A COMPARISON OF DIFFERENT METHODS FOR MODELLING RARE EVENTS DATA

Bart Van der Paal

Promotor: Prof. dr. Dries F. Benoit

Thesis submitted in fulfillment of the requirements for the degree of Master of Statistical Data Analysis
The author and promoter give the permission to use this thesis for consultation and to copy parts of it for personal use. Every other use is subject to the copyright laws, more specifically the source must be extensively specified when using from this thesis.

Ghent, August 2014

The promoter

Prof. Dr. Dries F. Benoit

The Author

Bart Van der Paal
Foreword

This thesis was written in 2014 in fulfillment of the requirements for the degree of Master of Science in Statistical Data Analysis. My work examines different methods for modelling rare events data using binomial linear regression. As such it touches upon theory and practice taught in several courses of the master program, notably Principles of Statistics, Analysis of Continuous Data, Statistical Inference, Statistical Computing, Categorical Data Analysis, Data Mining, Computer Intensive Methods, and last but not least Bayesian Statistics.

Preparing a thesis, doing research and writing the report represent an important workload. But it has offered me great satisfaction. I have enjoyed combining insights from different fields of statistics to optimally address the subject. I hope the readers of this work find it equally enjoyable to read it. And I hope it offers some interesting ideas that can be useful for future research.

I would like to thank my promoter Prof. Dr. Dries F. Benoit, for proposing this subject which proved to be highly interesting and challenging, as well as for his relentless support and access to unpublished papers from his own field of research. Additionally I would like to thank Prof. Dr. Xia Wang, from the University of Cincinatti, for sharing the R-code accompanying the paper “Generalized Extreme Value Regression for Binary Response data”.

Bart Van der Paal
Ghent, August 31st, 2014
## Contents

1 Introduction .......................................................... 1  
  1.1 Scope ........................................................................ 1  
  1.2 Content ...................................................................... 1  

2 Linear Regression ....................................................... 2  
  2.1 Model and estimation ................................................. 2  
    2.1.1 Basic regression model ........................................ 2  
    2.1.2 Normal error regression model ............................ 3  
  2.2 Remedial measures for influential outliers: robust regression ........................................... 3  
    2.2.1 LAR regression .................................................. 3  
    2.2.2 Other robust regression models ............................ 4  
    2.2.3 Conclusions on error distributions ........................ 4  
  2.3 Remedial measures for multicollinearity: penalized regression ........................................... 5  
    2.3.1 Multicollinearity .................................................. 5  
    2.3.2 Ridge and lasso .................................................. 5  

3 Binomial Linear Regression .............................................. 6  
  3.1 Model ........................................................................ 6  
  3.2 Estimation and calculation .......................................... 7  
  3.3 Loss functions .......................................................... 8  
  3.4 Evaluation of basic link functions ............................... 9  

4 Bayesian approach ....................................................... 12  
  4.1 Bayesian statistics ..................................................... 12  
  4.2 Our approach in this study .......................................... 13  
  4.3 Stan .......................................................................... 14  

5 Rare Events ................................................................... 15  
  5.1 Problem setting ......................................................... 15  
    5.1.1 Definition .......................................................... 15  
    5.1.2 Bias ...................................................................... 16  
    5.1.3 Infinite estimates .................................................. 17  
  5.2 Remedies for rare events data problems ...................... 17  
    5.2.1 Sampling methods ............................................... 17  
    5.2.2 Penalization or priors on $\beta$ ............................ 18  
    5.2.3 Skewed link function ............................................ 19
## Contents

### 6 Priors

6.1 Firth’s method .......................... 20
6.2 Gelman’s informative prior .......................... 20
  6.2.1 Centering and scaling .......................... 21
  6.2.2 A weakly informative t-family of prior distributions .......................... 21
  6.2.3 Implementation .......................... 22

### 7 Link functions

7.1 Quantile link functions .......................... 23
  7.1.1 Quantile regression .......................... 23
  7.1.2 Binary quantile regression .......................... 25
  7.1.3 Practical use .......................... 29
  7.1.4 Extension .......................... 33
7.2 Generalized extreme value link .......................... 35
  7.2.1 Generalized extreme value distribution .......................... 36
  7.2.2 GEV based models .......................... 36
  7.2.3 Evaluation and practical use for this study .......................... 38
7.3 Skewed t-Link .......................... 39
  7.3.1 Skewed t-Link distribution .......................... 39
  7.3.2 Practical use .......................... 41
7.4 Flexible student-t link .......................... 42
  7.4.1 The model .......................... 42
  7.4.2 Evaluation and practical use .......................... 44
7.5 Symmetric power link family .......................... 45
  7.5.1 Power link functions .......................... 45
  7.5.2 Practical use .......................... 47

### 8 Results

8.1 Models and method .......................... 50
8.2 Data sets .......................... 50
  8.2.1 Car .......................... 52
  8.2.2 Yeast .......................... 52
  8.2.3 Segment .......................... 53
  8.2.4 KRRP .......................... 53

### 9 Conclusion .......................... 56

### A Calculations .......................... 58

A.1 Section 7.1.3: correction factor on the default prior scale .......................... 58
A.2 Section 7.1.3: location for the prior on the intercept .......................... 58
A.3 Section 7.1.4: correction factor on the default prior scale .......................... 59
A.4 Section 7.5.2: location for the prior on the intercept .......................... 59

### B Stan model code .......................... 61

B.1 bqr .......................... 61
B.2 bqra .......................... 62
B.3 alr .......................... 64
B.4 alra .......................... 66
B.5 splogit .......................... 67

### Bibliography .......................... 69
Chapter 1

Introduction

1.1 Scope

The subject of modelling rare events data is possibly very large, including supervised and unsupervised machine learning methods, non-linear curve fitting, non-parametric methods, etc. The scope of this thesis is limited to generalized linear models, notably binomial linear regression. This subject is addressed from a frequentist and a Bayesian perspective. The focus is on the use of different priors and link functions. The priors help to obtain reasonable and stable point estimates for the regression coefficients. Different link functions are investigated and their effect on obtaining good fit to imbalanced binary data. The report contains an elaboration of the theory and an application of various methods on real data using cross-validation.

1.2 Content

As the subject is binomial linear regression, chapters 2 offers a quick rehearsal of the standard linear regression model and chapter 3 a short intro to binomial linear regression. Chapter 4 explains the difference between the frequentist and Bayesian perspective, and how both are useful for this subject. Chapter 5 describes what we understand as rare events data and the specific problems that must be solved when modelling these. Chapter 6 discusses the use of priors for remedying rare events data problems. Chapter 7 looks deeper into various link functions for the binomial regression model. Chapter 8 reports the results of the application of various models on some real-life data sets. Chapter 9 rounds up the conclusions of this study. In the Annex the Stan the Stan model code is attached.
Chapter 2
Linear Regression

2.1 Model and estimation

2.1.1 Basic regression model

The model can be stated as follows:

\[ y_i = X_i' \beta + \varepsilon_i \quad i = 1, \ldots, n \]  

(2.1)

Where:

- \( y_i \) is the value of the response in the \( i \)th observation
- \( \beta \) is a vector of \( p \) coefficients, the first of which \( \beta_0 \) is the intercept
- \( X_i \) is a vector containing 1 as first element and the values of the \( p - 1 \) covariates of the \( i \)th observation as other elements
- \( \varepsilon_i \) is a random error term, with mean \( E[\varepsilon_i] = 0 \), and variance \( \sigma^2[\varepsilon_i] = \sigma^2 \); \( \varepsilon_i \) and \( \varepsilon_j \) are uncorrelated so that their covariance is zero, i.e. \( \sigma[\varepsilon_i, \varepsilon_j] = 0 \) for all \( i, j \); \( i \neq j \).

Important assumptions of this model are:

- **Linearity**: the expected value of the response variable is a linear combination of the predictor variables
- **Constancy of variance** or homoskedasticity: the response variables have the same variance in their errors regardless of the values of the predictor variables
- **Lack of correlation** in the errors
- **No multicollinearity** in the predictors, the design matrix \( X \) must have full rank \( p \), a predictor is not allowed to be a linear combination of other predictors
Chapter 2. Linear Regression

The method of ordinary least squares (OLS) is used to find good estimators for the regression coefficients $\beta$. The estimates for the regression coefficients are the values that minimize the sum of squared residuals (SSR).

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} (y_i - X_i' \beta)^2$$  \hspace{1cm} (2.2)

Or equivalently in matrix form:

$$\hat{\beta} = (X'X)^{-1}X'y$$  \hspace{1cm} (2.3)

The OLS estimators are unbiased and have minimum variance among unbiased estimators.

2.1.2 Normal error regression model

The normal error regression model is the same as the basic regression model with the additional assumption that the errors are independent and identically distributed (i.i.d.) following a normal distribution: $\varepsilon_i \sim N(0, \sigma^2)$.

This model is fully parametric and it requires more than the assumption that errors are uncorrelated. It needs the stronger assumption of statistical independence. Hence it can be estimated using the method of maximum likelihood (ML).

As $y_i \sim N(X_i' \beta, \sigma^2)$, the ML estimate $\hat{\beta}$ is found as the vector that maximizes the likelihood:

$$\hat{\beta} = \underset{\beta}{\text{argmax}} \left[ \prod_{i=1}^{n} \frac{1}{\sigma \sqrt{2\pi}} \exp\left(\frac{-(y_i - X_i' \beta)^2}{2\sigma^2}\right) \right]$$  \hspace{1cm} (2.4)

Or equivalently as the vector that maximizes the log likelihood with constant terms left out:

$$\hat{\beta} = \underset{\beta}{\text{argmax}} \left[ -n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - X_i' \beta)^2 \right]$$  \hspace{1cm} (2.5)

From the latter equation it is easy to see that the ML estimator $\hat{\beta}$ is exactly the same as the OLS estimator $\hat{\beta}$. Only the inference is different. Under the assumptions of the normal error regression model the OLS estimator $\hat{\beta}$ has a student-t distribution with $(n - p)$ degrees of freedom. The MLE uses asymptotic inference: $\hat{\beta}$ is normally distributed as $n \rightarrow \infty$.

2.2 Remedial measures for influential outliers: robust regression

2.2.1 LAR regression

Occasionally data sets contain observations that have a strong influence on fitted values or regression coefficients. In the linear regression diagnostics toolbox there are several valuable tools to detect such observations, e.g. DFFITS, Cook’s distance and DFBETAS measures. The OLS method is particularly sensitive to these cases.
Or formulated differently: the assumption of normal errors is not necessarily applicable to the data. The normal distribution is not tolerant of outliers as it has a light tailed probability distribution function (pdf). Occasional influential outliers can exert a strong influence effectively pulling the multidimensional regression plane towards themselves.

Formulated differently again: the linear regression model estimates the mean conditional outcome \( \mathbb{E}[y_i|X_i] = X_i' \beta \). The mean as a statistic is sensitive to outliers. It proves that estimating the conditional median outcome \( \text{Median}[y_i|X_i] \) is more robust. That is done using least absolute residuals (LAR) regression. Whereas OLS minimizes the \( L_2 \) norm of the residuals, the LAR criterion minimizes the \( L_1 \) norm:

\[
\hat{\beta}_{\text{LAR}} = \arg\min_{\beta} \sum_{i=1}^{n} |y_i - X_i' \beta| \tag{2.6}
\]

The fully parametric equivalent estimated with the ML method and yielding the same point estimates for the coefficients is a regression model assuming Laplace distributed errors: \( y_i \sim \text{Laplace}(X_i' \beta, s) \), where \( s \) is the scale. The heavier tails of the Laplace distribution can accommodate larger outliers.

### 2.2.2 Other robust regression models

Several other robust regression models exist. It is e.g. perfectly possible to assume an error distribution with even heavier tails, e.g. the student-t distribution. Especially the student-t(1) or Cauchy distribution is very forgiving on outliers. The probability density function (pdf) of the Cauchy distribution with location \( \mu \) and scale \( s \) is:

\[
f(x|\mu, s) = \frac{1}{\pi s \left(1 + \left(\frac{x-\mu}{s}\right)^2\right)} \tag{2.7}
\]

The criterion minimized when assuming the Cauchy distribution for the error term is:

\[-n \log s + \sum_{i=1}^{n} \log \left(s^2 + (y_i - X_i' \beta)^2\right),\] which can accommodate occasionally very large outliers.

### 2.2.3 Conclusions on error distributions

We draw the following conclusion from the previous sections:

- the choice of a distribution for the errors is not so much about being close to the true data generation process, which is unknown in practice. Rather, different error distributions imply a different line fitting and a different interpretation of the coefficient estimates and fitted values.

- Assuming normally distributed errors results in finding the regression line that minimizes the SSR. The fitted values can be considered the conditional mean outcome values.

- Assuming Laplace distributed errors results in finding the line that minimizes the LAR. The fitted values can be considered the conditional median outcome values.
2.3 Remedial measures for multicollinearity: penalized regression

2.3.1 Multicollinearity

An assumption of the regression model is absence of multicollinearity. If some of the predictor variables are perfectly correlated, the solution of the linear regression model is not unique and \((X'X)\) in equation (2.3) cannot be inverted. Perfect correlation is easily solved by removing predictors which are a linear combination of other predictors. A harder to solve problem is the presence of strongly but not perfectly correlated predictors. These result in variance inflation on the estimated coefficients and very large values for several of the coefficient estimates.

2.3.2 Ridge and lasso

This problem is usually tackled by minimizing a penalized SSR. The estimates thus obtained are not unbiased anymore. The idea behind penalization is to introduce a small amount of bias to obtain more stable estimates and thus better out of sample predictions. Ridge regression minimizes the RSS with a \(L^2\) penalty term applied to the coefficients excluding the intercept:

\[
\hat{\beta}_{\text{ridge}} = \arg\min_\beta \left[ \sum_{i=1}^{n} (y_i - X_i'\beta)^2 + \lambda \sum_{j=1}^{p-1} \beta_j^2 \right] \tag{2.8}
\]

Where \(\lambda\) is a constant tuning parameter that determines the magnitude of the penalty, which is usually learned from the data using cross-validation.

Lasso regression minimizes the RSS with a \(L^1\) penalty term applied:

\[
\hat{\beta}_{\text{lasso}} = \arg\min_\beta \left[ \sum_{i=1}^{n} (y_i - X_i'\beta)^2 + \lambda \sum_{j=1}^{p-1} |\beta_j| \right] \tag{2.9}
\]

The lasso is an alternative to ridge regression and it has the ability to shrink parameters all the way to 0, thus applying automated variable selection.
Chapter 3

Binomial Linear Regression

3.1 Model

When the outcome variable is binary, classical linear regression is inadequate and binomial regression is preferred. Without loss of generality, the outcome variable $y$ is assumed to take values 1 (usually denoted as “success” or “positive”) or 0 (“failure” or “negative”). In binomial regression the outcome variable is not predicted directly as a linear combination of the covariates. Rather an outcome value is considered to be the result of a Bernoulli trial. Binomial regression predicts the probability $P(y_i = 1)$ by applying the inverse of a link function $g^{-1}$ to a linear combination of covariates. That leads to the following model:

$$P(y_i = 1) = g^{-1}(X_i'\beta) \quad i = 1, ..., n$$  \hspace{1cm} (3.1)

The inverse link function $g^{-1}$ can be any monotonically increasing function that maps an input in range $(-\infty, \infty)$ to $[0, 1]$, and it has the typical sigmoid shape. In practice the cumulative distribution functions (cdf) of well-known random distributions are used as inverse link functions, e.g. the normal cdf $\Phi$ or the logistic cdf.

The binomial regression model can also be understood as a binary choice model assuming a latent variable $u_i$, often called the utility. The utility is modelled as a linear combination of covariates with a random error term $\varepsilon_i$ added. The outcome variable $y_i$ is the result of censoring the utility: it takes value 1 if the utility exceeds 0, otherwise 0.

$$u_i = X_i'\beta + \varepsilon_i \quad y_i = \begin{cases} 0, & \text{if } u_i \leq 0 \\ 1, & \text{if } u_i > 0 \end{cases}$$  \hspace{1cm} (3.2)

We can calculate that the probability $P(y_i = 1)$ as

$$P(y_i = 1) = 1 - F_{\varepsilon}(-X_i'\beta)$$  \hspace{1cm} (3.3)

Where $F_{\varepsilon}$ is the cdf of the distribution of $\varepsilon$. For symmetric distributions that simplifies to:

$$P(y_i = 1) = F_{\varepsilon}(X_i'\beta)$$  \hspace{1cm} (3.4)

Two common link functions are:
• the logit link, which corresponds to standard logistically distributed errors $\varepsilon_i \sim \text{Logistic}(\mu = 0, s = 1)$.

• the probit link, which corresponds to standard normally distributed errors $\varepsilon_i \sim N(\mu = 0, \sigma^2 = 1)$.

3.2 Estimation and calculation

Binomial regression models belong to the class of generalized linear models (GLM) and are estimated in the ML framework, or using Bayesian methods. The likelihood function $L(\beta | y, X)$ is derived from the binomial distribution likelihood:

$$L(\beta | y, X) = \prod_{i=1}^{n} g^{-1}(X_i' \beta)^{y_i} (1 - g^{-1}(X_i' \beta))^{1-y_i}$$

(3.5)

The log likelihood $\ell(\beta | y, X)$ is:

$$\ell(\beta | y, X) = \sum_{i=1}^{n} y_i \log g^{-1}(X_i' \beta) + \sum_{i=1}^{n} (1 - y_i) \log(1 - g^{-1}(X_i' \beta))$$

(3.6)

There is no closed form for the ML estimator. The estimates are usually found with an iteratively reweighted least squares method such as the Newton(-Raphson) algorithm. The algorithm requires the vector of derivatives $\ell'(\beta)$ and the matrix of the second derivatives $\ell''(\beta)$ of the log likelihood function.

