approach just transfers the problem arising during calculation of expectations. If this infinite sum could be well approximated by a finite sum, the calculation would be manageable.

Obviously, robust generalised linear models are not only of interest as stand-alone approaches. Many algorithms rely on standard generalised linear models. There, the extension by the robustification might be beneficial. An example could be the approaches proposed by Houseman et al. (2014) and Zou et al. (2014) to analyse high-throughput DNA methylation data accounting for the cell type distribution in the sample tissues. This cell type distribution consideration is necessary because the cell type distribution is a possible confounder in such an association analysis due to its possible association with both DNA methylation and the trait of interest. In real data applications, one observed that there was an excess of small p-values when using the approach by Houseman et al. (2014) and that there was only a small overlap in the results of these two approaches (Kesselmeier et al., accepted; Kesselmeier and Scherag, in preparation). This can be caused by extreme observations that are expected in high-dimensional data. Hence, the use of a robust linear (mixed) model might be worth to consider for more consistent results.

In conclusion, the successful implementation and application of the Hampel function into the robust logistic regression framework proposed by Cantoni and Ronchetti (2001) suggests further research on this topic with respect to, e.g., different distributions and weighting functions for a broader application range and the chance of a power superiority for robust regression methods compared to standard regression approaches.

6 Summary

In genetic studies, data under investigation exhibit a high-dimensionality, i.e., there are many more independent variables than measured individuals. In high-dimensional data, one expects observations departing from the majority of the data (so-called outliers). Such outliers can seriously affect statistical results because applied approaches using maximum likelihood estimation can be strongly biased by outliers. Robust approaches account for such outliers by assigning a weight to each observation, thus controlling their impact. However, these approaches are only rarely used in genetic studies.

In this thesis, benefits and limitations of robust (generalised) linear models in comparison to the standard maximum likelihood approaches were investigated. For this purpose, an existing robust generalised linear model framework was generalised to incorporate another weighting function.

In a first set of analyses, several already existing standard and robust approaches for linear, Poisson as well as logistic regression were compared. There, the attention was drawn to model selection consistency and prediction accuracy, the influence of a single outlier and the influence of genotyping errors on estimates. The prediction accuracy was similar for (robust) linear and (robust) Poisson regression models in a real data application. In view of model selection consistency, Poisson regression selected two or three independent variables whereas linear regression always included the same single independent variable, which was, however, in common for all regression methods. These results were complemented by an inclusion of different independent variables into the standard and robust logistic regression models in a second real data application. Within this application, it was observed that robust logistic regression better controlled the outlier influence. A simulation study revealed a decreasing influence of genotyping errors on estimates with increasing causal allele carrier frequencies. Furthermore, there was an indication of a possible benefit of robust logistic regression.

At the time of method application, the robust generalised linear model framework only provided the bounded Huber function for observation weighting. In this thesis, the re-descending Hampel function was incorporated into this framework for logistic and Poisson regression by explicit calculations for the Fisher consistency correction and for the asymptotic variance as well as by adaptation of the existing source code. In a second set of analyses, the developed approach for robust logistic regression was compared against the standard and the existing robust logistic regression methods based on simulated and real data – both dealing with an (indirect) association analysis. In the simulation study, several populations were simulated assuming different penetrance models, minor allele frequencies, genotyping error rates and linkage between causal and marker allele locus. In the analysis, the attention was drawn to several statistical properties comprising mean squared error of the estimates, statistical power and type I error rate. In the simulation study, all approaches controlled the type I error rate. Based on the results of the statistical properties investigation, a method recommendation must depend on the aim of the analysis. To reach a large power for variant identification, standard logistic regression would be an adequate choice. If a small mean squared error probably avoiding a strong effect overestimation was the goal, robust logistic regression represented a valuable alternative to the standard approach. This especially held when analysing rare variants or assuming a recessive penetrance model both leading to a low probability to observe the causal genotype. If extreme outliers are expected in the data, the re-descending Hampel function should be favoured.

The aim for future work should be the examination of statistical properties (mean squared error, statistical power, type I error rate) of robust Poisson regression and of the robust hurdle model arising by the combination of the logistic and the truncated Poisson model – both applying the Hampel function. Additionally, an inclusion of further weighting functions as well as additional distributions would be of great interest for a broader application range and the chance of a power gain for robust regression methods.

Summarising, the coincidence of expected outliers and observed rare events in high-dimensional data challenges the analysis of genetic data. The results of this thesis indicate that these analyses can benefit from the application of robust logistic regression models to narrow down the winner's curse of rare and recessive susceptibility variants.

A On calculations to adapt the robust hurdle model to the use of the Hampel function

To build the logistic regression part of the hurdle model based on the Hampel function, one applies the formulas for the logistic regression developed in section 3.1.1. For the truncated Poisson regression part, one must calculate the three expectations needed to weight observations in the truncated Poisson model similarly to the Poisson model in section 3.1.2.

For the purpose of deducing the truncated Poisson part, let u_i be the covariate vector used within the truncated Poisson model. This vector can equal the covariate vector of the logistic regression part but it is not mandatory. Then, it holds for the expectation of the truncated-Poisson distributed random variable Y_i given the covariate vector u_i that

$$\mathrm{E}[Y_i|u_i] = \mu_i = \frac{\lambda_i}{1-e^{-\lambda_i}}$$

with Poisson parameter λ_i (Cantoni and Zedini, 2011). Note that

$$\begin{aligned} j\,\mathrm{P}[Y_i = j|j>0] &= \frac{\lambda_i}{1-e^{-\lambda_i}}\,\mathrm{P}[\tilde{Y_i} = j-1] \\ &= \mu_i\,\mathrm{P}[\tilde{Y_i} = j-1] \end{aligned}$$

and

$$\begin{aligned} j(j-1)\,\mathrm{P}[Y_i = j|j>0] &= \lambda_i\,\frac{\lambda_i}{1-e^{-\lambda_i}}\,\mathrm{P}[\tilde{Y_i} = j-2] \\ &= \lambda_i\,\mu_i\,\mathrm{P}[\tilde{Y_i} = j-2] \end{aligned}$$

with $Y_i \sim \mathrm{Poi}_{\mathrm{trunc}}(\lambda_i)$ (truncated Poisson distribution) and $\tilde{Y_i} \sim \mathrm{Poi}(\lambda_i)$. Then, the expectation for the Fisher consistency correction equals

$$\begin{aligned}
&\mathrm{E}\left[\psi_{\text{Hampel}}\left(\frac{Y_i-\mu_i}{V^{1/2}(\mu_i)}\right)\right] \\
&\quad= \frac{\mu_i}{V^{1/2}(\mu_i)}\left(\mathrm{P}\left[j_{a_1} \leq \tilde{Y}_i \leq j_{a_2}-1\right] - \mathrm{P}\left[j_{a_1}+1 \leq Y_i \leq j_{a_2}\right]\right) \\
&\qquad + a\left(\mathrm{P}\left[j_{a_2}+1 \leq Y_i \leq j_{b_2}\right] - \mathrm{P}\left[j_{b_1}+1 \leq Y_i \leq j_{a_1}\right]\right) \\
&\qquad + \frac{a}{c-b}\Big\{c\left(\mathrm{P}\left[j_{b_2}+1 \leq Y_i \leq j_{c_2}\right] - \mathrm{P}\left[j_{c_1}+1 \leq Y_i \leq j_{b_1}\right]\right) \\
&\qquad - \frac{\mu_i}{V^{1/2}(\mu_i)}\Big(\mathrm{P}\left[j_{b_2} \leq \tilde{Y}_i \leq j_{c_2}-1\right] - \mathrm{P}\left[j_{b_2}+1 \leq Y_i \leq j_{c_2}\right] \\
&\qquad + \mathrm{P}\left[j_{c_1} \leq \tilde{Y}_i \leq j_{b_1}-1\right] - \mathrm{P}\left[j_{c_1}+1 \leq Y_i \leq j_{b_1}\right]\Big)\Big\}
\end{aligned}$$

Besides Fisher consistency correction, one needs the two expectations for the asymptotic variance. They equal

$$\begin{aligned}
&\mathrm{E}\left[\psi^2_{\text{Hampel}}\left(\frac{Y_i-\mu_i}{V^{1/2}(\mu_i)}\right)\right] \\
&\quad= \frac{\mu_i}{V(\mu_i)}\Big(\mu_i\,\mathrm{P}\left[j_{a_1}+1 \leq Y_i \leq j_{a_2}\right] + \lambda_i\,\mathrm{P}\left[j_{a_1}-1 \leq \tilde{Y}_i \leq j_{a_2}-2\right] \\
&\qquad +(1-2\mu_i)\,\mathrm{P}\left[j_{a_1} \leq \tilde{Y}_i \leq j_{a_2}-1\right]\Big) \\
&\qquad + a^2\left(\mathrm{P}\left[j_{a_2}+1 \leq Y_i \leq j_{b_2}\right] + \mathrm{P}\left[j_{b_1}+1 \leq Y_i \leq j_{a_1}\right]\right) \\
&\qquad + \frac{a^2}{(c-b)^2}\Big\{c^2\left(\mathrm{P}\left[j_{b_2}+1 \leq Y_i \leq j_{c_2}\right] + \mathrm{P}\left[j_{c_1}+1 \leq Y_i \leq j_{b_1}\right]\right) \\
&\qquad - \frac{2c\mu_i}{V^{1/2}(\mu_i)}\Big(\mathrm{P}\left[j_{b_2} \leq \tilde{Y}_i \leq j_{c_2}-1\right] - \mathrm{P}\left[j_{b_2}+1 \leq Y_i \leq j_{c_2}\right] \\
&\qquad - \mathrm{P}\left[j_{c_1} \leq \tilde{Y}_i \leq j_{b_1}-1\right] + \mathrm{P}\left[j_{c_1}+1 \leq Y_i \leq j_{b_1}\right]\Big) \\
&\qquad + \frac{\mu_i}{V(\mu_i)}\Big[\mu_i\left(\mathrm{P}\left[j_{b_2}+1 \leq Y_i \leq j_{c_2}\right] + \mathrm{P}\left[j_{c_1}+1 \leq Y_i \leq j_{b_1}\right]\right) \\
&\qquad + \lambda_i\left(\mathrm{P}\left[j_{b_2}-1 \leq \tilde{Y}_i \leq j_{c_2}-2\right] + \mathrm{P}\left[j_{c_1}-1 \leq \tilde{Y}_i \leq j_{b_1}-2\right]\right) \\
&\qquad +(1-2\mu_i)\left(\mathrm{P}\left[j_{b_2} \leq \tilde{Y}_i \leq j_{c_2}-1\right] + \mathrm{P}\left[j_{c_1} \leq \tilde{Y}_i \leq j_{b_1}-1\right]\right)\Big]\Big\}
\end{aligned}$$

and

$$
\begin{aligned}
&\mathrm{E}\left[\psi_{\text{Hampel}}\left(\frac{Y_i-\mu_i}{V^{1/2}(\mu_i)}\right)\frac{Y_i-\mu_i}{V(\mu_i)}\right]\\
&\quad=\frac{\mu_i}{V^{3/2}(\mu_i)}\Big\{\mu_i\,\mathrm{P}\left[j_{a_1}+1\le Y_i\le j_{a_2}\right]+\lambda_i\,\mathrm{P}\Big[j_{a_1}-1\le\tilde{Y}_i\le j_{a_2}-2\Big]\\
&\qquad+(1-2\mu_i)\,\mathrm{P}\Big[j_{a_1}\le\tilde{Y}_i\le j_{a_2}-1\Big]\Big\}\\
&\qquad+\frac{a\mu_i}{V(\mu_i)}\Big(\mathrm{P}\Big[j_{a_2}\le\tilde{Y}_i\le j_{b_2}-1\Big]-\mathrm{P}\left[j_{a_2}+1\le Y_i\le j_{b_2}\right]\\
&\qquad-\mathrm{P}\Big[j_{b_1}\le\tilde{Y}_i\le j_{a_1}-1\Big]+\mathrm{P}\left[j_{b_1}+1\le Y_i\le j_{a_1}\right]\Big)\\
&\qquad+\frac{a\mu_i}{(c-b)V(\mu_i)}\Big\{c\Big(\mathrm{P}\Big[j_{b_2}\le\tilde{Y}_i\le j_{c_2}-1\Big]-\mathrm{P}\left[j_{b_2}+1\le Y_i\le j_{c_2}\right]\\
&\qquad-\mathrm{P}\Big[j_{c_1}\le\tilde{Y}_i\le j_{b_1}-1\Big]+\mathrm{P}\left[j_{c_1}+1\le Y_i\le j_{b_1}\right]\Big)\\
&\qquad-\frac{\mu_i}{V^{1/2}(\mu_i)}\Big[\mu_i\left(\mathrm{P}\left[j_{b_2}+1\le Y_i\le j_{c_2}\right]-\mathrm{P}\left[j_{c_1}+1\le Y_i\le j_{b_1}\right]\right)\\
&\qquad+\lambda_i\Big(\mathrm{P}\Big[j_{b_2}-1\le\tilde{Y}_i\le j_{c_2}-2\Big]-\mathrm{P}\Big[j_{c_1}-1\le\tilde{Y}_i\le j_{b_1}-2\Big]\Big)\\
&\qquad+(1-2\mu_i)\Big(\mathrm{P}\Big[j_{b_2}\le\tilde{Y}_i\le j_{c_2}-1\Big]-\mathrm{P}\Big[j_{c_1}\le\tilde{Y}_i\le j_{b_1}-1\Big]\Big)\Big]\Big\}
\end{aligned}
$$

B Supplemental tables

This chapter provides supplemental tables of the investigation of the statistical properties with standard and robust logistic regression methods with both influence functions. The following tables are provided:

- Type I error rates (table S1 on page 124)
- Bias, variance, MSE and statistical power . . .
    - . . . of the main scenarios – one table per method and tuning constant (tables S2-S5 on pages 125-131)
    - . . . of the additional scenario “rare variants” – one table per method and tuning constant (tables S6-S9 on pages 133-136)
    - . . . of the additional scenario “underlying recessive penetrance model” – one table per method and tuning constant (tables S10-S13 on pages 137-140)
- SNP and sample identifier of the real data application (table S14 on page 141 and table S15 on page 142)

Table S1: Type I error rates from standard and robust logistic regression analyses applying the Huber as well as the Hampel function for observation weighting under the null. The assumed parameters were uniformly distributed minor allele frequency between 0.05 and 0.50 and a dominant genotype relative risk of 1. In the reference simulation scenario, a fitted dominant penetrance model was used to evaluate 400 simulated studies with 1000 cases and 1000 controls. Tuning constants for robust methods are shown in brackets in the table header.

Investigated	Parameter	Standard	Huber	Hampel	Hampel
scenario	changed	[–]	[1.345]	$[(1.5, 3.5, 8) \cdot 0.9]$	$[(2, 4, 8) \cdot 0.7]$
Reference	–	0.054 (0.047, 0.061)	0.051 (0.044, 0.058)	0.050 (0.043, 0.057)	0.048 (0.041, 0.055)
Fitted penetrance	Additive	0.053 (0.046, 0.060)	0.052 (0.045, 0.059)	0.051 (0.044, 0.058)	0.049 (0.042, 0.056)
model	Recessive	0.044 (0.038, 0.050)	0.046 (0.040, 0.052)	0.045 (0.039, 0.051)	0.044 (0.038, 0.050)
Number of	100	0.048 (0.035, 0.061)	0.051 (0.037, 0.065)	0.049 (0.036, 0.062)	0.046 (0.033, 0.059)
simulated studies	200	0.053 (0.043, 0.063)	0.055 (0.045, 0.065)	0.053 (0.043, 0.063)	0.050 (0.040, 0.060)
	300	0.055 (0.047, 0.063)	0.053 (0.045, 0.061)	0.052 (0.044, 0.060)	0.050 (0.042, 0.058)
	500	0.053 (0.047, 0.059)	0.050 (0.044, 0.056)	0.049 (0.043, 0.055)	0.048 (0.042, 0.054)
	600	0.051 (0.045, 0.057)	0.050 (0.044, 0.056)	0.049 (0.044, 0.054)	0.047 (0.042, 0.052)
	700	0.051 (0.046, 0.056)	0.051 (0.046, 0.056)	0.050 (0.045, 0.055)	0.048 (0.043, 0.053)
	800	0.051 (0.046, 0.056)	0.050 (0.045, 0.055)	0.049 (0.044, 0.054)	0.047 (0.042, 0.052)
	900	0.051 (0.046, 0.056)	0.050 (0.045, 0.055)	0.049 (0.045, 0.053)	0.047 (0.043, 0.051)
	1000	0.051 (0.047, 0.055)	0.050 (0.046, 0.054)	0.049 (0.045, 0.053)	0.047 (0.043, 0.051)
Study size ⋆	200	0.057 (0.050, 0.064)	0.056 (0.049, 0.063)	0.055 (0.048, 0.062)	0.054 (0.047, 0.061)
	400	0.055 (0.048, 0.062)	0.050 (0.043, 0.057)	0.049 (0.042, 0.056)	0.049 (0.042, 0.056)
	600	0.050 (0.043, 0.057)	0.052 (0.045, 0.059)	0.051 (0.044, 0.058)	0.050 (0.043, 0.057)
	800	0.048 (0.041, 0.055)	0.048 (0.041, 0.055)	0.047 (0.040, 0.054)	0.047 (0.040, 0.054)
	1000	0.043 (0.037, 0.049)	0.043 (0.037, 0.049)	0.042 (0.036, 0.048)	0.042 (0.036, 0.048)
	1200	0.050 (0.043, 0.057)	0.045 (0.039, 0.051)	0.044 (0.038, 0.050)	0.043 (0.037, 0.049)
	1400	0.049 (0.042, 0.056)	0.050 (0.043, 0.057)	0.049 (0.042, 0.056)	0.048 (0.041, 0.055)
	1600	0.053 (0.046, 0.060)	0.051 (0.044, 0.058)	0.050 (0.043, 0.057)	0.049 (0.042, 0.056)
	1800	0.053 (0.046, 0.060)	0.050 (0.043, 0.057)	0.049 (0.042, 0.056)	0.048 (0.041, 0.055)

⋆ Study size = number of cases + number of controls (balanced groups)

Table S2: Standard logistic regression: Bias, variance and mean squared error (MSE) of estimated genotype relative risk (GRR) and statistical power. The assumed parameters under the reference simulation scenario were minor allele frequency (MAF) = 0.05, dominant GRR = 1.43 age-independent, $D' = 1$, $r^2 = 1$, no genotyping errors, 400 simulated studies with 1000 cases and 1000 controls and a fitted dominant penetrance model.

Investigated scenario	Parameter changed	Bias	Variance	MSE	Power
Reference	–	0.0095	0.0307	0.0308	0.598
Fitted penetrance model	Additive	−0.0089	0.0292	0.0293	0.578
	Recessive	0.4635	19.2920	19.5068	0.020
MAF	0.001	0.4688	23.521	23.7408	0.018
	0.005	0.0113	0.2342	0.2343	0.090
	0.010	0.0148	0.1299	0.1301	0.162
	0.100	0.0124	0.0183	0.0185	0.800
	0.150	0.0030	0.0122	0.0122	0.890
	0.200	0.0093	0.0114	0.0115	0.932
	0.250	0.0109	0.0109	0.0110	0.928
r^2	0.9	−0.0153	0.0279	0.0281	0.548
	0.8	−0.0470	0.0252	0.0274	0.495
	0.7	−0.0803	0.0265	0.0329	0.438
	0.6	−0.1107	0.0192	0.0315	0.435
GRR	Age-dependent*	0.4667	0.0372	0.2550	0.992
Genotyping error rate	0.01	−0.0180	0.0297	0.0300	0.548
	0.02	−0.0399	0.0280	0.0296	0.478
	0.03	−0.0609	0.0246	0.0283	0.488
	0.04	−0.0799	0.0236	0.0300	0.445
	0.05	−0.0957	0.0233	0.0325	0.422
Number of simulated studies	100	0.0060	0.0377	0.0377	0.580
	200	0.0119	0.0313	0.0314	0.590
	300	0.0128	0.0313	0.0315	0.597
	500	0.0057	0.0305	0.0305	0.580
	600	0.0100	0.0308	0.0309	0.583
	700	0.0110	0.0299	0.0300	0.584

	800	0.0093	0.0304	0.0305	0.585
	900	0.0108	0.0312	0.0313	0.586
	1000	0.0089	0.0309	0.0310	0.585
Study size⋆⋆	200	0.0068	0.2965	0.2965	0.068
	400	−0.0215	0.1446	0.1451	0.128
	600	−0.0175	0.0948	0.0951	0.185
	800	−0.0132	0.0726	0.0728	0.242
	1000	−0.0026	0.0562	0.0562	0.300
	1200	−0.0030	0.0485	0.0485	0.358
	1400	−0.0031	0.0426	0.0426	0.400
	1600	−0.0027	0.0387	0.0387	0.465
	1800	0.0014	0.0350	0.0350	51.7

⋆ Age-dependent dominant GRR as given by the tuples (age [years], GRR): $(35, 20), (40, 15), (45, 10), (50, 5), (55, 1.57), (60, 1), (65, 1), (70, 1), (75, 1)$, reflecting decreasing genetics effects with increasing age in agreement with an overall dominant GRR of 1.43 (reference scenario).

⋆⋆ Study size = number of cases + number of controls (balanced groups)

Table S3: Robust logistic regression with the Huber function [$c = 1.345$]: Bias, variance and mean squared error (MSE) of estimated genotype relative risk (GRR) and statistical power. The assumed parameters under the reference simulation scenario were minor allele frequency (MAF) = 0.05, dominant GRR = 1.43 age-independent, $D' = 1$, $r^2 = 1$, no genotyping errors, 400 simulated studies with 1000 cases and 1000 controls and a fitted dominant penetrance model.

Investigated scenario	Parameter changed	Bias	Variance	MSE	Power
Reference	–	0.0055	0.0304	0.0304	0.568
Fitted penetrance model	Additive	−0.0119	0.0289	0.0290	0.562
	Recessive	−0.1200	0.9000	0.9144	0.020
MAF	0.001	−0.0856	0.9631	0.9704	0.012
	0.005	0.0035	0.2327	0.2327	0.072
	0.010	0.0051	0.1293	0.1293	0.145
	0.100	0.0117	0.0186	0.0187	0.792
	0.150	0.0033	0.0123	0.0123	0.895
	0.200	0.0096	0.0120	0.0121	0.920
	0.250	0.0116	0.0116	0.0117	0.930
r^2	0.9	−0.0173	0.0276	0.0279	0.528
	0.8	−0.0503	0.0247	0.0272	0.480
	0.7	−0.0817	0.0265	0.0332	0.412
	0.6	−0.1130	0.0192	0.0320	0.385
GRR	Age-dependent⋆	0.5176	0.0501	0.3180	0.990
Genotyping error rate	0.01	−0.0206	0.0298	0.0302	0.528
	0.02	−0.0432	0.0279	0.0298	0.465
	0.03	−0.0635	0.0249	0.0289	0.445
	0.04	−0.0820	0.0234	0.0301	0.410
	0.05	−0.0969	0.0234	0.0328	0.392
Number of simulated studies	100	0.0036	0.0370	0.0370	0.500
	200	0.0086	0.0309	0.0310	0.545
	300	0.0095	0.0313	0.0314	0.567
	500	0.0018	0.0301	0.0301	0.556
	600	0.0064	0.0308	0.0308	0.562
	700	0.0072	0.0298	0.0299	0.564

	800	0.0055	0.0303	0.0303	0.564
	900	0.0065	0.0314	0.0314	0.566
	1000	0.0049	0.0310	0.0310	0.562
Study size[⋆⋆]	200	0.0042	0.3061	0.3061	0.072
	400	−0.0166	0.1509	0.1512	0.122
	600	−0.0157	0.0971	0.0973	0.182
	800	−0.0130	0.0736	0.0738	0.225
	1000	−0.0043	0.0569	0.0569	0.287
	1200	−0.0049	0.0491	0.0491	0.358
	1400	−0.0048	0.0428	0.0428	0.382
	1600	−0.0049	0.0387	0.0387	0.445
	1800	−0.0017	0.0346	0.0346	0.498

[⋆] Age-dependent dominant GRR as given by the tuples (age [years], GRR): $(35, 20), (40, 15), (45, 10), (50, 5), (55, 1.57), (60, 1), (65, 1), (70, 1), (75, 1)$, reflecting decreasing genetics effects with increasing age in agreement with an overall dominant GRR of 1.43 (reference scenario).

[⋆⋆] Study size = number of cases + number of controls (balanced groups)

Table S4: Robust logistic regression with the Hampel function $[(a, b, c) = (1.5, 3.5, 8) \cdot 0.9]$: Bias, variance and mean squared error (MSE) of estimated genotype relative risk (GRR) and statistical power. The assumed parameters under the reference simulation scenario were minor allele frequency (MAF) = 0.05, dominant GRR = 1.43 age-independent, $D' = 1$, $r^2 = 1$, no genotyping errors, 400 simulated studies with 1000 cases and 1000 controls and a fitted dominant penetrance model.

Investigated scenario	Parameter changed	Bias	Variance	MSE	Power
Reference	–	0.0040	0.0302	0.0302	0.565
Fitted penetrance model	Additive	−0.0132	0.0287	0.0289	0.555
	Recessive	−0.1032	1.2416	1.2523	0.020
MAF	0.001	−0.0767	1.5965	1.6024	0.014
	0.005	0.0045	0.2352	0.2352	0.072
	0.010	0.0047	0.1294	0.1294	0.145
	0.100	0.0100	0.0185	0.0186	0.790
	0.150	0.0015	0.0122	0.0122	0.895
	0.200	0.0077	0.0119	0.0120	0.918
	0.250	0.0097	0.0114	0.0115	0.930
r^2	0.9	−0.0187	0.0274	0.0277	0.528
	0.8	−0.0517	0.0245	0.0272	0.478
	0.7	−0.0829	0.0262	0.0331	0.412
	0.6	−0.1142	0.0190	0.0320	0.380
GRR	Age-dependent*	0.5196	0.0520	0.3220	0.990
Genotyping error rate	0.01	−0.0220	0.0295	0.0300	0.528
	0.02	−0.0446	0.0277	0.0297	0.465
	0.03	−0.0648	0.0247	0.0289	0.442
	0.04	−0.0833	0.0232	0.0301	0.410
	0.05	−0.0981	0.0232	0.0328	0.388
Number of simulated studies	100	0.0020	0.0367	0.0367	0.500
	200	0.0071	0.0306	0.0307	0.545
	300	0.0080	0.0311	0.0312	0.563
	500	0.0002	0.0299	0.0299	0.554
	600	0.0049	0.0306	0.0306	0.560
	700	0.0057	0.0296	0.0296	0.563

	800	0.0040	0.0301	0.0301	0.562
	900	0.0050	0.0312	0.0312	0.564
	1000	0.0034	0.0308	0.0308	0.560
Study size⋆⋆	200	0.0064	0.3123	0.3123	0.070
	400	−0.0171	0.1512	0.1515	0.120
	600	−0.0167	0.0965	0.0968	0.180
	800	−0.0142	0.0730	0.0732	0.225
	1000	−0.0056	0.0565	0.0565	0.287
	1200	−0.0063	0.0488	0.0488	0.358
	1400	−0.0062	0.0425	0.0425	0.382
	1600	−0.0064	0.0384	0.0384	0.442
	1800	−0.0032	0.0343	0.0343	0.495

⋆ Age-dependent dominant GRR as given by the tuples (age [years], GRR): $(35, 20), (40, 15), (45, 10), (50, 5), (55, 1.57), (60, 1), (65, 1), (70, 1), (75, 1)$, reflecting decreasing genetics effects with increasing age in agreement with an overall dominant GRR of 1.43 (reference scenario).

