Miriam Kesselmeier
Dr. sc. hum.

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

Fach/Fachrichtung: Medizinische Biometrie und Informatik
Doktorvater: apl. Prof. Dr.sc.agr. Justo Lorenzo Bermejo

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.