The mode-finding algorithm proceeds by choosing a starting value $\beta^0$. Then for $t = 1, 2, 3, \ldots$, a new iteration $\beta^t$ is calculated until convergence:

$$\beta^t = \beta^{t-1} + [\ell''(\beta^{t-1})]^{-1} \ell'(\beta^{t-1})$$

(3.7)

The advantage of this method is that convergence is extremely fast once the iterates are close to the solution, where the quadratic approximation is accurate.

Another standard algorithm to find ML point estimates or posterior modes is the expectation-maximization (EM) algorithm. Mathematically finding ML estimates or posterior modes is done by taking the partial derivatives of the likelihood or posterior distribution and solve a set of equations. Usually the set of equations cannot be solved directly. Latent variables are added to the model which formulate the model more easily, notably they make it easy to estimate the parameters given the latent variables. Latent variables and parameters cannot be estimated simultaneously. The EM algorithm estimates the latent variables and parameters alternatingly until convergence. It can be proven that each iteration of the EM algorithm increases the log likelihood or log posterior density. The latent variable representation of the binomial regression model makes EM widely applicable.
3.3 Loss functions

In binomial regression there is no elegant equivalent for a loss function that can be expressed as a norm on the residuals, as is the case in classical regression where OLS minimizes the $L_2$ norm and median regression minimizes $L_1$ norm of the residuals.

The loss function $L(y, X, \beta)$ is equal to the negative log likelihood, so for the binomial model the loss $L(y, X, \beta)$ is:

$$L(y, X, \beta) = - \sum_{\{y_i=1\}} \log g^{-1}(X_i' \beta) - \sum_{\{y_i=0\}} \log(1 - g^{-1}(X_i' \beta)) \quad (3.8)$$

Which doesn’t correspond to a norm or other well known loss function on the residuals indeed. For symmetric link functions equation (3.8) simplifies to

$$L(y^*, X, \beta) = \sum_{i=1}^{n} L_i(y^*, X, \beta) = \sum_{i=1}^{n} - \log g^{-1}(y_i^* X_i' \beta) \quad (3.9)$$

Where $y_i^*$ is the $(-1, 1)$ encoding of the binary outcome variable $y_i$, e.g. by applying transformation $y_i^* = 2y_i - 1$. The loss $L_i$ as a function of $y_i^* X_i' \beta$ is visualised for some link functions in figure 3.1.

![Figure 3.1: Binomial regression loss $L_i$ as a function of $y_i^* X_i' \beta$ for the logit, probit and cauchit link, as well as the student-t(8) link. For $y_i^* X_i' \beta \to -\infty$, the logit has a linear asymptotic and the probit a quadratic. The student-t based loss functions are not convex.](image)

However when looking at the value of the logit and probit loss functions for outliers, we find that they approximate a simple norm function again. Outliers in binomial regression are data points that will easily be misclassified, as the outcome $y_i$ has low fitted probability $P(y_i = y_i)$. In the binary choice model outliers can be understood as data points with a large absolute error $|\varepsilon_i|$ in the direction opposite to $\text{sgn}(X_i' \beta)$, so that the utility $u_i$ ends up on the other side of 0 even if $|X_i' \beta|$ is large. These are the points with large negative values for $y_i^* X_i' \beta$. 
Chapter 3. Binomial Linear Regression

Below is our elaboration of equation (3.9) for the logit and probit regression and their approximation for outliers as \( y_i^* X_i' \beta \to -\infty \).

In the logistic regression, the contribution \( L_i(y_i^*, X_i, \beta) \) of a data point to the overall loss is:

\[
L_i(y_i^*, X_i, \beta) = -\log \left( \frac{\exp(y_i^* X_i' \beta)}{1 + \exp(y_i^* X_i' \beta)} \right)
\] (3.10)

Resulting in \( L_i(y_i^*, X_i, \beta) \approx |X_i' \beta| \) for outliers in the logistic regression, i.e. as \( y_i^* X_i' \beta \to -\infty \).

For the standard normal cdf \( \Phi \) we have the following definition, which we expand using partial integration:

\[
\Phi(x) = \int_{-\infty}^{x} e^{-t^2/2} dt = \frac{1}{\sqrt{2}} \left[ e^{-x^2/2} \right]_{-\infty}^{x} - \int_{-\infty}^{x} e^{-t^2/2} \frac{1}{\sqrt{2}} dt
\] (3.11)

Partial integration can be repeated, but in the limit \( x \to -\infty \) taking the first term is a good approximation, so that:

\[
\log \left( \lim_{x \to -\infty} \Phi(x) \right) \approx -\frac{x^2}{2} + \log(-x) \approx -\frac{x^2}{2}
\] (3.12)

Resulting in \( L_i(y_i^*, X_i, \beta) \approx \frac{(X_i' \beta)^2}{2} \) for outliers in the probit regression.

So we can conclude:

- outliers in the logistic regression contribute to the loss with approximately the \( L_1 \) norm of the link value
- outliers in the probit regression contribute to loss with approximately half the \( L_2 \) norm of the link value

That makes the probit regression more sensitive to outliers than the logistic regression.

3.4 Evaluation of basic link functions

The link function in the GLM is fixed by the choice of the noise term distribution in the equivalent binary choice model. Logit and probit links are the most commonly used symmetric link functions and they are available in many statistical packages. Given the logit link function \( \text{logit}(p) = \log \left( \frac{1}{p} \right) \), logistic regression is often preferred over probit regression because of the convenient interpretation of a regression coefficient (excluding the slope), as the unit impact of the corresponding predictor on the log-odds of the outcome variable. That is especially important when the goal is to obtain an explanatory model. As seen in the previous section, it also leads to more stable estimates because logit regression is less sensitive to outliers than probit regression.

We have done a qualitative evaluation of different link functions corresponding to a binary choice noise term with different distributions: standard normal (probit), standard logistic
(logit), student-t where the optimal degrees of freedom is estimated from the data and student-
t allowing for a different degrees of freedom in its left and right tail. The latter two models
with student-t link are not available in standard R-packages such as \texttt{glm}, so we have done
our own implementation in \textit{Stan}. Models with such link functions were applied to simulated data with a large number of ob-
servations $n = 1001$ with one covariate. Data were generated according to the logit model,
probit model and cauchit model (i.e. the binary choice model with student-t(1) noise). We
have evaluated the ability to retrieve the true coefficients, the fit as measured by the logscore
$\frac{1}{n} \sum_{i=1}^{n} - \log(|y_i - y_{ipred}|)$ or the mean squared errors (MSE) $\frac{1}{n} \sum_{i=1}^{n} (y_i - y_{ipred})^2$ and the classi-
fication rate. That results in the following findings:

- as can be expected, coefficients estimates are close to the true values if the same model
  was used for fitting as for data generation
- in practice the sigmoid response curves of logit and probit regression are usually very
close to one another
- binomial regression with the student-t link finds very high values for the degrees of
  freedom when the noise is standard normal, leading to an identical fit as probit
- binomial regression with the student-t link results in a superior fit when the noise is
  Cauchy distributed. However it finds various values for the degrees of freedom, mostly
  between 0.5 and 3.
- binomial regression with the student-t link with different degrees of freedom in left and
  right tail allows for skewness. We have found that it was not superior to the student-t
  model with a single degree of freedom, at least on simulated data generated from a
  symmetric model.
- the classification rate of various models is comparable, there is no consistently better
  performing model.
Figure 3.2: Example of GLM with 3 different link functions fitted to binary data generated from a binary choice model $y_i = I(x_i + \varepsilon_i > 0)$. $n = 1001$ observations were generated with $X$ a regular sequence in $[-5,5]$ with increment 0.01 and $\varepsilon_i$ i.i.d. samples from the t-distribution. Logistic and probit regression find response curves that are very close to one another. Using the student-t link results in a sigmoid with tails that approach 0 and 1 much slower. The misclassification rate of the three methods happens to be identical. Using the student-t link results in the best fit measured using the logscore or the MSE.
Chapter 4

Bayesian approach

4.1 Bayesian statistics

Bayesian statistics is entirely based on the Bayes’ rule:

\[ p(\beta|D) = \frac{p(D|\beta)p(\beta)}{p(D)} \]  

(4.1)

In the context of binomial linear regression, \( \beta \) is the coefficients vector and \( D \) are the observed data \( D = \{y, X\} \).

Bayesian statistics searches the posterior distribution of the coefficients \( p(\beta|D) \) given the data, based on the likelihood of the data given the coefficients \( p(D|\beta) \), and the prior distribution \( p(\beta) \) representing our prior knowledge on \( \beta \).

The posterior distribution can usually not be expressed as a known random variable probability distribution, except in very simple models and on the condition that conjugate priors are used. The posterior distribution properties are estimated by gathering a sufficiently large sample from the posterior distribution using Monte Carlo methods, such as Markov Chain Monte Carlo (MCMC).

The inference in frequentist and Bayesian statistics is different. Frequentist inference:

- assumes that there is a true value of the coefficients \( \beta \), which is unknown
- assumes that the data \( D \) are the result of a sampling experiment which is infinitely repeatable
- uses confidence intervals: a 95% confidence interval is an interval around the point estimate that has a 95% probability of including the true value, if the experiment was repeated a large number of times

Bayesian inference:

- doesn’t necessarily assume that the coefficients \( \beta \) have a unique true value
• assumes instead that $\beta$ is a vector of random variables, but we can use the data at hand to update our knowledge on $\beta$

• uses credible intervals: a 95% credible interval is an interval (usually around the mode if the posterior is unimodal) that contains 95% of the posterior probability

The Bayesian and frequentist approach to linear regression problems is strongly related:

• MLE is equivalent to finding the mode of the posterior using improper flat priors on $\beta$

• The penalty in penalized MLE can be interpreted as the log of priors on $\beta$:
  – the ridge penalty is equivalent to a normal prior on $\beta$ excluding the intercept $\beta_0$
  – the lasso penalty is equivalent to a laplace prior on $\beta$ excluding the intercept $\beta_0$
  – the magnitude of the penalty $\lambda$ is a function of the variance of the prior distribution

4.2 Our approach in this study

Bayesian and frequentist statistics each have advantages and drawbacks, some of which are listed below (non-exhaustively).

Frequentist estimation:

• +: yields point estimates that are easy to use and to make predictions

• +: is fast to calculate, hence it is suitable for large data sets

• +: is easy to set-up, because many aspects don’t need to be specified such as prior distributions, initial values for numerical approximation. Most frequentist models are standardized procedures usable out-of-the-box

• -: produces point estimates and a variance-covariance matrix relying on an asymptotic normal distribution for the estimators

Bayesian estimation

• +: is ideal for getting insight in the full coefficient distribution

• +: offers more intuitive inference, hypothesis testing and intervals

• +: allows for the setup of advanced models, e.g. hierarchical models

• +: with proper priors is immune to singularities and near-singularities with matrix inversions, unlike frequentist inference.

• -: is more time-consuming to calculate especially on large sample sizes, as it requires MC sampling
Chapter 4. Bayesian approach

In this study the approach we have chosen is to use the best of both worlds. We will use the Bayesian approach of informative priors to impose penalization and avoid (near-)singularities. We are looking for methods with good default settings so that the procedures can get standardized and used out-of-the-box. We will calculate point estimates, i.e. the mode of the posterior distribution. Searching the mode of the posterior is much faster than the MCMC sampling which is required to get a full picture of the posterior distribution. Point estimates are convenient as the purpose of this study is to evaluate predictive performance of various methods.

4.3 Stan

MCMC sampling from the posterior distribution is usually done using a Gibbs sampler and the Metropolis algorithm. Configuring the jumping rules for the Metropolis algorithm is usually a challenge in itself, as endless possibilities exist for fixed or adaptive jumping rules. Sometimes the Metropolis algorithm can be avoided altogether if a random sample can be generated directly from a conditional distribution, e.g. using inversion sampling or slice sampling. But an inherent inefficiency of the Gibbs sampler remains which is revealed as random walk behaviour: the MCMC chain take a long time slowly exploring the range of the target distribution. The chains suffer from bad mixing. The cause of this behaviour is strong correlation in the posterior distribution.

Hamiltonian Monte Carlo (HMC) borrows an idea from physics to suppress the random walk behavior, thus allowing it to move much more rapidly through the target distribution. In addition to the posterior density (which, as usual, needs to be computed only up to a multiplicative constant), HMC also requires the gradient of the log-posterior density. The package Stan (Sampling Through Adaptive Neighbourhouds) implements HMC. It comes with a probabilistic programming language similar to the ones used by Bugs or Jags. Whereas the latter use interpreted languages, Stan transforms the model into C++ code which is then compiled into an efficient executable program. The gradient of the log-posterior, consisting of its partial derivatives, is calculated using algorithmic differentiation. In this study we have chosen to use Stan for the implementation of GLM binomial regression models with different priors and link functions, after having experimented with several implementations of MCMC with Gibbs sampling. Our experience is that Stan converges faster than basic Markov chain simulation and that it avoids the dreaded random walk behaviour. In addition to HMC, Stan offers optimization which can be used to calculate point estimates such as posterior modes and (penalized) MLE. Stan implements BGFS optimization, a Newton-like (see 3.2) method which relies on derivatives and second-derivatives. The possibility of using the same models for sampling as for optimization proves to be very convenient for our study. The approach we have chosen is to search for point estimates to make predictions, but in case of slow or difficult convergence, it helps to get an insight into the full distribution obtained by HMC sampling.
Chapter 5

Rare Events

5.1 Problem setting

5.1.1 Definition

According to King & Zeng (2001), rare events data consists of binary dependent variables with dozens to thousands of times fewer ones (events) than zeros (non-events). It is of course the rare events that are most interesting and that we would like to be able to predict with great precision. Some examples from different fields:

- finance: fraudulent transaction, borrower bankruptcy
- political science: war between states
- engineering: component failure
- marketing: response to a mass-mailing campaign, churn
- medicine: diagnosis of rare diseases

We can make the distinction between two types of rareness:

1. relative rareness, also called unbalanced or imbalanced data: a data set is said to be imbalanced when one class, i.e. the minority class or class of interest is much smaller than the majority class. Classification algorithms in general suffer when data is skewed towards one class.

2. absolute rareness, is essentially a small sample problem. Frequentist inference such as MLE breaks down when sample sizes get too small. Therefore Allison (2012) suggest that only absolute rareness matters, in the context of logistic regression. Agresti (2013) suggests to look at relative rareness compared to the number of predictors.


## 5.1.2 Bias

ML estimators have very desirable asymptotic properties, but are known to be biased in small samples. Sometimes exact inference using permutation methods is proposed instead of MLE. It is computationally expensive and therefore not ideal for the problem of rare events data, where the data set is not necessarily small, only the number of events is small compared to the number of predictors.

Bias reduction in MLE is proposed by King & Zeng (2001). Bias reduction in MLE was also extensively studied by Firth (1993) and Heinze & Schemper (2002). King & Zeng claim that logistic regression sharply underestimates the probability of rare events, i.e. the bias on the intercept is negative, so that the estimated intercept $\hat{\beta}_0$ is too small and as a result $P(y=1)$ is underestimated. In this study we have evaluated that claim on simulated data. It proves that the claim holds on data generated from a linear discriminant model, as King & Zeng have done. On data generated from a logistic model however, we find the opposite to be true: the bias on the estimated intercept $\hat{\beta}_0$ proves to be positive, as shown in figure 5.1.

![Histograms showing estimated intercepts and slopes for balanced and unbalanced data sets.](image)

**Figure 5.1:** Data sets with $n=200$ samples were generated from logistic regression model $\text{logit}(P(y_i = 1)) = x_i$, where the $x_i$ are drawn from a mixture of two standard normal distributions, with locations at $-3$ and $3$. In the unbalanced data only 2% of the $x_i$ were drawn from the normal centered around 3. The balanced data has 50% of the $x_i$ drawn from a normal around 3. A logistic regression model was fitted and coefficient estimates calculated. That procedures was repeated 10,000 times. The resulting distribution of the coefficients shows that the largest bias is on the intercept of the unbalanced data set.
5.1.3 Infinite estimates

The problem of infinite ML estimates for the binomial regression coefficients can be due to multicollinearity i.e. (near-)singularity of the design matrix. That problem can be tackled using penalization, e.g. ridge or lasso, just as for ordinary linear regression.

In addition discrete data regression can suffer from a different cause of infinite MLE estimates: complete separation. Complete separation occurs when a linear combination of the predictors is perfectly predictive of the outcome. In that case the MLE doesn’t exist. Quasi-complete separation occurs when data are close to perfectly separated. It leads to very large values for the coefficients, as illustrated in figure 5.2. It is essentially an occurrence of overfitting.

![Figure 5.2: Quasi-complete separated data and the logistic regression response curve. The slope and the intercept estimates have large absolute values.](image)

Separation is surprisingly common in applied binomial regression, especially with binary predictors. Obviously this problem can expected to be even more common in unbalanced data or if absolute rareness occurs.

5.2 Remedies for rare events data problems

5.2.1 Sampling methods

The basic sampling methods include under-sampling and over-sampling. Under-sampling eliminates majority-class observations while over-sampling duplicates minority-class observations. Both of these sampling techniques decrease the overall level of class imbalance, but they have several drawbacks, according to Weiss (2004). Under-sampling discards potentially useful majority-class examples and thus can degrade classifier performance. It is clear that any under-sampling scheme leads to a loss of information, even if the information content of
Chapter 5. Rare Events

the majority class observations is much lower than that of the majority class. Undersampling
is only useful for cutting down on processing time. Statistically there is no good reason to do
it, therefore we have discarded this method in our study.