⋆⋆ Study size = number of cases + number of controls (balanced groups)

Table S5: Robust logistic regression with the Hampel function $[(a, b, c) = (2, 4, 8) \cdot 0.7]$: Bias, variance and mean squared error (MSE) of estimated genotype relative risk (GRR) and statistical power. The assumed parameters under the reference simulation scenario were minor allele frequency (MAF) = 0.05, dominant GRR = 1.43 age-independent, $D' = 1$, $r^2 = 1$, no genotyping errors, 400 simulated studies with 1000 cases and 1000 controls and a fitted dominant penetrance model.

Investigated scenario	Parameter changed	Bias	Variance	MSE	Power
Reference	–	0.0024	0.0299	0.0299	0.565
Fitted penetrance model	Additive	−0.0146	0.0285	0.0287	0.552
	Recessive	−0.0846	1.2028	1.2100	0.022
MAF	0.001	−0.0506	1.4955	1.4981	0.014
	0.005	0.0047	0.2364	0.2364	0.070
	0.010	0.0044	0.1294	0.1294	0.145
	0.100	0.0084	0.0182	0.0183	0.790
	0.150	−0.0003	0.0121	0.0121	0.890
	0.200	0.0059	0.0117	0.0117	0.918
	0.250	0.0077	0.0113	0.0114	0.928
r^2	0.9	−0.0202	0.0271	0.0275	0.525
	0.8	−0.0530	0.0243	0.0271	0.472
	0.7	−0.0841	0.0259	0.0330	0.410
	0.6	−0.1153	0.0188	0.0321	0.378
GRR	Age-dependent*	0.5241	0.0536	0.3283	0.992
Genotyping error rate	0.01	−0.0235	0.0292	0.0298	0.522
	0.02	−0.0459	0.0274	0.0295	0.465
	0.03	−0.0661	0.0244	0.0288	0.438
	0.04	−0.0847	0.0229	0.0301	0.410
	0.05	−0.0994	0.0230	0.0329	0.390
Number of simulated studies	100	0.0002	0.0363	0.0363	0.510
	200	0.0056	0.0303	0.0303	0.545
	300	0.0065	0.0308	0.0308	0.563
	500	−0.0014	0.0296	0.0296	0.554
	600	0.0033	0.0303	0.0303	0.558
	700	0.0041	0.0293	0.0293	0.561

	800	0.0024	0.0298	0.0298	0.560
	900	0.0035	0.0309	0.0309	0.560
	1000	0.0019	0.0305	0.0305	0.556
Study size⋆⋆	200	0.0098	0.3228	0.2339	0.072
	400	−0.0168	0.1528	0.1531	0.112
	600	−0.0176	0.0960	0.0963	0.182
	800	−0.0155	0.0724	0.0726	0.220
	1000	−0.0068	0.0561	0.0561	0.282
	1200	−0.0076	0.0484	0.0485	0.350
	1400	−0.0076	0.0422	0.0423	0.385
	1600	−0.0078	0.0380	0.0381	0.442
	1800	−0.0048	0.0340	0.0340	0.490

⋆ Age-dependent dominant GRR as given by the tuples (age [years], GRR): $(35, 20), (40, 15), (45, 10), (50, 5), (55, 1.57), (60, 1), (65, 1), (70, 1), (75, 1)$, reflecting decreasing genetics effects with increasing age in agreement with an overall dominant GRR of 1.43 (reference scenario).

⋆⋆ Study size = number of cases + number of controls (balanced groups)

Table S6: Standard logistic regression: Bias, variance and mean squared error (MSE) of estimated genotype relative risk (GRR) and statistical power of the additional scenario “rare variants”. The fixed simulation parameters were $D' = 1$, $r^2 = 1$, 400 simulated studies with a fitted dominant penetrance model. MAF = minor allele frequency. # = number of.

MAF # Cases / # Controls Dominant GRR	Genotyping Error Rate	Bias	Variance	MSE	Power
0.010	0.00	0.0156	0.1072	0.1074	0.602
1000/1000	0.01	−0.0888	0.0934	0.1013	0.540
2.07	0.02	−0.1803	0.0792	0.1117	0.492
	0.03	−0.2469	0.0731	0.1341	0.455
	0.04	−0.3038	0.0646	0.1569	0.395
	0.05	−0.3486	0.0564	0.1779	0.348
0.005	0.00	0.0229	0.1926	0.1931	0.598
1000/1000	0.01	−0.2077	0.1499	0.1930	0.512
2.65	0.02	−0.3665	0.1181	0.2524	0.410
	0.03	−0.4731	0.0977	0.3215	0.368
	0.04	−0.5486	0.0858	0.3868	0.305
	0.05	−0.6064	0.0729	0.4406	0.260
0.001	0.00	0.0563	0.2047	0.2079	0.598
5000/5000	0.01	−0.7111	0.0461	0.5518	0.175
2.53	0.02	−0.7957	0.0257	0.6588	0.142
	0.03	−0.8318	0.0164	0.7083	0.100
	0.04	−0.8543	0.0129	0.7427	0.092
	0.05	−0.8703	0.0108	0.7682	0.058

Table S7: Robust logistic regression with the Huber function [$c = 1.345$]: Bias, variance and mean squared error (MSE) of estimated genotype relative risk (GRR) and statistical power of the additional scenario "rare variants". The fixed simulation parameters were $D' = 1$, $r^2 = 1$, 400 simulated studies with a fitted dominant penetrance model. MAF = minor allele frequency. # = number of.

MAF # Cases / # Controls Dominant GRR	Genotyping Error Rate	Bias	Variance	MSE	Power
0.010	0.00	−0.0058	0.1038	0.1038	0.555
1000/1000	0.01	−0.1054	0.0898	0.1009	0.488
2.07	0.02	−0.1934	0.0790	0.1164	0.435
	0.03	−0.2580	0.0696	0.1362	0.422
	0.04	−0.3122	0.0650	0.1625	0.365
	0.05	−0.3533	0.0558	0.1806	0.310
0.005	0.00	−0.0101	0.1829	0.1830	0.528
1000/1000	0.01	−0.2288	0.1403	0.1926	0.465
2.65	0.02	−0.3803	0.1197	0.2643	0.385
	0.03	−0.4842	0.0930	0.3274	0.322
	0.04	−0.5566	0.0866	0.3964	0.270
	0.05	−0.6084	0.0725	0.4427	0.235
0.001	0.00	0.0175	0.1900	0.1903	0.538
5000/5000	0.01	−0.7115	0.0466	0.5528	0.190
2.53	0.02	−0.7970	0.0262	0.6614	0.130
	0.03	−0.8348	0.0171	0.7140	0.092
	0.04	−0.8563	0.0131	0.7463	0.082
	0.05	−0.8715	0.0111	0.7706	0.055

Table S8: Robust logistic regression with the Hampel function $[(a, b, c) = (1.5, 3.5, 8) \cdot 0.9]$: Bias, variance and mean squared error (MSE) of estimated genotype relative risk (GRR) and statistical power of the additional scenario "rare variants". The fixed simulation parameters were $D' = 1$, $r^2 = 1$, 400 simulated studies with a fitted dominant penetrance model. MAF = minor allele frequency. # = number of.

MAF # Cases / # Controls Dominant GRR	Genotyping Error Rate	Bias	Variance	MSE	Power
0.010	0.00	−0.0063	0.1045	0.1045	0.555
1000/1000	0.01	−0.1063	0.0901	0.1014	0.488
2.07	0.02	−0.1945	0.0790	0.1168	0.430
	0.03	−0.2593	0.0694	0.1366	0.422
	0.04	−0.3134	0.0649	0.1631	0.362
	0.05	−0.3544	0.0555	0.1811	0.308
0.005	0.00	−0.0083	0.1875	0.1876	0.525
1000/1000	0.01	−0.2286	0.1420	0.1943	0.465
2.65	0.02	−0.3809	0.1205	0.2656	0.382
	0.03	−0.4853	0.0930	0.3285	0.320
	0.04	−0.5576	0.0864	0.3973	0.270
	0.05	−0.6095	0.0723	0.4438	0.235
0.001	0.00	0.0194	0.1948	0.1952	0.535
5000/5000	0.01	−0.7125	0.0462	0.5539	0.190
2.53	0.02	−0.7977	0.0259	0.6622	0.130
	0.03	−0.8353	0.0169	0.7146	0.092
	0.04	−0.8567	0.0129	0.7468	0.080
	0.05	−0.8718	0.0110	0.7710	0.055

Table S9: Robust logistic regression with the Hampel function $[(a, b, c) = (2, 4, 8) \cdot 0.7]$: Bias, variance and mean squared error (MSE) of estimated genotype relative risk (GRR) and statistical power of the additional scenario "rare variants". The fixed simulation parameters were $D' = 1$, $r^2 = 1$, 400 simulated studies with a fitted dominant penetrance model. MAF = minor allele frequency. # = number of.

MAF # Cases / # Controls Dominant GRR	Genotyping Error Rate	Bias	Variance	MSE	Power
0.010	0.00	−0.0076	0.1049	0.1050	0.555
1000/1000	0.01	−0.1079	0.0901	0.1017	0.480
2.07	0.02	−0.1959	0.0787	0.1171	0.425
	0.03	−0.2608	0.0690	0.1370	0.415
	0.04	−0.3146	0.0646	0.1636	0.365
	0.05	−0.3556	0.0553	0.1818	0.305
0.005	0.00	−0.0086	0.1901	0.1902	0.522
1000/1000	0.01	−0.2296	0.1429	0.1956	0.468
2.65	0.02	−0.3817	0.1210	0.2667	0.378
	0.03	−0.4868	0.0928	0.3298	0.320
	0.04	−0.5588	0.0861	0.3984	0.273
	0.05	−0.6104	0.0721	0.4447	0.235
0.001	0.00	0.0210	0.2021	0.2025	0.530
5000/5000	0.01	−0.7135	0.0456	0.5547	0.188
2.53	0.02	−0.7982	0.0256	0.6627	0.128
	0.03	−0.8356	0.0167	0.7149	0.092
	0.04	−0.8569	0.0127	0.7470	0.080
	0.05	−0.8719	0.0109	0.7711	0.052

Table S10: Standard logistic regression: Bias, variance and mean squared error (MSE) of estimated genotype relative risk (GRR) and statistical power of the additional scenario "underlying recessive penetrance model". The fixed simulation parameters were minor allele frequency (MAF) = 0.05, recessive GRR = 6.32 age-independent, $D' = 1$, $r^2 = 1$ and 400 simulated studies with 1000 cases and 1000 controls.

Fitted Penetrance Model	Genotyping Error Rate	Bias	Variance	MSE	Power
Recessive	0.00	1.1856	15.0893	16.4949	0.600
	0.01	−0.8478	0.2540	0.9728	0.470
	0.02	−1.1460	0.1697	1.4830	0.370
	0.03	−1.3053	0.1180	1.8218	0.290
	0.04	−1.4043	0.1083	2.0804	0.235
	0.05	−1.4564	0.0963	2.2174	0.238
Additive	0.00	−1.8377	0.0356	3.4127	0.055
	0.01	−1.8359	0.0356	3.4061	0.055
	0.02	−1.8345	0.0338	3.3992	0.062
	0.03	−1.8345	0.0320	3.3974	0.065
	0.04	−1.8365	0.0313	3.4040	0.065
	0.05	−1.8395	0.0313	3.4151	0.062
Dominant	0.00	−1.7147	0.0335	2.9737	0.108
	0.01	−1.7261	0.0325	3.0119	0.128
	0.02	−1.7328	0.0305	3.0331	0.108
	0.03	−1.7412	0.0268	3.0586	0.088
	0.04	−1.7493	0.0254	3.0855	0.090
	0.05	−1.7538	0.0246	3.1004	0.090

Table S11: Robust logistic regression with the Huber function [$c = 1.345$]: Bias, variance and mean squared error (MSE) of estimated genotype relative risk (GRR) and statistical power of the additional scenario "underlying recessive penetrance model". The fixed simulation parameters were minor allele frequency (MAF) = 0.05, recessive GRR = 6.32 age-independent, $D' = 1$, $r^2 = 1$ and 400 simulated studies with 1000 cases and 1000 controls.

Fitted Penetrance Model	Genotyping Error Rate	Bias	Variance	MSE	Power
Recessive	0.00	−0.1074	0.5232	0.5347	0.510
	0.01	−0.8649	0.2315	0.9796	0.420
	0.02	−1.1434	0.1717	1.4791	0.328
	0.03	−1.3000	0.1267	1.8167	0.278
	0.04	−1.4341	0.1054	2.1620	0.210
	0.05	−1.4901	0.0902	2.3106	0.160
Additive	0.00	−1.8434	0.0367	3.4348	0.050
	0.01	−1.8378	0.0351	3.4126	0.045
	0.02	−1.8381	0.0329	3.4115	0.048
	0.03	−1.8392	0.0327	3.4154	0.052
	0.04	−1.8355	0.0315	3.4006	0.050
	0.05	−1.8358	0.0293	3.3995	0.035
Dominant	0.00	−1.7153	0.0335	2.9758	0.105
	0.01	−1.7248	0.0321	3.0070	0.102
	0.02	−1.7344	0.0286	3.0367	0.095
	0.03	−1.7431	0.0274	3.0658	0.098
	0.04	−1.7530	0.0272	3.1002	0.108
	0.05	−1.7567	0.0220	3.1080	0.075

Table S12: Robust logistic regression with the Hampel function $[(a,b,c) = (1.5, 3.5, 8) \cdot 0.9]$: Bias, variance and mean squared error (MSE) of estimated genotype relative risk (GRR) and statistical power of the additional scenario "underlying recessive penetrance model". The fixed simulation parameters were minor allele frequency (MAF) = 0.05, recessive GRR = 6.32 age-independent, $D' = 1$, $r^2 = 1$ and 400 simulated studies with 1000 cases and 1000 controls.

Fitted Penetrance Model	Genotyping Error Rate	Bias	Variance	MSE	Power
Recessive	0.00	0.0624	1.2398	1.2437	0.438
	0.01	−0.8267	0.3113	0.9947	0.405
	0.02	−1.1349	0.1795	1.4675	0.330
	0.03	−1.3159	0.1390	1.8706	0.268
	0.04	−1.3999	0.0986	2.0583	0.232
	0.05	−1.4560	0.0883	2.2082	0.225
Additive	0.00	−1.8436	0.0364	3.4353	0.047
	0.01	−1.8383	0.0339	3.4132	0.052
	0.02	−1.8417	0.0344	3.4263	0.058
	0.03	−1.8338	0.0328	3.3956	0.055
	0.04	−1.8390	0.0313	3.4132	0.038
	0.05	−1.8379	0.0297	3.4076	0.038
Dominant	0.00	−1.7160	0.0331	2.9778	0.105
	0.01	−1.7242	0.0306	3.0035	0.098
	0.02	−1.7361	0.0296	3.0436	0.100
	0.03	−1.7412	0.0285	3.0603	0.108
	0.04	−1.7478	0.0254	3.0802	0.088
	0.05	−1.7508	0.0229	3.0882	0.080

Table S13: Robust logistic regression with the Hampel function $[(a, b, c) = (2, 4, 8) \cdot 0.7]$: Bias, variance and mean squared error (MSE) of estimated genotype relative risk (GRR) and statistical power of the additional scenario "underlying recessive penetrance model". The fixed simulation parameters were minor allele frequency (MAF) = 0.05, recessive GRR = 6.32 age-independent, $D' = 1$, $r^2 = 1$ and 400 simulated studies with 1000 cases and 1000 controls.

Fitted Penetrance Model	Genotyping Error Rate	Bias	Variance	MSE	Power
Recessive	0.00	0.0543	1.1527	1.1556	0.399
	0.01	−0.8214	0.3313	1.0060	0.408
	0.02	−1.1342	0.1822	1.4686	0.328
	0.03	−1.3175	0.1383	1.8741	0.262
	0.04	−1.4007	0.0993	2.0613	0.232
	0.05	−1.4575	0.0876	2.2119	0.228
Additive	0.00	−1.8436	0.0359	3.4348	0.044
	0.01	−1.8383	0.0335	3.4128	0.052
	0.02	−1.8417	0.0340	3.4259	0.058
	0.03	−1.8338	0.0324	3.3952	0.052
	0.04	−1.8390	0.0310	3.4129	0.038
	0.05	−1.8379	0.0294	3.4073	0.038
Dominant	0.00	−1.7166	0.0327	2.9794	0.105
	0.01	−1.7249	0.0303	3.0056	0.092
	0.02	−1.7367	0.0293	3.0454	0.098
	0.03	−1.7417	0.0281	3.0616	0.105
	0.04	−1.7483	0.0251	3.0817	0.088
	0.05	−1.7512	0.0226	3.0893	0.080

Table S14: SNP identifiers.

rs3934834, rs6687776, rs9442373, rs2298217, rs9442380, rs11260549, rs2887286, rs3813199, rs7515488, rs6675798, rs6685064, rs2649588, rs819980, rs2031709, rs880051, rs2296716, rs6603811, rs7531583, rs16825336, rs6681938, rs10907192, rs4648592, rs7525092, rs2474460, rs2459994, rs884080, rs908742, rs4648808, rs3107151, rs3128291, rs3753242, rs424079, rs2257182, rs2460000, rs263526, rs10797417, rs10910047, rs12119470, rs2017143, rs903919, rs884940, rs10910050, rs903916, rs2279702, rs2173049, rs2645065, rs903904, rs2843143, rs2843142, rs2055204, rs7527871, rs4648831, rs2840528, rs903914, rs2643891, rs2840538, rs10910061, rs2279703, rs7519807, rs2843160, rs903901, rs2643901, rs2843127, rs903903, rs1123571, rs3736330, rs2840532, rs3001336, rs2494428, rs12022929, rs4531246, rs4648843, rs6659405, rs10910078, rs2494626, rs13376356, rs11588930, rs12049628, rs17373634, rs2477703, rs3762444, rs7535528, rs6667605, rs734999, rs3748816, rs12138909, rs11590198, rs3890745, rs2377041, rs10909890, rs4648482, rs10797342, rs897634, rs2045331, rs2045332, rs2606411, rs4648441, rs10797368, rs10909845, rs11583804, rs878201, rs2485945, rs12046158, rs1572657, rs10909852, rs12562637, rs7534897, rs3795263, rs7412983, rs2142569, rs2297829, rs1569419, rs926244, rs2993493, rs1890336, rs4648453, rs2817178, rs10797380, rs7538096, rs2817185, rs731031, rs2651899, rs10752733, rs10737190, rs10909901, rs12124147, rs2651906, rs16823542, rs6424069, rs2455118, rs10797386, rs3002685, rs3002686, rs10492940, rs16823802, rs905135, rs12562988, rs10909918, rs12757342, rs1553291, rs4648377, rs2455144, rs2483260, rs16824089, rs1108600, rs2483274, rs6683273, rs4415513, rs4648380, rs946758, rs12748963, rs2500286, rs12073172, rs17399569, rs2500262, rs4648487, rs4648489, rs2493310, rs12085231, rs868688, rs10492938, rs17399998, rs2493275, rs871822, rs6424074, rs11578011, rs12024847, rs870124, rs2493292, rs2493285, rs1984069, rs870171, rs2493272, rs2487670, rs2487680, rs12562167, rs2493314, rs4648505, rs2821040, rs947344, rs4648392, rs12119711, rs4648524, rs10737192, rs878063, rs9628616, rs2821063, rs947354, rs4648527, rs6697749, rs7544357, rs2821025, rs2821023, rs4648398, rs4276857, rs2821007, rs7528494, rs4648545, rs7523732, rs3765703, rs3765705, rs3765731, rs3765736, rs3765761, rs3765766, rs747827, rs12731705, rs12117836, rs3737589, rs1181888, rs1181883, rs1181877, rs1181875, rs10910025, rs2275819, rs1175549, rs2799182, rs6663840, rs2275831, rs4648426, rs10797348, rs7367066, rs4131373, rs12082157, rs11589102, rs6695346, rs12724233, rs11583257, rs7519349, rs7519458, rs4654479, rs4654480, rs6661168, rs4654482, rs11590912, rs10799202, rs12119556, rs10915433, rs6681347, rs7522140, rs12031557, rs11587331, rs6691155, rs12135298, rs12749761

Table S15: Sample identifiers.

hu00147A, hu002B3C, hu016B28, hu0199C8, hu019BBA, hu025CEA, hu0515BA, hu05FD49, hu066C78, hu11603C, hu14ECAE, hu155D20, hu16360E, hu16A1B3, hu1712BC, hu19C09F, hu1AF744, hu1BD549, hu1BDBA5, hu2331A5, hu25BD97, hu27FD1F, hu297562, hu2BC187, hu2D53F2, hu2DEBA7, hu2E413D, hu2FEC01, hu30888B, hu30F119, hu345185, hu3458D8, hu34D5B9, hu35071E, hu352868, hu35389E, hu363FD6, hu3696DA, hu394092, hu3B89BD, hu3D355A, hu3F864B, hu41D90F, hu44DCFF, hu459AD0, hu4753BA, hu48C4EB, hu499ED5, hu4AEB32, hu4B07B3, hu4B11A3, hu4BE378, hu4BE6F2, hu4CA5B9, hu4D2239, hu4E03BC, hu524B5B, hu56B3B6, hu57C8BE, hu589D0B, hu59141C, hu594129, hu5A2074, hu5D9DE3, hu5F0DCB, hu5FCE15, hu5FF6B0, hu60AB7C, hu619F51, hu63A000, hu63DA55, hu654B61, hu67B84E, hu6D1115, hu6E37AB, hu6ED94A, hu72110E, hu75BE2C, hu775356, hu77AB33, hu77CC58, hu781EE2, hu787E67, hu7DE7FD, hu82436A, hu84B706, hu8602F1, hu868880, hu8A5FBF, hu8B4E43, hu8D99F6, hu90B053, hu91BD69, hu925B56, hu939B7C, hu96713F, hu993257, hu9A0F06, huA35014, huA4F281, huA5FD8B, huA720D3, huAC827A, huAD719C, huAF3C63, huB2C416, huB59C05, huB5A0DF, huB63C0C, huB7EC37, huBAA265, huBC03A7, huBD9C9B, huBE28C7, huBFEDCE, huC1C7D0, huC92BC9, huCCA261, huD0449C, huD0D79A, huD3E181, huD4F7DB, huD50D1C, huD52556, huD57BBF, huD58ABC, huD7960A, huD87BFC, huD9D625, huDB1635, huDD6E7A, huDDEC1D, huE31062, huE4CA90, huE9E777, huEAA57B, huEBD467, huED0F40, huF06AD0, huF7E042, huF9E138, huFE71F3, huFF6AB4, huFFAD8

C Supplemental figures

This chapter provides supplemental figures of the investigation of the statistical properties with standard and robust logistic regression methods with both influence functions. These figures are the boxplots of the estimates ...

- ... of the main scenarios – one figure per changed parameter (figures S1-S7 on pages 144-150)
- ... of the additional scenario “rare variants” (figure S8 on page 151)
- ... of the additional scenario “underlying recessive penetrance model” (figure S9 on page 152)

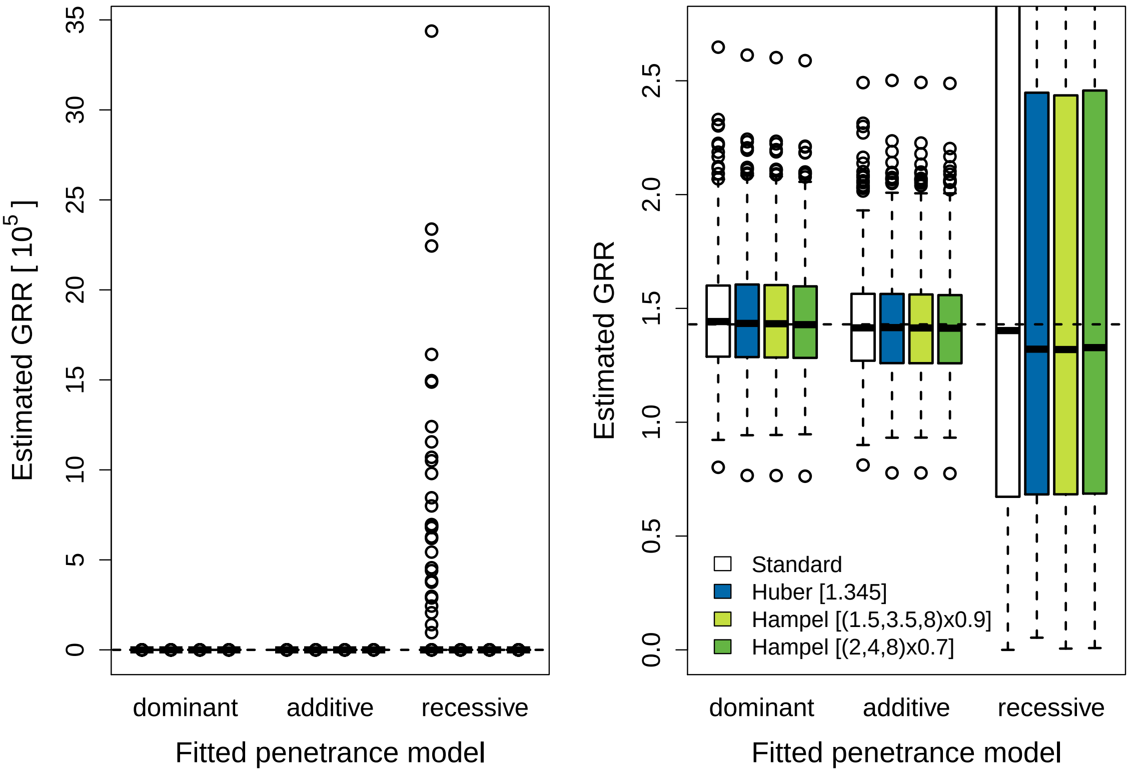


Figure S1: Boxplots of estimated genotype relative risk (GRR) according to different penetrance models fitted to the simulated data (minor allele frequency (MAF) = 0.05, dominant GRR = 1.43 age-independent, $D' = 1$, $r^2 = 1$, no genotyping errors, 400 simulated studies with 1000 cases and 1000 controls). The left panel displays the complete domain of estimated GRRs. The y-axis is limited to $0 - 2.75$ in the right panel. Tuning constants for the robust logistic regression models are shown in brackets in the legend.