Further according to Weiss (2004), over-sampling increases the sample size and thus the time
required for estimation. Worse yet, over-sampling involves making exact copies of examples,
so it may lead to overfitting. That claims is surely true for certain machine-learning algo-
rithms such as decision trees. But it is not valid for the logistic regression. Oversampling
only shifts the intercept, it doesn’t affect the slopes. That is easy to understand for logistic
regression applied to two-way contingency tables, where the $\beta_1$ represents the log-odds ratio:
multiplying the counts in one column, doesn’t affect this ratio. That property appears to
hold for continuous predictors too. Therefore we have not used the method of oversampling
in this study.

Owen (2007) has studied infinitely imbalanced logistic regression and he explains that imbal-
anced data are not necessarily a problem for logistic regression. He proves that the estimate
of the intercept: $\hat{\alpha} \to -\infty$ like $-\log(N)$, where $N$ is the number of $\{y=0\}$. But under mild
conditions, the logistic regression slopes vector $\hat{\beta}$ converges as $N \to \infty$. It can be found as
the solution of equation:

$$\bar{x} = \frac{\int e^{x^\prime \beta} dF_0(x)}{\int e^{x^\prime \beta} dF_0(x)}$$  \hspace{1cm} (5.1)

where $F_0$ is the distribution of $X$ given $y = 0$ and $\bar{x}$ is the average of the sample $x_i$ values
for which $y = 1$. The limiting solution is the exponential tilt required to bring the population
mean of $X$ given $y = 0$ onto the sample mean of $X$ given $y = 1$.

That results in a special case if $F_0$ is the normal distribution:

if \quad $F_0 = N(\mu, \Sigma)$ \quad then $\lim_{N \to \infty} \beta(N) = \Sigma^{-1}(\bar{x} - \mu)$ \hspace{1cm} (5.2)

Equation 5.1 means that, in the infinitely imbalanced limit, logistic regression only uses the
$\{y_i = 1\}$ data points through their average feature vector $\bar{x}$. That finding leads to interesting
insights. Logistic regression only has as many parameters as covariates plus the intercept, so
it is clear that it cannot be as flexible as some other machine learning methods. But knowing
that those parameters are strongly tied to $\bar{x}$, offers insight into how logistic regression works
on imbalanced problems. It is reasonable to expect better results from logistic regression
when the $x_i$ where $y_i = 1$ are in a single tight cluster near $\bar{x}$ than when there are outliers, or
when the $x_i$ where $y_i = 1$ are in several well separated clusters in different directions. Owen
thus suggests to cluster $x_i$ where $y_i = 1$ and then fit a logistic regression per cluster. That
is a highly interesting idea, but we have not included it in our scope. It could be a topic for
further research.

5.2.2 Penalization or priors on $\beta$

Penalization which is equivalent to using proper priors on $\beta$ solves the problem of infinite
coefficient estimates. As said in subsection 5.1.3 penalty methods such as ridge or lasso
work equally well on binomial regression. In addition they solve the issue of (quasi-)complete
separation too. Some priors prevent the small sample bias which is typical for MLE. The subject of priors is elaborated in the next chapter.

5.2.3 Skewed link function

The logit and probit CDF are symmetric link functions: they approach 0 at the same rate as they approach 1. Several authors suggest that a skewed link function should be able to fit better to unbalanced data, as shown in figure 5.3.

![Figure 5.3](image.png)

**Figure 5.3:** The full line shows the logistic regression response curve, the dashed line the response line of the binomial regression with the complementary log-log link, a skewed link-function. The skewed response curve approaches 1 faster than 0 and happens to have a better fit, as measured by the log score, on this unbalanced binary data set.

A lot of research has been done on alternative link functions with variable skewness, which is ideally learned from the data itself. We elaborate on the subject of skewed link functions in chapter 7.
Chapter 6

Priors

6.1 Firth’s method

Firth (1993) explains that the bias in MLE is due to the combination of (i) unbiasedness in the score function at the true value of the parameter vector and (ii) the curvature of the score function. Firth proposes a general method to reduce bias in MLE by introducing a small bias in the score function. His method is bias-preventive, not bias-corrective. It removes the bias of order $n^{-1}$, so the bias is reduced to order $n^{-2}$. He shows that this method is equivalent to the use of a Jeffreys’ invariant prior, if the target parameter is the canonical parameter of an exponential family as is the case with GLM and most link functions such as logit, probit and cloglog. When the model matrix is of full rank, the penalty is strictly concave. Maximizing the penalized likelihood yields a maximum penalized likelihood that always exists and is unique. Heinze & Schemper (2002) show that Firth’s method is an ideal solution to the issue of separation.

King & Zeng (2001) propose an alternative estimation method to reduce the bias, available in package relogit. Their method is very similar to Firth’s method, which is more widely available.

Allison (2012) sees no drawbacks to using the Firth correction, and he even suggests that a case could be made to always use penalized likelihood with the Firth penalty rather than conventional MLE. Georg Heinze has extensively compared the Firth method with ordinary MLE on small samples, and he has found that the Firth method was consistently superior: point estimates are more precise, confidence intervals are more reliable in terms of coverage probabilities.

Several packages implementing Firth’s method exist in R: logist, brlr, brglm.

6.2 Gelman’s informative prior

Gelman et al. propose a weakly informative default prior distribution for logistic and other regression models.
Gelman proposes a true prior on the coefficients $\beta$ in the Bayesian sense. The proposed prior is informative. Gelman relies on the following assumption or rather general knowledge about coefficients in the logistic regression: a typical change in an input variable is unlikely to correspond to a change as large as 5 on the logistic scale. Such a change would move the probability from 0.01 to 0.50 or from 0.50 to 0.99.

6.2.1 Centering and scaling

Applying this prior knowledge requires to get the input variable’s scale right. For example, an input representing age expressed in years needs a different prior if age were expressed in months. So before a default prior can be applied, inputs must be properly scaled, except binary covariates which have their own natural scale, i.e. a change of 1. Centering is required too because otherwise it’s not possible to put a prior on the intercept. Centering and scaling is done as follows:

- binary inputs are centered and scaled to have a mean of 0 and range 1
- other (discrete as well as continuous) inputs are shifted to have a mean of 0 and scaled to have a standard deviation of 0.5. This scaling puts the variables on the same scale as symmetric binary input which, taking on values $\pm\frac{1}{2}$, have standard deviation 0.5.

Gelman makes the distinction between inputs and predictors. For example, in a regression on age, sex and their interactions, there are four predictors (the constant term, age, sex and age $\times$ sex), but just two inputs, age and sex. It is the input variables, not the predictors that are standardized. An appealing byproduct of applying the model to rescaled inputs, is that it automatically implies more stringent restrictions on interactions. For example, consider three symmetric binary inputs $x_1, x_2, x_3$. From the rescaling, each will take on values $\pm\frac{1}{2}$. Then any two-way interaction will take on the values $\pm\frac{1}{4}$ and the three-way interactions $\pm\frac{1}{8}$. But all their coefficients will get the same default prior distribution, so the total contribution of a three-way interaction is $\frac{1}{4}$ that of the main effect.

6.2.2 A weakly informative t-family of prior distributions

For each coefficient a student-t prior distribution is foreseen with mean 0, degrees-of-freedom parameter $\nu$, and scale $s$, with $\nu$ and $s$ chosen to provide the minimal prior information required to constrain the coefficients to lie in a reasonable range. Gelman prefers the t-family because flat-tailed distributions allow for robust inference and it allows easy and stable computation in logit or probit regression. Computation with a normal prior distribution is even easier but the flexibility of the student-t family is preferred. In a cross-validation experiment on a large number of data sets, Gelman has found that the Cauchy is a conservative choice that gives consistently good results, although fine-tuning of $\nu$ can occasionally result in better predictive performance on some data sets.

The default choice is scale $s = 2.5$ for all of the coefficients of the logistic regression except the intercept. The student-t distribution with scale 2.5 favours values below 5 in absolute
value. The fat tails of the Cauchy used in the default model allow the occasional probability of much larger values. Combined with the standardization in subsection 6.2.1 it implies that the absolute difference in logit probability should be less than 5 when moving from one standard deviation below the mean, to one standard deviation above the mean, in any input variable.

For the intercept of the logistic regression a weaker default prior distribution is applied: a Cauchy with center 0 and scale 10, which implies that the success probability $P(y_i = 1)$ for an average case is expected to be between $10^{-9}$ and $1 - 10^{-9}$.

For the probit regression, the default scales should be divided with a factor 1.6, as the normal cdf is 1.6 times steeper at the intersection with $P(y = 0.5)$ than the logit. (In fact in R function `bayesglm` there is an error in the implementation which uses a default prior scale that is 1.6 times larger for the probit regression.)

### 6.2.3 Implementation

Gelman offers his *weakly informative default prior distribution* for logistic and other regression models in function `bayesglm` of package `arm`, an extension of the `glm` function in R.

The function calculates point estimates for the coefficients as the mode of the posterior using an iteratively reweighted least squares method.

Gelman explains that modifying the logistic regression estimation algorithm to include a normal prior on the coefficients is simple: the prior information can effortlessly be included in the MLE algorithm by adding a pseudo-data point for each coefficient.

Including a student-t distributed prior can be done by expressing the t-distribution as a normal with unknown variance $\sigma^2_j$, which is itself inverse chi-squared distributed, so $\beta_j \sim \text{studentT}(\nu_j, \mu_j, s_j)$ is equivalent to:

$$
\beta_j \sim N(\mu_j, \sigma^2_j), \quad \sigma^2_j \sim \text{Inv}\chi^2(\nu_j, s_j^2)
$$

The $\sigma_j$ are treated as missing parameters in an EM algorithm. The algorithm uses iterations of the iteratively reweighted least squares method to estimate $\beta$ followed by an EM step to estimate $\sigma_j$, until convergence.

Function `bayesglm` offers all of the `glm` functionality, augmented with the possibility to overrule the default values for the student-t prior properties: the prior mean, scale and degrees of freedom.
Chapter 7

Link functions

7.1 Quantile link functions

Binary Quantile regression is of interest for our study of GLM models for rare events data because it uses skewed links, which are said to offer a better fit to imbalanced data, see subsection 5.2.3. Furthermore Kordas (2006) suggests that choosing higher quantiles is appropriate in imbalanced data where \( y = 1 \) events are rare. In such data the information is mainly contained in the ones, so choosing a high quantile places more weight on the rare \( y=1 \) event, and centres estimation around the informative part of the sample. So in the next subsections we look first at continuous quantile regression, before going deeper into binary quantile regression. Later we re-use the quantile idea to construct skewed link function for other distributions than the asymmetric Laplace distribution.

7.1.1 Quantile regression

The basic regression model estimates the mean conditional outcome as a linear combination of the covariates: \( E[y_i|X_i] = X_i'\beta \). That is done with the OLS method, which is equivalent to ML estimation assuming normally distributed errors. The estimator is unbiased if the model assumptions are satisfied, one of which is homoskedasticity of the error distribution. As seen in chapter 1, this model is sensitive to outliers. LAR regression or median regression minimizes the \( L_1 \) norm of the residuals, see (2.6). Median regression is less sensitive to outliers and it proves to be resistant to heteroskedasticity too. The ML equivalent of median regression is a linear regression model assuming i.i.d. Laplace distributed errors.

As explained by Benoit & Van den Poel (2009) and Koenker & Hallock (2001) quantile regression extends the median regression to other conditional quantiles. Knowing that the \( \tau^{th} \) sample quantile of a sample data set \( x_i \), with \( i = 1, ..., n \) is calculated as the minimizer:

\[
\xi(\tau) = \arg\min_{\xi} \sum_{i=1}^{n} \rho_{\tau}(x_i - \xi)
\]

(7.1)

where \( \rho_{\tau}(z) = z(\tau - I(z < 0)) \) is the quantile loss function, estimates of the quantile regression

23
coefficients $\beta(\tau)$ are found as the solution of:

$$\hat{\beta}(\tau) = \argmin_{\beta} \sum_{i=1}^{n} \rho_{\tau}(y_i - X_i'\beta) \quad (7.2)$$

Yu & Moyeed (2001) show that the minimization problem in (7.2) is equivalent to ML estimation of a linear regression model with i.i.d. asymmetric Laplace distributed errors. Benoit & Van den Poel (2009) use therefore the three-parameter asymmetric Laplace distribution (ALD) introduced by Yu & Zhang, which has the following pdf:

$$f(x|\mu, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma} \exp\left(-\frac{(x-\mu)}{\sigma} [\tau - I(x \leq \mu)]\right) \quad (7.3)$$

and cdf:

$$F(x|\mu, \sigma, \tau) = \begin{cases} \tau \exp\left(\frac{1-\tau}{\sigma}(x-\mu)\right), & \text{if } x \leq \mu \\ 1 - (1-\tau) \exp\left(-\frac{\tau}{\sigma}(x-\mu)\right), & \text{if } x > \mu \end{cases} \quad (7.4)$$

which are displayed in figure 7.1.

Figure 7.1: pdf and cdf of the asymmetric Laplace distribution with $\mu = 0$ and $\sigma = 1$

We see that the ALD pdf consists of a left tail that is exponentially distributed with scale $\sigma/\tau$ and an exponentially distributed right tail with scale $\sigma/(1-\tau)$. The left tail contains a fraction $\tau$ of the probability mass, so that the cdf has value $\tau$ at $x = \mu$.

Yu & Zhang explain that the ALD has a direct link to the estimation of quantiles and quantile regression.

Several implementations of quantile regression are available in R. Koenker offers the function $r\eta$ in package quantreg to estimate the quantile regression model with frequentist inference.
Benoit has taken a Bayesian approach with function \textit{bayesQR} in a R-package with the same name. The estimation of a linear regression model at different quantiles offers a much richer insight in the relation between predictors and outcome than the standard regression model, which produces just one coefficient vector estimate $\hat{\beta}$ and otherwise relies on the assumption that the variance of the error is constant over the entire range of the outcome variable. Quantile regression is illustrated with an example on simulated data.

$N = 200$ data points are generated from a heteroskedastic linear regression model:

\[ u_i = 1.5x_i + \varepsilon_i \quad \text{with} \ x_i \sim \text{Unif}(-5,5) \quad \text{and} \ \varepsilon_i \sim N(0,2(x_i+5)) \quad (7.5) \]

As seen in figure 7.2 the OLS and median regression lines are quite close, although the OLS overestimates the true value of the slope because of the violation of the assumption of homoskedasticity. Doing the regression at multiple quantiles $\tau \in (0.05,0.25,0.5,0.75,0.95)$ reveals the heteroskedasticity, especially in this case where the variance of the errors is linearly dependent on the predictor.

![Figure 7.2: Data generated from a linear regression model with heteroskedastic normal errors. Quantile regression detects the heteroskedasticity and offers a rich insight in the relation between predictor and outcome. Quantile regression coefficients were estimated with R-package \textit{bayesQR}\(](image)

### 7.1.2 Binary quantile regression

Manski (1975) introduced the Maximum Score Indicator, which imposes very weak assumptions on the distribution of $\varepsilon_i$ in the binary choice model of (3.2): that the $\tau^{th}$ quantile of
\( \varepsilon_i \) conditional on \( X_i \) is zero. That means that the conditional quantile of the latent variable (utility) \( u_i \) in a binary choice model can be written as:

\[
Q_{\tau}(u_i|X_i) = X_i' \beta_{\tau}
\]

That expression can be transformed into a function of the observed outcome variable \( y_i \) by applying the monotone-increasing function \( I(x \leq 0) \), so that the conditional quantile function of the observed outcome variable becomes:

\[
Q_{\tau}(y_i|X_i) = I(X_i' \beta_{\tau} \leq 0)
\]

Benoit & Van den Poel (2012) propose a Bayesian approach to binary quantile regression, assuming the ALD distribution for \( \varepsilon_i \), which obviously satisfies Manski’s assumption, and diffuse normal priors on the coefficients. The implementation uses Gibbs sampling, alternatingly sampling from the fully conditional latent-variable distribution and from the fully conditional posterior distribution of the coefficients vector \( \beta \). Benoit & Van den Poel provide an efficient implementation that draws directly from both conditional distributions, requiring no Metropolis step. It is available as function \textit{BayesQR} in the R-package with the same name.

Care must be taken when interpreting the coefficient estimates, as illustrated by an example that builds on the simulated data specified in (7.5). The binary variable \( y_i \) is observed by censoring the latent variable \( u_i \) as usual in the binary choice model of (3.2), so: \( y_i = I(u_i > 0) \). Fitting the binary quantile model results in estimates conveniently visualized in the quantile plot in figure 7.3.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{quantile_plot.png}
\caption{Quantile plot of the intercept (beta1) and slope (beta2) of a binary quantile regression, estimated with R-package \textit{bayesQR}}
\end{figure}

The quantile plot reveals the following:
Chapter 7. Link functions

- The intercept increases with the quantile $\tau$ giving the intercept quantile plot a typical tangent-like curve, approaching $-\infty$ as $\tau \to 0$ and approaching $\infty$ as $\tau \to 1$.

- The slope as a function of the quantile $\tau$ has the typical U-form curve, approaching $\infty$ at both extremes of the quantile range.