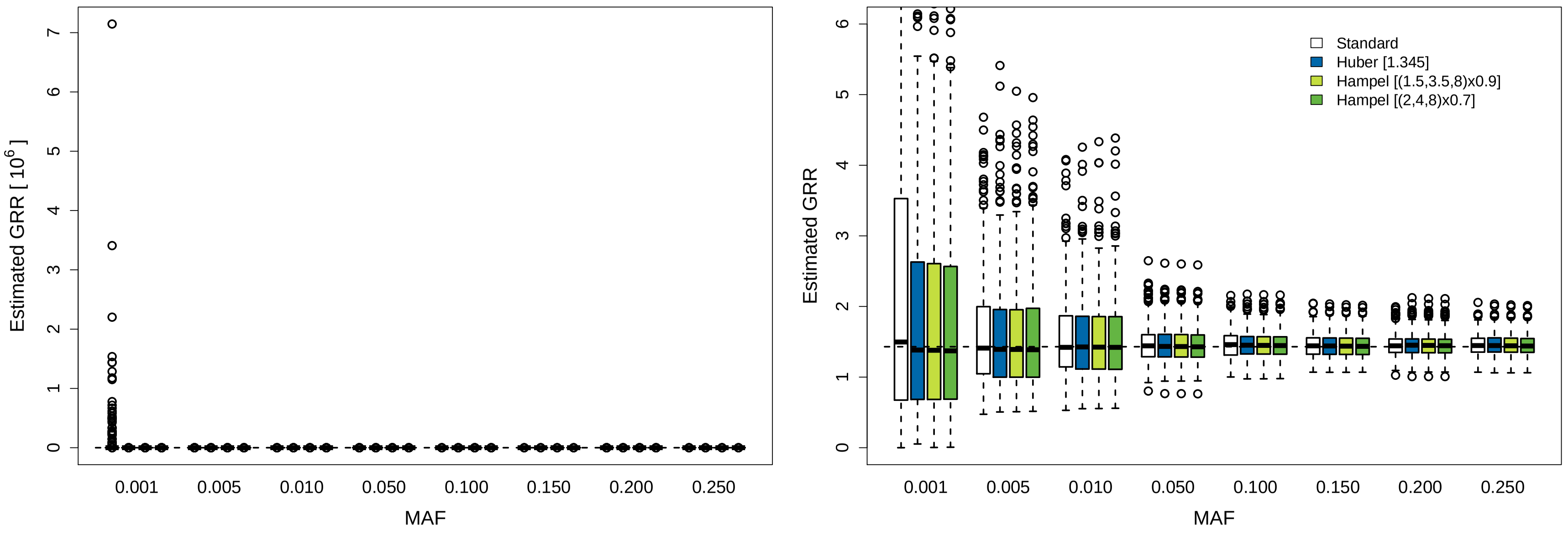


Figure S2: Boxplots of the estimated genotype relative risk (GRR) according to different minor allele frequencies (MAFs) (dominant GRR = 1.43 age-independent, $D' = 1$, $r^2 = 1$, no genotyping errors, 400 simulated studies with 1000 cases and 1000 controls and a fitted dominant penetrance model). The left panel displays the complete domain of estimated GRRs. The display is limited to $0 - 6$ in the right panel. The tuning constants for the robust logistic regression methods are given in brackets in the legend.

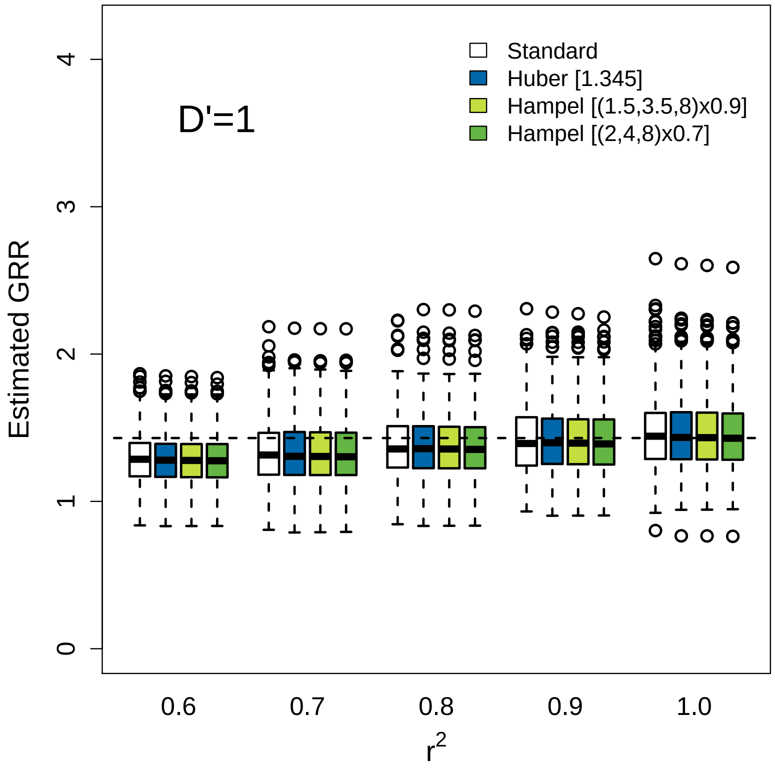


Figure S3: Boxplots of the estimated genotype relative risk (GRR) according to different r^2 (minor allele frequency (MAF) = 0.05, dominant GRR = 1.43 age-independent, $D' = 1$, no genotyping errors, 400 simulated studies with 1000 cases and 1000 controls and a fitted dominant penetrance model). The tuning constants for the robust logistic regression methods are given in brackets in the legend.

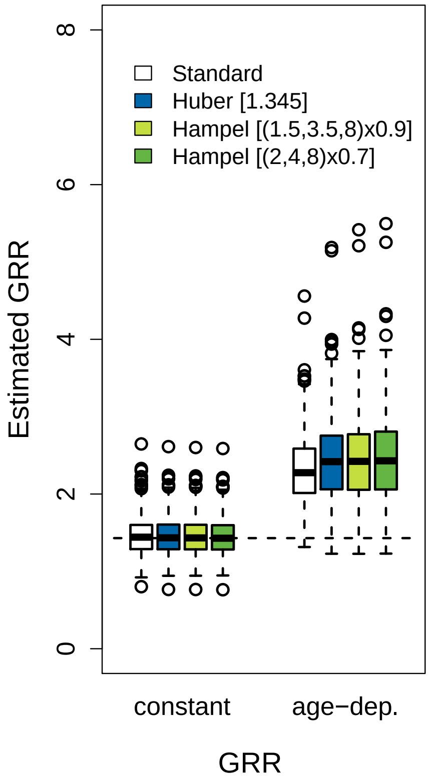


Figure S4: Boxplots of the estimated genotype relative risk (GRR) according to an age-independent dominant GRR and age-dependent dominant GRRs. The age-dependent GRRs are given by the tuples (age [years], GRR): $(35, 20)$, $(40, 15)$, $(45, 10)$, $(50, 5)$, $(55, 1.57)$, $(60, 1)$, $(65, 1)$, $(70, 1)$, $(75, 1)$, reflecting decreasing genetics effects with increasing age in agreement with an overall dominant GRR of 1.43 (age-independent reference scenario). The fixed simulation parameters are minor allele frequency (MAF) $= 0.05$, $D' = 1$, $r^2 = 1$, no genotyping errors and 400 simulated studies with 1000 cases and 1000 controls and a fitted dominant penetrance model. The tuning constants for the robust logistic regression methods are given in brackets in the legend.

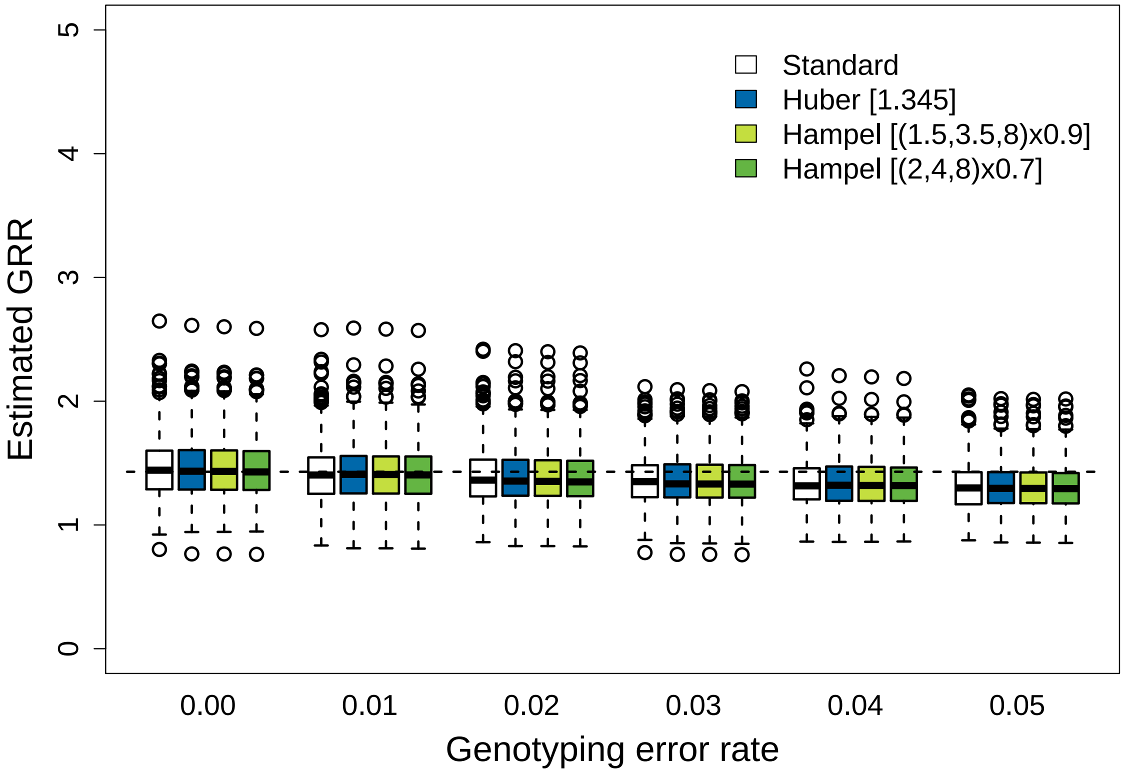


Figure S5: Boxplots of the estimated genotype relative risk (GRR) according to different genotyping error rates (minor allele frequency (MAF) = 0.05, dominant GRR = 1.43 age-independent, $D' = 1$, $r^2 = 1$, 400 simulated studies with 1000 cases and 1000 controls and a fitted dominant penetrance model). The tuning constants for the robust logistic regression methods are given in brackets in the legend.

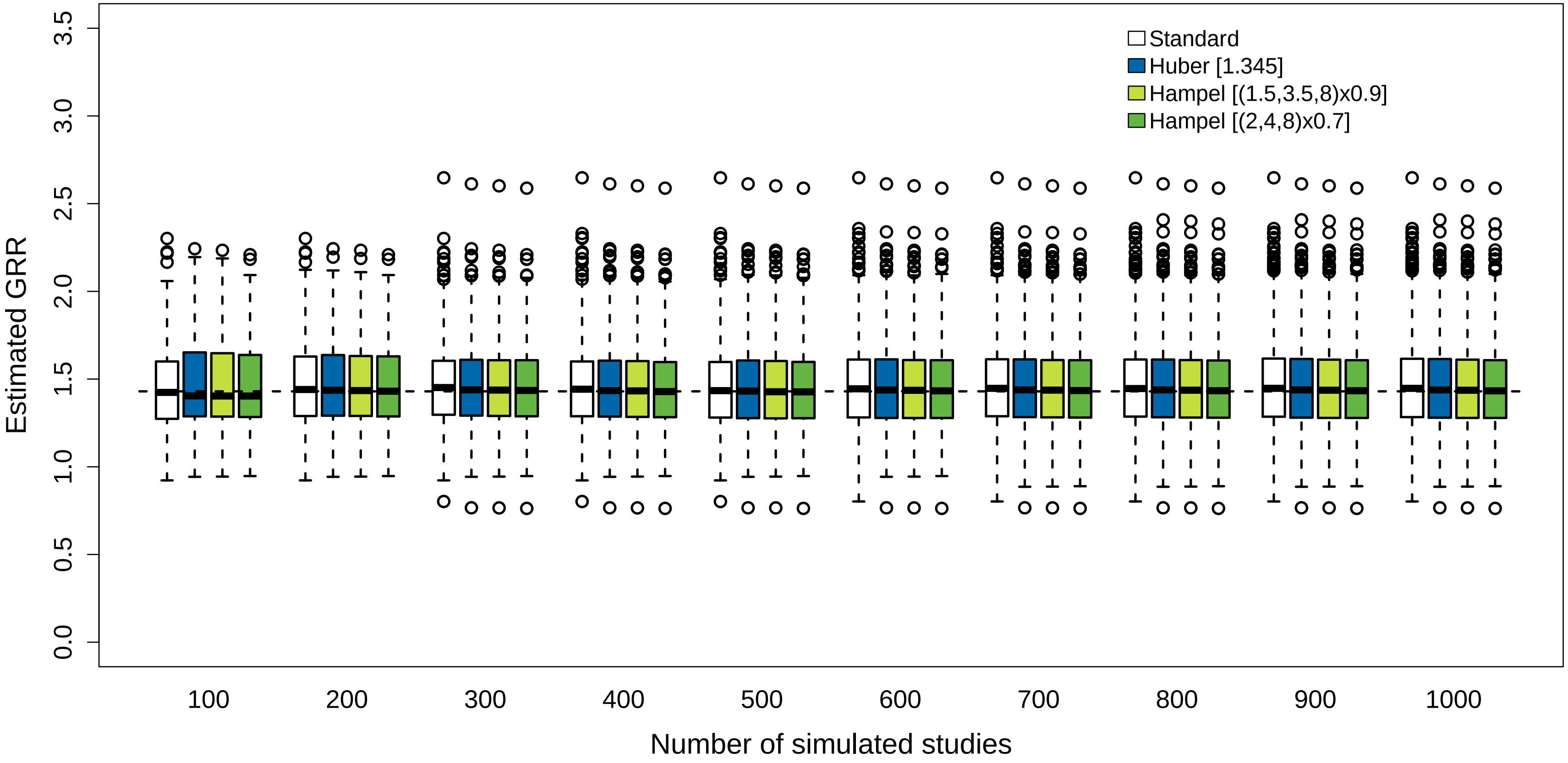


Figure S6: Boxplots of the estimated genotype relative risk (GRR) according to different numbers of simulated studies (minor allele frequency (MAF) = 0.05, dominant GRR = 1.43 age-independent, $D' = 1$, $r^2 = 1$, no genotyping errors, simulated studies with 1000 cases and 1000 controls and a fitted dominant penetrance model). The tuning constants for the robust logistic regression methods are given in brackets in the legend.

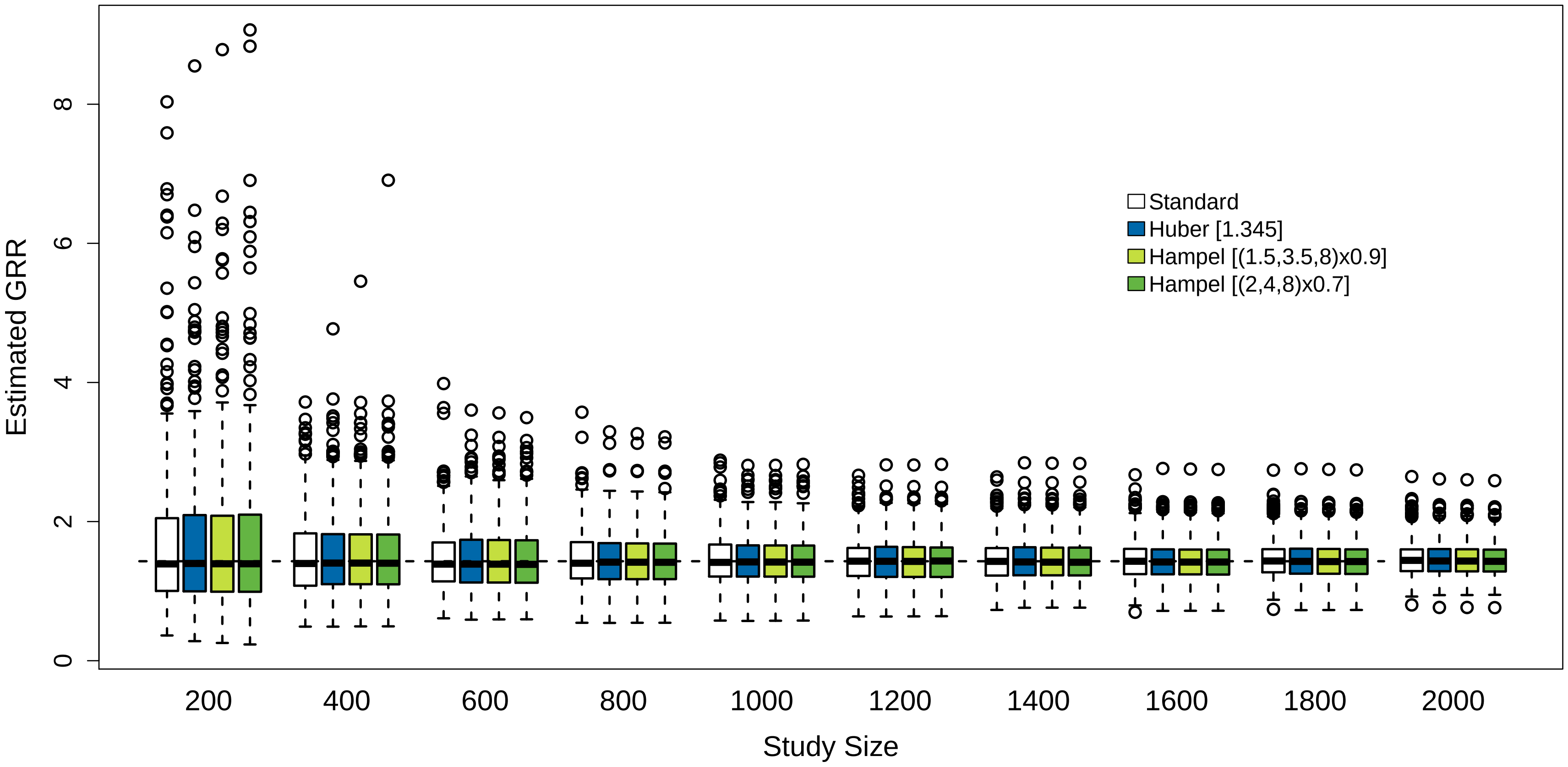


Figure S7: Boxplots of the estimated genotype relative risk (GRR) according to different study sizes (minor allele frequency (MAF) = 0.05, dominant GRR = 1.43 age-independent, $D' = 1$, $r^2 = 1$, no genotyping errors, 400 simulated studies with a fitted dominant penetrance model). Study size denotes the number of individuals in the study (number of cases = number of controls). The tuning constants for the robust logistic regression methods are given in brackets in the legend.

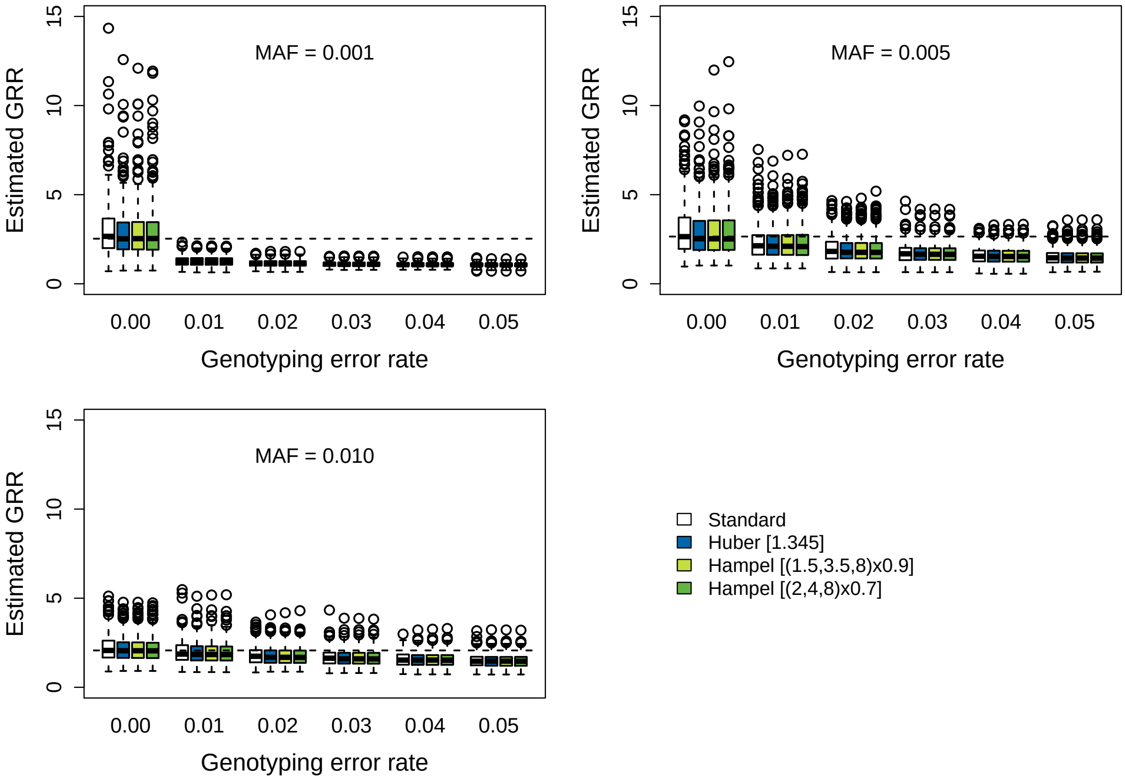


Figure S8: Boxplots of the estimated genotype relative risk (GRR) according to different minor allele frequencies (MAFs) for rare variants (0.001, 0.005, 0.010). The fixed simulation parameters were $D' = 1$, $r^2 = 1$, 400 simulated studies with a fitted dominant penetrance model. The simulated age-independent dominant GRR and the study size depended on the MAF: MAF = 0.001 with GRR = 2.53 and 5000 cases/5000 controls, MAF = 0.005 with GRR = 2.65 and 1000 cases/1000 controls, MAF = 0.010 with GRR = 2.05 and 1000 cases/1000 controls. The tuning constants for the robust logistic regression methods are given in brackets in the legend. The MAF is given on the plot area.

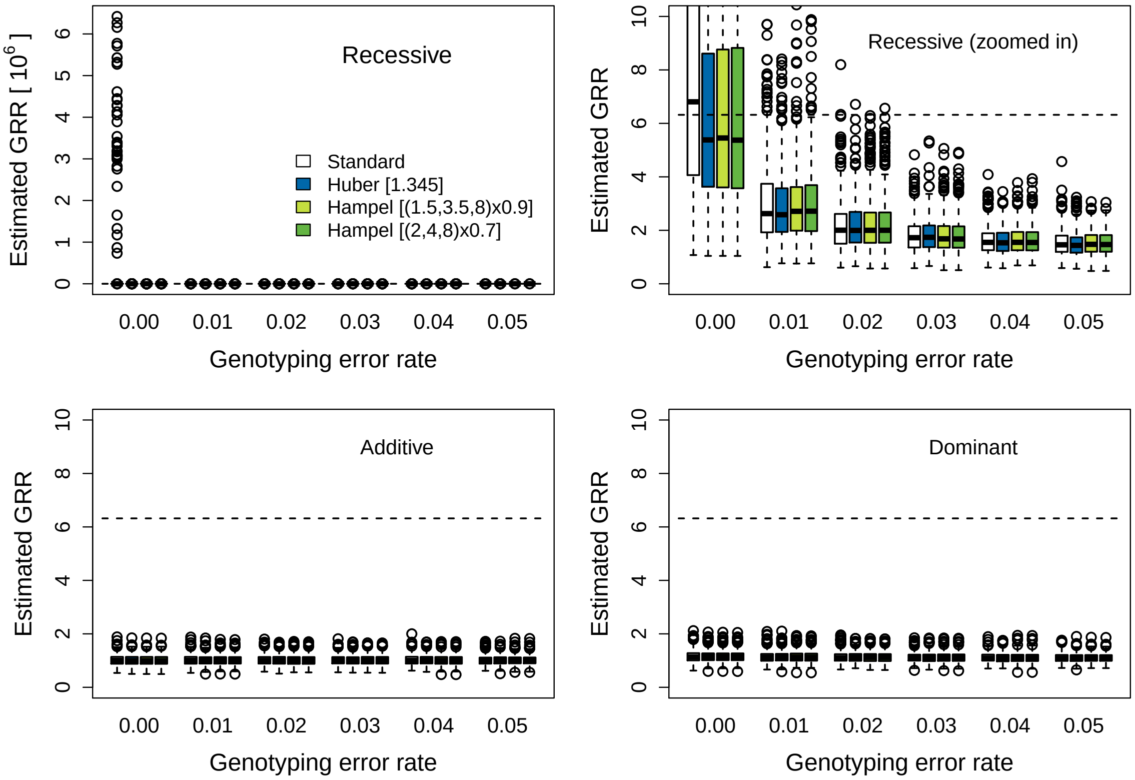


Figure S9: Boxplots of the estimated genotype relative risk (GRR) according to the fitted penetrance model to recessive simulated data. The fixed simulation parameters were minor allele frequency (MAF) = 0.05, recessive GRR = 6.32 age-independent, $D' = 1$, $r^2 = 1$, 400 simulated studies with 1000 cases and 1000 controls. The tuning constants for the robust logistic regression methods are given in brackets in the legend. The fitted penetrance model is given on the plot area. Note the two top panels for a fitted recessive penetrance model with different axis scaling (left: complete domain, right: limited to $0 - 10$).

D Supplemental source code

This section provides installation instructions for the extended R package, example source code for the simulation study in the reference scenario including the calculation of the statistical properties and the source code for the real data application. Proper working directory definitions in each R script are mandatory. Be aware of expensive calculations.

D.1 Installation of the extended R package

The instructions are exactly orientated on the glmrobMqle.R file of the package `robustbase` version 0.9-8 (Date: 14/06/2013). All code sections have to be placed exactly where they can be found in the original file for the Huber function. Afterwards, the package has again to be installed as a whole under a modified name, e.g. `robustbaseAdj`.

Installation: With respect to the considered R version, the installation of the 32-bit version only is mandatory. Otherwise the installation will not work. Then:

1. Open the command line.
2. Go to directory `R/bin`.
3. Write into command line `R CMD INSTALL path` with `path` indicating the path to folder "robustbaseAdj" containing the R package.

D.2 Simulation study for statistical properties evaluation

This section provides code examples for the reference scenario of the simulation study comprising

- the random number generation to sample individuals from the population (section D.2.1),
- the calculation of allele frequencies in the population (section D.2.2),
- the population simulation (sections D.2.3 (marker locus) and D.2.4 (null marker loci),
- the standard and robust logistic regression analysis and the calculation of the statistical properties (type I error rate, bias, variance, MSE, statistical power) – section D.2.5 for the marker locus and section D.2.6 for the null marker loci.

D.2.1 Random samples

Random numbers to draw 1000 cases and 1000 controls from the population comprising 3,500,000 cases and 3,500,000 controls for 400 studies

```
# Working directories
dir.save <- "Directory to save results"

# Settings
set.seed(12061950)
repetitions <- 400
pop.size <- 3500000
no.cases <- 1000
no.controls <- 1000

# Define 400 samples
stichprobe <- data.frame(reps=1:repetitions,
                probe=matrix(NA, ncol=(no.cases+no.controls),
                nrow=repetitions))
# Draw samples
for (i in 1:repetitions){
  sample.ca <- sample(1:pop.size, no.cases)
  sample.co <- sample((pop.size+1):(2*pop.size), no.controls)
  stichprobe[i, 2:(no.cases+1)] <- sort(sample.ca)
  stichprobe[i, (no.cases+2):(no.cases+no.controls+1)] <-
    sort(sample.co)
}
# Save the results
setwd(dir.save)
write.csv2(stichprobe, "Stichprobe_Daten_JenaFinal.csv",
  row.names=FALSE)
```

D.2.2 Allele frequencies

Structure:

1. Preparations
2. Marker
   a) Read data and prepare it
   b) Function definitions
   c) Calculate allele frequencies
3. Null marker
   a) Calculate allele frequencies

```
# -------------------------------------------- #
# -------------------------------------------- #
# 1. Preparations
# -------------------------------------------- #
# -------------------------------------------- #

# Working directory
dir.read <- "Directory to read data"
dir.save <- "Directory to save results"

# Constant values
# Study
no.cases      <- 1000
no.controls   <- 1000
repetition    <- 400
# Marker
p.c              <- 0.05
```

```
penetrance.model <- "dominant
Dp.r2            <- data.frame(dp=1, r2=1)
# Null marker
no.null <- 10
GRR1 <- 1
GRR2 <- 1

# ---------------------------------------------- #
# ---------------------------------------------- #
# 2. Marker
# ---------------------------------------------- #
# ---------------------------------------------- #

# ---------------------------------------------- #
# a. Read data and prepare it
# ---------------------------------------------- #

# Read data: Incidence rates and age distribution
setwd(dir.read)
alter.inzidenz <- read.table("Globocan_Inzidenz.txt",
                    header=TRUE, sep="\t")
alter.distr <- read.table("EU25Pop.txt", header=TRUE,
                    sep=" ")

# Select colorectal cancer for incidence rates
alter.inzidenz <- alter.inzidenz[alter.inzidenz$Cancer==
                    "Colorectum", 4:12]
rownames(alter.inzidenz) <- "inzidenz"
alter.inzidenz <- t(alter.inzidenz)
inzidenz <- data.frame(alter=seq(35, 75, 5),
              inzidenz=alter.inzidenz)
rownames(inzidenz) <- 1:nrow(inzidenz)

# Select age intervals
alter.distr <- alter.distr[, 9:17]
alter.distr <- t(alter.distr)
alter <- data.frame(alter=seq(35, 75, 5), distr=alter.distr)
rownames(alter) <- 1:nrow(alter)

# GRR
grr.c.hom.mat     <- data.frame(age.int=alter[, 1], grr.parts=1.43)
```

```
# Combine age and incidence rates, calculate cummulative
# incidence
inzidenz <- data.frame(inzidenz, kumm=cumsum(inzidenz$inzidenz))
# Calculate percentages incidence
inzidenz <- data.frame(inzidenz, proz=inzidenz$kumm*5/100000)
# Incidence according to case-control status
inzidenz <- data.frame(inzidenz, fall=alter$distr*inzidenz$proz,
               kontrolle=alter$distr*(1-inzidenz$proz))
# Scaling within cases and controls to 100%
inzidenz <- data.frame(inzidenz,
               fall.proz=inzidenz$fall/sum(inzidenz$fall),
               kontrolle.proz=inzidenz$kontrolle/
               sum(inzidenz$kontrolle))
# Define age categories
inzidenz <- data.frame(inzidenz,
              fall.int=cumsum(inzidenz$fall.proz),
              kontrolle.int=cumsum(inzidenz$kontrolle.proz))

# Prevalence matrix
prevalence    <- inzidenz[, c(1, 4)]
colnames(prevalence) <- c("age.int", "prev.parts")

# -------------------------------------------- #
# b. Function definitions
# -------------------------------------------- #

# Calculate heterozygotic GRR
grr.c.het.expr <- expression({
  grr.c.het <- ifelse(pen.mod=="dominant", grr.c.hom,
                  ifelse(pen.mod == "recessive", 1,
                          0.5*(grr.c.hom+1)
                )
  )
})

# Calculate kappe_0
kappa.0.expr <- expression({
  kappa.0   <- prev/(pc*pc*grr.c.hom + 2*pc*(1-pc)*grr.c.het +
                       (1-pc)*(1-pc))
})

# Calculate p_M
```

```
p.m.expr <- expression({
  pm <- 1/(korr*(1-pc)/pc/dp/dp + 1)
  d   <- dp*(1-pm)*pc
})

# Calculate GRR according to Lorenzo Bermejo et al. (2011)
grr.m.expr <- expression({
  PrCM <- pm*pc+d
  PrCm <- (1-pm)*pc-d
  PrcM <- pm*(1-pc)-d
  Prcm <- (1-pm)*(1-pc)+d

  PrGCMgCM <- grr.c.hom*PrCM*PrCM
  PrGCMgCm <- grr.c.hom*2*PrCM*PrCm
  PrGCmgCm <- grr.c.hom*PrCm*PrCm
  PrGCMgcM <- grr.c.het*2*PrCM*PrcM
  PrGCMgcm <- grr.c.het*2*PrCM*Prcm
  PrGCmgcM <- grr.c.het*2*PrCm*PrcM
  PrGCmgcm <- grr.c.het*2*PrCm*Prcm
  PrGcMgcM <- PrcM*PrcM
  PrGcMgcm <- 2*PrcM*Prcm
  PrGcmgcm <- Prcm*Prcm

  kMMnum  <- PrGCMgCM+PrGCMgcM+PrGcMgcM
  kMMden  <- PrCM*PrCM+2*PrCM*PrcM+PrcM*PrcM
  kMM     <- kMMnum/kMMden

  kMmnum  <- PrGCMgCm+PrGCMgcm+PrGCmgcM+PrGcMgcm
  kMmden  <- 2*PrCM*PrCm+2*PrCM*Prcm+2*PrCm*PrcM+2*PrcM*Prcm
  kMm     <- kMmnum/kMmden

  kmmnum  <- PrGCmgCm+PrGCmgcm+PrGcmgcm
  kmmden  <- PrCm*PrCm+2*PrCm*Prcm+Prcm*Prcm
  kmm     <- kmmnum/kmmden

  grr.m.hom <- kMM/kmm
  grr.m.het <- kMm/kmm

})

# GRR age-dependent
# pA = P(M)
# Result as dataframe

# Allele frequency within cases and controls
# based on script by Justo Lorenzo Bermejo
allelfrequenz <- function(pA, GRR1, GRR2, prev){
  ausgabe <- data.frame(geno=c("aa", "Aa", "AA"), fall=NA,
                        kontrolle=NA)

  k <- prev
  enda=0
  f=0.00001
  while (enda==0){
    f=f+0.00001;
    #Genotypes among cases
    uno1=(1-pA)*(1-pA)*f;
    uno2=2*pA*(1-pA)*f*GRR2;
    uno3=pA*pA*f*GRR1;
    den=uno1+uno2+uno3;
    kp=den;
    pAA_case=uno1/den;
    pAB_case=uno2/den;
    pBB_case=uno3/den;

    #Genotypes among controls
    dos1=(1-pA)*(1-pA)*(1-f);
    dos2=2*pA*(1-pA)*(1-f*GRR2);
    dos3=pA*pA*(1-f*GRR1);

    den=dos1+dos2+dos3;
    pAA_cont=dos1/den;
    pAB_cont=dos2/den;
    pBB_cont=dos3/den;

    if (k <= kp) {enda=1};
  }

  ausgabe$fall <- matrix(c(pAA_case, pAB_case, pBB_case),
                         ncol=1)
  ausgabe$kontrolle <- matrix(c(pAA_cont, pAB_cont, pBB_cont),
                              ncol=1)

  return(ausgabe)
```

```
}

vergleich <- function(geno){
  geno.new <- ifelse(rn.gte<0.5 & geno!="MM", "MM",
                ifelse(rn.gte>=0.5 & geno!="mm", "mm", "Mm"))
  return(geno.new)
}

# -------------------------------------------- #
# c. Calculate allele frequencies
# -------------------------------------------- #

rownumber <- nrow(grr.c.hom.mat)*length(p.c)*
               length(penetrance.model)*nrow(Dp.r2)*3
allelfrequenzen <- data.frame(alter=rep(NA, rownumber),
                      prev=rep(NA, rownumber),
                      grr=rep(NA, rownumber),
                      pen.mod=rep(NA, rownumber),
                      pc=rep(NA, rownumber),
                      dp=rep(NA, rownumber),
                      r2=rep(NA, rownumber),
                      geno=rep(NA, rownumber),
                      fall=rep(NA, rownumber),
                      kontrolle=rep(NA, rownumber))
idx <- 1
for (dpr2 in 1:nrow(Dp.r2)){
  dp <- Dp.r2$dp[dpr2]
  korr <- Dp.r2$r2[dpr2]
  for(pen.mod in penetrance.model){
    for (pc in p.c){
      eval(p.m.expr)
      for (i in 1:length(grr.c.hom.mat$grr.parts)){
        grr.c.hom <- grr.c.hom.mat$grr.parts[i]
        eval(grr.c.het.expr)
        eval(grr.m.expr)
        prev <- prevalence$prev.parts[i]
        eval(kappa.0.expr)
        z1 <- allelfrequenz(pA=pm, GRR1=grr.m.hom, GRR2=grr.m.het,
                prev=prev)
        allelfrequenzen[idx:(idx+2),
          (length(allelfrequenzen)-1):(length(allelfrequenzen))] <-
          z1[, 2:3]
        allelfrequenzen$geno[idx:(idx+2)] <- c(0,1,2)
        allelfrequenzen$prev[idx:(idx+2)] <- prev
        allelfrequenzen$alter[idx:(idx+2)] <- prevalence$age.int[i]
        allelfrequenzen$grr[idx:(idx+2)] <- grr.c.hom
        allelfrequenzen$pen.mod[idx:(idx+2)] <- pen.mod
        allelfrequenzen$pc[idx:(idx+2)] <- pc
        allelfrequenzen$dp[idx:(idx+2)] <- dp
        allelfrequenzen$r2[idx:(idx+2)] <- korr
        idx <- idx+3
      }
    }
  }
}
setwd(dir.save)
titel <- "Allelfrequenzen_Grundlage_Fall_Kontrolle_EU.csv"
write.csv2(allelfrequenzen, titel)

# -------------------------------------------- #
# -------------------------------------------- #
# 3. NULL MARKER
# -------------------------------------------- #
# -------------------------------------------- #

set.seed(687541)

# MAF at locus j, j=1, ..., no.null
p0 <- runif(no.null, 0.05, 0.5)

# -------------------------------------------- #
# a. Calculate allele frequencies
# -------------------------------------------- #

rownumber <- length(p0)*nrow(prevalence)*3
allelfrequenzen <- data.frame(prev=rep(NA, rownumber),
                              p0=rep(NA, rownumber),
                              id=rep(NA, rownumber),
                              geno=rep(NA, rownumber),
                              fall=rep(NA, rownumber),
                              kontrolle=rep(NA, rownumber))
idx <- 1
idx.p0 <- 0
for (pc in p0){
  idx.p0 <- idx.p0+1
  for(prev in prevalence$prev.parts){
    z1 <- allelfrequenz(pA=pc, GRR1=GRR1, GRR2=GRR2, prev=prev)
    allelfrequenzen[idx:(idx+2),
      (length(allelfrequenzen)-1):(length(allelfrequenzen))] <-
      z1[, 2:3]
    allelfrequenzen$geno[idx:(idx+2)] <- z1$geno
    allelfrequenzen$prev[idx:(idx+2)] <- prev
    allelfrequenzen$p0[idx:(idx+2)] <- pc
    allelfrequenzen$id[idx:(idx+2)] <- idx.p0
    idx <- idx+3
  }
}
setwd(dir.save)
titel <- "Allelfrequenzen_Grundlage_Fall_Kontrolle_Nullmarker_EU.csv"
write.csv2(allelfrequenzen, titel)
```

D.2.3 Populations at marker locus

Structure:

1. Preparations
2. Marker
   a) Read data and prepare it
   b) Function definitions
   c) Age, GRR and genotypes

```
# -------------------------------------------- #
# -------------------------------------------- #
# 1. Preparations
# -------------------------------------------- #
# -------------------------------------------- #

# Working directories
dir.read <- "Directory to read data"
dir.save <- "Directory to save results"

# Constant values
# Number cases and controls
no.cases <- 3500000
no.controls <- 3500000
# Marker
pc <- 0.05
Dp.r2 <- data.frame(dp=1, r2=1)
pen.mod <- "dominant"

# Seed
set.seed(341950)

# ------------------------------------------- #
# ------------------------------------------- #
# 2. Marker
# ------------------------------------------- #
# ------------------------------------------- #

# ------------------------------------------- #
# a. Read data and prepare it
# ------------------------------------------- #

# Read data: Incidence rates and age distribution
setwd(dir.read)
alter.inzidenz <- read.table("Globocan_Inzidenz.txt",
                    header=TRUE, sep="\t")
alter.distr <- read.table("EU25Pop.txt", header=TRUE, sep=" ")

# Select colorectal cancer for incidence rates
alter.inzidenz <- alter.inzidenz[alter.inzidenz$Cancer==
                    "Colorectum", 4:12]
rownames(alter.inzidenz) <- "inzidenz"
alter.inzidenz <- t(alter.inzidenz)
inzidenz <- data.frame(alter=seq(35, 75, 5),
              inzidenz=alter.inzidenz)
rownames(inzidenz) <- 1:nrow(inzidenz)

# Select age intervals
alter.distr <- alter.distr[, 9:17]
alter.distr <- t(alter.distr)
alter <- data.frame(alter=seq(35, 75, 5), distr=alter.distr)
rownames(alter) <- 1:nrow(alter)

# Combine age and incidence rates, calculate cummulative
# incidence
```

```
inzidenz <- data.frame(inzidenz, kumm=cumsum(inzidenz$inzidenz))
# Calculate percentages incidence
inzidenz <- data.frame(inzidenz, proz=inzidenz$kumm*5/100000)
# Incidence according to case-control status
inzidenz <- data.frame(inzidenz,
               fall=alter$distr*inzidenz$proz,
               kontrolle=alter$distr*(1-inzidenz$proz))
# Scaling within cases and controls to 100%
inzidenz <- data.frame(inzidenz,
               fall.proz=inzidenz$fall/sum(inzidenz$fall),
               kontrolle.proz=inzidenz$kontrolle/sum(inzidenz$kontrolle))
# Define age categories
inzidenz <- data.frame(inzidenz, fall.int=cumsum(inzidenz$fall.proz),
               kontrolle.int=cumsum(inzidenz$kontrolle.proz))

# Prevalence matrix
prevalence    <- inzidenz[, c(1, 4)]
colnames(prevalence) <- c("age.int", "prev.parts")

# GRR
grr.c.hom.mat     <- data.frame(age.int=alter[, 1], grr.parts=1.43)

# Allele frequencies
allelfrequenzen <-
   read.csv2("Allelfrequenzen_Grundlage_Fall_Kontrolle_EU.csv")
allelfrequenzen <- allelfrequenzen[, 2:length(allelfrequenzen)]

# -------------------------------------------- #
# b. Function definitions
# -------------------------------------------- #

# Age calculation
alter.func <- function(zufall){
  if (zufall[2]=="control"){
     if (as.numeric(zufall[1])<min(inzidenz$kontrolle.int)){
        alter.aus <- min(inzidenz$alter)
     }
     if (as.numeric(zufall[1])>min(inzidenz$kontrolle.int)){
        alter.aus <- inzidenz$alter[max(which(inzidenz$kontrolle.int <
          as.numeric(zufall[1])))+1]
     }
     if (as.numeric(zufall[1])==1){
        alter.aus <- max(inzidenz$alter)
     }
  }
  if (zufall[2]=="case"){
    if (as.numeric(zufall[1])<min(inzidenz$fall.int)){
        alter.aus <- min(inzidenz$alter)
     }
     if (as.numeric(zufall[1])>min(inzidenz$fall.int) &
            as.numeric(zufall[1])<1){
        alter.aus <- inzidenz$alter[max(which(inzidenz$fall.int <
            as.numeric(zufall[1])))+1]
     }
     if (as.numeric(zufall[1])==1){
        alter.aus <- max(inzidenz$alter)
     }
  }
  return(alter.aus)
}

# Calculate heterozygotic GRR
grr.c.het.expr <- expression({
  grr.c.het <- ifelse(rep(pen.mod, no.cases+no.controls) ==
                 "dominant", grr.c.hom,
      ifelse(rep(pen.mod, no.cases+no.controls) == "recessive",
         rep(1, no.cases+no.controls), 0.5*(grr.c.hom+1)
      )
  )
})

# Calculate kappe_0
kappa.0.expr <- expression({
  kappa.0   <- prev.1000/(pc*pc*grr.c.hom + 2*pc*(1-pc)*grr.c.het +
      (1-pc)*(1-pc))
})

# Calculate p_M
p.m.expr <- expression({
  pm <- 1/(korr*(1-pc)/pc/dp/dp + 1)
  d   <- dp*(1-pm)*pc
})

# Calculate GRR according to Lorenzo Bermejo et al. (2011)
```

```
grr.m.expr <- expression({
  PrCM <- pm*pc+d
  PrCm <- (1-pm)*pc-d
  PrcM <- pm*(1-pc)-d
  Prcm <- (1-pm)*(1-pc)+d

  PrGCMgCM <- grr.c.hom*PrCM*PrCM
  PrGCMgCm <- grr.c.hom*2*PrCM*PrCm
  PrGCmgCm <- grr.c.hom*PrCm*PrCm
  PrGCMgcM <- grr.c.het*2*PrCM*PrcM
  PrGCMgcm <- grr.c.het*2*PrCM*Prcm
  PrGCmgcM <- grr.c.het*2*PrCm*PrcM
  PrGCmgcm <- grr.c.het*2*PrCm*Prcm
  PrGcMgcM <- PrcM*PrcM
  PrGcMgcm <- 2*PrcM*Prcm
  PrGcmgcm <- Prcm*Prcm

  kMMnum  <- PrGCMgCM+PrGCMgcM+PrGcMgcM
  kMMden  <- PrCM*PrCM+2*PrCM*PrcM+PrcM*PrcM
  kMM     <- kMMnum/kMMden

  kMmnum  <- PrGCMgCm+PrGCMgcm+PrGCmgcM+PrGcMgcm
  kMmden  <- 2*PrCM*PrCm+2*PrCM*Prcm+2*PrCm*PrcM+2*PrcM*Prcm
  kMm     <- kMmnum/kMmden

  kmmnum  <- PrGCmgCm+PrGCmgcm+PrGcmgcm
  kmmden  <- PrCm*PrCm+2*PrCm*Prcm+Prcm*Prcm
  kmm     <- kmmnum/kmmden

  grr.m.hom <- kMM/kmm
  grr.m.het <- kMm/kmm

})

# -------------------------------------------- #
# c. Age, GRR and genotypes
# -------------------------------------------- #

# Age
zufallszahlen <- data.frame(zahl=runif((no.cases+no.controls), 0, 1),
                   status=c(rep("case", no.cases),
                           rep("control", no.controls)))
```

```
alter <- apply(as.matrix(zufallszahlen), 1, alter.func)
daten <- data.frame(ca.co=c(rep(1, no.cases), rep(0, no.controls)),
           alter=alter, geno=NA)

# GRR
prev.1000 <- numeric(no.cases+no.controls)
for (prev.age in 1:length(alter)){
    prev.1000[prev.age] <- prevalence$prev.parts[
                    prevalence$age.int == alter[prev.age]]
}
grr.c.hom <- numeric(no.cases+no.controls)
for (grr.age in 1:length(alter)){
    grr.c.hom[grr.age] <- grr.c.hom.mat$grr.parts[
        grr.c.hom.mat$age.int == alter[grr.age]]
}
eval(grr.c.het.expr)
eval(kappa.0.expr)

# Remove objects that are not needed anymore
rm("inzidenz")
rm("vergleich")
rm("zufallszahlen")

# Genotypes
for (dp.r2 in 1:nrow(Dp.r2)){
    daten <- data.frame(ca.co=c(rep(1, no.cases),
        rep(0, no.controls)), alter=alter, geno=NA)
    dp <- Dp.r2$dp[dp.r2]
    korr <- Dp.r2$r2[dp.r2]
    eval(p.m.expr)    # p_M, d
    eval(grr.m.expr)  # GRR_M,hom , GRR_M,het
    rn <- runif(no.cases+no.controls, 0, 1)
    referenz <- data.frame(alter=alter, aa=NA, Aa=NA, AA=NA)
    bed01 <- allelfrequenzen$pen.mod==pen.mod &
      as.character(allelfrequenzen$pc)==as.character(pc) &
      as.character(allelfrequenzen$dp)==as.character(dp) &
      as.character(allelfrequenzen$r2)==as.character(korr)
    referenz.ca <- referenz[1:no.cases, ]
    for (a.schleife in prevalence$age.int){
      if (dim(referenz.ca[referenz.ca$alter==a.schleife, 2:4])[1]
           >0){
        referenz.ca[referenz.ca$alter==a.schleife, 2:4] <-
        data.frame(matrix(rep(allelfrequenzen$fall[bed01 &
          allelfrequenzen$grr ==
          grr.c.hom.mat$grr.parts[grr.c.hom.mat$age.int==
          a.schleife] &
          as.character(allelfrequenzen$prev) ==
          as.character(prevalence$prev.parts[
          prevalence$age.int == a.schleife])],
          nrow(referenz.ca[referenz.ca$alter ==
            a.schleife, 2:4])),
          nrow=nrow(referenz.ca[referenz.ca$alter ==
            a.schleife, 2:4]), byrow=TRUE))
   }
}
referenz.co <- referenz[(no.cases+1):(no.cases+no.controls), ]
for (a.schleife in prevalence$age.int){
  if(dim(referenz.co[referenz.co$alter==a.schleife, 2:4])[1]
        >0){
      referenz.co[referenz.co$alter==a.schleife, 2:4] <-
        data.frame(matrix(rep(allelfrequenzen$kontrolle[bed01 &
          allelfrequenzen$grr ==
            grr.c.hom.mat$grr.parts[grr.c.hom.mat$age.int==
            a.schleife] &
            as.character(allelfrequenzen$prev) ==
            as.character(prevalence$prev.parts[
            prevalence$age.int == a.schleife])],
            nrow(referenz.co[referenz.co$alter ==
            a.schleife, 2:4])),
            nrow=nrow(referenz.co[referenz.co$alter ==
            a.schleife, 2:4]), byrow=TRUE))
   }
}
referenz <- rbind(referenz.ca, referenz.co)
rm("referenz.co")
rm("referenz.ca")
z01 <- which(rn < referenz$AA)
daten$geno[z01] <- "MM"
z01 <- which((rn >= referenz$AA) & (rn < referenz$Aa + referenz$AA))
daten$geno[z01] <- "Mm"
z01 <- which(rn >= referenz$Aa + referenz$AA)
daten$geno[z01] <- "mm"
rm("referenz")
rm("rn")
```

```
    rm("z01")

    # Numerical genotypes according to a dominant penetrance model
    daten$geno[daten$geno=="MM"] <- 1
    daten$geno[daten$geno=="Mm"] <- 1
    daten$geno[daten$geno=="mm"] <- 0

    setwd(dir.save)
    titel <- "Daten_Bevoelkerung_EU_Geno_7Mio.csv"
    write.csv2(daten, titel, row.names=FALSE)
    rm("daten")
    rm("titel")
    gc()
    print(dp.r2)
}
```

D.2.4 Populations at null marker loci

Structure:

1. Preparations
2. Null marker
   a) Read data and prepare it
   b) Function definitions
   c) Age, GRR and genotypes

```
# -------------------------------------------- #
# -------------------------------------------- #
# 1. Preparations
# -------------------------------------------- #
# -------------------------------------------- #

# Working directories
dir.read <- "Directory to read data"
dir.save <- "Directory to save results"

# Constant values
# Number cases and controls
no.cases <- 3500000
no.controls <- 3500000
```

```
# Null marker
no.null <- 10
pen.mod <- "dominant"
# Seed
set.seed(687541)
# MAF at locus j, j=1, ..., no.null
p0 <- runif(no.null, 0.05, 0.5)

# Seed
set.seed(18041985)


# --------------------------------------------- #
# --------------------------------------------- #
# 2. Null marker
# --------------------------------------------- #
# --------------------------------------------- #

# --------------------------------------------- #
# a. Read data and prepare it
# --------------------------------------------- #

# Read data: Incidence rates and age distribution
setwd(dir.read)
alter.inzidenz <- read.table("Globocan_Inzidenz.txt", header=TRUE,
                    sep="\t")
alter.distr <- read.table("EU25Pop.txt", header=TRUE,
                    sep=" ")

# Select colorectal cancer for incidence rates
alter.inzidenz <- alter.inzidenz[alter.inzidenz$Cancer=="Colorectum",
                     4:12]
rownames(alter.inzidenz) <- "inzidenz"
alter.inzidenz <- t(alter.inzidenz)
inzidenz <- data.frame(alter=seq(35, 75, 5),
               inzidenz=alter.inzidenz)
rownames(inzidenz) <- 1:nrow(inzidenz)

# Select age intervals
alter.distr <- alter.distr[, 9:17]
alter.distr <- t(alter.distr)
```

```
alter <- data.frame(alter=seq(35, 75, 5), distr=alter.distr)
rownames(alter) <- 1:nrow(alter)

# GRR
grr.c.hom.mat      <- data.frame(age.int=alter[, 1], grr.parts=a.)

# Combine age and incidence rates, calculate cummulative incidence
inzidenz <- data.frame(inzidenz, kumm=cumsum(inzidenz$inzidenz))
# Calculate percentages incidence
inzidenz <- data.frame(inzidenz, proz=inzidenz$kumm*5/100000)
# Incidence according to case-control status
inzidenz <- data.frame(inzidenz, fall=alter$distr*inzidenz$proz,
              kontrolle=alter$distr*(1-inzidenz$proz))
# Scaling within cases and controls to 100%
inzidenz <- data.frame(inzidenz,
             fall.proz=inzidenz$fall/sum(inzidenz$fall),
             kontrolle.proz=inzidenz$kontrolle/
             sum(inzidenz$kontrolle))
# Define age categories
inzidenz <- data.frame(inzidenz,
             fall.int=cumsum(inzidenz$fall.proz),
             kontrolle.int=cumsum(inzidenz$kontrolle.proz))

# Prevalence matrix
prevalence    <- inzidenz[, c(1, 4)]
colnames(prevalence) <- c("age.int", "prev.parts")

# Allele frequencies
allelfrequenzen <-
  read.csv2(
  "Allelfrequenzen_Grundlage_Fall_Kontrolle_Nullmarker_EU.csv")
allelfrequenzen <- allelfrequenzen[, 2:length(allelfrequenzen)]

# -------------------------------------------- #
# b. Function definitions
# -------------------------------------------- #

# Age calculation
alter.func <- function(zufall){
  if (zufall[2]=="control"){
     if (as.numeric(zufall[1])<min(inzidenz$kontrolle.int)){
        alter.aus <- min(inzidenz$alter)
     }
     if (as.numeric(zufall[1])>min(inzidenz$kontrolle.int)){
        alter.aus <- inzidenz$alter[max(which(inzidenz$kontrolle.int <
                       as.numeric(zufall[1])))+1]
     }
     if (as.numeric(zufall[1])==1){
        alter.aus <- max(inzidenz$alter)
     }
  }
  if (zufall[2]=="case"){
    if (as.numeric(zufall[1])<min(inzidenz$fall.int)){
        alter.aus <- min(inzidenz$alter)
     }
     if (as.numeric(zufall[1])>min(inzidenz$fall.int) &
             as.numeric(zufall[1])<1){
        alter.aus <- inzidenz$alter[max(which(inzidenz$fall.int <
                        as.numeric(zufall[1])))+1]
     }
     if (as.numeric(zufall[1])==1){
        alter.aus <- max(inzidenz$alter)
     }
  }
  return(alter.aus)
}

# -------------------------------------------- #
# c. Age, GRR and genotypes
# -------------------------------------------- #

# Age
zufallszahlen <- data.frame(zahl=runif((no.cases+no.controls), 0, 1),
                   status=c(rep("case", no.cases),
                     rep("control", no.controls)))
alter <- apply(as.matrix(zufallszahlen), 1, alter.func)
rm("alter.distr")
rm("alter.func")
rm("alter.inzidenz")

# Prevalence
prev.1000 <- numeric(no.cases+no.controls)
for (prev.age in 1:length(alter)){
    prev.1000[prev.age] <- prevalence$prev.parts[prevalence$age.int ==
                             alter[prev.age]]
}
rm("inzidenz")
rm("zufallszahlen")

# Genotypes
daten <- data.frame(ca.co=c(rep(1, no.cases), rep(0, no.controls)),
                     alter=alter, geno=matrix(NA, ncol=no.null,
                       nrow=(no.cases+no.controls)))
idx <- 0
for (pc in p0){
  idx <- idx+1
  rn <- runif(no.cases+no.controls, 0, 1)
  referenz <- data.frame(alter=alter, aa=NA, Aa=NA, AA=NA)
  bed01 <- as.character(allelfrequenzen$p0)==as.character(pc)
  referenz.ca <- referenz[1:no.cases, ]
  for (a.schleife in prevalence$age.int){
    if (dim(referenz.ca[referenz.ca$alter==a.schleife, 2:4])[1]
          >0){
      referenz.ca[referenz.ca$alter==a.schleife, 2:4] <-
        data.frame(matrix(rep(allelfrequenzen$fall[bed01 &
        as.character(allelfrequenzen$prev) ==
        as.character(prevalence$prev.parts[prevalence$age.int ==
        a.schleife])],
        nrow(referenz.ca[referenz.ca$alter==a.schleife, 2:4])),
        nrow=nrow(referenz.ca[referenz.ca$alter==a.schleife, 2:4]),
        byrow=TRUE))
    }
  }
  referenz.co <- referenz[(no.cases+1):(no.cases+no.controls), ]
  for (a.schleife in prevalence$age.int){
    if(dim(referenz.co[referenz.co$alter==a.schleife, 2:4])[1]
         >0){
      referenz.co[referenz.co$alter==a.schleife, 2:4] <-
        data.frame(matrix(rep(allelfrequenzen$kontrolle[bed01 &
        as.character(allelfrequenzen$prev) ==
        as.character(prevalence$prev.parts[prevalence$age.int ==
        a.schleife])], nrow(referenz.co[referenz.co$alter ==
        a.schleife, 2:4])),
        nrow=nrow(referenz.co[referenz.co$alter==a.schleife, 2:4]),
        byrow=TRUE))
    }
  }
  referenz <- rbind(referenz.ca, referenz.co)
  z01 <- which(rn < referenz$AA)
  daten[z01, idx+2] <- "MM"
  z01 <- which((rn >= referenz$AA) & (rn < referenz$Aa + referenz$AA))
  daten[z01, idx+2] <- "Mm"
  z01 <- which(rn >= referenz$Aa + referenz$AA)
  daten[z01, idx+2] <- "mm"

  # Numerical genotypes according to a dominant penetrance model
  daten[daten[, idx+2]=="MM", idx+2] <- 1
  daten[daten[, idx+2]=="Mm", idx+2] <- 1
  daten[daten[, idx+2]=="mm", idx+2] <- 0

  print(idx)

}
# Save results
setwd(dir.save)
titel <- paste("Daten_Bevoelkerung_EU_Geno_7Mio_Nullmarker.csv", sep="")
write.csv2(daten, titel, row.names=FALSE)
```