The reason is easy to understand when considering the inverse link function i.e. the response curve of the binary choice model with ALD distributed $\varepsilon_i$. Following (3.3) we see that the inverse link function equals the ALD cdf with $\mu = 0$ and $\sigma = 1$ mirrored around $x = 0$ and $y = 0.5$, resulting in sigmoid curves as in figure 7.4. The binary quantile regression sigmoid curves increase slower for quantiles $\tau$ further away from the median, leading to higher slope coefficients and the typical U-form in the quantile plot. The lower quantile regression sigmoids are shifted to the left, and the higher quantiles to the right, which explains the large negative intercept for lower quantile and the large positive intercept for higher quantiles. The response curves fitted to the simulated data are shown in figure 7.5.

The absolute magnitude of the slope coefficients is thus not relevant, which is sometimes overlooked for example in Miguelis et al. 2012. Neither is the binary quantile regression able to retrieve the original heteroskedasticity of the latent variable, as this information is completely lost by censoring. So we cannot expect that the heteroskedasticity in the latent data is revealed as binary quantile regression slopes increasing with the quantile, as is the
case for continuous quantile regression with heteroskedasticity in the observed data, as seen in figure 7.2. Kordas normalizes the slope coefficients with a division by $||\beta||$, the Euclidean norm of the slope vector. In a binomial regression model the slope vector, i.e. the coefficients vector $\beta$ excluding the intercept $\beta_0$, can be interpreted as the direction in $(p-1)$-dimensional space of increasing $P(y=1)$. By normalizing the slope vectors they are shrunken to fall on an $(p-1)$-dimensional sphere, allowing for easy comparison of the relative weights of the coefficients. A slope vector of the lower quantiles corresponds to an error distribution in the binary choice model with a small left tail and large right tail, so the multidimensional plane will fit the data with small tolerance of negative errors and large tolerance of positive errors. Hence the estimation of the quantile regression model leads to different normalized slope vectors for different quantiles $\tau$. Occasionally it can even lead to sign reversal, as Kordas (2006) has found. The cause is interaction between the predictors, which is clarified with an example. We generate a sample of $N = 1000$ of logistically distributed data from the model with a non-linear interaction term:

$$\text{logit}(P(y=1)) = 4x_1 + x_2 - 4\text{sign}(x_1)x_1\sqrt{|x_2|}$$

(7.8)

with $x_1 \sim \text{Unif}(-1,1)$ and $x_2 \sim \text{Unif}(-1,1)$. The logodds of $P(y = 1)$ is a non-linear function of the predictors, resulting in the logodds surface visualized in figure 7.6. In a logistic regression model leaving the interaction term out, the curved surface is approximated by a plane. We have fitted the binary quantile regression model $g_{\tau}(P(y=1)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, where $g_{\tau}()$ is the quantile link function with $\tau \in (0.05, 0.25, 0.5, 0.75, 0.95)$. The lowest quantile regression gives a large weight to negative outliers while it is relatively
Chapter 7. Link functions

tolerant of positive outliers, so it finds a coefficients vector with a negative component \( \beta_2 \), as is seen in the quantile plot of the normalized coefficients in figure 7.7. Hence the quantile plots can be used to detect if interaction terms should be added. A normalized coefficient that decreases with \( \tau \) while another one decreases implies an interaction. Adding the interaction term \( x_1x_2 \) to the model removes most of this effect and leads to almost flat quantile plots, even if the interaction term differs from the true interaction.

7.1.3 Practical use

The skewed link functions of binary quantile regression are attractive for our application on rare events data. Indeed Kordas (2006) suggests that choosing higher quantiles is appropriate in imbalanced data where \( y = 1 \) events are rare. A natural choice for the ideal quantile would be \( \tau = 1 - \bar{y} \), where \( \bar{y} \) is the overall sample probability \( P(y = 1) \). If the binomial linear model is appropriate, i.e. if there is a direction in \((p-1)\)-dimensional space where \( P(y = 1) \) increases, we may expect an average data point to have \( P(y = 1) = \bar{y} \). So, using the quantile link with \( \tau = 1 - \bar{y} \) on a model with centered predictors \( X_i \), the intercept could be zero, so that 
\[
P(y = 1) = 1 - F_\tau(0) = 1 - \tau = \bar{y}.
\]

However, it is not sure that this choice is always optimal. Say we rebalance a rare-events data set by over-sampling the ones. That doesn’t change the information content, so using a lower \( \tau \) due to higher \( \bar{y} \) is not justified. Therefore we prefer to make \( \tau \) a parameter to be estimated along with coefficient vector \( \beta \).
Chapter 7. Link functions

We also need to apply stronger priors to make binary quantile regression stable on real data. We prefer to apply the weakly informative default prior of chapter 6.2, which Gelman has applied with great success to the logistic regression. But several adjustments are required:

- As seen in figure 7.4, the slope of the response curve at $P(y=1) = 0.5$ is much lower for the quantiles further away from the median. For the median, the rate of increase of the response curve at $P(y=1) = 0.5$ is the density $f(x=0, \tau = 0.5) = 0.25$, where $f(x, \tau)$ is the ALD pdf with location 0 and scale 1. That is the same value as for the logistic regression, so for binary median regression we can use the same student-t prior with default scale $s = 2.5$ as proposed in chapter 6.2. For quantiles further away from the median, we need a larger default scale for the priors to allow for larger coefficients. The correction factor to be applied to the default scale is $0.25 / f(Q(0.5, \tau), \tau)$, where $Q(x, \tau)$ is the ALD quantile function with location 0 and scale 1. After some algebra, we find that the correction to be applied is $0.5/\tau$ for $\tau < 0.5$ and $0.5/(1 - \tau)$ for $\tau \geq 0.5$. The elaboration is given in appendix A.1.

- As seen in figure 7.4, the response curves have their intersection with the horizontal line at $P(y_i=1) = 0.5$ at increasingly large $|X'\beta|$ for quantiles $\tau$ further away from the median. That is corrected by using a student-t prior on the intercept with a location different from zero. We find that the location for the prior on the intercept should
Chapter 7. Link functions

be \(-Q(x = 0.5, \tau)\), or if we also want to take the imbalance of the data into account
\(-Q(x = 1 - \bar{y}, \tau)\). The elaboration is given in appendix A.2.

We get additional insight in the effect of these corrections, by translating the response curve
with the proposed location shift of the prior on the intercept, and by scaling the response
curve with the correction factor proposed on the prior scale on the slope coefficients. The
result is displayed in figure 7.8. Due to the nature of the exponential distribution, and the way
it was used to build the three-parameter ALD, scaling and translation results in a remarkable
response curve. The left tail of the sigmoid is the same as that of the median for the lower
quantiles and the right tail of the sigmoid is the same as that of the median for the upper
quantiles. The different quantiles only affect the rate at which the response curve approaches
0 or 1 on one side.

![1-F(-x) translated](image1)

![1-F(-x) translated and scaled](image2)

**Figure 7.8:** scaled and translated ALD response curves. After scaling and translation, the left tail
of the sigmoid is identical for the lower quantiles and the right tail of the sigmoid is
identical for the upper quantiles.

In our study we have found that the proposed corrections are not just required to make the
weakly informative prior work. They are also essential to avoid excessive correlation between
\( \tau \) and the coefficient vector \( \beta \). It is important to avoid such correlation when doing MCMC sampling from the posterior. But we find that it also makes the mode-finding algorithm more stable.

Finally we also need a prior for \( \tau \), which must be constrained to range \((0,1)\). The beta distribution seems natural for that purpose. We choose the meta-parameters \( \alpha \) and \( \beta \) such that:

- one of the \( \alpha \) and \( \beta \) is equal to 2 and the other one large. We don’t want \( \alpha \) or \( \beta \) lower than 2, as zero prior probability is required at the extremes \( \tau = 0 \) and \( \tau = 1 \)
- the prior distribution mean \( \alpha/(\alpha + \beta) \) is equal to \( 1 - \bar{y} \), which is the natural choice for the quantile
- however we never use values for \( \alpha \) or \( \beta \) large than 50, no matter how imbalanced the data set is, to avoid an overly peaked prior distribution

Some examples of beta priors are displayed in figure 7.9

![Beta distribution for some values of \( \alpha \) and \( \beta \)](image)

**Figure 7.9:** Beta distribution for some values of \( \alpha \) and \( \beta \)

We have implemented two Bayesian models in Stan, using priors as described above and a likelihood based on a binary choice model with ALD distributed noise:

- **bqr**, which uses fixed \( \tau = 1 - \bar{y} \)
- **bqra**, which uses \( \tau \) as model parameter
7.1.4 Extension

The way that the ALD pdf (7.3) and the ALD cdf (7.4) have been constructed from the symmetric Laplace distribution, can be generalized to define a new class of asymmetric distributions based on any symmetric probability distribution from the location scale family. We haven’t found any paper or reference mentioning the construction of asymmetric families based on symmetric location scale distributions, so we add our own definition of the asymmetric distribution hereunder. The cdf of the asymmetric family is:

\[
F_{\text{asymmetric}}(x|\mu, \sigma, \tau) = \begin{cases} 
2\tau F(x|\mu, \frac{\sigma}{\tau(1-\tau)}), & \text{if } x \leq \mu \\
1 - 2(1-\tau)F(-x|\mu, \frac{\sigma}{\tau}), & \text{if } x > \mu 
\end{cases} 
\]  

(7.9)

Where \(F(x|\mu, \sigma)\) is a symmetric cumulative distribution function from the location scale family. The resulting class of asymmetric distributions \(F_{\text{asymmetric}}(x|\mu, \sigma, \tau)\) has the properties that it includes the symmetric distribution at \(\tau = 0\), and \(F_{\text{asymmetric}}(\mu|\mu, \sigma, \tau) = \tau\) or said differently: precisely \(\tau\) of the probability mass is in the left tail, and \((1-\tau)\) in the right tail.

The pdf of the asymmetric distribution is consequently:

\[
f_{\text{asymmetric}}(x|\mu, \sigma, \tau) = \begin{cases} 
\frac{4(1-\tau)}{\sigma} f(x|\mu, \frac{\sigma}{\tau(1-\tau)}), & \text{if } x \leq \mu \\
\frac{4(1-\tau)}{\sigma} f(x|\mu, \frac{\sigma}{\tau}), & \text{if } x > \mu 
\end{cases} 
\]  

(7.10)

Additionally Manski’s maximum score indicator doesn’t require that the ALD is used for quantile regression. On the contrary it imposes only the very weak assumption of (7.6). That motivates us to define a new three-parameter asymmetric logistic regression distribution, with pdf:

\[
f(x|\mu, \sigma, \tau) = \begin{cases} 
\frac{4\tau(1-\tau)}{\sigma} \frac{\exp[-2(1-\tau)(x-\mu)/\sigma]}{1+\exp[-2(1-\tau)(x-\mu)/\sigma]^2}, & \text{if } x \leq \mu \\
\frac{4\tau(1-\tau)}{\sigma} \frac{\exp[2\tau(x-\mu)/\sigma]}{1+\exp[2\tau(x-\mu)/\sigma]^2}, & \text{if } x > \mu 
\end{cases} 
\]  

(7.11)

and cdf:

\[
F(x|\mu, \sigma, \tau) = \begin{cases} 
\frac{2\tau}{1+\exp[-2(1-\tau)(x-\mu)/\sigma]}, & \text{if } x \leq \mu \\
1 - \frac{2\tau}{1+\exp[2\tau(x-\mu)/\sigma]}, & \text{if } x > \mu 
\end{cases} 
\]  

(7.12)

which are shown in figure 7.10

We expected that using this alternative distribution function for binary quantile regression would lead to easier optimization. As explained in chapter 3.2, optimization algorithms usually require the first and second-order derivative of the log likelihood. A discontinuous first-order derivative is a serious problem, which typically leads to convergence problems in Newton-like algorithms. The continuous quantile regression suffers from this problem. That is easy to understand when looking at the ADL pdf in figure 7.1. The Newton algorithm searches for the mode of the curve, as the point where the first-order derivative becomes zero. That doesn’t happen for the ALD pdf. Its first-order derivative is discontinuous in the mode. Additionally the ALD pdf is not logconcave, but piecewise linear, resulting in a multimodal
**Figure 7.10:** pdf and cdf of the asymmetric logistic distribution with \( \mu = 0 \) and \( \sigma = 1 \)

log likelihood for the continuous quantile regression. Benoit & Van den Poel’s choice for a Bayesian approach is thus completely justified.

In the binary quantile regression the first-order derivative of the log likelihood is continuous. The second-order derivative however isn’t, but that proves to be no problem. Contrary to continuous quantile regression, binary quantile regression is well-behaved, as the log likelihood is unimodal or equivalently, the loss function is convex, as can be seen in figure [7.11](#). Doing a calculation similar as in section [3.3](#), we find that outliers at \( y = 1 \) contribute \( \tau |X_i'\beta| \) to the loss, while outliers at \( y = 0 \) contribute \( (1 - \tau) |X_i'\beta| \). So also in the approximative contribution to the loss, we see the benefit of an asymmetric link function for fitting to imbalanced data. In a rare events data set the zeroes are abundant and the ones are rare. So we can expect many outliers at \( y = 0 \), which shouldn’t contribute too much to the overall loss. Outliers at \( y = 1 \) however should contribute a lot, so that a good fit is obtained to the rare class. So a rare events data set will benefit from \( \tau \) close to 1. This intuition corresponds with Kordas’ statements and with our findings when estimating the optimal \( \tau \) from the data.

Comparing the median binomial regression to logistic regression, we find that outliers contribute approximately \( \frac{1}{2} |X_i'\beta| \) and \( |X_i'\beta| \) to the loss respectively. So in the limit both loss functions rely on the \( L_1 \) norm of the link value. The order of the norm is the same, but median binomial regression is twice as tolerant of outliers as the logistic regression.

We want to apply the weakly informative default prior of chapter [6.2](#) again, so we must apply adjustments similar to those for the ADL quantile regression:

- the correction factor to be applied to the default scale of the student-t prior on the
coefficients except the intercept is \( 0.25/f(Q(0.5, \tau), \tau) \), where \( f(x, \tau) \) is the asymmetric logistic pdf and \( Q(x, \tau) \) is the asymmetric logistic quantile function, both with location 0 and scale 1. After some algebra, we find that the correction to be applied is \( \frac{(1 - \tau)}{\tau(3 - 4\tau)} \) for \( \tau < 0.5 \) and \( \frac{\tau}{(1 - \tau)(4\tau - 1)} \) for \( \tau \geq 0.5 \). The elaboration is found in appendix A.3.

- the student-t prior on the intercept should have location \( -Q(x = 0.5, \tau) \), or if we also want to take the imbalance of the data into account \( -Q(x = 1 - \bar{y}, \tau) \). The elaboration identical as for the ADL quantile regression and found in appendix A.2.

We have implemented two Bayesian models in Stan, using priors as described above and a likelihood based on a binary choice model with asymmetric logistically distributed noise:

- \( alr \), which uses fixed \( \tau = 1 - \bar{y} \)
- \( alra \), which uses \( \tau \) as model parameter

### 7.2 Generalized extreme value link

Wang & Dey (2008) have first proposed the use of a link function based on the generalized extreme value (GEV) distribution as a "new flexible skewed link function for binary response
data”. In the next subsections we look into the GEV distribution, and we evaluate the models using this link.

### 7.2.1 Generalized extreme value distribution

The GEV distribution is a family of continuous probability distributions developed within the extreme value theory to combine the Gumbel, Fréchet and Weibull families. It has the following cdf:

\[
F(x | \mu, \sigma, \xi) = \exp \left\{ - \left[ 1 + \xi \left( \frac{x - \mu}{\sigma} \right) \right]^{-\frac{1}{\xi}} \right\} \tag{7.13}
\]

for \(1 + \xi \left( \frac{x - \mu}{\sigma} \right) > 0\), where \(\mu \in \mathbb{R}\) is the location parameter, \(\sigma > 0\) scale parameter and \(\xi \in \mathbb{R}\) the shape parameter.

For \(\xi = 0\) expression (7.13) is formally undefined and is replaced by the Gumbel distribution, the result obtained by taking the limit as \(\xi \to 0\):

\[
F(x | \mu, \sigma, 0) = \exp \left\{ - \exp \left( - \frac{x - \mu}{\sigma} \right) \right\} \tag{7.14}
\]

The density function is, consequently:

\[
f(x | \mu, \sigma, \xi) = \frac{1}{\sigma} \left[ 1 + \xi \left( \frac{x - \mu}{\sigma} \right) \right]^{-\frac{1}{\xi} - 1} \exp \left\{ - \left[ 1 + \xi \left( \frac{x - \mu}{\sigma} \right) \right]^{-\frac{1}{\xi}} \right\} \tag{7.15}
\]

again for \(1 + \xi \left( \frac{x - \mu}{\sigma} \right) > 0\), and the density is zero outside of the relevant range.

For \(\xi = 0\) the density becomes:

\[
f(x | \mu, \sigma, 0) = \frac{1}{\sigma} \left[ - \left( \frac{x - \mu}{\sigma} \right) \right] \exp \left\{ - \exp \left( - \frac{x - \mu}{\sigma} \right) \right\} \tag{7.16}
\]

, which is positive for \(x \in \mathbb{R}\).

The GEV pdf and cdf are displayed in figure 7.12

### 7.2.2 GEV based models

Wang & Dey (2008) propose a GEV distribution based link function for the binomial regression model. That means that in the binary choice model (3.2) it is assumed that \(\varepsilon_i \sim f_{GEV}(\mu = 0, \sigma = 1, \xi)\), so that: \(P(Y_i = 1) = 1 - F_{GEV}(-X_i'\beta)\).