D.2.5 Logistic regression at the marker locus and the statistical properties bias, variance, mean squared error and statistical power

Structure:

1. Preparations
2. Logistic regression
   a) Standard
   b) Huber
   c) Hampel
3. Statistical properties
   a) Power
   b) MSE

```
# -------------------------------------------- #
# -------------------------------------------- #
# 1. Preparations
# -------------------------------------------- #
# -------------------------------------------- #

# Working directory
dir.read <- "Directory to read data"
dir.save <- "Directory to save results"

# -------------------------------------------- #
# -------------------------------------------- #
# 2. Logistic regression
# -------------------------------------------- #
# -------------------------------------------- #

# -------------------------------------------- #
# a. Standard
# -------------------------------------------- #

# Read data
setwd(dir.read)
dominant <- read.csv2("Daten_Bevoelkerung_EU_Geno_7Mio.csv ")

# Read matrix with randomly drawn individuals from the population
stichprobe <- read.csv2("Stichprobe_Daten_JenaFinal.csv")

# Settings
repetitions <- 400
pop.size <- 3500000
no.cases <- 1000
no.controls <- 1000

speicher.stand <- data.frame(reps=1:repetitions, pVal.dom=NA,
                    koeff.dom=NA, se.dom=NA, ci.low.dom=NA,
                    ci.up.dom=NA)
for (i in 1:repetitions){
    # Get case control status for sample i
    ca.co <- dominant$ca.co[as.vector(unlist(stichprobe[i,
              2:length(stichprobe)]))]
    # Get age for sample i
    alter <- dominant$alter[as.vector(unlist(stichprobe[i,
              2:length(stichprobe)]))]
    # Genotype for sample i
    geno.dom <- dominant$geno[as.vector(unlist(stichprobe[i,
              2:length(stichprobe)]))]

    # Regression
    zwischen1 <- glm(ca.co ~ geno.dom + alter,
                      family=binomial(link = "logit"))

    # Saving
    speicher.stand$pVal.dom[i] <-
      summary(zwischen1)$coefficients[2,4]
    speicher.stand$koeff.dom[i] <-
      summary(zwischen1)$coefficients[2,1]
    speicher.stand$se.dom[i] <- summary(zwischen1)$coefficients[2,2]
    speicher.stand$ci.low.dom[i] <- speicher.stand$koeff.dom[i] -
                                    1.96*speicher.stand$se.dom[i]
    speicher.stand$ci.up.dom[i] <- speicher.stand$koeff.dom[i] +
                                    1.96*speicher.stand$se.dom[i]

    cat(paste("Rep ", i, "\n", sep=""))
}
# Save results
setwd(dir.save)
write.csv2(speicher.stand, "Standard_DomSimu.csv", row.names=FALSE)

# -------------------------------------------- #
# b. Huber
# -------------------------------------------- #

# Read data
setwd(dir.read)
dominant <- read.csv2(
    "Daten_Bevoelkerung_EU_7Mio_341950_domRef_domNum.csv")

# Read matrix with randomly drawn individuals from the population
stichprobe <- read.csv2("Stichprobe_Daten_JenaFinal.csv")

# Settings
repetitions <- 400
pop.size <- 3500000
no.cases <- 1000
no.controls <- 1000

library(robustbase)
tun.huber <- 1.345
speicher.huber <- data.frame(reps=1:repetitions, pVal.dom=NA,
     koeff.dom=NA, se.dom=NA, ci.low.dom=NA, ci.up.dom=NA)
for (i in 1:repetitions){
    # Get case control status for sample i
    ca.co <- dominant$ca.co[as.vector(unlist(stichprobe[i,
               2:length(stichprobe)]))]
    # Get age for sample i
    alter <- dominant$alter[as.vector(unlist(stichprobe[i,
              2:length(stichprobe)]))]
    # Genotype for sample i
    geno.dom <- dominant$geno[as.vector(unlist(stichprobe[i,
              2:length(stichprobe)]))]

    # Regression
    zwischen1 <- try(glmrob(ca.co ~ geno.dom + alter,
                   family=binomial("logit"), weights.on.x = "hat",
                   tcc=tun.huber),
                   silent=TRUE)

    # Saving
    if (sum(class(zwischen1)=="try-error")==0){
      speicher.huber$pVal.dom[i] <-
        summary(zwischen1)$coefficients[2,4]
      speicher.huber$koeff.dom[i] <-
        summary(zwischen1)$coefficients[2,1]
      speicher.huber$se.dom[i] <-
        summary(zwischen1)$coefficients[2,2]
      speicher.huber$ci.low.dom[i] <-
        speicher.huber$koeff.dom[i] - 1.96*speicher.huber$se.dom[i]
      speicher.huber$ci.up.dom[i] <-
        speicher.huber$koeff.dom[i] + 1.96*speicher.huber$se.dom[i]
    }

    cat(paste("Rep ", i, "\n", sep=""))
}
# Save results
setwd(dir.save)
write.csv2(speicher.huber, "Huber_DomSimu.csv", row.names=FALSE)
```

```
detach("package:robustbase")

# ---------------------------------------------- #
# c. Hampel
# ---------------------------------------------- #

# Read data
setwd(dir.read)
dominant <- read.csv2(
    "Daten_Bevoelkerung_EU_7Mio_341950_domRef_domNum.csv")

# Read matrix with randomly drawn individuals from the population
stichprobe <- read.csv2("Stichprobe_Daten_JenaFinal.csv")

# Settings
repetitions <- 400
pop.size <- 3500000
no.cases <- 1000
no.controls <- 1000

library(robustbaseAdj)
tun.hampel1 <- c(1.5, 3.5, 8)*0.9016085
tun.hampel2 <- c(2, 4, 8)*0.690794

speicher.hampel <- data.frame(reps=1:repetitions, pVal.dom.09=NA,
                              koeff.dom.09=NA,
                              se.dom.09=NA,
                              ci.low.dom.09=NA,
                              ci.up.dom.09=NA,
                              pVal.dom.69=NA,
                              koeff.dom.69=NA,
                              se.dom.69=NA,
                              ci.low.dom.69=NA,
                              ci.up.dom.69=NA)
for (i in 1:repetitions){
    # Get case control status for sample i
    ca.co <- dominant$ca.co[as.vector(unlist(stichprobe[i,
      2:length(stichprobe)]))]
    # Get age for sample i
    alter <- dominant$alter[as.vector(unlist(stichprobe[i,
      2:length(stichprobe)]))]
    # Genotype for sample i
```

```
geno.dom <- dominant$geno[as.vector(unlist(stichprobe[i,
  2:length(stichprobe)]))]

# Regression with 1st tuning constant
zwischen1 <- try(glmrob(ca.co ~ geno.dom + alter,
               family=binomial("logit"),
               weights.on.x = "hat", tcc=tun.hampel1), silent=TRUE)

# Saving
if (sum(class(zwischen1)=="try-error")==0){
  speicher.hampel$pVal.dom.09[i] <-
    summary(zwischen1)$coefficients[2,4]
  speicher.hampel$koeff.dom.09[i] <-
    summary(zwischen1)$coefficients[2,1]
  speicher.hampel$se.dom.09[i] <-
      summary(zwischen1)$coefficients[2,2]
  speicher.hampel$ci.low.dom.09[i] <-
      speicher.hampel$koeff.dom.09[i] -
      1.96*speicher.hampel$se.dom.09[i]
  speicher.hampel$ci.up.dom.09[i] <-
      speicher.hampel$koeff.dom.09[i] +
      1.96*speicher.hampel$se.dom.09[i]
}

# Regression with 2nd tuning constant
zwischen1 <- try(glmrob(ca.co ~ geno.dom + alter,
     family=binomial("logit"),
     weights.on.x = "hat", tcc=tun.hampel2), silent=TRUE)

# Saving
if (sum(class(zwischen1)=="try-error")==0){
  speicher.hampel$pVal.dom.69[i] <-
    summary(zwischen1)$coefficients[2,4]
  speicher.hampel$koeff.dom.69[i] <-
    summary(zwischen1)$coefficients[2,1]
  speicher.hampel$se.dom.69[i] <-
      summary(zwischen1)$coefficients[2,2]
  speicher.hampel$ci.low.dom.69[i] <-
      speicher.hampel$koeff.dom.69[i] -
      1.96*speicher.hampel$se.dom.69[i]
  speicher.hampel$ci.up.dom.69[i] <-
      speicher.hampel$koeff.dom.69[i] +
          1.96*speicher.hampel$se.dom.69[i]
    }

    cat(paste("Rep ", i, "\n", sep=""))
}
# Save results
setwd(dir.save)
write.csv2(speicher.hampel, "Hampel_DomSimu.csv", row.names=FALSE)
detach("package:robustbaseAdj")


# -------------------------------------------- #
# -------------------------------------------- #
# 2. Statistical properties
# -------------------------------------------- #
# -------------------------------------------- #

rm(list=ls())

# -------------------------------------------- #
# a. Power
# -------------------------------------------- #

setwd(dir.save)
speicher.stand <- read.csv2("Standard_DomSimu.csv")
speicher.huber <- read.csv2("Huber_DomSimu.csv")
speicher.hampel <- read.csv2("Hampel_DomSimu.csv")

# Standard
power.speicher <- data.frame(Method=c("Standard", "Huber",
    "Hampel", ""),
    Tuning=c("none", "1.345", "(1.5,3.5,8)*0.9016085",
    "(2,4,8)*0.690794"), Dominant=NA)
power.speicher$Dominant[1] <-
  round(length(speicher.stand$pVal.dom[
  is.na(speicher.stand$pVal.dom)==F &
  speicher.stand$pVal.dom<0.05]) /
  length(speicher.stand$pVal.dom[
  is.na(speicher.stand$pVal.dom)==F])*100,
  digits=1)

# Huber
```

```
power.speicher$Dominant[2] <-
  round(length(speicher.huber$pVal.dom[
  is.na(speicher.huber$pVal.dom)==F &
  speicher.huber$pVal.dom<0.05]) /
  length(speicher.huber$pVal.dom[
  is.na(speicher.huber$pVal.dom)==F])*100,
  digits=1)

# Hampel with 1st tuning constant
power.speicher$Dominant[3] <-
  round(length(speicher.hampel$pVal.dom.09[
  is.na(speicher.hampel$pVal.dom.09)==F &
  speicher.hampel$pVal.dom.09<0.05]) /
  length(speicher.hampel$pVal.dom.09[
  is.na(speicher.hampel$pVal.dom.09)==F])*100,
  digits=1)

# Hampel with 2nd tuning constant
power.speicher$Dominant[4] <-
  round(length(speicher.hampel$pVal.dom.69[
  is.na(speicher.hampel$pVal.dom.69)==F &
  speicher.hampel$pVal.dom.69<0.05]) /
  length(speicher.hampel$pVal.dom.69[
  is.na(speicher.hampel$pVal.dom.69)==F])*100,
  digits=1)

# Save results
write.table(power.speicher, "Power_SimuDom.txt",
  row.names=F, quote=F, sep=";", dec=".")

# ------------------------------------------- #
# b. MSE
# ------------------------------------------- #

ref <- log(1.43)
mse.speicher <- data.frame(Method=c("Standard", "Huber",
  "Hampel", ""),
  Tuning=c("none", "1.345", "(1.5,3.5,8)*0.9016085",
  "(2,4,8)*0.690794"),
  Dominant.MeanBias=NA, Dominant.Variance=NA,
  Dominant.MeanSE=NA)
```

```
# Mean bias
mse.speicher$Dominant.MeanBias[1] <-
  round(mean(speicher.stand$koeff.dom)-ref,
  digits=4)
mse.speicher$Dominant.MeanBias[2] <-
  round(mean(speicher.huber$koeff.dom)-ref,
  digits=4)
mse.speicher$Dominant.MeanBias[3] <-
  round(mean(speicher.hampel$koeff.dom.09)-ref,
  digits=4)
mse.speicher$Dominant.MeanBias[4] <-
  round(mean(speicher.hampel$koeff.dom.69)-ref,
  digits=4)

# Variance
mse.speicher$Dominant.Variance[1] <-
  round(var(speicher.stand$koeff.dom),
  digits=4)
mse.speicher$Dominant.Variance[2] <-
  round(var(speicher.huber$koeff.dom),
  digits=4)
mse.speicher$Dominant.Variance[3] <-
  round(var(speicher.hampel$koeff.dom.09),
  digits=4)
mse.speicher$Dominant.Variance[4] <-
  round(var(speicher.hampel$koeff.dom.69),
  digits=4)

# Mean squared error
mse.speicher$Dominant.MeanSE[1] <-
  round((mse.speicher$Dominant.MeanBias[1])^2 +
  mse.speicher$Dominant.Variance[1],
  digits=4)
mse.speicher$Dominant.MeanSE[2] <-
  round((mse.speicher$Dominant.MeanBias[2])^2 +
  mse.speicher$Dominant.Variance[2],
  digits=4)
mse.speicher$Dominant.MeanSE[3] <-
  round((mse.speicher$Dominant.MeanBias[3])^2 +
  mse.speicher$Dominant.Variance[3],
  digits=4)
mse.speicher$Dominant.MeanSE[4] <-
```

```
  round((mse.speicher$Dominant.MeanBias[4])^2 +
  mse.speicher$Dominant.Variance[4],
  digits=4)

# Save results
write.table(mse.speicher, "MSE_SimuDom.txt", row.names=F,
  quote=F, sep=";", dec=".")
```

D.2.6 Logistic regression at the null marker loci and the type I error rate

Structure:

1. Preparations
2. Logistic regression
   a) Standard
   b) Huber
   c) Hampel
3. Statistical properties
   a) Type I error rate

```
# ------------------------------------------- #
# ------------------------------------------- #
# 1. Preparations
# ------------------------------------------- #
# ------------------------------------------- #

# Working directory
dir.read <- "Directory to read data"
dir.save <- "Directory to save results"

# ------------------------------------------- #
# ------------------------------------------- #
# 2. Logistic regression
# ------------------------------------------- #
# ------------------------------------------- #

# ------------------------------------------- #
# a. Standard
# ------------------------------------------- #

# Read data
setwd(dir.read)
daten <- read.csv2(
    "Daten_Bevoelkerung_EU_Geno_7Mio_Nullmarker.csv")

# Read matrix with randomly drawn individuals from the population
stichprobe <- read.csv2("Stichprobe_Daten_JenaFinal.csv")

# Settings
repetitions <- 400
pop.size <- 3500000
no.cases <- 1000
no.controls <- 1000

speicher.stand <- data.frame(reps=rep(1:repetitions, 10),
  p0=sort(rep(1:10, 400)), pVal.dom=NA, koeff.dom=NA,
  se.dom=NA, ci.low.dom=NA, ci.up.dom=NA)
idx <- 0
for (pc in 1:10){
   for (i in 1:repetitions){
      idx <- idx+1
      # Get case control status for sample i
      ca.co <- daten$ca.co[as.vector(unlist(stichprobe[i,
        2:length(stichprobe)]))]
      # Get age for sample i
      alter <- daten$alter[as.vector(unlist(stichprobe[i,
        2:length(stichprobe)]))]
      # Genotype for sample i
      geno <- daten[as.vector(unlist(stichprobe[i,
        2:length(stichprobe)])), pc+2]
      geno <- as.character(geno)
      geno.dom <- geno
      geno.dom[geno.dom=="MM"] <- "1"
      geno.dom[geno.dom=="Mm"] <- "1"
      geno.dom[geno.dom=="mm"] <- "0"
      geno.dom <- as.numeric(geno.dom)

      # Regression
      zwischen1 <- glm(ca.co ~ geno.dom + alter,
        family=binomial(link = "logit"))

      # Saving
      speicher.stand$pVal.dom[idx] <-
        summary(zwischen1)$coefficients[2,4]
      speicher.stand$koeff.dom[idx] <-
        summary(zwischen1)$coefficients[2,1]
      speicher.stand$se.dom[idx] <-
        summary(zwischen1)$coefficients[2,2]
      speicher.stand$ci.low.dom[idx] <-
        speicher.stand$koeff.dom[idx] -
        1.96*speicher.stand$se.dom[idx]
      speicher.stand$ci.up.dom[idx] <-
        speicher.stand$koeff.dom[idx] +
        1.96*speicher.stand$se.dom[idx]
   }
   print(pc)
}
# Save results
setwd(dir.save)
write.csv2(speicher.stand, "Standard_DomSimu_Nullmarker.csv",
  row.names=FALSE)

# -------------------------------------------- #
# b. Huber
# -------------------------------------------- #

# Read data
setwd(dir.read)
daten <- read.csv2("Daten_Bevoelkerung_EU_Geno_7Mio_Nullmarker.csv")

# Read matrix with randomly drawn individuals from the population
stichprobe <- read.csv2("Stichprobe_Daten_JenaFinal.csv")

# Settings
repetitions <- 400
pop.size <- 3500000
no.cases <- 1000
no.controls <- 1000

library(robustbase)
tun.huber <- 1.345
```

```
speicher.huber <- data.frame(reps=rep(1:repetitions, 10),
  p0=sort(rep(1:10, 400)), pVal.dom=NA, koeff.dom=NA,
  se.dom=NA, ci.low.dom=NA, ci.up.dom=NA)
idx <- 0
for (pc in 1:10){
   for (i in 1:repetitions){
      idx <- idx+1
      # Get case control status for sample i
      ca.co <- daten$ca.co[as.vector(unlist(stichprobe[i,
        2:length(stichprobe)]))]
      # Get age for sample i
      alter <- daten$alter[as.vector(unlist(stichprobe[i,
        2:length(stichprobe)]))]
      # Genotype for sample i
      geno <- daten[as.vector(unlist(stichprobe[i,
        2:length(stichprobe)])), pc+2]
      geno <- as.character(geno)
      geno.dom <- geno
      geno.dom[geno.dom=="MM"] <- "1"
      geno.dom[geno.dom=="Mm"] <- "1"
      geno.dom[geno.dom=="mm"] <- "0"
      geno.dom <- as.numeric(geno.dom)

      # Regression
      zwischen1 <- try(glmrob(ca.co ~ geno.dom + alter,
        family=binomial("logit"),
        weights.on.x = "hat",tcc=tun.huber),
        silent=TRUE)

      # Saving
      if (sum(class(zwischen1)=="try-error")==0){
         speicher.huber$pVal.dom[idx] <-
          summary(zwischen1)$coefficients[2,4]
         speicher.huber$koeff.dom[idx] <-
          summary(zwischen1)$coefficients[2,1]
         speicher.huber$se.dom[idx] <-
          summary(zwischen1)$coefficients[2,2]
         speicher.huber$ci.low.dom[idx] <-
          speicher.huber$koeff.dom[idx] -
          1.96*speicher.huber$se.dom[idx]
         speicher.huber$ci.up.dom[idx] <-
          speicher.huber$koeff.dom[idx] +
          1.96*speicher.huber$se.dom[idx]
      }
   }
print(pc)
}
# Save results
setwd(dir.save)
write.csv2(speicher.huber, "Huber_DomSimu_Nullmarker.csv",
  row.names=FALSE)
detach("package:robustbase")

# -------------------------------------------- #
# c. Hampel
# -------------------------------------------- #

# Read data
setwd(dir.read)
daten <- read.csv2("Daten_Bevoelkerung_EU_Geno_7Mio_Nullmarker.csv")

# Read matrix with randomly drawn individuals from the population
stichprobe <- read.csv2("Stichprobe_Daten_JenaFinal.csv")

# Settings
repetitions <- 400
pop.size <- 3500000
no.cases <- 1000
no.controls <- 1000

library(robustbaseAdj)
tun.hampel1 <- c(1.5, 3.5, 8)*0.9016085
tun.hampel2 <- c(2, 4, 8)*0.690794

speicher.hampel <- data.frame(reps=rep(1:repetitions, 10),
  p0=sort(rep(1:10, 400)), pVal.dom.09=NA, koeff.dom.09=NA,
  se.dom.09=NA, ci.low.dom.09=NA, ci.up.dom.09=NA,
  pVal.dom.69=NA, koeff.dom.69=NA,
  se.dom.69=NA, ci.low.dom.69=NA, ci.up.dom.69=NA)
idx <- 0
for (pc in 1:10){
   for (i in 1:repetitions){
      idx <- idx+1
      # Get case control status for sample i
```

```
ca.co <- daten$ca.co[as.vector(unlist(stichprobe[i,
    2:length(stichprobe)]))]
# Get age for sample i
alter <- daten$alter[as.vector(unlist(stichprobe[i,
    2:length(stichprobe)]))]
# Genotype for sample i
geno <- daten[as.vector(unlist(stichprobe[i,
    2:length(stichprobe)])), pc+2]
geno <- as.character(geno)
geno.dom <- geno
geno.dom[geno.dom=="MM"] <- "1"
geno.dom[geno.dom=="Mm"] <- "1"
geno.dom[geno.dom=="mm"] <- "0"
geno.dom <- as.numeric(geno.dom)

# Regression with 1st tuning constant
zwischen1 <- try(glmrob(ca.co ~ geno.dom + alter,
  family=binomial("logit"), weights.on.x = "hat",
  tcc=tun.hampel1), silent=TRUE)

# Saving
if (sum(class(zwischen1)=="try-error")==0){
   speicher.hampel$pVal.dom.09[idx] <-
     summary(zwischen1)$coefficients[2,4]
   speicher.hampel$koeff.dom.09[idx] <-
     summary(zwischen1)$coefficients[2,1]
   speicher.hampel$se.dom.09[idx] <-
     summary(zwischen1)$coefficients[2,2]
   speicher.hampel$ci.low.dom.09[idx] <-
     speicher.hampel$koeff.dom.09[idx] -
     1.96*speicher.hampel$se.dom.09[idx]
   speicher.hampel$ci.up.dom.09[idx] <-
     speicher.hampel$koeff.dom.09[idx] +
     1.96*speicher.hampel$se.dom.09[idx]
}

# Regression with 2nd tuning constant
zwischen1 <- try(glmrob(ca.co ~ geno.dom + alter,
  family=binomial("logit"), weights.on.x = "hat",
  tcc=tun.hampel2), silent=TRUE)

# Saving
```

```
      if (sum(class(zwischen1)=="try-error")==0){
         speicher.hampel$pVal.dom.69[idx] <-
           summary(zwischen1)$coefficients[2,4]
         speicher.hampel$koeff.dom.69[idx] <-
           summary(zwischen1)$coefficients[2,1]
         speicher.hampel$se.dom.69[idx] <-
           summary(zwischen1)$coefficients[2,2]
         speicher.hampel$ci.low.dom.69[idx] <-
           speicher.hampel$koeff.dom.69[idx] -
           1.96*speicher.hampel$se.dom.69[idx]
         speicher.hampel$ci.up.dom.69[idx] <-
           speicher.hampel$koeff.dom.69[idx] +
           1.96*speicher.hampel$se.dom.69[idx]
      }
   }
   print(pc)
}
# Save results
setwd(dir.save)
write.csv2(speicher.hampel, "Hampel_DomSimu_Nullmarker.csv",
  row.names=FALSE)
detach("package:robustbaseAdj")


# -------------------------------------------- #
# -------------------------------------------- #
# 3. Statistical properties
# -------------------------------------------- #
# -------------------------------------------- #

rm(list=ls())

# -------------------------------------------- #
# a. Type I error rate
# -------------------------------------------- #

setwd(dir.save)
speicher.stand <- read.csv2("Standard_DomSimu_Nullmarker.csv")
speicher.huber <- read.csv2("Huber_DomSimu_Nullmarker.csv")
speicher.hampel <- read.csv2("Hampel_DomSimu_Nullmarker.csv")

fpr.speicher <- data.frame(Method=c("Standard", "Huber",
                   "Hampel", ""),
  Tuning=c("none", "1.345", "(1.5,3.5,8)*0.9016085",
    "(2,4,8)*0.690794"),
  FPR.dom=NA)

# Standard
fpr <- round(length(speicher.stand$pVal.dom[
  is.na(speicher.stand$pVal.dom)==F &
  speicher.stand$pVal.dom<0.05]) /
  length(speicher.stand$pVal.dom[
  is.na(speicher.stand$pVal.dom)==F])*100,
  digits=1)
low <- round((fpr/100-1.96*sqrt(fpr/100*(1-fpr/100)/4000))*100,
  digits=1)
up <- round((fpr/100+1.96*sqrt(fpr/100*(1-fpr/100)/4000) )*100,
  digits=1)
fpr.speicher$FPR.dom[1] <- paste(fpr, " (", low, ", ", up, ")",
  sep="")
# Huber
fpr <- round(length(speicher.huber$pVal.dom[
  is.na(speicher.huber$pVal.dom)==F &
  speicher.huber$pVal.dom<0.05]) /
  length(speicher.huber$pVal.dom[
  is.na(speicher.huber$pVal.dom)==F])*100,
  digits=1)
low <- round((fpr/100-1.96*sqrt(fpr/100*(1-fpr/100)/4000))*100,
  digits=1)
up <- round((fpr/100+1.96*sqrt(fpr/100*(1-fpr/100)/4000) )*100,
  digits=1)
fpr.speicher$FPR.dom[2] <- paste(fpr, " (", low, ", ", up, ")",
  sep="")
# Hampel with 1st tuning constant
fpr <- round(length(speicher.hampel$pVal.dom.09[
  is.na(speicher.hampel$pVal.dom.09)==F &
  speicher.hampel$pVal.dom.09<0.05]) /
  length(speicher.hampel$pVal.dom.09[
  is.na(speicher.hampel$pVal.dom.09)==F])*100,
  digits=1)
low <- round((fpr/100-1.96*sqrt(fpr/100*(1-fpr/100)/4000))*100,
  digits=1)
up <- round((fpr/100+1.96*sqrt(fpr/100*(1-fpr/100)/4000) )*100,
  digits=1)
```

```
fpr.speicher$FPR.dom[3] <- paste(fpr, " (", low, ", ", up, ")",
  sep="")
# Hampel with 2nd tuning constant
fpr <- round(length(speicher.hampel$pVal.dom.69[
  is.na(speicher.hampel$pVal.dom.69)==F &
  speicher.hampel$pVal.dom.69<0.05]) /
  length(speicher.hampel$pVal.dom.69[
  is.na(speicher.hampel$pVal.dom.69)==F])*100,
  digits=1)
low <- round((fpr/100-1.96*sqrt(fpr/100*(1-fpr/100)/4000))*100,
  digits=1)
up <- round((fpr/100+1.96*sqrt(fpr/100*(1-fpr/100)/4000) )*100,
  digits=1)
fpr.speicher$FPR.dom[4] <- paste(fpr, " (", low, ", ", up, ")",
  sep="")

# Save results
write.table(fpr.speicher, "FPR_SimuDom_Nullmarker.txt", row.names=F,
  quote=F, sep=";", dec=".")
```

D.3 Real data application for statistical properties evaluation

Code for the analysis of the real data (section D.3.1) as well as the visualisation of their results (section D.3.2)

D.3.1 Analysis of the real data

Structure:

1. Input formats
2. Settings
3. Preparations
4. Logistic regression

```
# ------------------------------------------- #
# ------------------------------------------- #
# 1. Input formats
# -------------------------------------------- #
# -------------------------------------------- #

# Genotype data:
#   .csv-file
#   as given by PGP

# Demographic data:
#   .csv-file
#     column names:
#     id            : given PGP
#     age.yrs       : age in years
#     height.cm     : height in cm
#     DataAvailable : is genotype data available?


# -------------------------------------------- #
# -------------------------------------------- #
# 2. Settings
# -------------------------------------------- #
# -------------------------------------------- #

genotype.dir    <- "Directory of genotype data"
demographic.dir <- "Directory of demographic information"
save.dir        <- "Directory to save analysis results"

# Number of SNPs to read
no.snps <- 1000


# -------------------------------------------- #
# -------------------------------------------- #
# 3. Preparations
# -------------------------------------------- #
# -------------------------------------------- #

# Get a list of genotype files
setwd(genotype.dir)
dateien <- dir()

# If there is more than one file for an individual,
# take the latest one
# Which individuals do have several?
personen <- substr(dateien, start=1, stop=8)
mehrere <- which(duplicated(personen)==T)
mehrere <- unique(personen[mehrere])
# Manual exclusion
dateien.auswahl <- dateien[dateien !=
  "hu11603C_genome_Angela_Harris_Full_20120618075158.txt"]
dateien.auswahl <- dateien.auswahl[dateien.auswahl !=
  "hu2A4D22_genome_Stephan_George_Full_20130210221109.txt"]
dateien.auswahl <- dateien.auswahl[dateien.auswahl !=
  "hu394092_genome_Paul_Conroy_Full_20110111011125.txt"]
dateien.auswahl <- dateien.auswahl[dateien.auswahl !=
  "hu3D355A_20110727023010.txt"]
dateien.auswahl <- dateien.auswahl[dateien.auswahl !=
  "hu459AD0_genome_Bernard_Moscia_Full_20110116053218.txt"]
dateien.auswahl <- dateien.auswahl[dateien.auswahl !=
  "hu5A1D5F_genome_Matthew_Kelty_Mito_20110331040943.txt"]
  # no mitochondrial or Y chromosome
dateien.auswahl <- dateien.auswahl[dateien.auswahl !=
  "hu5A1D5F_genome_Matthew_Kelty_Y_20110331040918.txt"]
  # no mitochondrial or Y chromosome
dateien.auswahl <- dateien.auswahl[dateien.auswahl !=
  "hu6ED94A_genome_Norman_Megill_Full_20100527043516.txt"]
dateien.auswahl <- dateien.auswahl[dateien.auswahl !=
  "hu6FECE9_genome_jim_berry_Full_20110112103611.txt"]
dateien.auswahl <- dateien.auswahl[dateien.auswahl !=
  "hu840B0B_genome_Brandon_Galbraith_Full_20110124131334.txt"]
dateien.auswahl <- dateien.auswahl[dateien.auswahl !=
  "huAC827A_genome_Jim_Turner_Full_20110324084155.txt"]
dateien.auswahl <- dateien.auswahl[dateien.auswahl !=
  "huBC03A7_genome_Debra_Patek_Full_20110107081441.txt"]
dateien.auswahl <- dateien.auswahl[dateien.auswahl !=
  "huC4A276_genome_Anastasia_Webber_Full_20110910195237.txt"]
dateien.auswahl <- dateien.auswahl[dateien.auswahl !=
  "huC92BC9_genome_William_Ramey_Mito_20120508065539.txt"]
  # no mitochondrial or Y chromosome
dateien.auswahl <- dateien.auswahl[dateien.auswahl !=
  "huC92BC9_genome_William_Ramey_Y_20120508065802.txt"]
  # no mitochondrial or Y chromosome
dateien.auswahl <- dateien.auswahl[dateien.auswahl !=
  "huD57BBF_genome_James_Vick_Full_20101216062019.txt"]
```

```
dateien.auswahl <- dateien.auswahl[dateien.auswahl !=
  "huDB1635_20110727031252.txt"]
dateien.auswahl <- dateien.auswahl[dateien.auswahl !=
  "huDD1522_genome_Ken_Mortimer_Full_20140223140859.txt"]
dateien.auswahl <- dateien.auswahl[dateien.auswahl !=
  "huF06AD0_genome_Beau_Gunderson_Full_20110307220402.txt"]

# Read clinical data and exclude individuals without
# information about age, genotype or height
# Read
setwd(demographic.dir)
klin <- read.csv2(
    "KlinischeDaten_perHandAusgelesen_23anMe_Verwendet.csv")
# Exclusion: no age or no genotype data
klin <- klin[is.na(klin$age.yrs)==F, ]
ausschluss <- grep("no", klin$DataAvailable)
klin <- klin[-ausschluss, ]
# Exclusion: no height
probanden <- klin$id[is.na(klin$height.cm)==F]

# Create genotype matrix
setwd(paste(basedir, "23andMe/Text", sep=""))
daten <- data.frame(snp="rs", chr=0, pos=0, geno="XY", pers="hu")
for (i in probanden){
   datei <- grep(i, dateien.auswahl)
   # Exclusion of individuals with inconsistent files
   if (!(i %in% c("hu52E130", "huD00199", "huBB5257"))){
      daten.pers <- read.table(dateien.auswahl[datei], sep="\t",
        nrows=no.snps, header=F)
      colnames(daten.pers) <- c("snp", "chr", "pos", "geno")
      daten.pers <- data.frame(daten.pers, pers=i)
      daten <- rbind(daten, daten.pers)
   }
}
# Delete first row that was created for initialisation
daten <- daten[2:nrow(daten), ]
# Drop unused levels in the genotype matrix
library(gdata)
daten <- drop.levels(daten)

# SNPs that were represented by 208 individuals
snps.max <- c("rs10492938", "rs10492940", "rs10737190",
  "rs10737192", "rs10752733", "rs10797342", "rs10797348",
  "rs10797368", "rs10797380", "rs10797386", "rs10797417",
  "rs10799202", "rs10907192", "rs10909845", "rs10909852",
  "rs10909890", "rs10909901", "rs10909918", "rs10910025",
  "rs10910047", "rs10910050", "rs10910061", "rs10910078",
  "rs10915433", "rs1108600", "rs1123571", "rs11260549",
  "rs11578011", "rs11583257", "rs11583804", "rs11587331",
  "rs11588930", "rs11589102", "rs11590198", "rs11590912",
  "rs1175549", "rs1181875", "rs1181877", "rs1181883",
  "rs1181888", "rs12022929", "rs12024847", "rs12031557",
  "rs12046158", "rs12049628", "rs12073172", "rs12082157",
  "rs12085231", "rs12117836", "rs12119470", "rs12119556",
  "rs12119711", "rs12124147", "rs12135298", "rs12138909",
  "rs12562167", "rs12562637", "rs12562988", "rs12724233",
  "rs12731705", "rs12748963", "rs12749761", "rs12757342",
  "rs13376356", "rs1553291", "rs1569419", "rs1572657",
  "rs16823542", "rs16823802", "rs16824089", "rs16825336",
  "rs17373634", "rs17399569", "rs17399998", "rs1890336",
  "rs1984069", "rs2017143", "rs2031709", "rs2045331",
  "rs2045332", "rs2055204", "rs2142569", "rs2173049",
  "rs2257182", "rs2275819", "rs2275831", "rs2279702",
  "rs2279703", "rs2296716", "rs2297829", "rs2298217",
  "rs2377041", "rs2455118", "rs2455144", "rs2459994",
  "rs2460000", "rs2474460", "rs2477703", "rs2483260",
  "rs2483274", "rs2485945", "rs2487670", "rs2487680",
  "rs2493272", "rs2493275", "rs2493285", "rs2493292",
  "rs2493310", "rs2493314", "rs2494428", "rs2494626",
  "rs2500262", "rs2500286", "rs2606411", "rs263526",
  "rs2643891", "rs2643901", "rs2645065", "rs2649588",
  "rs2651899", "rs2651906", "rs2799182", "rs2817178",
  "rs2817185", "rs2821007", "rs2821023", "rs2821025",
  "rs2821040", "rs2821063", "rs2840528", "rs2840532",
  "rs2840538", "rs2843127", "rs2843142", "rs2843143",
  "rs2843160", "rs2887286", "rs2993493", "rs3001336",
  "rs3002685", "rs3002686", "rs3107151", "rs3128291",
  "rs3736330", "rs3737589", "rs3748816", "rs3753242",
  "rs3762444", "rs3765703", "rs3765705", "rs3765731",
  "rs3765736", "rs3765761", "rs3765766", "rs3795263",
  "rs3813199", "rs3890745", "rs3934834", "rs4131373",
  "rs424079", "rs4276857", "rs4415513", "rs4531246",
  "rs4648377", "rs4648380", "s4648381", "rs4648392",
  "rs4648398", "rs4648426", "rs4648441", "rs4648453",
  "rs4648482", "rs4648487", "rs4648489", "rs4648505",
  "rs4648524", "rs4648527", "rs4648545", "rs4648592",
  "rs4648808", "rs4648831", "rs4648843", "rs4654479",
  "rs4654480", "rs4654482", "rs6424069", "rs6424074",
  "rs6603811", "rs6659405", "rs6661168", "rs6663840",
  "rs6667605", "rs6675798", "rs6681347", "rs6681938",
  "rs6683273", "rs6685064", "rs6687776", "rs6691155",
  "rs6695346", "rs6697749", "rs731031", "rs734999",
  "rs7367066", "rs7412983", "rs747827", "rs7515488",
  "rs7519349", "rs7519458", "rs7519807", "rs7522140",
  "rs7523732", "rs7525092", "rs7527871", "rs7528494",
  "rs7531583", "rs7534897", "rs7535528", "rs7538096",
  "rs7544357", "rs819980", "rs868688", "rs870124",
  "rs870171", "rs871822", "rs878063", "rs878201",
  "rs880051","rs884080", "rs884940", "rs897634",
  "rs903901", "rs903903", "rs903904", "rs903914",
  "rs903916", "rs903919", "rs905135", "rs908742",
  "rs926244", "rs9442373", "rs9442380", "rs946758",
  "rs947344", "rs947354", "rs9628616")

# Reduce genotype matrix to the needed SNPs
daten.max.snps <- daten[daten$snp %in% snps.max, ]

# Drop unused levels
library(gdata)
daten.max.snps <- drop.levels(daten.max.snps)
# 208 individuals with 245 SNPs

# Genotype coding according to MAF: identification of the
# minor allele

# Matrix
# 1 Individual per row, 1 SNP per column, 1 column per
# clinical information

# ID of individuals and SNPs
personen.unique <- as.character(unique(daten.max.snps$pers))
snps.unique <- as.character(unique(daten.max.snps$snp))

# Initialisation
daten.matrix <- matrix(NA, ncol=(length(snps.unique)+2),
  nrow=length(personen.unique))
```

```
daten.matrix <- data.frame(daten.matrix)
colnames(daten.matrix) <- c("age", "height",
  sort(snps.unique))
rownames(daten.matrix) <- personen.unique

# Fill matrix
for (i in 1:nrow(daten.matrix)){
   id <- rownames(daten.matrix)[i]
   daten.matrix$age[i] <- klin$age.yrs[klin$id==id]
   daten.matrix$height[i] <-
     klin$height.cm[klin$id==id]
   inter <- data.frame(rs=colnames(daten.matrix)[
     3:ncol(daten.matrix)])
   inter.2 <- merge(inter,
     daten.max.snps[daten.max.snps$pers==id,
     c("snp", "geno")], by.x="rs", by.y="snp",
     all.x=T, all.y=F)
   inter.2 <- inter.2[order(inter.2$rs),]
   daten.matrix[i, 3:ncol(daten.matrix)] <- inter.2$geno
}

# Code "--" as NA
for (i in 3:ncol(daten.matrix)){
   daten.matrix[daten.matrix[, i]=="--", i] <- NA
}

# Delete individuals with missing values
daten.matrix.2 <- na.omit(daten.matrix)
daten.matrix <- daten.matrix.2

# Calculate MAF per SNP
# Create table with rs, chromosome, position, minor allele,
# MAF and genotype frequency
tab.snps <- data.frame(snp=rep(NA, length(snps.unique)),
  chr=NA, pos=NA, min.allel=NA, maf=NA, maf.min.allel=NA,
  geno.min.allel=NA, geno.hetero=NA, geno.max.allel=NA)
# Preparation step: matrix with rs, chromosome and position
tab.inter <- data.frame(snp=snps.unique, chr=NA, pos=NA)
for (i in 1:nrow(tab.inter)){
  tab.inter$chr[i] <- as.numeric(daten$chr[daten$snp ==
    as.character(tab.inter$snp[i])][1])
  tab.inter$pos[i] <- as.numeric(daten$pos[daten$snp ==
    as.character(tab.inter$snp[i])][1])
}
tab.inter <- tab.inter[order(tab.inter$pos), ]
tab.snps[, 1:3] <- tab.inter
# Main step: MAF calculation
for (i in 1:nrow(tab.snps)){
  vektor <- daten.matrix[, colnames(daten.matrix) ==
    as.character(tab.snps$snp[i])]
  summ <- summary(as.factor(vektor))
  # 3 different genotypes
  if (length(names(summ))==3){
    maf.1 <- (2*summ[[1]]+summ[[2]])/(2*sum(summ))
    maf.2 <- (2*summ[[3]]+summ[[2]])/(2*sum(summ))
    selten <- ifelse(maf.1 <= maf.2, 1, 2)
    minor.allel <- ifelse(selten==1,
      substr(names(summ)[1], start=1, stop=1),
      substr(names(summ)[3], start=1, stop=1))
    genoFrequ.min.allel <- ifelse(selten==1,
      paste(round(100*summ[[1]]/sum(summ), digits=0),
      " (", names(summ)[1], ")", sep=""),
      paste(round(100*summ[[3]]/sum(summ), digits=0),
      " (", names(summ)[3], ")", sep=""))
    genoFrequ.hetero <- paste(round(100*summ[[2]]/sum(summ),
      digits=0), " (", names(summ)[2], ")", sep="")
    genoFrequ.max.allel <- ifelse(selten==1,
      paste(round(100*summ[[3]]/sum(summ), digits=0),
      " (", names(summ)[3], ")", sep=""),
      paste(round(100*summ[[1]]/sum(summ), digits=0),
      " (", names(summ)[1], ")", sep=""))
  # 2 different genotypes
  } else if (length(names(summ))==2){
    print(paste("2 genotypes: ", i, ". ", sep=""))
    if (substr(names(summ)[1], start=1, stop=1) ==
      substr(names(summ)[1], start=2, stop=2)){
       if (substr(names(summ)[2], start=1, stop=1) !=
         substr(names(summ)[2], start=2, stop=2)){
          maf.1 <- (2*summ[[1]]+summ[[2]])/(2*sum(summ))
           maf.2 <- summ[[2]]/(2*sum(summ))
           selten <- ifelse(maf.1 <= maf.2, 1, 2)
           if (selten==1){
             minor.allel <- substr(names(summ)[1],
               start=1, stop=1)
```

```
     a <- substr(names(summ)[2], start=1, stop=1)
     b <- substr(names(summ)[2], start=2, stop=2)
     major.allel <- c(a, b)[which(!(c(a,b))
       %in% minor.allel)]
   } else if (selten==2){
     major.allel <- substr(names(summ)[1],
       start=1, stop=1)
     a <- substr(names(summ)[2], start=1, stop=1)
     b <- substr(names(summ)[2], start=2, stop=2)
     minor.allel <- c(a, b)[which(!(c(a,b))
       %in% major.allel)]
   }
   genoFrequ.min.allel <- ifelse(selten==1,
     paste(round(100*summ[[1]]/sum(summ), digits=0),
     " (", names(summ)[1], ")", sep=""),
     paste("0 (", minor.allel, minor.allel, ")",
     sep=""))
   genoFrequ.hetero <- paste(round(100*summ[[2]]/sum(summ),
     digits=0), " (", names(summ)[2], ")", sep="")
   genoFrequ.max.allel <- ifelse(selten==1,
     paste("0 (", major.allel, major.allel, ")", sep=""),
     paste(round(100*summ[[1]]/sum(summ), digits=0),
     " (", names(summ)[1], ")", sep=""))
} else if (substr(names(summ)[2], start=1, stop=1) ==
  substr(names(summ)[2], start=2, stop=2)){
   maf.1 <- (2*summ[[1]])/(2*sum(summ))
   maf.2 <- (2*summ[[2]])/(2*sum(summ))
   selten <- ifelse(maf.1 <= maf.2, 1, 2)
   if (selten==1){
     minor.allel <- substr(names(summ)[1], start=1, stop=1)
   } else if (selten==2){
     minor.allel <- substr(names(summ)[2], start=1, stop=1)
   }
   genoFrequ.min.allel <- ifelse(selten==1,
     paste(round(100*summ[[1]]/sum(summ), digits=0),
     " (", names(summ)[1], ")", sep=""),
     paste(round(100*summ[[2]]/sum(summ), digits=0),
     " (", names(summ)[2], ")", sep=""))
   genoFrequ.hetero <- paste("0 (",
     substr(names(summ)[1], start=1, stop=1),
     substr(names(summ)[2], start=1, stop=1), ")", sep="")
   genoFrequ.max.allel <- ifelse(selten==1,
          paste(round(100*summ[[2]]/sum(summ), digits=0),
          " (", names(summ)[2], ")", sep=""),
          paste(round(100*summ[[1]]/sum(summ), digits=0),
          " (", names(summ)[1], ")", sep=""))
     }
  } else if (substr(names(summ)[1], start=1, stop=1) !=
    substr(names(summ)[1], start=2, stop=2)){
     maf.1 <- summ[[1]]/(2*sum(summ))
     maf.2 <- (summ[[1]]+2*summ[[2]])/(2*sum(summ))
     selten <- ifelse(maf.1 <= maf.2, 1, 2)
     if (selten==2){
       minor.allel <- substr(names(summ)[2], start=1,
         stop=1)
       a <- substr(names(summ)[2], start=1, stop=1)
       b <- substr(names(summ)[2], start=2, stop=2)
       major.allel <- c(a, b)[which(!(c(a,b)) %in% minor.allel)]
     } else if (selten==1){
       a <- substr(names(summ)[1], start=1, stop=1)
       b <- substr(names(summ)[1], start=2, stop=2)
       minor.allel <- c(a, b)[which(!(c(a,b)) %in%
         substr(names(summ)[2], start=1, stop=1))]
       major.allel <- c(a, b)[which(!(c(a,b)) %in% minor.allel)]
     }
     genoFrequ.min.allel <- ifelse(selten==1,
       paste("0 (", minor.allel, minor.allel, ")", sep=""),
       paste(round(100*summ[[2]]/sum(summ), digits=0),
       " (", names(summ)[2], ")", sep=""))
     genoFrequ.hetero <- paste(round(100*summ[[1]]/sum(summ),
       digits=0), " (", names(summ)[1], ")", sep="")
     genoFrequ.max.allel <- ifelse(selten==1,
       paste(round(100*summ[[2]]/sum(summ), digits=0),
       " (", names(summ)[2], ")", sep=""),
       paste("0 (", major.allel, major.allel, ")", sep=""))
  }
# 1 genotype
} else if (length(names(summ))==1){
  print(paste("1 genoytpe: ", i, ". ", sep=""))
  if (substr(names(summ)[1], start=1, stop=1) ==
    substr(names(summ)[1], start=2, stop=2)){
     maf.1 <- summ[[1]]/(sum(summ))
     maf.2 <- 0
     selten <- ifelse(maf.1 <= maf.2, 1, 2)
```

```
        minor.allel <- "-"
        genoFrequ.min.allel <- "0 (--)"
        genoFrequ.hetero <- "0 (--)"
        genoFrequ.max.allel <- paste(round(100*summ[[1]]/sum(summ),
          digits=0), " (", names(summ)[1], ")", sep="")
     } else if (substr(names(summ)[1], start=1, stop=1) !=
       substr(names(summ)[1], start=2, stop=2)){
        maf.1 <- summ[[1]]/(2*sum(summ))
        maf.2 <- summ[[1]]/(2*sum(summ))
        selten <- 1
        minor.allel <- substr(names(summ)[1], start=1, stop=1)
        genoFrequ.min.allel <- paste("0 (", minor.allel,
          minor.allel, ")", sep="")
        genoFrequ.hetero <- paste(round(100*summ[[1]]/sum(summ),
          digits=0), " (", names(summ)[1], ")", sep="")
        genoFrequ.max.allel <- paste("0 (", substr(names(summ)[1],
          start=2, stop=2), substr(names(summ)[1], start=2,
          stop=2), sep="")
     }
   }
   frequ <- ifelse(selten==1, round(maf.1*100, digits=0),
     round(maf.2*100, digits=0))
   minor.allel.frequ <- paste(frequ, " (", minor.allel, ")",
     sep="")

   tab.snps$min.allel[i] <- as.character(minor.allel)
   tab.snps$maf[i] <- frequ
   tab.snps$maf.min.allel[i] <- minor.allel.frequ
   tab.snps$geno.min.allel[i] <- genoFrequ.min.allel
   tab.snps$geno.hetero[i] <- genoFrequ.hetero
   tab.snps$geno.max.allel[i] <- genoFrequ.max.allel
}
# Save results
titel <- paste(save.dir, "/SNPs_Uebersicht.csv", sep="")
write.csv2(tab.snps, titel, row.names=F)

# Function for numerical coding of SNPs
kodierung <- function(vektor, rs.no){
  minor <- tab.snps$min.allel[tab.snps$snp==rs.no]
  geno <- paste(minor, minor, sep="")
  vektor <- as.character(vektor)
  vektor[vektor==geno] <- "2"
  vektor[grep(minor, vektor)] <- "1"
  vektor[!(vektor %in% c("1", "2"))] <- "0"
  return(vektor)
}

# Coding
for (i in 3:ncol(daten.matrix)){
   daten.matrix[, i] <- as.numeric(kodierung(daten.matrix[, i],
     colnames(daten.matrix)[i]))
}



# Response - median-dichotomised
med <- median(daten.matrix$height)
daten.matrix <- data.frame(response=rep(NA, nrow(daten.matrix)),
  daten.matrix)
daten.matrix$response[daten.matrix$height <= med] <- 0
daten.matrix$response[daten.matrix$height > med] <- 1

# Overview table clinical data
tab.klin <- data.frame(vari=c("No. Persons", "Height [cm]",
  "Median (Q1, Q3)", "Age [years]", "Median (Q1, Q3)",
  "No. SNPs", "MAF [%]", "Median (Q1, Q3)"), value=NA)
idx <- 1
tab.klin$value[idx] <- nrow(daten.matrix)
idx <- idx+1
tab.klin$value[idx] <- ""
idx <- idx+1
tab.klin$value[idx] <- paste(median(daten.matrix$height),
  " (", quantile(daten.matrix$height, probs=0.25),
  ", ", quantile(daten.matrix$height, probs=0.75), ")",
  sep="")
idx <- idx+1
tab.klin$value[idx] <- ""
idx <- idx+1
tab.klin$value[idx] <- paste(median(daten.matrix$age),
  " (", quantile(daten.matrix$age, probs=0.25),
  ", ", quantile(daten.matrix$age, probs=0.75), ")",
  sep="")
idx <- idx+1
tab.klin$value[idx] <- ncol(daten.matrix)-3
idx <- idx+1
```

```
tab.klin$value[idx] <- ""
idx <- idx+1
tab.klin$value[idx] <- paste(median(tab.snps$maf),
  " (", quantile(tab.snps$maf, probs=0.25), ", ",
  quantile(tab.snps$maf, probs=0.75), ")", sep="")
titel <- paste(save.dir, "/Datenuebersicht.csv", sep="")
write.csv2(tab.klin, titel, row.names=F)

# -------------------------------------------- #
# -------------------------------------------- #
# 4. Logistic regression
# -------------------------------------------- #
# -------------------------------------------- #

# Regression - for a fitted recessive penetrance
# model
tab.rezessiv <- data.frame(snp=tab.snps$snp,
  chr=tab.snps$chr, pos=tab.snps$pos,
  p.stand=NA, koeff.stand=NA, p.huber=NA,
  koeff.huber=NA, p.hampel.09=NA,
  koeff.hampel.09=NA, p.hampel.06=NA,
  koeff.hampel.06=NA)

# Standard
for (i in 1:nrow(tab.rezessiv)){
   response <- daten.matrix$response
   alter <- daten.matrix$age
   geno <- daten.matrix[, colnames(daten.matrix) ==
     as.character(tab.rezessiv$snp[i])]
   geno[geno==1] <- 0
   geno[geno==2] <- 1
   zwischen1 <- glm(response ~ geno + alter,
     family=binomial(link = "logit"))
   if (sum(dim(summary(zwischen1)$coeff)==c(3,4))==2){
      est <- summary(zwischen1)$coeff["geno", ]
      tab.rezessiv$p.stand[i] <- est[[4]]
      tab.rezessiv$koeff.stand[i] <- est[[1]]
   }
}

# Huber
library(robustbase)
```

```
tun.huber <- 1.345
for (i in 1:nrow(tab.rezessiv)){
   response <- daten.matrix$response
   alter <- daten.matrix$age
   geno <- daten.matrix[, colnames(daten.matrix) ==
     as.character(tab.rezessiv$snp[i])]
   geno[geno==1] <- 0
   geno[geno==2] <- 1
   zwischen1 <- try(glmrob(response ~ geno + alter,
     family=binomial("logit"), weights.on.x = "hat",
     tcc=tun.huber), silent=TRUE)
   if (sum(class(zwischen1)=="try-error")==0){
         if (sum(dim(summary(zwischen1)$coeff)==c(3,4))==2){
         est <- summary(zwischen1)$coeff["geno", ]
         tab.rezessiv$p.huber[i] <- est[[4]]
         tab.rezessiv$koeff.huber[i] <- est[[1]]
      }
   }
}
detach("package:robustbase")

# Hampel
library(robustbaseAdj)
tun.hampel1 <- c(1.5, 3.5, 8)*0.9016085
tun.hampel2 <- c(2, 4, 8)*0.690794
for (i in 1:nrow(tab.rezessiv)){
   response <- daten.matrix$response
   alter <- daten.matrix$age

   # 1st tuning constant
   geno <- daten.matrix[, colnames(daten.matrix) ==
     as.character(tab.rezessiv$snp[i])]
   geno[geno==1] <- 0
   geno[geno==2] <- 1
   zwischen1 <- try(glmrob(response ~ geno + alter,
     family=binomial("logit"), weights.on.x = "hat",
     tcc=tun.hampel1), silent=TRUE)
   if (sum(class(zwischen1)=="try-error")==0){
         if (sum(dim(summary(zwischen1)$coeff)==c(3,4))==2){
         est <- summary(zwischen1)$coeff["geno", ]
         tab.rezessiv$p.hampel.09[i] <- est[[4]]
         tab.rezessiv$koeff.hampel.09[i] <- est[[1]]
      }
   }

   # 2nd tuning constant
   geno <- daten.matrix[, colnames(daten.matrix) ==
     as.character(tab.rezessiv$snp[i])]
   geno[geno==1] <- 0
   geno[geno==2] <- 1
   zwischen1 <- try(glmrob(response ~ geno + alter,
     family=binomial("logit"), weights.on.x = "hat",
     tcc=tun.hampel2), silent=TRUE)
   if (sum(class(zwischen1)=="try-error")==0){
         if (sum(dim(summary(zwischen1)$coeff)==c(3,4))==2){
         est <- summary(zwischen1)$coeff["geno", ]
         tab.rezessiv$p.hampel.06[i] <- est[[4]]
         tab.rezessiv$koeff.hampel.06[i] <- est[[1]]
      }
   }
}
detach("package:robustbaseAdj")

# Save results
titel <- paste(save.dir, "/Ergebnisse/Regression_komplErg_Rez.csv",
  sep="")
write.csv2(tab.rezessiv, titel, row.names=F)
```