Wang & Dey explain that the commonly used complementary log-log (cloglog) link is prone to misspecification because of its positive and fixed skewness. The GEV link however is flexible in fitting the skewness in the data with the unknown shape parameter, and it supports a much wider range of skewness. The GEV model includes the binomial regression with the cloglog link at \(\xi = 0\). The GEV distribution also approaches the standard normal distribution for specific values of \(\mu \approx -0.35579, \sigma \approx 0.99903\) and \(\xi \approx -0.27760\). A QQ-plot reveals that the GEV with these parameters coincides with standard normal between the 0.02 and 0.98 quantiles. The discrepancy lies mainly in the tails.
A Bayesian approach is proposed for estimation. The authors explain that it can be shown that when \( \xi < 0.5 \), the regularity conditions required by MLE are not satisfied, hence the standard asymptotic likelihood results are not applicable. Hence a Bayesian method is favoured as Bayesian inference does not depend on the regularity conditions required by the asymptotic theory of MLE.

The propriety of the posterior is proven for proper and improper priors on the coefficients \( \beta \). As a prior for \( \xi \) the authors find that it is reasonable to assume \( -1 \leq \xi \leq 1 \), suggesting a uniform prior. Positive large \( \xi \)'s result in improper posterior under many improper priors for \( \beta \), including the uniform prior. Negative large \( \xi \)'s are said to rarely occur in practice.

The authors have applied this model in Wang & Dey (2010) on

- simulated data from the cloglog model and the probit model. The authors explain that the GEV model is suitable. The cloglog model is retrieved: with increasing samples size \( \xi \) tends to 0. The probit model is retrieved too: for different sample size, the 95% HPD intervals always includes \( \xi \approx -0.27760 \). The GEV model has consistently the lowest in-sample deviance, as measured by 2 times the negative log likelihood, \( D(y, \theta) = -2 \log p(y|\theta) \) where \( \theta = c(\beta, \xi) \), when compared to the logit, probit and cloglog. In our opinion that shouldn’t be surprising as the GEV is the most flexible model.
model also gives good results for the out-of-sample deviance, a measure of predictive performance. It scores not as well as the true model, but quite close, and almost always better than other less flexible models.

- B2B electronic payments system adoption data, a data set with 15,175 observations, 70% of which are positives. The GEV model proves to have a better predictive performance than logit, probit and cloglog. The GEV model was fitted yielding a positive skewness estimate $\hat{\xi} = 1.4$. Inverting the 0 and 1 observations, results in a negative skewness. That proves the superiority of the GEV over cloglog which performed poorly on the inverted data, as expected, due to its fixed positive skewness.

Wang & Dey haven’t packaged their method in an R-package. I have received their R-code, implementing a simple Metropolis with Gibbs sampler with improper uniform priors.

Raffaella Calabrese is an author that has extensively published on GEV regression. In Calabrese & Osmetti (2011) logistic regression is compared with GEV regression on a rare events data set: bankruptcy in Italian SMEs. In this application false positives aren’t too bad: it leads to the bank not granting the loan, hence losing a small revenue, the margin on the interest payments. A false negative has far more costly consequences: the bank loses whole or part of the credit exposure. Therefore new metrics of predictive performance are proposed $\text{MSE}^+$ and $\text{MAE}^+$, which are essentially Mean Squared Error (MSE) and Mean Absolute Error (MAE) metrics taking only the positive errors into account. The authors find that the GEV outperforms logistic regression on these measures, obviously at the expense of a large number of false negatives. In our opinion it would have been better to use a different metric of predictive performance such as the area under the ROC-curve.

In Calabrese (2012) logistic regression is compared with GEV regression again on bankruptcy data in Italian SMEs. The paper explains that the common measure of predictive performance the AUC of the ROC curve is not appropriate in this context, because it doesn’t take the cost of misclassification into account. Therefore the authors propose an alternative: the curve of Misclassification Error Loss (MEL) and a new classifier performance measure.

In Calabrese et al. (2013) the GEV model is extended from binomial regression on a linear combination of predictors $\beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ji}$ to a linear combination of smoothed predictors $\beta_0 + \sum_{j=1}^{p-1} f_j(x_{ji})$, where $f_j(\cdot)$ are unknown one-dimensional smooth functions of the continuous covariates $x_{ji}$.

Calabrese offers her algorithms in R-package $bgeva$. It doesn’t estimate $\xi$. Rather the user should try different values of $\xi$ to find out which gives the best fit.

### 7.2.3 Evaluation and practical use for this study

In our opinion the GEV is not suitable as a link function for GLM because of the range limitation it imposes: $1 + \xi \left(\frac{x - \mu}{\sigma}\right) > 0$, except where $\xi = 0$. We believe it makes the GEV
link overly sensitive to outliers. That will lead to overfitting and hence inferior predictions. No matter how large the sample size $n$, a single outlying point can force the estimation of a completely different value of $\xi$. That is surely an undesirable property which we don’t expect in a good classifier. A different $\xi$ is required to shift the point where the response curve reaches $P(y=1) = 1$ for negative $\xi$ or $P(y=1) = 0$ for positive $\xi$. The different $\xi$ thus affects also the entire skewness. Estimating $\sigma$ along with $\beta$ and $\xi$ could be of help, but that is unusual in binomial linear regression. Usually it makes the problem overdefined as the coefficients and $\sigma$ are perfectly correlated: $\beta_j/\sigma$ is constant for all coefficients. Also note that the out-of-sample deviance measure can reach infinity because of a single data point, even if it is only moderately outlying.

We are also skeptical on the claim that the GEV model approximately includes the probit regression. The discrepancy is in the tails, says Wang, but the tail behaviour is of great importance because it determines the contribution of outliers to the overall loss, as explained in section 3.3. Wang has fitted the GEV model on simulated probit data with 4 predictors, for several sample sizes, $N = 200$, $N = 1000$ and $N = 5000$ and found 95% HPD intervals for $\xi$: $(-0.8, 0.17)$, $(-0.36, -0.03)$ and $(-0.31, -0.13)$, all three including $\xi \approx -0.27760$ at which the GEV distribution approximates the standard normal distribution. However, given the shape of the GEV pdf, and its limited support, we can be sure that $\xi \rightarrow 0$ as $N$ is increased further, and the 95% HPD intervals won’t include $\xi \approx -0.27760$ anymore.

### 7.3 Skewed t-Link

In [Kim (2002)] the skewed t-distribution is introduced, a parametric class of distributions that have the properties (i) strict inclusion of the student-t distribution, (ii) mathematical tractability, (iii) wide range of the indices of skewness. Furthermore the class approximately includes well known symmetric and asymmetric distributions, such as the standard normal, the logistic and the skew-normal distributions.

#### 7.3.1 Skewed t-Link distribution

The skewed t-distribution is defined as the distribution of a random variable $T$:

$$T = \frac{\theta}{\sqrt{1 + \theta^2}} |U| + \frac{1}{\sqrt{1 + \theta^2}} V$$  \hspace{1cm} (7.17)

Where $U \sim N(0, 1/\lambda)$ and $V \sim N(0, 1/\lambda)$ and $\lambda \sim Gamma(\nu/2, 2/\nu)$ and $\nu > 0$.

Then the distribution of $T$ is the skewed t distribution, $T \sim St(\theta, \nu)$, with probability density function:

$$f(x|\theta, \nu) = 2f_{\nu}(x)F_{\nu+1} \left( \frac{\theta x \sqrt{\nu + 1}}{\sqrt{\nu + x^2}} \right), \quad -\infty < x < \infty$$  \hspace{1cm} (7.18)
Chapter 7. Link functions

The degrees of freedom $\nu$ regulates the heaviness of the tails and $\theta$ configures the amount of skewness. Some examples of skewed t probability density function are displayed in figure 7.13.

![Figure 7.13: pdf of the skewed t distribution](image)

Kim uses it as the distribution for the noise term $\varepsilon_i$ in the binary choice model of (3.2). For the estimation of the model a Bayesian and a frequentist approach are proposed. The Bayesian approach uses a multivariate normal prior on $\beta$ and a normal prior on $\theta$. A Gibbs sampling algorithm is then proposed alternatingly sampling from the fully conditional posteriors on the latent variables, $\beta, \theta$ and finally $\nu$. To sample from $\nu$'s conditional posterior, two ways are proposed:

1. $\nu$ is chosen from a finite set, so that the conditional distribution of $\nu$ becomes discrete. It is then easy to simulate from the discrete distribution

2. a standard uniform prior is assumed on $1/\nu$, so $\nu$ has the inverse uniform distribution in range $(1, \infty)$

The frequentist approach proposes a Monte-Carlo EM algorithm for the ML estimation of the skewed-link model with known degrees of freedom $\nu$, or otherwise $\nu$ unknown from a finite set.

Kim has compared the performance of his model in several ways:

- On simulated data from a true skewed t link model: the MLEs are calculated for the logit, probit and cloglog model and the skewed-t model with the degrees of freedom fixed
at the true value $\nu = 10$ and the skewed probit model. Unsurprisingly the skewed-t model with the degrees of freedom fixed at the true value performs best: it is able to retrieve the true value of the coefficient $\beta$ and skewness parameter $\theta$ it has the best fit as measured by the deviance. The skewed probit is second best both in terms of the precision of the estimates and fit. The symmetric link models, logit and probit, are not adequate for fitting the simulated data. The 95% confidence intervals of $\beta$ of each model include 0.

- by doing a comprehensive Bayesian analysis on a simulated data generated from the skewed link model, with vague priors for $\beta$ and $\theta$ and the degrees of freedom was assumed unknown from a finite set (5,10,20,40) with equal prior probabilities assigned. The true value for the degrees of freedom $\nu = 10$ comes out as the most probable. Using a conditional Bayes factor procedure, the Bayesian test rejects the symmetric t-link model in favor of the true skewed t link model.

- on real data: the body fat data set. The data set contains a binary response and 3 continuous covariates. There are $N = 36$ observations, 32 of which are positives. The logistic, probit and cloglog model were fitted using the ML method. The skewed probit and the skewed t link model with $\nu = 8$ were fitted using the MCMC method, using vague priors on $\beta$ and $\theta$. The goodness of fit and predictive ability are assessed with the sum of squares of Pearson residuals and Somers’ D index. The three ML methods estimate regression coefficients with large standard error, so that the 95% confidence intervals include 0 for all predictors. The coefficient estimates by the skewed models, are however significant and their 95% HDP intervals don’t include 0. The conditional Bayes factor points to overwhelming evidence for the skewed models. The goodness-of-fit statistics Pearson’s $X^2$ and Somers’ D advocate that the skewed t link model is the best fitting model.

7.3.2 Practical use

We haven’t implemented the skewed t link model for several reasons. Although the skewness can be regulated by $\theta$, the amount of skewness configurable by this model remains limited. There are far more skewed models as we will see later. Additionally we prefer a model that is easily usable out of the box, so ideally the degrees of freedom $\tau$ should be estimated from the data. We have attempted an implementation of binomial regression with a simple non-skewed student-t link, however with limited success. The implementation works fine on simulated data, but convergence proves problematic on real data: $\nu$ is hard to estimate, even with strong informative priors, and the optimization program gets stuck. The reason is that the inverse student-t link is not logconcave, or equivalently the loss function is not convex as seen in figure 3.1 leading to a multi-modal log likelihood function or posterior distribution.

We don’t like statements that the skewed t link model encompasses the usual symmetric and asymmetric link models. E.g. the student-t distribution with $\nu = 8$ is often said to approximate the logistic distribution. The discrepancy is in the tails, however the tail behaviour is
a great importance as it determines the robustness against binomial outliers. In figure 3.1 we see that the student-t(8) loss deviates largely from the logistic loss in its asymptotic behaviour. The student-t loss is not convex. It is much more tolerant of outliers and it doesn’t have a linear asymptotic for large negative $y_i^*X_i^\prime\beta$, whereas the logistic loss converges to the $L_1$ norm of the link value for large negative $y_i^*X_i^\prime\beta$.

In the context of this study we are surely interested in student-t link binomial regression, because of its promise of high robustness against outliers. But it proves not to fit in our approach of looking for a point estimate being the mode of a posterior distribution, because the student-t link leads to multimodal log likelihoods and posteriors. It could be the topic of further research.

7.4 Flexible student-t link

In Eyheramendy & Madigan (2007) a new flexible family of link functions is proposed for the generalized linear model. The family of link functions is based on the student-t cdf and said to approximately include the logistic regression (when the degrees of freedom is 8) and the probit regression when the degrees of freedom is large. The authors acknowledge that links based on symmetric distributions are limited: their corresponding cdf approaches 1 at the same rate as they approach 0, which is not always reasonable. Therefore they start of a class of skewed densities and extend these.

7.4.1 The model

Fernandez and Steel propose a class of skewed densities indexed by a scalar $\delta \in (0, \infty)$ of the form:

$$f(x|\delta) = \frac{2}{\delta + 1/\delta} \left[ f\left(\frac{x}{\delta}\right)I(x \geq 0) + f(x\delta)I(x < 0)\right], \quad x \in \mathbb{R}$$  \hspace{1cm} (7.19)

Where $f(.)$ is a univariate pdf with the mode at 0 and symmetry around the mode. The parameter $\delta$ determines the amount of mass at each side of the mode, and hence the skewness of the densities.

This way to define a class of skewed densities is applied to the student-t distribution, which is then assumed as probability distribution for the noise term $\varepsilon_i$ in the binary choice model of (3.2). That forms the basis of a Bayesian hierarchical model that contains, in addition to the parameter $\delta$ that controls the skewness of the density, and the student-t degrees of freedom which controls the thickness of the tail, a third parameter that controls the sparseness of the model, i.e. the number of regression parameters with zero posterior mode. In studies where there are a large number of predictors, this methodology is said to give a way of discriminating and selecting relevant predictors.

So Eyheramendy & Madigan propose the following priors:

- the distribution of the regression coefficients $\beta_j$ is normal with mean 0 and variance $\tau_j$: $\beta_j \sim N(0, \tau_j)$
• the distribution of the hyperparameters $\tau_j$ is exponential: $\tau_j \sim \exp(2/\gamma)$

Integrating with respect of $\tau_j$, one finds that the distribution of $\beta$ conditional on $\gamma$ is the Laplace or double exponential distribution: $P(\beta|\gamma) = \sqrt{\gamma/2} \exp(-\sqrt{\gamma}|\beta_j|)$.

The model is defined from a Bayesian perspective, but the estimation uses frequentist inference. The authors propose a mode-finding procedure based on the EM algorithm, treating the latent variables in the binary choice model as missing data. Additionally they use the decomposition of the Laplace distribution into a two-level Bayesian hierarchical model to estimate $\beta$ treating the $\tau_j$ parameters as missing data. A similar decomposition is done for the student-t distribution of the latent variables $u_i$: $u_i \sim \text{StudentT}(X_i^T\beta, \nu)$ is decomposed in the familiar two-level hierarchical model: $u_i \sim N(X_i^T\beta, \lambda_i)$ and $\lambda_i \sim \Gamma(\nu/2, 2/\nu)$. The EM treats $\lambda_i$ as additional missing data.

The parameters $\nu$, $\gamma$ and $\delta$ are not estimated. Rather multiple models are fitted on combinations of these parameters selected from a finite set.

Eyheramendy & Madigan have applied the model to:

• simulated data sets all with binary predictors, leading to the following conclusions:
  
  – the flexible student-t model consistently gives a better performance than logit and probit as measured by the in-sample misclassification rate.
  
  – Most of the best models choose the degrees of freedom $\nu$ of the Student-t distribution to be equal to 1, i.e. they prefer fat-tailed distributions.
  
  – the flexible student-t model also outperforms the elastic net model on 8 out of 10 data sets.
  
  – $\delta = 1.2$ is usually preferred. Our interpretation is that this value of $\delta$ corresponds to $\tau = 0.4$ in assymetric model of section 7.1.4 which comes as no surprise as the simulated has $\bar{y} = 0.6$, so the rule of thumb that $\tau = 1 - \bar{y}$ proves to hold in this case
  
  – small values of $\delta$ are preferred which means a preference for good fit and little regularization. Our evaluation is that such is not surprising as the in-sample fit is evaluated.
  
  – flexible probit (approximated by Student-t with $\nu = 30$) and flexible logit (approximated by Student-t with $\nu = 8$) usually prefer the same value for parameter $\delta$ and $\gamma$ as the flexible student-t model

• text categorization: the data set is split in a training, validation and testing set. Several model are fitted on the training set for various values of the three tuning parameters. The validation set is used for choosing the best model, which is subsequently used to make predictions on testing set (using a fixed threshold probability of 0.5). Predictive performance is measured using averages of precision (the positive predicted value), recall (the true positive rate or sensitivity), and the $F_1$ score (the harmonic mean of recall and precision). That is done for 5 different splits of the data set. The flexible student-t
model outperforms the logistic and probit regression models on recall and $F_1$ score. But probit followed by logistic regression prove to perform better on sensitivity.

7.4.2 Evaluation and practical use

The class of distributions proposed by Fernandez and Steel is in fact identical to the asymmetric family defined in section 7.1.4. The pdf in (7.19) and the one in (7.10) are the same, only the way the skewness is regulated by parameters $\tau$ vs $\delta$ is slightly different: $\tau$ is directly interpretable as the amount of mass in the left tail, whereas $\delta$ is the precision (i.e. the inverse scale) of the left tail. The authors don’t take into account that values of $\delta \neq 1$ imply shifted and less steep response curves resulting in larger absolute values for the intercept and slope coefficients. These can of course be allowed for by less penalization, but then care must be taken when interpreting the amount of penalization applied, as it is correlated with $\delta$.

The Laplace prior on the coefficients is nothing else than a lasso penalization and the parameter $\gamma$ that controls the sparseness of the model is comparable to parameter $\lambda$ that regulates the magnitude of the lasso penalization in (2.9). The only difference is that the flexible student-t model also applies a penalty to the intercept, which lasso doesn’t do.