D.3.2 Visualisation of the real data analysis results

Structure:

1. Settings
2. Input
3. Manhattan and OR plots
4. Diagnostic plots

```
# -------------------------------------------- #
# -------------------------------------------- #
# 1. Settings
# -------------------------------------------- #
# -------------------------------------------- #
```

```
data.dir <- "Directory to analysis results"
save.dir <- "Directory to save plot"



# ------------------------------------------ #
# ------------------------------------------ #
# 2. Input
# ------------------------------------------ #
# ------------------------------------------ #

# Read results
titel <- paste(data.dir, "/Regression_komplErg_Rez.csv",
  sep="")
tab.rezessiv <- read.csv2(titel)

# Titles to save plots
titel.manhattan.or <- paste(save.dir,
  "/Manhattan_OR_Plots_Rezessiv.png", sep="")
titel.diagnostics <- paste(save.dir,
  "/Cook_rs7519458_rs2500262_Rezessiv.png", sep="")



# ------------------------------------------ #
# ------------------------------------------ #
# 3. Manhattan and OR plots
# ------------------------------------------ #
# ------------------------------------------ #

# Create plot as .png with resolution equal to 300dpi

# Open graphic device
bitmap(titel.manhattan.or, res=300, width=25, height=40)

# Layout of plot
  par(mar=c(5,5.5,5,1))
  layout(mat=matrix(c(1,1,2,2,3,3,4,4,5,5,6,6,7,7,8,8,9,9,
    10,10,11,11,12,12), ncol=6, byrow=T), widths=rep(4, 6),
    heights=rep(10, 4))

#1: Row name
  plot(1,1, bty="n", col="white", xaxt="n", yaxt="n", xlab="",
    ylab="", main="", ylim=c(0,3))
  text(1,1.5,"Standard", cex=2)

#2: Manhattan
  plot(1:nrow(tab.rezessiv), -log10(tab.rezessiv$p.stand),
    pch=19, col="black", cex=2, ylim=c(0,3), xlab="SNP",
    ylab=expression(-log[10]~"(p-value)"), main="",
    cex.lab=2, cex.axis=1.8, xaxt="n", yaxt="n")
  lines(x=c(0, nrow(tab.rezessiv)+1), y=rep(-log10(0.05), 2),
    col="black", lwd=4, type="l", lty="dotted")
  axis(2, at=c(0,1,2,3), labels=c(0,1,2,3), cex.axis=1.8)

#3: estimated ORs
  plot(1:nrow(tab.rezessiv), exp(tab.rezessiv$koeff.stand),
    pch=19, col="black", cex=2, xlab="SNP",
    ylab=expression("Estimated OR ["~10^7~"]"),
    main="", cex.lab=2, cex.axis=1.8, xaxt="n",
    yaxt="n")
  lines(x=c(0, nrow(tab.rezessiv)+1), y=rep(1, 2), col="black",
    lwd=4, type="l", lty="dotted")
  axis(2, at=c(0,0.5,1,1.5)*10000000, labels=c(0,0.5,1.0,1.5),
    cex.axis=1.8)

#4: Row name
  plot(1,1, bty="n", col="white", xaxt="n", yaxt="n", xlab="",
    ylab="", main="", ylim=c(0,3))
  text(1,2,"Huber", cex=2)
  text(1,1,"[1.345]", cex=2)

#5: Manhattan
  plot(1:nrow(tab.rezessiv), -log10(tab.rezessiv$p.huber),
    pch=19, col="darkgrey", cex=2, ylim=c(0,3), xlab="SNP",
    ylab=expression(-log[10]~"(p-value)"), main="", cex.lab=2,
    cex.axis=1.8, cex.main=2.5, xaxt="n", yaxt="n")
  lines(x=c(0, nrow(tab.rezessiv)+1), y=rep(-log10(0.05), 2),
    col="black", lwd=4, type="l", lty="dotted")
  axis(2, at=c(0,1,2,3), labels=c(0,1,2,3), cex.axis=1.8)

#6: estimated ORs
  plot(1:nrow(tab.rezessiv), exp(tab.rezessiv$koeff.huber), pch=19,
    col="darkgrey", cex=2, ylim=c(0,12), xlab="SNP",
    ylab="Estimated OR", main="", cex.lab=2, cex.axis=1.8,
    cex.main=2.5, xaxt="n", yaxt="n")
  lines(x=c(0, nrow(tab.rezessiv)+1), y=rep(1, 2), col="black",
    lwd=4, type="l", lty="dotted")
  axis(2, at=c(0,4,8,12), labels=c(0,4,8,12), cex.axis=1.8)

#7: Row name
  plot(1,1, bty="n", col="white", xaxt="n", yaxt="n", xlab="",
    ylab="", main="", ylim=c(0,3))
  text(1,2,"Hampel", cex=2)
  text(1,1,"[(1.5, 3.5, 8)x0.9]", cex=2)

#8: Manhattan
  plot(1:nrow(tab.rezessiv), -log10(tab.rezessiv$p.hampel.09),
    pch=19, col="grey", cex=2, ylim=c(0,3), xlab="SNP",
    ylab=expression(-log[10]~"(p-value)"), main="", cex.lab=2,
    cex.axis=1.8, cex.main=2.5, xaxt="n", yaxt="n")
  lines(x=c(0, nrow(tab.rezessiv)+1), y=rep(-log10(0.05), 2),
    col="black", lwd=4, type="l", lty="dotted")
  axis(2, at=c(0,1,2,3), labels=c(0,1,2,3), cex.axis=1.8)

#9: estimated ORs
  plot(1:nrow(tab.rezessiv), exp(tab.rezessiv$koeff.hampel.09),
    pch=19, col="grey", cex=2, ylim=c(0,12), xlab="SNP",
    ylab="Estimated OR", main="", cex.lab=2, cex.axis=1.8,
    cex.main=2.5, xaxt="n", yaxt="n")
  lines(x=c(0, nrow(tab.rezessiv)+1), y=rep(1, 2),
    col="black", lwd=4, type="l", lty="dotted")
  axis(2, at=c(0,4,8,12), labels=c(0,4,8,12), cex.axis=1.8)

#10: Row name
  plot(1,1, bty="n", col="white", xaxt="n", yaxt="n", xlab="",
    ylab="", main="", ylim=c(0,3))
  text(1,2,"Hampel", cex=2)
  text(1,1,"[(2, 4, 8)x0.7]", cex=2)

#11: Manhattan
  plot(1:nrow(tab.rezessiv), -log10(tab.rezessiv$p.hampel.06),
    pch=19, col="lightgrey", cex=2, ylim=c(0,3), xlab="SNP",
    ylab=expression(-log[10]~"(p-value)"), main="", cex.lab=2,
    cex.axis=1.8, cex.main=2.5, xaxt="n", yaxt="n")
  lines(x=c(0, nrow(tab.rezessiv)+1), y=rep(-log10(0.05), 2),
    col="black", lwd=4, type="l", lty="dotted")
  axis(2, at=c(0,1,2,3), labels=c(0,1,2,3), cex.axis=1.8)

#11: estimated ORs
  plot(1:nrow(tab.rezessiv), exp(tab.rezessiv$koeff.hampel.06),
    pch=19, col="lightgrey", cex=2, ylim=c(0,12), xlab="SNP",
    ylab="Estimated OR", main="", cex.lab=2, cex.axis=1.8,
    cex.main=2.5, xaxt="n", yaxt="n")
  lines(x=c(0, nrow(tab.rezessiv)+1), y=rep(1, 2), col="black",
    lwd=4, type="l", lty="dotted")
  axis(2, at=c(0,4,8,12), labels=c(0,4,8,12), cex.axis=1.8)

# close graphic device
dev.off()



# -------------------------------------------- #
# -------------------------------------------- #
# 4. Diagnostic plots
# -------------------------------------------- #
# -------------------------------------------- #

# for rs2500262 and rs2500262

# Indices of these SNPs in the analysis loop
# rs2500262
i <- 155

# rs7519458
i <- 230

# Save the standard logistic regression models for these SNPs
zwischen.rs7519458 <-
  "standard logistic regression model for SNP rs7519458"
zwischen.rs2500262 <-
  "standard logistic regression model for SNP rs2500262"

# Create plot as .png with resolution equal to 300dpi

# Open graphic device
bitmap(titel.diagnostics, res=300, width=10, height=5)

# Layout
```

```
  par(mfrow=c(1,2), mar=c(5,7,5,1))

# Plots
  plot(zwischen.rs7519458, which=4, main="rs7519458",
    caption='Observation number', sub.caption = "",
    lwd=2, cex=2, cex.axis=1.8, cex.lab=2, cex.main=2.5)
  plot(zwischen.rs2500262, which=4, main="rs2500262",
    caption='Observation number', sub.caption = "", lwd=2,
    cex=2, cex.axis=1.8, cex.lab=2, cex.main=2.5)

# Close device
dev.off()
```

Bibliography

Adewale, A. J. and Xu, X. (2010). Robust designs for generalized linear models with possible overdispersion and misspecified link functions. *Computational Statistics & Data Analysis*, 54:875–890.

Adimari, G. and Ventura, L. (2001). Robust inference for generalized linear models with application to logistic regression. *Statistics & Probability Letters*, 55:413–419.

Alamgir, Ali, A., Khan, S. A., Khan, D. M., and Khalil, U. (2013). A new efficient redescending M-estimator: Alamgir redescending M- estimator. *Research Journal of Recent Sciences*, 2:79–91.

Ali, S., Akhter, S., Neubauer, H., Scherag, A., Kesselmeier, M., Khan, I., Azam, A., Qadeer, S., and Ali, Q. (in preparation). Brucellosis in pregnant women from Pakistan: An observational study.

Almasy, L., Dyer, T. D., Peralta, J. M., Jun, G., Fuchsberger, C., Almeida, M. A., Kent Jr., J. W., Fowler, S., Duggirala, R., and Blangero, J. (2014). Data for Genetic Analysis Workshop 18: Human whole genome sequence, blood pressure, and simulated phenotypes in extended pedigrees. *BMC Proceedings*, 8 (suppl 2):S2.

Altman, E. I. (1968). Financial ratios, discriminate analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23:589–609.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972). *Robust Estimates of Location – Survey and Advances*. Princeton University Press, Princeton (New Jersey, USA).

Arora, N. and Biegler, L. T. (2001). Redescending estimators for data reconciliation and parameter estimation. *Computers and Chemical Engineering*, 25:1585–1599.

Beaton, A. E. and Tukey, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16:147–185.

Bednarski, T. (2002). Estimation in the generalized Poisson model via robust testing. *Metrika*, 55:27–36.

Bianco, A. M. and Yohai, V. J. (1996). Robust estimation in the logistic regression model. In *Robust Statistics, Data Analysis, and Computer Intensive Methods: In Honor of Peter Huber's 60th Birthday*, pages 17–34. Springer, New York, USA.

Bortz, J. and Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler*. Springer, Heidelberg, Germany, 7th edition. Chapter 2.2.6.

Breuhahn, K. (2010). Molekulare Progressionsmechanismen der humanen Hepatokarzinogenese. *Der Pathologe*, 31 (Suppl. 2):170–176.

Burga, A., Casanueva, M. O., and Lehner, B. (2011). Predicting mutation outcome from early stochastic variation in genetic interaction partners. *Nature*, 480:250–U133.

Cantoni, E. (2004). Analysis of robust quasi-deviances for generalized linear models. *Journal of Statistical Software*, 10.

Cantoni, E. and Ronchetti, E. (2001). Robust inference for generalized linear models. *Journal of the American Statistical Association*, 96:1022–1030.

Cantoni, E. and Ronchetti, E. (2006). A robust approach for skewed and heavy-tailed outcomes in the analysis of health care expenditures. *Journal of Health Economics*, 25:198–213.

Cantoni, E. and Zedini, A. B. Y. (2011). A robust version of the hurdlemodel. *Journal of Statistical Planning and Inference*, 141:1214–1223.

Çetin, M. and Erar, A. (2006). A simulation study on classic and robust variable selection in linear regression. *Applied Mathematics and Computation*, 175:1629–1643.

CHEK2 Breast Cancer Case-Control Consortium (2004). CHEK2*1100delC and susceptibility to breast cancer: A collaborative analysis involving 10,860 breast cancer cases and 9,065 controls from 10 studies. *American Journal of Human Genetics*, 74:1175–1182.

Chen, L., Yu, G., Langefeld, C. D., Miller, D. J., Guy, R. T., Raghuram, J., Yuan, X., Herrington, D. M., and Wang, Y. (2011). Comparative analysis of methods for detecting interacting loci. *BMC Genomics*, 12:344.

Church, G. M. (2005). The Personal Genome Project. *Molecular Systems Biology*, 1.

Croux, C. and Haesbroeck, G. (2003). Implementing the Bianco and Yohai estimator for logistic regression. *Computational Statistics and Data Analysis*, 44:273–295.

de Lope, C. R., Tremosini, S., Forner, A., Reig, M., and Bruix, J. (2012). Management of HCC. *Journal of Hepatology*, 56 (Suppl. 1):S75–S87.

De Maesschalck, R., Jouan-Rimbaud, D., and Massart, D. L. (2000). The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 50:1–18.

Dixon, W. J. (1950). Analysis of extreme values. *Annals of Mathematical Statistics*, 21:488–506.

Duan, N., Manning Jr., W. G., Morris, C. N., and Newhouse, J. P. (1983). A comparison of alternative models for the demand for medical care. *Journal of Business and Economic Statistics*, 1:115–126.

Duggirala, R., Blangero, J., Almasy, L., Dyer, T. D., Williams, K. L., Leach, R. J., O'Connell, P., and Stern, M. P. (1999). Linkage of type 2 diabetes mellitus and of age at onset to a genetic location on chromosome 10q in Mexican Americans. *American Journal of Human Genetics*, 64:1127–1140.

Ehret, G. B. (2010). Genome-wide association studies: Contribution of genomics to understanding blood pressure and essential hypertension. *Current Hypertension Reports*, 12:17–25.

El-Serag, H. B. and Rudolph, K. I. (2007). Hepatocellular carcinoma: Epidemiology and molecular carcinogenesis. *Gastroenterology*, 132:2557–2576.

Elçioglu, N. H., Pawlik, B., Colak, B., Beck, M., and Wollnik, B. (2009). A novel loss-of-function mutation in the GNS gene causes Sanfilippo syndrome type D. *Genetic Counseling*, 20:133–139.

Engholm, G., Ferlay, J., Christensen, N., Bray, F., Gjerstorff, M. L., Klint, A., Køtlum, J. E., Olafsdóttir, E., Pukkala, E., and Storm, H. H. (2010). Nordcan – A nordic tool for cancer information, planning, quality control and research. *Acta Oncologica*, 49(5):725–736.

Engholm, G., Ferlay, J., Christensen, N., Johannesen, T. B., Klint, Å., Køtlum, J. E., Milter, M. C., Ólafsdóttir, E., Pukkala, E., and Storm, H. H. (2012). Nordcan: Cancer incidence, mortality, prevalence and survival in the nordic countries, version 5.2. Technical report, Association of the Nordic Cancer Registries. Danish Cancer Society, `http://www.ancr.nu`. accessed on 23/01/2013.

Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D. M., Forman, D., and Bray, F. (2013). Globocan 2012 v1.0, cancer incidence and mortality worldwide: IARC CancerBase No. 11 [internet]. Technical report, International Agency for Research on Cancer (Lyon, France), `http://globocan.iarc.fr`. accessed on 17/07/2014.

Ferretti, N., Kelmansky, D., Yohai, V. J., and Zamar, R. H. (1999). A class of locally and globally robust regression estimates. *Journal of the American Statistical Association*, 94:174–188.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222:309–368.

Fox, J. and Weisberg, S. (2011). Robust regression in R. In *An R Companion to Applied Regression*, pages 1–8. SAGE Publications, Thousand Oaks (CA, USA), 2nd edition. Appendix.

Friedman, J. H. (1997). On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1:55–77.

Garner, C. (2007). Upward bias in odds ratio estimates from genome-wide association studies. *Genetic Epidemiology*, 31:288–295.

Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias variance dilemma. *Neural Computation*, 4:1–58.

Ghosh, A., Zou, F., and Wright, F. A. (2008). Estimating odds ratios in genome scans: An approximate conditional likelihood approach. *American Journal of Human Genetics*, 82:1064–1074.

Göring, H. H. H., Terwilliger, J. D., and Blangero, J. (2001). Large upward bias in estimation of locus-specific effects from genomewide scans. *American Journal of Human Genetics*, 69:1357–1369.

Hagel, S., Reischke, J., Kesselmeier, M., Winning, J., Gastmeier, P., Brunkhorst,

F. M., Scherag, A., and Pletz, M. W. (2015). Quantifying the Hawthorne effect in hand hygiene compliance through comparing direct observation with automated hand hygiene monitoring. *Infection Control & Hospital Epidemiology*, 36:957–962.

Hampel, F. R. (1968). *Contributions to the theory of robust estimation.* PhD thesis, University of California, Berkeley, USA. Unpublished.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions.* John Wiley & Sons, New York, USA. Chapter 1, Chapter 2.6.

Haukka, J. K. (1995). Correction for covariate measurement error in generalized linear models – a bootstrap approach. *Biometrics*, 51:1127–1132.

Hauser, R. P. and Booth, D. (2011). Predicting bankruptcy with robust logistic regression. *Journal of Data Science*, 9:565–584.

Hawkins, D. M. (1980). *Identification of Outliers.* Monographs on Applied Probability and Statistics. Chapman & Hall, London, UK. Chapter 1.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47:153–161.

Hein, R., Beckmann, L., and Chang-Claude, J. (2008). Sample size requirements for indirect association studies of gene–environment interactions (gxe). *Genetic Epidemiology*, 32:235–245.

Hemminki, K. and Lorenzo Bermejo, J. (2007). Constraints for genetic association studies imposed by attributable fraction and familial risk. *Carcinogenesis*, 28:648–656.

Heritier, S., Cantoni, E., Copt, S., and Victoria-Feser, M.-P. (2009). *Robust Methods in Biostatistics.* John Wiley & Sons Ltd., West Sussex, UK. Chapters 2.5.5 and 5.3.1.

Heritier, S. and Ronchetti, E. (1994). Robust bounded-influence tests in general parametric models. *Journal of the American Statistical Association*, 89:897–904.

Hill, W. G. and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, 38:226–231.

Hirschhorn, J. N., Lohmueller, K., Byrne, E., and Hirschhorn, K. (2002). A comprehensive review of genetic association studies. *Genetics in Medicine*, 4:45–61.

Ho, J. E., Levy, D., Rose, L., Johnson, A. D., Ridker, P. M., and Chasman, D. I. (2011). Discovery and replication of novel blood pressure genetic loci in the women's genome health study. *Journal of Hypertension*, 29:62–69.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (1983). *Understanding robust and exploratory data analysis.* John Wiley & Sons Ltd., New York, USA. Chapter 11D.

Hong, E. P. and Park, J. W. (2012). Sample size and statistical power calculation in genetic association studies. *Genomics & Informatics*, 10:117–122.

Hong, H., Xu, L., Liu, J., Jones, W. D., Su, Z., Ning, B., Perkins, R., Ge, W., Miclaus, K., Zhang, L., Park, K., Green, B., Han, T., Fang, H., Lambert, C. G., Vega, S. C., Lin, S. M., Jafari, N., Czika, W., Wolfinger, R. D., Goodsaid, F., Tong, W., and Shi, L. (2012). Technical reproducibility of genotyping SNP arrays used in genome-wide association studies. *PLoS ONE*, 7:e44483.

Hoshida, Y., Nijman, S. M. B., Kobayashi, M., Chan, J. A., Brunet, J.-P., Chiang, D. Y., Villanueva, A., Newell, P., Ikeda, K., Hashimoto, M., Watanabe, G., Gabriel, S., Friedman, S. L., Kumada, H., Llovet, J. M., and Golub, T. R. (2009). Integrative transcriptome analysis reveals common molecular subclasses of human hepatocellular carcinoma. *Cancer Research*, 69:7385–7392.

Hosseinian, S. and Morgenthaler, S. (2011). Robust binary regression, *Journal of Statistical Planning and Inference*, 141:1497–1509.

Houseman, E. A., Molitor, J., and Marsit. C. J. (2014). Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics*, 30:1431–1439.

Howey, R. and Cordell, H. J. (2014). Imputation without doing imputation: A new method for the detection of non-genotyped causal variants. *Genetic Epidemiology*, 38:173–190.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35:73–101.

Hunt, K. J., Lehman, D. M., Arya, R., Fowler, S., Leach, R. J., Göring, H. H. H., Almasy, L., Blangero, J., Dyer, T. D., Duggirala, R., and Stern, M. P. (2005). Genome-wide linkage analyses of type 2 diabetes in Mexican Americans: The San Antonio family diabetes/gallbladder study. *Diabetes*, 54:2655–2662.

Jajo, N. K. (2005). A review of robust regression and diagnostic procedures in linear regression. *Acta Mathematicae Applicatae Sinica*, 21:209–224.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2015). *An Introduction to Statistical Learning – with Applications in R.* Springer, New York, USA, 6th edition. Chapters 2.2, 5.1., and 6.1.

Jonas, B. S., Franks, P., and Ingram, D. D. (1997). Are symptoms of anxiety and depression risk factors for hypertension? Longitudinal evidence from the national health and nutrition examination survey I epidemiologic follow-up study. *Archives of Family Medicine*, 6:43–49.

Katzmarzyk, P. T., Rankinen, T., Pérusse, L., Rao, D. C., and Bouchard, C. (2001). Familial risk of high blood pressure in the Canadian population. *American Journal of Human Biology*, 13:620–625.

Kennedy, G. C., Matsuzaki, H., Dong, S., Liu, W. M., Huang, J., Liu, G., Su, X., Cao, M., Chen, W., Zhang, J., Liu, W., Yang, G., Di, X., Ryder, T., He, Z., Surti, U., Phillips, M. S., Boyce-Jacino, M. T., Fodor, S. P., and Jones, K. W. (2003). Large-scale genotyping of complex DNA. *Nature Biotechnology*, 21:1233–1237.

Kesselmeier, M., Legrand, C., Peil, B., Kabisch, M., Fischer, C., Hamann, U., and Lorenzo Bermejo, J. (2014). Practical investigation of the performance of robust logistic regression to predict the genetic risk of hypertension. *BMC Proceedings*, 8(Suppl 1):S65.

Kesselmeier, M. and Lorenzo Bermejo, J. (in preparation). Robust logistic regression to narrow down the winner's curse for rare and recessive susceptibility variants.

Kesselmeier, M., Neumann, O., Longerich, T., Geffers, R., Zucman-Rossi, J., Imbeaud, S., Derambure, C., Schirmacher, P., and Lorenzo Bermejo, J. (in preparation). Aetiology specific methylation and gene expression patterns in human hepatocellular carcinoma.

Kesselmeier, M., Pütter, C., Volckmar, A.-L., Baurecht, H., Grallert, H., Illig, T., Ismail, K., Ollikainen, M., Silén, Y., Keski-Rahkonen, A., Bulik, C. M., Collier, D. A., Zeggini, E., Hebebrand, J., Scherag, A., Hinney, A., GCAN, and WTCCC3. (accepted). High-throughput DNA methylation analysis in anorexia nervosa confirms *TNXB* hypermethylation in starvation. *The World Journal of*

Biological Psychiatry.

Kesselmeier, M. and Scherag, A. (in preparation). High-throughput DNA methylation association analyses with reference-free cell type adjustment: A method comparison.

Kohavi, R. and Wolpert, D. H. (1996). Bias plus variance decomposition for zero-one loss functions. In *Machine Learning: Proceedings of the Thirteenth International Conference*, pages 275–283.

Koller, M. and Mächler, M. (2014). *Definitions of* Ψ*-Functions Available in Robustbase.*

Koller, M. and Stahel, W. A. (2011). Sharpening Wald-type inference in robust regression for small samples. *Computational Statistics & Data Analysis*, 55:2504–2515.

Kordzakhia, N., Mishra, G. D., and Reiersølmoen, L. (2001). Robust estimation in the logistic regression model. *Journal of Statistical Planning and Inference*, 98:211–223.

Kraft, P., Cox, D. G., Paynter, R. A., Hunter, D., and De Vivo, I. (2005). Accounting for haplotype uncertainty in matched association studies: A comparison of simple and flexible techniques. *Genetic Epidemiology*, 28:261–272.

Laird, N. M. and Lange, C. (2011). *The Fundamentals of Modern Statistical Genetics.* Springer, Heidelberg, Germany. Chapter 1.2.

Levy, D., Ehret, G. B., Rice, K., Verwoert, G. C., Launer, L. J., Dehghan, A., Glazer, N. L., Morrison, A. C., Johnson, A. D., Aspelund, T., Aulchenko, Y., Lumley, T., Köttgen, A., Vasan, R. S., Rivadeneira, F., Eiriksdottir, G., Guo, X., Arking, D. E., Mitchell, G. F., Mattace-Raso, F. U. S., Smith, A. V., Taylor, K., Scharpf, R. B., Hwang, S.-J., Sijbrands, E. J. G., Bis, J., Harris, T. B., Ganesh, S. K., O'Donnell, C. J., Hofman, A., Rotter, J. I., Coresh, J., Benjamin, E. J., Uitterlinden, A. G., Heiss, G., Fox, C. S., Witteman, J. C. M., Boerwinkle, E., Wang, T. J., Gudnason, V., Larson, M. G., Chakravarti, A., Psaty, B. M., and van Duijn, C. M. (2009). Genome-wide association study of blood pressure and hypertension. *Nature Genetics*, 41:677–687.

Lewontin, R. C. (1964). The interaction of selection and linkage. I. general considerations; heterotic models. *Genetics*, 49:49–67.

Li, T. and Hsiao, C. (2004). Robust estimation of generalized linear models with measurement errors. *Journal of Econometrics*, 118:51–65.

Lin, W. and Schaid, D. J. (2009). Power comparisons between similarity-based multilocus association methods, logistic regression, and score tests for haplotypes. *Genetic Epidemiology*, 33:183–197.

Lorenzo Bermejo, J., Perez, A. G., Brandt, A., Hemminki, K., and Matthews, A. G. (2011). Comparison of six statistics of genetic association regarding their ability to discriminate between causal variants and genetically linked markers. *Human Heredity*, 72:142–152.

Lv, J., Yang, H., and Guo, C. (2015). An efficient and robust variable selection method for longitudinal generalized linear models. *Computational Statistics and Data Analysis*, 82:74–88.

Majumdar, A., Bhattacharya, S., Basu, A., and Ghosh, S. (2013). A novel Bayesian semiparametric algorithm for inferring population structure and adjusting for case-control association tests. *Biometrics*, 69:164–173.

McCullagh, P. and Nelder, J. A. (1996). *Generalized Linear Models*. Monographs on Statistics & Applied Probability 37. Chapman & Hall, London, UK, 2nd edition. Chapter 2.

Meijers-Heijboer, H., van den Ouweland, A., Klijn, J., Wasielewski, M., de Snoo, A., Oldenburg, R., Hollestelle, A., Houben, M., Crepin, E., van Veghel-Plandsoen, M., Elstrodt, F., van Duijn, C., Bartels, C., Meijers, C., Schutte, M., McGuffog, L., Thompson, D., Easton, D. F., Sodha, N., Seal, S., Barfoot, R., Mangion, J., Chang-Claude, J., Eccles, D., Eeles, R., Evans, D. G., Houlston, R., Murday, V., Narod, S., Peretz, T., Peto, J., Phelan, C., Zhang, H. X., Szabo, C., Devilee, P., Goldgar, D., Futreal, P. A., Nathanson, K. L., Weber, B. L., Rahman, N., and Stratton, M. R. (2002). Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nature Genetics*, 31:55–59.

Mersmann, O. (2014). *microbenchmark: Accurate Timing Functions*.

Min, Y. and Agresti, A. (2002). Modeling nonnegative data with clumping at zero: A survey. *Journal of Iranian Statistical Society*, 1:7–33.

Mitchell, B. D., Kammerer, C. M., Blangero, J., Mahaney, M. C., Rainwater, D. L., Dyke, B., Hixson, J. E., Henkel, R. D., Sharp, R. M., Comuzzie, A. G., VandeBerg, J. L., Stern, M. P., and MacCluer, J. W. (1996). Genetic and environmental contributions to cardiovascular risk factors in Mexican Americans: The San Antonio family heart study. *Circulation*, 94:2159–2170.

Mittal, S. and El-Serag, H. B. (2013). Epidemiology of hepatocellular carcinoma: Consider the population. *Journal of Clinical Gastroenterology*, 47(Supp. 1):S2–S6.

Montgomery, G. W., Campbell, M. J., Dickson, P., Herbert, S., Siemering, K., Ewen-White, K. R., Visscher, P. M., and Martin, N. G. (2005). Estimation of the rate of SNP genotyping errors from DNA extracted from different tissues. *Twin Research and Human Genetics*, 8:346–352.

Muhlbauer, A., Spichtinger, P., and Lohmann, U. (2009). Application and comparison of robust linear regression methods for trend estimation. *Journal of Applied Meteorology and Climatology*, 48:1961–1970.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33:341–365.

Müller, C. H. (2004). Redescending M-estimators in regression analysis, cluster analysis and image analysis. *Discussiones Mathematicae Probability and Statistics*, 24:59–75.

Muthukrishnan, R. and Radha, M. (2010). M-estimators in regression models. *Journal of Mathematics Research*, 2:23–27.

Nakamura, T. (1990). Corrected score function for errors-in-variables models: Methodology and application to generalized linear models. *Biometrika*, 77:127–137.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, 135:370–384.

Neumann, O., Kesselmeier, M., Geffers, R., Pellegrino, R., Radlwimmer, B., Hoffmann, K., Ehemann, V., Schemmer, P., Schirmacher, P., Lorenzo Bermejo, J., and Longerich, T. (2012). Methylome analysis and integrative profiling of human HCCs identify novel protumorigenic factors. *Hepatology*, 56:1817–1827.

Newton-Cheh, C., Johnson, T., Gateva, V., Tobin, M. D., Bochud, M., Coin, L., Najjar, S. S., Zhao, J. H., Heath, S. C., Eyheramendy, S., Papadakis, K., Voight,

B. F., Scott, L. J., Zhang, F., Farrall, M., Tanaka, T., Wallace, C., Chambers, J. C., Khaw, K.-T., Nilsson, P., van der Harst, P., Polidoro, S., Grobbee, D. E., Onland-Moret, N. C., Bots, M. L., Wain, L. V., Elliott, K. S., Teumer, A., Luan, J., Lucas, G., Kuusisto, J., Burton, P. R., Hadley, D., McArdle, W. L., Wellcome Trust Case Control Consortium, Brown, M., Dominiczak, A., Newhouse, S. J., Samani, N. J., Webster, J., Zeggini, E., Beckmann, J. S., Bergmann, S., Lim, N., Song, K., Vollenweider, P., Waeber, G., Waterworth, D. M., Yuan, X., Groop, L., Orho-Melander, M., Allione, A., Di Gregorio, A., Guarrera, S., Panico, S., Ricceri, F., Romanazzi, V., Sacerdote, C., Vineis, P., Barroso, I., Sandhu, M. S., Luben, R. N., Crawford, G. J., Jousilahti, P., Perola, M., Boehnke, M., Bonnycastle, L. L., Collins, F. S., Jackson, A. U., Mohlke, K. L., Stringham, H. M., Valle, T. T., Willer, C. J., Bergman, R. N., Morken, M. A., Döring, A., Gieger, C., Illig, T., Meitinger, T., Org, E., Pfeufer, A., Wichmann, H. E., Kathiresan, S., Marrugat, J., O'Donnell, C. J., Schwartz, S. M., Siscovick, D. S., Subirana, I., Freimer, N. B., Hartikainen, A.-L., McCarthy, M. I., O'Reilly, P. F., Peltonen, L., Pouta, A., de Jong, P. E., Snieder, H., van Gilst, W. H., Clarke, R., Goel, A., Hamsten, A., Peden, J. F., Seedorf, U., Syvänen, A.-C., Tognoni, G., Lakatta, E. G., Sanna, S., Scheet, P., Schlessinger, D., Scuteri, A., Dörr, M., Ernst, F., Felix, S. B., Homuth, G., Lorbeer, R., Reffelmann, T., Rettig, R., Völker, U., Galan, P., Gut, I. G., Hercberg, S., Lathrop, G. M., Zelenika, D., Deloukas, P., Soranzo, N., Williams, F. M., Zhai, G., Salomaa, V., Laakso, M., Elosua, R., Forouhi, N. G., Völzke, H., Uiterwaal, C. S., van der Schouw, Y. T., Numans, M. E., Matullo, G., Navis, G., Berglund, G., Bingham, S. A., Kooner, J. S., Connell, J. M., Bandinelli, S., Ferrucci, L., Watkins, H., Spector, T. D., Tuomilehto, J., Altshuler, D., Strachan, D. P., Laan, M., Meneton, P., Wareham, N. J., Uda, M., Jarvelin, M.-R., Mooser, V., Melander, O., Loos, R. J. F., Elliott, P., Abecasis, G. R., Caulfield, M., and Munroe, P. B. (2009). Genome-wide association study identifies eight loci associated with blood pressure. *Nature Genetics*, 41:666–676.

Office for Official Publications of the European Communities (2006). *Population statistics – Detailed Tables*. Luxembourg.

Padmanabhan, S., Newton-Cheh, C., and Dominiczak, A. F. (2012). Genetic basis of blood pressure and hypertension. *Trends in Genetics*, 28:397–408.

Pencina, M. J., D'Agostino Sr., R. B., D'Agostino Jr., R. B., and Vasan, R. S.

(2008). Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Statistics in Medicine*, 27:157–172.

Powers, S., Gopalakrishnan, S., and Tintle, N. (2011). Assessing the impact of non-differential genotyping errors on rare variant tests of association. *Human Heredity*, 72:153–160.

R Core Team (2013). *R: A language and environment for statistical computing.*

Ramsey, F. L. and Schafer, D. W. (2002). *The statistical sleuth – A course in methods of data analysis.* Brooks/Cole, Cengage Learning, Belmont (CA, USA), 2nd edition. Chapter 22.2.

Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., and Maechler, M. (2012). *robustbase: Basic Robust Statistics.*

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American statistical association*, 79:871–880.

Sandoval-Soto, L., Kesselmeier, M., Schmitt, V., Wild, A., and Kesselmeier, J. (2012). Observations of the uptake of carbonyl sulfide (COS) by trees under elevated atmospheric carbon dioxide concentrations. *Biogeosciences*, 9:2935–2945.

Sarkar, S. K., Midi, H., and Rana, S. (2011). Detection of outtiers and influential observations in binary logistic regression: An empirical study. *Journal of Applied Seiences*, 11:26–35.

Seyffert, W. (2003). *Lehrbuch der Genetik.* Spektrum-Verlag/Gustav-Fischer Verlag, Heidelberg, Germany, 2nd edition. Chapter 21.4.

Shevlyakov, G., Morgenthaler, S., and Shurygin, A. (2008). Redescending M-estimators. *Journal of Statistical Planning and Inference*, 138:2906–2917.

Stratton, M. R., Campbell, P. J., and Futreal, P. A. (2009). The cancer genome. *Nature*, 458:719–724.

Valdora, M. and Yohai, V. J. (2014). Robust estimators for generalized linear models. *Journal of Statistical Planning and Inference*, 146:31–48.

van Wieringen, W. N., Roś, B. P., and Wilting, S. M. (2013). Modeling the DNA copy number aberration patterns in observational high-throughput cancer data.

Statistical Applications in Genetics and Molecular Biology, 12:143–174.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, USA, 4th edition.

Vickers, A. J. and Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, 26:565–574.

Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J. (2012). Five years of GWAS discovery. *American Journal of Human Genetics*, 90:7–24.

Wang, Y., O'Connell, J. R., McArdle, P. F., Wade, J. B., Dorff, S. E., Shah, S. J., Shi, X., Pan, L., Rampersaud, E., Shen, H., Kim, J. D., Subramanya, A. R., Steinle, N., Parsa, A., Ober, C. C., Welling, P. A., Chakravarti, A., Weder, A. B., Cooper, R. S., Mitchell, B. D., Shuldiner, A. R., and Chang, Y.-P. C. (2009). Whole-genome association study identifies STK39 as a hypertension susceptibility gene. *Proceedings of the National Academy of Sciences of the United States of America*, 106:226–231.

Wen, Y.-W., Tsai, Y.-W., Wu, D. B.-C., and Chen, P.-F. (2013). The impact of outliers on net-benefit regression model in cost-effectiveness analysis. *PLOS ONE*, 8:e65930.

Wilcox, R. R. (1998). A note on the Theil–Sen regression estimator when the regressor is random and the error term is heteroscedastic. *Biometrical Journal*, 40:261–268.

Wray, N. R., Yang, J., Hayes, B. J., Price, A. L., Goddard, M. E., and Visscher, P. M. (2013). Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics*, 14:507–515.

Xiao, R. and Boehnke, M. (2009). Quantifying and correcting for the winner's curse in genetic association studies. *Genetic Epidemiology*, 33:453–462.

Yanek, L. R., Moy, T. F., Blumenthal, R. S., Raqueño, J. V., Yook, R. M., Hill, M. N., Becker, L. C., and Becker, D. M. (1998). Hypertension among siblings of persons with premature coronary heart disease. *Hypertension*, 32:123–128.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., and Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42:565–U131.

Yu, W., Gwinn, M., Clyne, M., Yesupriya, A., and Khoury, M. J. (2008). A navigator for human genome epidemiology. *Nature Genetics*, 40:124–125.

Zeileis, A., Kleiber, C., and Jackman, S. (2007). Regression models for count data in R. Research Report Series / Department of Statistics and Mathematics 53, Department of Statistics and Mathematics, WU Vienna University of Economics and Business, Wien, Austria.

Zhong, H. and Prentice, R. L. (2008). Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics*, 9:621–634.

Zöllner, S. and Pritchard, J. K. (2007). Overcoming the winner's curse: Estimating penetrance parameters from case-control data. *American Journal of Human Genetics*, 80:605–615.

Zou, J., Lippert, C., Heckerman, D., Aryee, M., and Listgarten, J. (2014). Epigenome-wide association studies without the need for cell-type composition. *nature Methods*, 11:309–311.

Own publications

Research articles

Published / accepted

Bergmann, L., Martini, S., **Kesselmeier, M.**, Armbruster, W., Notheisen, T., Adamzik, M., and Eichholz, R. (accepted). Phrenic Nerve Block Caused by Interscalene Brachial Plexus Block: Breathing Effects of Different Sites of Injection. *BMC Anesthesiology*.

Blurton, S. P., **Kesselmeier, M.**, and Gondan, M. (2012). Fast and accurate calculations for cumulative first-passage time distributions in Wiener diffusion models. *Journal of Mathematical Psychology*, 56:470–475.

Gondan, M., Blurton, S. P., and **Kesselmeier, M.** (2014). Even faster and even more accurate first-passage time densities and distributions for the Wiener diffusion model. *Journal of Mathematical Psychology*, 60:20–22.

Hagel, S., Reischke, J., **Kesselmeier, M.**, Winning, J., Gastmeier, P., Brunkhorst, F. M., Scherag, A., and Pletz, M. W. (2015). Quantifying the Hawthorne effect in hand hygiene compliance through comparing direct observation with automated hand hygiene monitoring. *Infection Control & Hospital Epidemiology*, 36:957–962.

Hinney, A., **Kesselmeier, M.**, Jall, S., Volckmar, A.-L., Föcker, M., Antel, J., GCAN, WTCCC3, Heid, I. M., Winkler, T. W., GIANT, Grant, S. F. A., EGG, Guo, Y., Bergen, A. W., Kaye, W., Berrettini, W., Hakonarson, H., Price Foundation Collaborative Group, Children's Hospital of Philadelphia/Price Foundation, Herpertz-Dahlmann, B., de Zwaan, M., Herzog, W., Ehrlich, S., Zipfel, S., Egberts, K. M., Adan, R., Brandys, M., van Elburg, A., Boraska Perica, V., Franklin, C. S., Tschöp, M. H., Zeggini, E., Bulik, C. M., Collier, D., Scherag, A., Müller, T. D., and Hebebrand, J. (2016). Evidence for three genetic loci involved in both anorexia nervosa risk and variation of body mass index. *Molecular Psychiatry.* [Epub ahead of print]

Kesselmeier, M., Legrand, C., Peil, B., Kabisch, M., Fischer, C., Hamann, U., and Lorenzo Bermejo, J. (2014). Practical Investigation of the Performance of Robust Logistic Regression to Predict the Genetic Risk of Hypertension. *BMC Proceedings*, 8(Suppl 1):S65.

Kesselmeier, M., Pütter, C., Volckmar, A.-L., Baurecht, H., Grallert, H., Illig, T., Ismail, K., Ollikainen, M., Silén, Y., Keski-Rahkonen, A., Bulik, C. M., Collier, D. A., Zeggini, E., Hebebrand, J., Scherag, A., Hinney, A., GCAN, and WTCCC3 (accepted). High-throughput DNA methylation analysis in anorexia nervosa confirms *TNXB* hypermethylation in starvation. *The World Journal of Biological Psychiatry.*

Neumann, O., **Kesselmeier, M.**, Geffers, R., Pellegrino, R., Radlwimmer, B., Hoffmann, K., Ehemann, V., Schemmer, P., Schirmacher, P., Lorenzo Bermejo, J., and Longerich, T. (2012). Methylome analysis and integrative profiling of human HCCs identify novel protumorigenic factors. *Hepatology*, 56:1817–1827.

Sandoval-Soto, L., **Kesselmeier, M.**, Schmitt, V., Wild, A., and Kesselmeier, J. (2012). Observations of the uptake of carbonyl sulfide (COS) by trees under elevated atmospheric carbon dioxide concentrations. *Biogeosciences*, 9:2935–2945.

Stein, C., Makarewicz, O., Bohnert, J. A., Pfeifer, Y., **Kesselmeier, M.**, Hagel, S., and Pletz, M. W. (2015). Three dimensional checkerboard synergy analysis of colistin, meropenem, tigecycline against multidrug-resistant clinical Klebsiella pneumonia isolates. *PLOS ONE*, 10:e0126479.

In preparation

Ali, S., Akhter, S., Neubauer, H., Scherag, A., **Kesselmeier, M.**, Khan, I., Azam, A., Qadeer, S., and Ali, Q. (in preparation). Brucellosis in Pregnant Women from Pakistan: An Observational Study.

Kesselmeier, M., and Lorenzo Bermejo, J. (in preparation). Robust logistic regression to narrow down the winner's curse for rare and recessive susceptibility

variants.

Kesselmeier, M., Neumann, O., Longerich, T., Geffers, R., Zucman-Rossi, J., Imbeaud, S., Derambure, C., Schirmacher, P., and Lorenzo Bermejo, J. (in preparation). Aetiology specific methylation and gene expression patterns in human hepatocellular carcinoma.

Kesselmeier, M., and Scherag, A. (in preparation). High-throughput DNA methylation association analyses with reference-free cell type adjustment: A method comparison.

Conference contributions

Blurton, S. P., Gondan, M., and **Kesselmeier, M.** (2016). Fast and accurate calculations for cumulative first-passage time distributions in Wiener diffusion models with drift variation. *The 2016 Meeting of the European Mathematical Psychology Group.*

Blurton, S. P., **Kesselmeier, M.**, and Gondan, M. (2013). Fast and accurate calculations for Wiener diffusion models. *46th Annual Meeting of the Society for Mathematical Psychology.*

Ganzinger, M., **Kesselmeier, M.**, Fabian, J., and Knaup, P. (2012). An Encapsulated R Application for the Guided Analysis of Genomic Data. *Studies in Health Technology and Informatics*, 180:1144–1146.

Hinney, A., **Kesselmeier, M.**, Volckmar, A.-L., Antel, J., GCAN, WTCCC3, Heid, I. M., Winkler, T. W., GIANT, Herpertz-Dahlmann, B., de Zwaan, M., Herzog, W., Ehrlich, S., Zipfel, S., Egberts, K. M., Adan, R., Brandys, M., Zeggini, E., Bulik, C., Collier, D., Scherag, A. and Hebebrand, J. (2015). Genetic variation at three genetic loci involved in anorexia nervosa are associated with body weight regulation. *16th International ESCAP Congress – From Research to Clinical Practice Linking the Expertise.*

Hinney, A., **Kesselmeier, M.**, Volckmar, A.-L., Antel, J., GCAN, WTCCC3, Heid, I. M., Winkler, T. W., GIANT, Herpertz-Dahlmann, B., de Zwaan, M., Herzog, W., Ehrlich, S., Zipfel, S., Egberts, K. M., Adan, R., Brandys, M., Zeggini, E., Bulik, C., Collier, D., Scherag, A. and Hebebrand, J. (2015). Genetic

variation at three genetic loci involved in anorexia nervosa are associated with body weight regulation. *ECO2015 – 22nd European Congress on Obesity.*

Hinney, A., Volckmar, A., **Kesselmeier, M.**, Antel, J., Heid, I. M., Herpertz-Dahlmann, B., Bulik, C., Collier, D., Scherag, A., and Hebebrand, J. (2015). Genetic variation at three genetic loci involved in anorexia nervosa are associated with body weight regulation. *European Child & Adolescent Psychiatry*, 24:S64.

Kesselmeier, M., Ganzinger, M., Knaup, P., Kieser, M., and Lorenzo Bermejo, J. (2011). Regression models to explore genomic instability and methylation in the TRR collective. *SFB/TRR77 Retreat 2011.*

Kesselmeier, M., Hinney, A., Hebebrand, J., and Scherag, A. (2015). High-throughput DNA methylation association analyses with reference-free cell type adjustment: A method comparison. *60. GMDS-Jahrestagung 2015.*

Kesselmeier, M., Longerich, T., Geffers, R., Ganzinger, M., and Lorenzo Bermejo, J. (2012). Comparison Of Count Models Regarding Their Ability To Capture The Relationship Between Genomic Instability, Methylation And Expression In Human Hepatocellular Carcinoma. *Genetic Epidemiology*, 36:157–157.

Kesselmeier, M., Longerich, T., Geffers, R., and Lorenzo Bermejo, J. (2012). Comparison of count models regarding their ability to capture the relationship between genomic instability and methylation in human hepatocellular carcinoma. *58. Biometrisches Kolloquium.*

Kesselmeier, M., Longerich, T., Neumann, O., Geffers, R., and Lorenzo Bermejo, J. (2012). Comparison of count models regarding their ability to capture the relationship between genomic instability and methylation in human hepatocellular carcinoma. *Annals of Human Genetics*, 76:417–418.

Kesselmeier, M., Longerich, T., Neumann, O., Geffers, R., and Lorenzo Bermejo, J. (2012). Comparing Count Regression Models to Investigate the Relationship Between Genomic Instability and Gene Methylation in Human Hepatocellular Carcinoma. *Genetic Epidemiology*, 36: 768–769.

Kesselmeier, M., Longerich, T., Neumann, O., Geffers, R., and Lorenzo Bermejo, J. (2012). Comparison of count models regarding their ability to capture the relationship between genomic instability and methylation in human hepatocellular carcinoma. *International Research conference on liver cancer – "From molecular*

pathogenesis to targeted therapies".

Kesselmeier, M., and Lorenzo Bermejo, J. (2013). Comparative performance of robust logistic regression in the framework of genetic risk prediction. *3rd joint Statistical Meeting of the DAGStat "Statistics under one Umbrella".*

Kesselmeier, M., and Lorenzo Bermejo, J. (2013). Hampel's function in robust logistic regression applied to genetic association analysis. *Human Heredity*, 76:104–105.

Kesselmeier, M., and Lorenzo Bermejo, J. (2016). Comparison of standard and robust logistic regression with different weighting functions. *4th joint Statistical Meeting of the DAGStat "Statistics under one Umbrella".*

Kesselmeier, M., Neumann, O., Geffers, R., Longerich, T., and Lorenzo Bermejo, J. (2012). Aetiology-specific gene methylation and expression in human hepatocellular carcinoma. *SFB/TRR77 Junior Retreat 2012.*

Kesselmeier, M., and Scherag, A. (2015). High-throughput DNA methylation association analyses with reference-free cell type adjustment: A method comparison. *Infection*, 43 (Supplement 1):S29.

Kesselmeier, M., and Scherag, A. (2016). Comparison of high-throughput DNA methylation association analyses approaches with reference-free cell type adjustment. *4th joint Statistical Meeting of the DAGStat "Statistics under one Umbrella".*

Lorenzo Bermejo, J., Kabisch, M., Fischer, C., Legrand, C., **Kesselmeier, M.**, Hamann, U., and Peil, B. (2012). Exploratory Investigation of the Accuracy of Genotype Imputation Relying on Different Approaches to Identify the Population of Reference. *Genetic Analysis Workshop 18.*

Lorenzo Bermejo, J., **Kesselmeier, M.**, Legrand, C., Kabisch, M., Fischer, C., Hamann, U., and Peil, B. (2013). Preliminary results on the potential of robust methods for three common applications in statistical genetics. *41st European Mathematical Genetics Meeting.*

Lorenzo Bermejo, J., **Kesselmeier, M.**, Legrand, C., Kabisch, M., Fischer, C., Hamann, U., and Peil, B. (2013). Robust methods in three common statistical genetics applications. *IGES 22nd annual conference.*

Neumann, O., **Kesselmeier, M.**, Geffers, R., Pellegrino, R., Radlwimmer, B.,

Hoffmann, K., Schemmer, P., Schirmacher, P., Lorenzo Bermejo, J., and Longerich, T. (2012). Integrative molecular profiling-based identification of protumorigenic factors of human hepatocarcinogenesis. *International Research conference on liver cancer – "From molecular pathogenesis to targeted therapies"*.

Neumann, O., **Kesselmeier, M.**, Radlwimmer, B., Schemmer, P., Lorenzo Bermejo, J., Schirmacher, P., and Longerich, T. (2012). Methylation profiling combined with genomic and transcriptomic data points out to new tumor suppressor genes in human hepatocarcinogenesis. *Zeitschrift für Gastroenterologie*, 50 – P5_42.

Scherag, A., **Kesselmeier, M.**, Hagel, S., Pletz, M., and Brunkhorst, F. M. (2016). Ergebnisse der krankenhausweiten ALERTS Studie zu nosokomialen Infektionen und Interventionseffekten nach Hygienemanahmen. *DACh-Epidemiologietagung 2016*.

Schöneweck, F., **Kesselmeier, M.**, and Scherag, A. (2013). Host genomics in sepsis: An update on the available evidence and the ongoing methodological challenges. *Human Heredity*, 76:103.

Curriculum vitae

Personal information

Name	Miriam Kesselmeier
Birth day, place of birth	May 28, 1983 in Köln
Nationality	German

School career

09/1994-03/2003	Rabanus-Maurus-Gymnasium Mainz
28/03/2003	Abitur

University career

04/2004-01/2011	Studies in Mathematics with minor Physics at the Johannes Gutenberg University of Mainz
28/04/2006	Diplom-Vorprüfung
31/01/2011	Diplom
02/2012	Acceptance of doctoral thesis proposal by the University of Heidelberg

Professional career

03/2008-05/2008	Scientific assistant at the faculty for Physics, Mathematics and Computer Science, Johannes Gutenberg University Mainz
10/2008-03/2010	Scientific assistant and Diploma student at Fraport AG, Frankfurt am Main
03/2011-12/2013	Research assistant at the Institute of Medical Biometry and Informatics, Ruprecht Karls University of Heidelberg
since 01/2014	Research assistant at the Center for Sepsis Control and Care, Clinical Epidemiology, Jena University Hospital

Acknowledgement

I acknowledge the support of my doctoral adviser Prof. Dr. Justo Lorenzo Bermejo. He offered me a fascinating task and his mentoring was very helpful. The fruitful discussions resulted in new ideas and were of great value for my work.

I gratefully acknowledge the support of this work by the programme SFB/TRR77 under a grant of the Deutsche Forschungsgemeinschaft (DFG, subproject Z2).

Special thanks to Prof. Dr. André Scherag from the Center for Sepsis Control and Care (CSCC) at the Jena University Hospital. He provided me the opportunity to continue my doctoral research on my new position in Jena. His comments giving support to me were always very helpful.

I want to thank Barbara, Carine, Franziska, Justine, Maria and Matthias for their friendship and being always open for conversation. Their support within these discussions was of great help to continue my work and to develop new approaches. Special thanks to Elke who always lent me her ears whenever necessary.

I would like to specifically note the continuous support of my parents who encouraged me since birth wherever needed.