We don’t like that a student-t with $\nu = 8$ is called approximately logistic and a probit with $\nu = 30$ approximately normal, in the context of binomial regression. It is true that the probability density distributions are close enough where most of the density is found, but the discrepancy is in the tails and the tail behaviour has a great impact on the fit.

The authors take the same approach as we do in this study: a Bayesian model is interpreted as a penalized likelihood model estimated with a mode-finding algorithm to obtain point estimates. Their current EM algorithm allows only categorical predictors, but the plan is to extend it for continuous predictors. The principle difference between flexible student-t and our preferred approach is in the priors used: we prefer the informative prior which Gelman has applied with great success on the logistic regression, whereas the flexible student-t uses a regularization method where the amount of regularization must be learned, typically by doing a cross validation requiring the splitting of the data set in a training, validation and testing set.

Another difference with our preferred approach is that we attempt to estimate all parameters together from the data, whereas the flexible student-t model requires the fitting of many models with different values for the parameters $\nu$, $\gamma$ and $\delta$, selected from a predefined finite set. The result is that the flexible student-t model is not usable out of the box, it must be set-up, the data set must be split and a finite set of parameter values must be defined.

The use of the student-t link is an attractive idea, because of its promise of high robustness against outliers. It leads however to multi-modal posteriors, a fact that is not mentioned in the paper. Our attempt to implement the student-t link, not even a skewed one, was successful on simulated data, but broke down when applied to real data: the degrees of freedom $\nu$ proved hard to estimate, and often the optimization program got stuck, probably
because because of the multi-modal posterior. So we have tried a different approach than optimization: MCMC sampling from the posterior, to find out how the model behaved and if signs of multi-modality could be detected from the posterior density plots. Our experience was that Stan experienced difficulties too when sampling from the posterior, resulting in long elapsed times. The MCMC trace plots showed a strong random walk behaviour, even if Stan is generally less sensitive to that problem. It occurred however without noticeable correlation in the coefficients, which is the usual cause of this behaviour. The chains don’t seem to stabilize around a mean, but rather keep exploring large ranges of the parameter space, even when applying highly restrictive informative priors. Our interpretation is that the relatively small loss contributed by outliers make the binomial regression with the student-t link and variable \( \nu \) a truly difficult model to estimate.

### 7.5 Symmetric power link family

\cite{jiang2013} explain that we need flexible link functions allowing the data to tell how much skewness must be incorporated. Several authors have proposed an additional power parameter on the cdf corresponding to standard link functions, e.g. the scobit model, which is a generalization of the logistic model by introduction of a power parameter. These link functions can accommodate flexible skewness in one direction, but the skewness in the other direction can be asymmetrically limited. Therefore they propose a new class of symmetric power link functions, based on the cdf corresponding to symmetric baseline link functions and its mirror reflection.

#### 7.5.1 Power link functions

In the binary choice model proposed in section 3.1 we find that \( P(y_i = 1) = F_\varepsilon(X_i' \beta) \) where \( F_\varepsilon \) is the cdf of the symmetric distribution of \( \varepsilon \). Here a general class of flexible link functions is proposed based on a symmetric baseline link function and its mirror reflection in the following manner. If \( F_0^{-1} \) is a baseline link function with its corresponding cdf \( F_0 \) for which the pdf is symmetric about zero, a symmetric power link family based on \( F_0 \) with power parameter \( r \) is proposed as:

\[
F(x, r) = \begin{cases} 
F_0^r(\xi), & \text{if } 0 < r \leq 1 \\
1 - F_0^{1/r}(-rx), & \text{if } 1 < r < \infty \end{cases} \tag{7.20}
\]

The intuition for the development of (7.20) is to utilize the fact that \( F_0^r(x) \) is a valid cdf and it achieves flexible left skewness when \( r < 1 \), while the same property holds for its mirror reflection \( 1 - F_0^{1/r}(-x) \) with skewness in the opposite direction. By combining the two in one single family of link functions, positive as well as negative skewness can be achieved. Also the \( x \) is scaled by the same parameter \( r \) to prevent the mode of the pdf to be too far away from zero and slope of the cdf becoming too flat. The baseline link \( F_0 \) is included as a special case as \( r = 1 \). The power link cdf and pdf are displayed in figure 7.14.

\cite{jiang2013} propose three symmetric power link functions based on different base link functions as follows:
Figure 7.14: Symmetric power link cdf and pdf for different values of \( r \) under logit, Laplace and normal baseline link functions

1. the symmetric power logit (\textit{splogit}) link family using the logistic distribution with location 0 and scale 1 as baseline \( F_0 \)

2. the symmetric power t (\textit{spt}) link family using the student-t distribution as baseline \( F_0 \). In addition to the power \( r \) this family has the degrees of freedom \( \nu \) as parameter, thus allowing to adjust both the skewness of the distribution and the heaviness of its tails, therefore accounting for an extremely rich class of link functions

3. the symmetric power exponential (\textit{spep}) link family using the exponential power (ep) distribution as baseline \( F_0 \). The ep distribution is symmetric with pdf

\[
f(x|\mu, \sigma, p) = \frac{p}{2\sigma p^{1/p} \Gamma(1/p)} \exp \left( -\left( \frac{|x-\mu|}{\sigma} \right)^p \right), \quad -\infty < x < \infty
\]  \hspace{1cm} (7.21)

and cdf

\[
F(x|\mu, \sigma, p) = \frac{1}{2} + \text{sgn}(x-\mu) \gamma \left( 1/p, -\left( \frac{|x-\mu|}{\sigma} \right)^p \right), \quad -\infty < x < \infty
\]  \hspace{1cm} (7.22)
with $\mu \in \mathbb{R}$, $\sigma \in (0, \infty)$ and $p \in (0, \infty)$ and where $\gamma$ denotes the normalized lower incomplete gamma function. Clearly the normal distribution and the Laplace distribution are special cases at $p = 2$ and $p = 1$ respectively. If we set $\mu = 0$ and $\sigma = 1$ the ep distribution is symmetric about zero with flexible tail properties as $p$ varies. Here $p$ is restricted to be within the range of $[1, 2]$ to avoid distributions with thinner tails than the normal distribution. Using the ep with these settings as baseline link function $F_0$ results in the symmetric power exponential link family, which through its parameters $p$ and $r$ allows for combinations of large ranges of skewness and heaviness of tails.

As prior distributions the authors propose:

- an improper uniform prior on $\beta$, i.e. $\pi(\beta) \propto 1$.
- for the power parameter $r$, a proper gamma prior with mean 1 and and reasonable large variance

It is proven that by introducing an additional power parameter $r$ in the sense of (7.20), the posterior propriety under the uniform prior on $\beta$ is unchanged with a proper prior for $r$. The same holds for the spt family with degrees of freedom $\nu > k$, where $k$ is the number of columns of $X$.

[Jiang et al.] propose MCMC sampling from the posterior, alternatingly sampling from the fully conditional distributions $[\beta| r, X, y]$, $[r| \beta, X, y]$ and also $[\nu| \beta, r, X, y]$ if under spt and $[p| \beta, r, X, y]$ under spep. Each draw is done using the Adaptive Rejection Metropolis algorithm implemented in JAGS.

The model was compared to the GLM with some other standard or flexible link functions on simulated data. [Jiang et al.] find that the power parameter $r$ captures the skewness of the true model very well. The symmetric power models also perform very well on two measures used to compare the performance. The first is the Deviance Information Criterion (DIC), which balances the fit of a model with its complexity. The other measure is the logarithm of the pseudo-marginal likelihood (LMPL), which measures the accuracy of prediction based on leave-one-out-cross-validation ideas. The splogit model also proves to be robust when fitted on skewed data generated from a generalized extreme value link binomial linear model. The model was also applied on real data: the Protea species co-occurrence data, with good results.

7.5.2 Practical use

The skewed link symmetric power link functions are attractive for our application on rare events data, so we have implemented our own version of splogit and spep in Stan. We haven’t attempted an implementation of spt given our difficulties with the simple non-skewed student-t link.

However we won’t use the improper uniform prior on $\beta$ neither the weakly informative normal
prior on $\beta$ used by the authors on the simulated and real data sets. To stabilize the binomial regression with the symmetric power link and make it resistant against collinearity and (quasi-)complete separation, we want to apply the weakly informative default prior of chapter 6.2 again.

That requires corrections similar to those for the ALD quantile regression. For the \textit{splogit} that implies:

- the correction to be applied to the default scale of the student-t prior on the coefficients except the intercept is $0.25/f(Q(0.5, r), r)$, where $f(x, r)$ is the symmetric power logistic pdf and $Q(x, r)$ is the symmetric power logistic quantile function. After some algebra, we find that the correction to be applied is $0.5/(1 - 0.5^{1/r})$ for $r \leq 1$ and $0.5/(1 - 0.5^r)$ for $r > 1$. The elaboration is given in appendix A.4.

- we won’t configure a location for the student-t prior on the intercept because the shift of the link functions to the left for $r < 1$ or to the right for $r > 1$ is minimal, due to the scale already applied in (7.20).

For the \textit{spep} we have attempted do something similar: the correction to be applied to the default scale of the student-t prior on the coefficients except the intercept is $0.25/f(Q(0.5, p, r), p, r)$, where $f(x, p, r)$ is the symmetric power exponential pdf and $Q(x, p, r)$ is the symmetric power exponential quantile function. There is unfortunately no way to simplify this expression. We can calculate it in R, but that’s not good enough, we need to calculate it in Stan, as the correction varies with $p$ and $r$ which are model parameters too. Unfortunately Stan doesn’t offer quantile functions of common probability distributions. And there is no closed form of the Gamma quantile function. Initial tests without correction on the slope coefficients scale showed that \textit{spep} doesn’t easily converge. It appears to produce multimodal posteriors even on simple simulated data, and even when apply strong priors. Given these difficulties we have finally decided to drop the \textit{spep} model from this study.

On the power parameter $r$ we find a lognormal prior with location 0 more natural than the Gamma prior proposed by Jiang \textit{et al.} as it gives equal probability to values of $r$ as to values $1/r$. We fix the variance of the lognormal at default value $\sigma = 2$, allowing for a wide range of skew.

Finally it is interesting to look at the loss function resulting from \textit{splogit}. The loss function is visualized for the logit link and the splogit link with two values of $r$ in figure 7.15. We have used values $r < 1$, so the link function is skewed to the left, implying that it is relatively tolerant of outliers at $y = 1$ and intolerant of outliers at $y = 0$. For $r = 1$ the splogit link is equal to the logistic regression. The figure shows that the loss function for $y = 1$ quickly converges to the logistic loss, while the loss for $y = 0$ becomes much larger for values of $r$ further away from 1. Doing a calculation similar as in section 3.3, we find:

- for $r < 1$: outliers at $y = 1$ contribute $|X_i'y|\beta$ to the loss, while outliers at $y = 0$ contribute $-\log(r) + \frac{1}{r}|X_i'y|\beta$
for $r > 1$: outliers at $y = 0$ contribute $|X_i'\beta|$ to the loss, while outliers at $y = 1$ contribute \( \log(r) + r|X_i'\beta| \)

The loss functions are convex so the log likelihood or the posterior are unimodal for fixed $r$. Estimating the right $r$ from the data is however less evident as for the quantile loss functions. The latter can easily reduce the overall loss: if the data contain more positive outliers than negative or vice versa, the binary quantile regression will push $\tau$ to one side to reduce the overall loss. The splogit is almost not capable of doing so, as values $r$ away from 1 seem to imply an inevitable increase in the loss. That explains why the mode finding algorithm becomes easily unstable when the optimal $r$ is to be estimated.

Figure 7.15: Loss functions of the logistic regression and the splogit link for different value of $r$. The splogit link is skewed, so two curves are displayed for each value of $r$, since the loss for $y_i^* = -1$ data points is different from the loss for $y_i^* = 1$ data points. In these cases where $r < 1$, the upper curve is for $y_i^* = -1$ and the lower curve for $y_i^* = 1$. For $r > 1$, the inverse would be true.
Chapter 8

Results

8.1 Models and method

As the goal of this study is to compare predictive performance, we test various models in a cross-validation exercise on several data sets. We have narrowed this study to an evaluation of the binomial linear regression model with different link functions and different priors, and their performance on rare events data.

So the following models are evaluated:

- \texttt{glm}: conventional logistic regression
- \texttt{brglm}: logistic regression with the Jeffreys prior (Firth’s method)
- \texttt{bayesglm}: Gelman’s extension of the glm package, logistic regression with a weakly informative default prior
- \texttt{bqr}: our implementation in Stan of binary quantile regression using an asymmetric Laplace based link function, and Gelman’s weakly informative default prior. The quantile parameter $\tau$ is fixed at $1 - \bar{y}$
- \texttt{bqra}: same as \texttt{bqr} but $\tau$ is estimated along with the regression coefficients
- \texttt{alr}: implementation of binary quantile regression using an asymmetric logit link function, and Gelman’s weakly informative default prior. The quantile parameter $\tau$ is fixed at $1 - \bar{y}$
- \texttt{alra}: same as \texttt{alr} but the optimal $\tau$ is estimated along with the coefficients
- \texttt{splogit}: binomial regression using the symmetric power logit link family and Gelman’s weakly informative default prior. The power parameter $r$ is estimated from the data

The choice for the binomial linear model implies the assumption that there is a direction in $(p - 1)$ dimensional space where the probability $P(y = 1)$ increases. This assumption does not
necessarily hold on real-life data sets. There may be interactions between covariates or higher order non-linear effects. Or the positives \( y = 1 \) can be appear in distant clusters in \((p-1)\) dimensional space, not allowing for the choice of a single direction. Therefore we compare the predictive performance of the GLM models with two other state of the art data mining methods:

- **RandomForest**: RandomForests uses the idea of bagging or bootstrap aggregation, a technique for reducing the variance of a prediction function by averaging it over a collection of bootstrap samples. Trees are ideal candidates for bagging, since they can capture complex interaction structures in the data, but trees are notoriously noisy. Bagging reduces the variance, and the variance reduction is improved by random feature selection which reduces the correlation between the trees. We have used the function `randomForest` in the R-package with the same name, to build forests of 5000 trees.

- **SVM**: support vector machines (SVM) are supervised learning models representing the feature data as points in space and searching for an optimal separation of points belonging to different categories. They try to divide the categories by a gap that is as wide as possible, and where not possible, they try to minimize the overlap. SVMs work on an expanded feature space resulting from the application of functions on the covariates. As there is an infinite choice of functions, that leads to an infinite-dimensional optimization problem. By choosing the functions from function spaces generated by kernels, the optimization problem becomes finite-dimensional and only requires the inner kernel product of the feature matrix. We have used function `ksvm` in R-package `kernlab` using the default Gaussian kernel and the cost of constraint violation set at \( C = 10 \).

The cross-validation exercise implies that the data set is split in a number of partitions \( K \), of which \( K-1 \) are used to build the model and the remaining partition to generate predictions. That is done alternatingly over all partitions until predictions have been obtained for all data. We have used \( K \) between 7 and 10, depending on the data set. Care should be taken to do it properly. The data set must be split in random divisions, each containing an equal or almost equal number of positives. The predictions are then used to evaluate the predictive performance. We have used the following measures, where \( p_i \) is the predicted outcome probability:

- the root mean square error (RMSE) or Brier score: \( \frac{1}{n} \sum_{i=1}^{n} (y_i - p_i)^2 \)
- the misclassification error rate (ER) using cut-off value 0.5: \( \frac{1}{n} \sum_{i=1}^{n} |y_i - I(p_i > 0.5)| \)
- the false positive rate (FPR) using cut-off value 0.5: \( \frac{1}{n_{(y=0)}} \sum_{i=1}^{n} (1 - y_i)I(p_i > 0.5) \)
- the false negative rate (FNR) using cut-off value 0.5: \( \frac{1}{n_{(y=1)}} \sum_{i=1}^{n} (1 - y_i)I(p_i <= 0.5) \)
- the area under the ROC curve (AUC)
Chapter 8. Results

- the logarithmic score (logscore) which is equivalent to the average negative loglikelihood of the out-of-sample data: 
  $$-\frac{1}{n} \sum_{i=1}^{n} [y_i \log p_i + (1-y_i) \log(1-p_i)]$$

8.2 Data sets

We have evaluated the models on 4 data sets from UCI Machine Learning Repository, some with categorical and some with continuous covariates. Where the outcome variable is multivalued, we have selected one event to make it binary. Three imbalanced data sets are thus obtained where the outcome \(y\) is rare. Finally the models have also been tested on a balanced data set to verify how they behave there, and if a skewed link function is preferred. The data sets are listed in table 8.1. A summary of the results of the GLM models is available in table 8.2. The results of the two data mining methods are listed in table 8.3.

<table>
<thead>
<tr>
<th>Name</th>
<th>Obs</th>
<th>Cat</th>
<th>Cont</th>
<th>Pred</th>
<th>Outcome</th>
<th>(\bar{y})</th>
</tr>
</thead>
<tbody>
<tr>
<td>car</td>
<td>1728</td>
<td>6</td>
<td>0</td>
<td>15</td>
<td>class=&quot;vgood&quot;</td>
<td>3.7%</td>
</tr>
<tr>
<td>yeast</td>
<td>1484</td>
<td>0</td>
<td>8</td>
<td>8</td>
<td>class=&quot;POX&quot;</td>
<td>1.34%</td>
</tr>
<tr>
<td>segment</td>
<td>2310</td>
<td>0</td>
<td>19</td>
<td>19</td>
<td>class=7</td>
<td>14,3%</td>
</tr>
<tr>
<td>krkp</td>
<td>3204</td>
<td>36</td>
<td>0</td>
<td>37</td>
<td>class=&quot;won&quot;</td>
<td>52,1%</td>
</tr>
</tbody>
</table>

Table 8.1: The 4 data sets from the UCI Machine Learning Repository. Each data set is described with its name, the number of observations in it (Obs), the number of categorical attributes in it (Cat), the number of continuous attributes (Cont), the number of predictors generated from the attributes (Pred), the event corresponding to the binary outcome (Outcome) and the percentage of observations having the positive outcome (\(\bar{y}\)).

8.2.1 Car

When applying the standard logistic regression (glm) to this data set, the problem of quasi-separation and multicollinearity is obvious. Several coefficients have very large values, often with large standard errors, so that only one of the predictors is significant at the 5% level. The Jeffreys prior solves this problem: it shrinks the coefficients and all predictors become significant.

The Gelman prior removes this problem too and shrinks the coefficients even more. When applying the adaptive asymmetric methods, we find that the preferred skewness corresponds to what is suggested by the imbalance in the data. Both \(bqr\) and \(alr\) find optimal \(\tau = 0.95\) which is only slightly lower that \(1 - \bar{y}\).

The \(splogit\) model estimates the power parameter at \(r = 12.1\), which corresponds with skewness in the same direction as the asymmetric methods. Of the GLM models, the \(bayesglm\) model comes out as the best performer, followed closely the \(splogit\) and the quantile link models. However all tested models are wildly outperformed by the SVM model which achieves 0 misclassification.
8.2.2 Yeast

When applying the standard logistic regression (\textit{glm}) to this data set, the problem of variance inflation is obvious: there are some very large coefficient estimates, most with large standard error, so only one predictor is significant at the 95\% level. Firth’s method \textit{brglm} fails unless predictor ‘erl’ is removed. It shrinks the coefficients, only one remains significant. Gelman’s weakly informative prior shrinks the coefficients too and only one remains significant. The adaptive asymmetric methods \textit{bqra} and \textit{alra} find values of $\tau = 0.995$ and $\tau = 0.992$ respectively, which is even higher than $1 - \bar{y}$.

The power parameter of \textit{splogit} must be constrained by using a sufficiently small scale for its lognormal prior. Otherwise the algorithm becomes unstable, encountering infinite gradients and terminating before convergence due to insufficient decrease in logprobability found. It also leads to an overfitted model producing 0 or 1 probabilities, leading to infinite logscores in case of misclassification. When sufficiently constrained the model finds a value for the power parameter $r = 1.39$. The \textit{bqr} model comes out as the best performer, closely followed the others quantile methods \textit{bqr}, \textit{alr} and \textit{alra}. The \textit{RandomForest} proves to have the highest AUC.

8.2.3 Segment

The standard logistic regression (\textit{glm}) finds that not one of the predictors is significant at the 95\% level. \textit{brglm} seems to breaks down: all predictors are suddenly significant, but with huge values for the coefficients. \textit{bayesglm} estimates strongly shrunken coefficients, 2 of which are significant. \textit{bqr} also estimates strongly shrunken predictors. Both \textit{bqra} and \textit{alra} find optimal $\tau = 0.89$ which is higher than $1 - \bar{y}$.

\textit{splogit} allows a large scale on the prior of the power parameter before it becomes unstable. We find a power parameter $r = 5.59$. The \textit{bqra} model comes out as the best performer of the GLM models. However the \textit{RandomForest} performs very well too, and actually has a much better fit as measured by the logscore.

8.2.4 KRKP

In some of the cross-validation partitions there are predictors with just one remaining value. Some algorithms require its removal. The standard logistic regression (\textit{glm}) finds many significant predictors, but also very large coefficients. That is due to quasi-complete separation which this data set is prone to, as all its inputs are binary, except one. Firth’s method \textit{brglm} results in less but still many significant predictors, some of whose coefficients have huge values. Gelman’s \textit{bayesglm} also finds large values for the coefficients, but already more shrunken than those found by Firth’s method. The quantile link methods with adaptive skewness \textit{bqra} and \textit{alra} estimate $\tau = 0.133$ and $\tau = 0.147$ respectively, both much lower than $1 - \bar{y}$, as this data set is close to balanced. It shows that skewed link functions are not just useful for imbalanced data. \textit{splogit} estimates the power parameter $r = 0.669$, thus adding skewness in the same direction as $\tau$. The \textit{bqr} model comes out as the best performer of the GLM models. However all models are completely outperformed by the \textit{SVM}. 
## Chapter 8. Results

### Car

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE</th>
<th>ER</th>
<th>FPR</th>
<th>FNR</th>
<th>AUC</th>
<th>logscore</th>
</tr>
</thead>
<tbody>
<tr>
<td>glm</td>
<td>8.76</td>
<td>1.10</td>
<td>0.60</td>
<td>13.8</td>
<td>99.8</td>
<td>2.34</td>
</tr>
<tr>
<td>brglm</td>
<td>8.61</td>
<td>1.10</td>
<td>0.60</td>
<td>13.8</td>
<td>99.8</td>
<td>2.42</td>
</tr>
<tr>
<td><strong>bayesglm</strong></td>
<td><strong>8.45</strong></td>
<td><strong>0.93</strong></td>
<td><strong>0.42</strong></td>
<td><strong>13.8</strong></td>
<td><strong>99.8</strong></td>
<td><strong>2.36</strong></td>
</tr>
<tr>
<td>bqr</td>
<td>8.70</td>
<td>1.10</td>
<td>0.54</td>
<td>15.4</td>
<td>99.8</td>
<td>2.41</td>
</tr>
<tr>
<td>bqra</td>
<td>8.69</td>
<td>1.10</td>
<td>0.54</td>
<td>15.4</td>
<td>99.8</td>
<td>2.42</td>
</tr>
<tr>
<td>alr</td>
<td>8.60</td>
<td>1.04</td>
<td>0.54</td>
<td>13.9</td>
<td>99.8</td>
<td>2.37</td>
</tr>
<tr>
<td>alra</td>
<td>8.59</td>
<td>1.04</td>
<td>0.54</td>
<td>13.9</td>
<td>99.8</td>
<td>2.37</td>
</tr>
<tr>
<td>splogit</td>
<td>8.56</td>
<td>0.98</td>
<td>0.48</td>
<td>13.8</td>
<td>99.8</td>
<td><strong>2.36</strong></td>
</tr>
</tbody>
</table>

### Yeast

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE</th>
<th>ER</th>
<th>FPR</th>
<th>FNR</th>
<th>AUC</th>
<th>logscore</th>
</tr>
</thead>
<tbody>
<tr>
<td>glm</td>
<td>9.51</td>
<td>1.01</td>
<td>0.20</td>
<td>60</td>
<td>79</td>
<td>4.97</td>
</tr>
<tr>
<td>brglm</td>
<td>9.47</td>
<td>1.01</td>
<td>0.21</td>
<td>60</td>
<td>80</td>
<td>4.85</td>
</tr>
<tr>
<td>bayesglm</td>
<td>9.43</td>
<td>0.94</td>
<td>0.20</td>
<td>55</td>
<td>78</td>
<td>4.88</td>
</tr>
<tr>
<td><strong>bqr</strong></td>
<td><strong>9.26</strong></td>
<td><strong>0.87</strong></td>
<td><strong>0.27</strong></td>
<td><strong>45</strong></td>
<td><strong>81</strong></td>
<td><strong>4.54</strong></td>
</tr>
<tr>
<td>bqra</td>
<td>9.26</td>
<td>0.88</td>
<td>0.27</td>
<td>45</td>
<td>82</td>
<td>4.55</td>
</tr>
<tr>
<td>alr</td>
<td>9.26</td>
<td>0.88</td>
<td>0.27</td>
<td>45</td>
<td>82</td>
<td>4.56</td>
</tr>
<tr>
<td>alra</td>
<td>9.26</td>
<td>0.88</td>
<td>0.27</td>
<td>45</td>
<td>82</td>
<td>4.57</td>
</tr>
<tr>
<td>splogit</td>
<td>9.50</td>
<td>0.94</td>
<td>0.20</td>
<td>55</td>
<td>77</td>
<td>5.09</td>
</tr>
</tbody>
</table>

### Segment

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE</th>
<th>ER</th>
<th>FPR</th>
<th>FNR</th>
<th>AUC</th>
<th>logscore</th>
</tr>
</thead>
<tbody>
<tr>
<td>glm</td>
<td>5.29</td>
<td>0.30</td>
<td>0.2</td>
<td>0.9</td>
<td>99.8</td>
<td>5.53</td>
</tr>
<tr>
<td>brglm</td>
<td>7.20</td>
<td>0.52</td>
<td>0.3</td>
<td>1.8</td>
<td>99.0</td>
<td>18.7</td>
</tr>
<tr>
<td>bayesglm</td>
<td>3.58</td>
<td>0.17</td>
<td>1.2</td>
<td>0.0</td>
<td>99.9</td>
<td>0.86</td>
</tr>
<tr>
<td>bqr</td>
<td>3.11</td>
<td>0.09</td>
<td>0.6</td>
<td>0.0</td>
<td>99.9</td>
<td>1.08</td>
</tr>
<tr>
<td><strong>bqra</strong></td>
<td><strong>3.03</strong></td>
<td><strong>0.09</strong></td>
<td><strong>0.6</strong></td>
<td><strong>0.0</strong></td>
<td><strong>99.9</strong></td>
<td><strong>1.08</strong></td>
</tr>
<tr>
<td>alr</td>
<td>3.52</td>
<td>0.12</td>
<td>0.9</td>
<td>0.0</td>
<td>99.7</td>
<td>8.06</td>
</tr>
<tr>
<td>alra</td>
<td>3.53</td>
<td>0.13</td>
<td>0.9</td>
<td>0.0</td>
<td>99.6</td>
<td>8.06</td>
</tr>
<tr>
<td>splogit</td>
<td>3.21</td>
<td>0.13</td>
<td>0.9</td>
<td>0.0</td>
<td>100</td>
<td>0.61</td>
</tr>
</tbody>
</table>

### Krkp

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE</th>
<th>ER</th>
<th>FPR</th>
<th>FNR</th>
<th>AUC</th>
<th>logscore</th>
</tr>
</thead>
<tbody>
<tr>
<td>glm</td>
<td>14.81</td>
<td>2.31</td>
<td>2.33</td>
<td>2.27</td>
<td>99.6</td>
<td>8.53</td>
</tr>
<tr>
<td>brglm</td>
<td>15.15</td>
<td>2.52</td>
<td>2.51</td>
<td>2.53</td>
<td>99.5</td>
<td>8.72</td>
</tr>
<tr>
<td>bayesglm</td>
<td>15.08</td>
<td>2.55</td>
<td>2.75</td>
<td>2.33</td>
<td>99.6</td>
<td>8.35</td>
</tr>
<tr>
<td><strong>bqr</strong></td>
<td><strong>14.73</strong></td>
<td><strong>2.36</strong></td>
<td><strong>2.57</strong></td>
<td><strong>2.14</strong></td>
<td><strong>99.6</strong></td>
<td><strong>8.12</strong></td>
</tr>
<tr>
<td>bqra</td>
<td>14.88</td>
<td>2.80</td>
<td>4.00</td>
<td>1.49</td>
<td>99.5</td>
<td>9.10</td>
</tr>
<tr>
<td>alr</td>
<td>15.03</td>
<td>2.55</td>
<td>2.75</td>
<td>2.33</td>
<td>99.6</td>
<td>8.27</td>
</tr>
<tr>
<td>alra</td>
<td>14.97</td>
<td>2.80</td>
<td>3.77</td>
<td>1.75</td>
<td>99.5</td>
<td>8.92</td>
</tr>
<tr>
<td>splogit</td>
<td>14.93</td>
<td>2.83</td>
<td>4.07</td>
<td>1.49</td>
<td>99.5</td>
<td>9.38</td>
</tr>
</tbody>
</table>

**Table 8.2:** Test results for the GLM models, all figures are shown as percentages.
### Car

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE</th>
<th>ER</th>
<th>FPR</th>
<th>FNR</th>
<th>AUC</th>
<th>logscore</th>
</tr>
</thead>
<tbody>
<tr>
<td>RandomForest</td>
<td>8.29</td>
<td>0.52</td>
<td>0.00</td>
<td>13.8</td>
<td>100</td>
<td>2.89</td>
</tr>
<tr>
<td>SVM</td>
<td>0.9</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>100</td>
<td>0.26</td>
</tr>
</tbody>
</table>

### Yeast

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE</th>
<th>ER</th>
<th>FPR</th>
<th>FNR</th>
<th>AUC</th>
<th>logscore</th>
</tr>
</thead>
<tbody>
<tr>
<td>RandomForest</td>
<td>9.61</td>
<td>1.08</td>
<td>0.27</td>
<td>60</td>
<td>84</td>
<td>4.84</td>
</tr>
<tr>
<td>SVM</td>
<td>9.46</td>
<td>1.07</td>
<td>0.13</td>
<td>70</td>
<td>68</td>
<td>5.17</td>
</tr>
</tbody>
</table>

### Segment

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE</th>
<th>ER</th>
<th>FPR</th>
<th>FNR</th>
<th>AUC</th>
<th>logscore</th>
</tr>
</thead>
<tbody>
<tr>
<td>RandomForest</td>
<td>3.04</td>
<td>0.13</td>
<td>0.00</td>
<td>0.90</td>
<td>100</td>
<td>0.37</td>
</tr>
<tr>
<td>SVM</td>
<td>3.60</td>
<td>0.13</td>
<td>0.00</td>
<td>0.90</td>
<td>99.9</td>
<td>0.80</td>
</tr>
</tbody>
</table>

### Krkp

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE</th>
<th>ER</th>
<th>FPR</th>
<th>FNR</th>
<th>AUC</th>
<th>logscore</th>
</tr>
</thead>
<tbody>
<tr>
<td>RandomForest</td>
<td>12.71</td>
<td>1.25</td>
<td>1.63</td>
<td>0.9</td>
<td>99.9</td>
<td>8.88</td>
</tr>
<tr>
<td>SVM</td>
<td>5.41</td>
<td>0.4</td>
<td>0.39</td>
<td>0.42</td>
<td>100</td>
<td>1.32</td>
</tr>
</tbody>
</table>

**Table 8.3:** Test results for two data mining methods, all figures are shown as percentages.
Chapter 9

Conclusion

The problems of variance inflation and (quasi-)complete separation prove to be common in conventional logistic regression on real life data sets. Firth’s method is said to resolve both problems but it performs rather poorly in the cross-validation. The second-order unbiasedness property of the Jeffreys’ prior can’t compete with logistic regression using the informative prior proposed by Gelman. Gelman’s method works fine, and it even outperform the other GLM methods on one data set. The reason why it outperforms the more flexible methods is not clear. Upon research we have found that the function bayesglm doesn’t work 100% as specified in the documentation. The priors seems to be slightly different than specified, leading to a lucky shot on one data set.

In general we can say however that the more flexible methods, using a skewed link function outperform the symmetric link function. The statement that skewed link functions are better on imbalanced data proves to hold. The additional flexibility even helped on a fairly balanced data set. We find that the symmetric power link functions perform worse than the quantile-based skewed link functions. The splogit proved to be less stable, easily producing multi-modal posteriors. That is due to its loss function which isn’t able reduce the overall loss as elegantly as the quantile methods, which can benefit by allowing a lower loss on the category with numerous outliers in exchange for a higher loss where outliers are rare.

Suprisingly the more flexible models using quantile-based link functions where the quantile is estimated from the data, do not generally outperform their less flexible equivalent, which fixes the quantile \( \tau = 1 - \bar{y} \). Usually their performance is equivalent, only on data set krkp the more flexible model was clearly worse. The explanation must be that the more flexible models lead to overfitting to the training data, hence inferior predictions on test data.

Finally when comparing the GLM models to some other state of the art data mining classification models, it is found that they can be easily outperformed. SVM gives spectacularly better results on two data sets, reaching 100% of AUC and a zero, or close to zero misclassification rate. On two other data sets we found that RandomForests could easily keep up with the best of the GLM models, even slightly outperforming these depending on the performance measure used. That finding shows that GLM models are not necessarily a good default choice. The
logistic regression remains attractive as an explanatory model, because of its convenient interpretation of the regression coefficients as the impact on the log-odds of the outcome variable. But the GLMs as a class of models all assume that the probability of $P(y = 1)$ increases in one direction of the feature space. It supports interactions if these are added explicitly as predictors, but it cannot discover these automatically as Random Forests do. Similarly non-linear effects can be incorporated into GLM models by explicitly adding transformed inputs as predictors. SVM however efficiently performs non-linear classification using its kernel method, implicitly mapping inputs into high-dimensional feature spaces. Additionally SVM thrives when (quasi-)separation of data points is possible, whereas GLM models suffer from it, leading to large or infinite values for the coefficients. That is in fact a form of extreme overfitting, that must be mitigated using informative priors or penalization.

A way to overcome these weaknesses of GLM especially when dealing with imbalanced data could be the suggestion by Owen. He has shown convincingly that the logistic regression coefficients are strongly tied to the average feature vector $\bar{x}$ of the $y=1$ data points. Clustering the $x_i$ where $y_i = 1$ and subsequently fitting GLM models to each cluster, could lead to an interesting mixture of GLM fits. Making predictions requires then to assign data points to a cluster first, or even better to assign a probability of belonging to the each cluster, followed by doing the GLM prediction. It is an interesting idea that could be a topic for further research.
Appendix A

Calculations

A.1 Section 7.1.3: correction factor on the default prior scale

The correction factor to be applied to the default prior scale is \(0.25/f(Q(0.5, \tau), \tau)\), where \(f(x, \tau)\) is the ALD pdf and \(Q(x, \tau)\) the ALD quantile function, both with location 0 and scale 1.

The nominator 0.25 is the rate of increase of the logistic response curve at \(P(y_i = 0.5)\).

The denominator is rate of increase of the ALD response curve at \(P(y_i = 0.5)\). That is easy to see as follows. The ALD response function is given by \(P(y_i = 1) = 1 - F(-X'_i\beta, \tau)\), where \(F(x, \tau)\) is the ALD cdf with location 0 and scale 1. The ALD response function takes value 0.5 where \(X'_i\beta = -Q(0.5, \tau)\). The rate of increase of the ALD response function is found as its first derivate \(f(-X'_i\beta, \tau)\). Hence the rate of increase of the ALD response curve at \(P(y_i = 0.5)\) is \(f(Q(0.5, \tau), \tau)\).

The rate of increase is lower than 0.25 for \(\tau\) further away from the median, we allow for larger coefficients \(\beta\) by correcting the prior scale with factor \(0.25/f(Q(0.5, \tau), \tau)\).

The ALD quantile function is given by:

\[
Q(x|\mu, \sigma, \tau) = \begin{cases} 
\mu + \frac{\sigma}{1-\tau} \log\left(\frac{x}{\tau}\right), & \text{if } 0 \leq x \leq \tau \\
\mu - \frac{\sigma}{\tau} \log\left(\frac{1-x}{1-\tau}\right), & \text{if } \tau < x \leq 1
\end{cases}
\]  

(A.1)

Evaluating the quantile function at \(x = 0.5\), \(\mu = 0\) and \(\sigma = 1\), and filling in the result in the pdf (7.3) results in \(0.5\tau\) for \(\tau < 0.5\) and \(0.5(1 - \tau)\) for \(\tau \geq 0.5\).

Using this value as the denominator, the correction to be applied to the scale is \(0.5/\tau\) for \(\tau < 0.5\) and \(0.5/(1 - \tau)\) for \(\tau \geq 0.5\).

A.2 Section 7.1.3: location for the prior on the intercept

As location for the prior on the intercept we propose \(-Q(x = 0.5, \tau)\) to offset the shift of the response curve for values of \(\tau\) different from the median. Using a symmetric prior distribution for the intercept \(\beta_0\), the most probably value of the intercept is thus \(-Q(x = 0.5, \tau)\).

When using scaled predictors, the mean feature vector is zero, hence the link value is equal
Appendix A. Calculations

59

to the intercept $\beta_0$. For this link value, the response corresponding to the most probable
value of $\beta_0$ is $P(y_i = 1) = 1 - F(Q(x = 0.5, \tau), \tau) = 0.5$, which is reasonable for balanced data
assuming that the binomial regression model is applicable, i.e. there is a direction in $(p-1)$
dimensional space where $P(y=1)$ increases.

To take the imbalance of the data into account the location of the prior on the intercept
should rather be $-Q(x = 1 - \bar{y}, \tau)$, following a similar reasoning as above.

**A.3 Section 7.1.4: correction factor on the default prior scale**

Similarly as in appendix A.1 the denominator of expression $0.25/f(Q(0.5, \tau), \tau)$ needs to be
calculated. The asymmetric logistic quantile function is given by:

$$Q(x|\mu, \sigma, \tau) = \begin{cases} 
\mu - \frac{\sigma}{2(1-\tau)} \log\left(\frac{2\tau}{x} - 1\right), & \text{if } 0 \leq x \leq \tau \\
\mu + \frac{\sigma}{2\tau} \log\left(\frac{2(1-\tau)}{1-x} - 1\right), & \text{if } \tau < x \leq 1
\end{cases}$$

(A.2)

Evaluating the quantile function at $x = 0.5$, $\mu = 0$ and $\sigma = 1$, and filling in the result in the
pdf (7.11) results in $\frac{(1-\tau)(3 - 4\tau)}{4(1-\tau)}$ for $\tau < 0.5$ and $\frac{4\tau(1-\tau)}{4\tau}$ for $\tau \geq 0.5$.

Hence the correction to be applied to the scale of the slope coefficients is $\frac{(1-\tau)}{\tau(3 - 4\tau)}$ for $\tau < 0.5$
and $\frac{\tau}{(1-\tau)(4\tau-1)}$ for $\tau \geq 0.5$.

**A.4 Section 7.5.2: location for the prior on the intercept**

As the logistic regression cdf with location 0 and scale 1 is given by $F_0(x) = \frac{1}{1+e^{-x}}$, the spllogit
cdf is

$$F(x, r) = \begin{cases} 
\frac{1}{(1+e^{-x/r})^r}, & \text{if } 0 \leq r \leq 1 \\
\frac{1}{(1+e^{rx})^r}, & \text{if } 1 < r \leq \infty
\end{cases}$$

(A.3)

Consequently the spllogit pdf is:

$$f(x, r) = \begin{cases} 
\frac{e^{-x/r}}{(1+e^{-x/r})^{r+1}}, & \text{if } 0 \leq r \leq 1 \\
\frac{e^{rx}}{(1+e^{rx})^{r+1}}, & \text{if } 1 < r \leq \infty
\end{cases}$$

(A.4)

And the spllogit quantile function is:

$$Q(x, r) = \begin{cases} 
-r \log\left(x^{-\frac{1}{r}} - 1\right), & \text{if } 0 \leq r \leq 1 \\
\frac{1}{r} \log\left((1-y)^{r} - 1\right), & \text{if } 1 < r \leq \infty
\end{cases}$$

(A.5)

Evaluating the quantile function at $x = 0.5$ and filling in the result in the pdf results in
$0.5(1 - 0.5^r)$ for $r \leq 1$ and $0.5(1 - 0.5^r)$ for $r > 1$. 
Using this value as the denominator, the correction to be applied to the scale is $\frac{0.5}{1 - 0.5^{1/r}}$ for $r \leq 1$ and $\frac{0.5}{1 - 0.5^r}$ for $r > 1$. 
Appendix B

Stan model code

B.1 bqr

data {
  int<lower=0> N; // number of data items
  int<lower=0> K; // number of predictors
  matrix[N,K] x; // predictor matrix
  int<lower=0,upper=1> y[N];
  real<lower=0,upper=1> tau;
  real<lower=0> priorScale;
  int<lower=1> priorDf;
  row_vector[K] center;
  vector[K] scale;
}

transformed data {
  real<lower=0> rateCorrection;
  real alphaBase;
  // helper variables
  real yRatio;

  # 1.0 in denominator to avoid implicit rounding to integer
  yRatio <- sum(y)/(1.0*N);

  if (1-yRatio <= tau) alphaBase <- -1/(1-tau)*log((1-yRatio)/tau);
  else alphaBase <- 1/tau*log(yRatio/(1-tau));

  // the factor to be applied to as a result of the quantile
  if (tau <= 0.5) rateCorrection <- 0.5/tau;
  else rateCorrection <- 0.5/(1-tau);
}

parameters {

real alpha; // intercept
vector[K] beta; // coefficients for predictors
}
transformed parameters {
  real<lower=0,upper=1> p[N];

  // to remove correlation, alpha is shift with alphaBase and because of the centering
  // the alphaOut to be used in predictions go into alpaOut
  real alphaOut;
  // to remove correlation, we keep the beta on the same scale, independent of tau
  // the beta to be used in predictions go into betaOut
  vector[K] betaOut;

  // alphaOut and betaOut are for using on the real scale
  betaOut <- beta*rateCorrection;
  alphaOut <- alphaBase + alpha - center * betaOut;

  // apply the ALD CDF
  for (n in 1:N) {
    real u;
    u <- -(alphaOut + x[n] * betaOut);
    if (u <= 0)
      p[n] <- 1 - tau*exp((1-tau)*u);
    else
      p[n] <- (1-tau)*exp(-tau*u);
    }
  }
}
model {
  alpha ~ cauchy(0, 10);

  for (k in 1:K) {
    beta[k] ~ student_t(priorDf, 0, priorScale/scale[k]);
  }
  y ~ bernoulli(p);
}

B.2 bqra

data {
  int<lower=0> N; // number of data items
  int<lower=0> K; // number of predictors
  matrix[N,K] x; // predictor matrix
}
Appendix B. Stan model code

```stan
int<lower=0,upper=1> y[N];
real<lower=0> priorScale;
int<lower=1> priorDf;
int<lower=1> shapeTau[2];
row_vector[K] center;
vector[K] scale;
}
transformed data {
  // helper variables
  real yRatio;

  # 1.0 in denominator to avoid implicit rounding to integer
  yRatio <- sum(y)/(1.0*N);
}
parameters {
  real alpha; // intercept
  vector[K] beta; // coefficients for predictors
  real<lower=0,upper=1> tau;
}
transformed parameters {
  real<lower=0,upper=1> p[N];

  // to remove correlation, we keep the beta on the same scale, independent of tau
  // the beta to be used in predictions go into betaOut
  vector[K] betaOut;
  real<lower=0> rateCorrection;
  real alphaBase;
  real alphaOut;

  if (1-yRatio <= tau) alphaBase <- -1/(1-tau)*log((1-yRatio)/tau);
  else alphaBase <- 1/tau*log(yRatio/(1-tau));

  // the factor to be applied to the scale as a result of the quantile
  if (tau <= 0.5) rateCorrection <- 0.5/tau;
  else rateCorrection <- 0.5/(1-tau);

  // alphaOut and betaOut are for using on the true scale
  betaOut <- beta*rateCorrection;
  alphaOut <- alphaBase + alpha - center * betaOut;

  // apply the ALD CDF
```
for (n in 1:N) {
    real u;
    u <- -(alphaOut + x[n] * betaOut);
    if (u <= 0)
        p[n] <- 1 - tau*exp((1-tau)*u);
    else
        p[n] <- (1-tau)*exp(-tau*u);
}
}
}

model {
    tau ~ beta(shapeTau[1],shapeTau[2]);
    alpha ~ cauchy(0, 10);
    for (k in 1:K) {
        beta[k] ~ student_t(priorDf, 0, priorScale/scale[k]);
    }
    y ~ bernoulli(p);
}

B.3 alr

data {
    int<lower=0> N; // number of data items
    int<lower=0> K; // number of predictors
    matrix[N,K] x; // predictor matrix
    int<lower=0,upper=1> y[N];
    real<lower=0,upper=1> tau;
    real<lower=0> priorScale;
    int<lower=1> priorDf;
    real<lower=0> priorScaleForIntercept;
    row_vector[K] center;
    vector[K] scale;
}

transformed data {
    real<lower=0> rateCorrection;
    real alphaBase;
    // helper variables
    real yRatio;

    # 1.0 in denominator to avoid implicit rounding to integer
    yRatio <- sum(y)/(1.0*N);

    if (1-yRatio <= tau)
Appendix B. Stan model code

\[
\text{alphaBase} \leftarrow 0.5/(1 - \text{tau}) \times \log(2 \cdot \text{tau} / (1 - \text{yRatio}) - 1);
\]

else
\[
\text{alphaBase} \leftarrow -0.5/\text{tau} \times \log(2 \cdot (1 - \text{tau}) / (\text{yRatio}) - 1);
\]

// the factor to be applied to as a result of the quantile
if (\text{tau} \leq 0.5)
\[
\text{rateCorrection} \leftarrow (1 - \text{tau}) / \text{tau} / (3 - 4 \cdot \text{tau});
\]
else
\[
\text{rateCorrection} \leftarrow \text{tau} / (1 - \text{tau}) / (4 \cdot \text{tau} - 1);
\]

\}

parameters {
  \text{real} \text{alpha}; // intercept
  \text{vector}[K] \text{beta}; // coefficients for predictors
}

transformed parameters {
  \text{real<lower=0,upper=1>} \text{p}[N];

  // to remove correlation, we shift alpha with alphaBase and because of the center
  // the alphaOut to be used in predictions go into alphaOut
  \text{real} \text{alphaOut};

  // to remove correlation, we keep the beta on the same scale, independent of tau
  // the beta to be used in predictions go into betaOut
  \text{vector}[K] \text{betaOut};

  \text{alphaOut} \leftarrow \text{alphaBase} + \text{alpha} - \text{center} \cdot \text{betaOut};

  // alphaOut and betaOut are for using on the real scale
  \text{betaOut} \leftarrow \text{beta} \cdot \text{rateCorrection};

  \text{alphaOut} \leftarrow \text{alphaBase} + \text{alpha} - \text{center} \cdot \text{betaOut};

  \}

// apply the Asymmetric Logistic CDF
\text{for} (n \text{ in } 1:N) \{
  \text{real} u;
  u \leftarrow -(\text{alphaOut} + \text{x}[n] \cdot \text{betaOut});
  \text{if} (u \leq 0)
  \text{p}[n] \leftarrow 1 - 2 \cdot \text{tau} / (1 + \exp(-2 \cdot (1 - \text{tau}) \cdot u));
  \text{else}
  \text{p}[n] \leftarrow 2 \cdot (1 - \text{tau}) / (1 + \exp(2 \cdot \text{tau} \cdot u));
\}

\}

model {
  \text{alpha} \sim \text{cauchy}(0, \text{priorScaleForIntercept});

  \text{for} (k \text{ in } 1:K) \{
    \text{beta}[k] \sim \text{student_t}(\text{priorDf}, 0, \text{priorScale} / \text{scale}[k]);
  \}
Appendix B. Stan model code

y ~ bernoulli(p);
}

B.4 alra

data {
  int<lower=0> N; // number of data items
  int<lower=0> K; // number of predictors
  matrix[N,K] x; // predictor matrix
  int<lower=0,upper=1> y[N];
  real<lower=0> priorScale;
  int<lower=1> priorDf;
  int<lower=1> shapeTau[2];
  real<lower=0> priorScaleForIntercept;
  row_vector[K] center;
  vector[K] scale;
}

transformed data {
  // helper variables
  real yRatio;

  # 1.0 in denominator to avoid implicit rounding to integer
  yRatio <- sum(y)/(1.0*N);
}

parameters {
  real alpha; // intercept
  vector[K] beta; // coefficients for predictors
  real<lower=0,upper=1> tau;
}

transformed parameters {
  real<lower=0,upper=1> p[N];
  // to remove correlation, we keep the beta on the same scale, independent of tau
  // the beta to be used in predictions go into betaOut
  vector[K] betaOut;
  real<lower=0> rateCorrection;
  real alphaBase;
  real alphaOut;

  // the factor to be applied to as a result of the quantile
  if (tau <= 0.5) rateCorrection <- (1-tau)/tau/(3-4*tau);
  else rateCorrection <- tau/(1-tau)/(4*tau-1);
// alphaOut and betaOut are for using on the real scale
betaOut <- beta*rateCorrection;

if (1-yRatio <= tau)
    alphaBase <- 0.5/(1-tau) * log(2*tau/(1-yRatio)-1);
else
    alphaBase <- -0.5/tau * log(2*(1-tau)/(yRatio)-1);

alphaOut <- alphaBase + alpha - center * betaOut;

// apply the Asymmetric Logistic CDF
for (n in 1:N) {
    real u;
    u <- -(alphaOut + x[n] * betaOut);
    if (u <= 0)
        p[n] <- 1 - 2*tau/(1+exp(-2*(1-tau)*u));
    else
        p[n] <- 2*(1-tau)/(1+exp(2*tau*u));
}

model {
    tau ~ beta(shapeTau[1], shapeTau[2]);
    alpha ~ cauchy(0, priorScaleForIntercept);
    for (k in 1:K) {
        beta[k] ~ student_t(priorDf, 0, priorScale/scale[k]);
    }
    y ~ bernoulli(p);
}

B.5 splogit

data {
    int<lower=0> N; // number of data items
    int<lower=0> K; // number of predictors
    matrix[N,K] x; // predictor matrix
    int<lower=0,upper=1> y[N];
    real<lower=0> priorScale;
    int<lower=1> priorDf;
    real<lower=0> priorScaleForIntercept;
    real<lower=0> priorScaleForLogr;
    row_vector[K] center;
    vector[K] scale;
Appendix B. Stan model code

```stan
parameters {
  real<lower=0> r;
  real alpha; // intercept
  vector[K] beta; // coefficients for predictors
}
transformed parameters {
  real<lower=0> rateCorrection;
  // the alphaOut to be used in predictions go into alphaOut
  real alphaOut;
  // to remove correlation, we keep the beta on the same scale, independent of r
  // the betas to be used in predictions go into betaOut
  vector[K] betaOut;

  real<lower=0,upper=1> p[N];

  // the factor to be applied to as a result of the power r
  if (r <= 1) rateCorrection <- 0.5/(1-pow(0.5, 1/r));
  else rateCorrection <- 0.5/(1-pow(0.5,r));

  // alphaOut and betaOut are for using on the real scale
  betaOut <- beta*rateCorrection;
  alphaOut <- alpha - center * betaOut;

  // apply the Symmetric Power Logistic
  for (n in 1:N) {
    if (r <= 1) p[n] <- pow(logistic_cdf((alphaOut + x[n] * betaOut)/r, 0, 1), r);
    else p[n] <- 1 - pow(logistic_cdf(-r*(alphaOut + x[n] * betaOut), 0, 1), 1/r);
  }
}
model {
  r ~ lognormal(0,priorScaleForLogr);
  alpha ~ cauchy(0, priorScaleForIntercept);
  for (k in 1:K) {
    beta[k] ~ student_t(priorDf, 0, priorScale/scale[k]);
  }
  y ~ bernoulli(p);
}
```
Bibliography


