Stefan Embacher, BSc

Regularized Ordinal Regression Applied on Team Performance Data

MASTER'S THESIS

to achieve the university degree of

Diplom-Ingenieur

Master's degree programme: Mathematics

submitted to

Graz University of Technology

Supervisor

Herwig Friedl, Ao. Univ.-Prof. Dipl.-Ing. Dr.techn
Institute for Statistics

Graz, March 2021

AFFIDAVIT

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master thesis.

Date

Signature

DANKSAGUNG

An dieser Stelle möchte ich mich bei einigen Personen bedanken, die direkt oder indirekt an der Erstellung dieser Arbeit beteiligt waren.

Allen voran möchte ich mich bei meiner Familie bedanken, die mir stets das größte mögliche Vertrauen entgegenbringt, mich in allen meinen Entscheidungen unterstützt und mir trotz allem genug Raum und Zeit gibt um aus meinen Fehlern zu lernen.

Außerdem gebührt mein großer Dank Professor Dipl.-Ing. Dr.techn. Herwig Friedl, der mit seinen Anregungen, stets freundlicher Kritik und vielen hilfreichen und geduldigen Antworten auf meine Fragen zu großen Teilen zur erfolgreichen Erstellung dieser Arbeit beigetragen hat. Des Weiteren möchte ich mich bei Hannes Biedermann bedanken, der mir die Möglichkeit gegeben hat die Konzepte dieser Arbeit auf Basis von Daten meines Herzensvereins zu entwickeln und damit entscheidend zum Wiederfinden meiner Freude an der Mathematik beigetragen hat.

Abschließend möchte ich mich bei meinen Freunden bedanken die mich immer ermutigen und stets Freude und Spaß in mein Leben bringen, allen voran bei meinen Studienkollegen Simon und besonders Flo. Ich bin sehr dankbar für die fachliche und persönliche Unterstützung in den letzten Jahren.

ABSTRACT

This thesis focuses on ordinal regression models, which extend multinomial response models by taking the ordinal structure of the response into account. Besides introducing the cumulative logit model, which is a prominent ordinal regression model, we discuss regularization methods. The presented regularization methods deal with the problems of multi-collinearity and overfitting resulting by a large number of available explanatory variables compared to the sample size. Despite being used for ordinal regression they could be applied to many other models as well, since they base on the idea to penalize the log-likelihood function. While ordinal regression models allow a great variety of applications, we discuss the developed theoretical concepts on a real data example in sports. We therefore model the match outcome as an ordinal variable and compare the different regularization methods with the unregularized ordinal regression model.

ZUSAMMENFASSUNG

Diese Arbeit konzentriert sich auf ordinale Regressionsmodelle, welche die multinomialen Response Modelle erweitern in dem sie die ordinale Struktur der Response berücksichtigen. Neben der Einführung des kumulativen Logit Modells, ein bekanntes ordinales Regressionsmodell, werden auch Regularisierungsmethoden diskutiert. Diese Regularisierungsmethoden adressieren die Probleme die durch Multi-Kollinearität und Overfitting auftreten, welche wiederum durch eine große Anzahl an möglichen Prädiktoren verglichen mit der verfügbaren Stichprobengröße entstehen. Da sie auf der Idee beruhen einen Strafterm in die log-likelihood Funktion hinzuzufügen, könnten sie, obwohl sie in dieser Arbeit nur auf ordinale Regressionsmodelle angewendet werden, auch auf viele andere Modellklassen angewendet werden. Während ordinale Regressionsmodelle eine große Anzahl an Anwendungsfeldern ermöglichen, werden wir die entwickelten theoretischen Konzepte auf ein Datenbeispiel aus dem Sportbereich anwenden. Daher modellieren wir den Ausgang eines Spiels als ordinale Variable und vergleichen die verschiedenen Regularisierungsmethoden mit den Ergebnissen des unregularisierten Modells.

Contents

1 Introduction

In many fields of studies categorical data and the corresponding methods are indispensable. For instance, in social studies numberless surveys measure attitude and opinion with a response of the possible form (totally disagree, disagree, neutral, agree, totally agree). Categorical variables are also important in the medical sector, where we could measure the severity of an injury, pain or the benefits of a treatment. Even though these are two obvious examples, they are not exclusive, allowing a great variety of applications. While there are different types of categorical variables we will focus on the ones with an underlying ordinal structure and discuss models which explicitly take this ordinal structure into account.

When we face samples where the number of possible explanatory variables is large compared to the number of observations, models might tend to overfit. We therefore introduce and discuss different regularization techniques, which all base on the idea to add a penalty term to the log-likelihood function when finding the maximum likelihood estimator.

In Chapter 2 we start with an introductory part, shortly describing linear regression models, extend them to the class of generalized linear models and discuss logistic regression, log-linear Poisson regression and multinomial response models as explicit examples. Section 2.3 provides some details on possible model selection criteria. Chapter 3 then introduces the cumulative logit model, which accounts for the ordinal structure of the response variable, followed by the derivation of the corresponding (log-)likelihood function, all necessary derivatives and a discussion of the implementation in R (R Core Team, 2019) in Chapter 4. Chapter 5 addresses the problems of multi-collinearity and overfitting, deals with regularization methods, ridge regression and lasso as prominent examples and provides details on how to choose the tuning parameters in the penalized log-likelihood function. To evaluate the introduced ordinal regression models we then describe three goodness of fit tests, found in Chapter 6.

Finally, in Chapter 7, we apply the developed models on a real data example, where the available data contains performance information on team basis from two ice hockey seasons. In Section 7.1 we model the expected number of goals scored by both teams using a log-linear Poisson model, discuss how this can be used to predict the outcome of a game and evaluate the performance of the fitted model. In Section 7.2 we then model the match outcome directly as an ordinal variable, applying the developed theoretical concepts, compare the fitting procedure for the unregularized version and different regularization methods and discuss possible adjustments. Lastly we discuss the results for ordinal regression and compare them to the results obtained by the log-linear Poisson model.

2 Introduction to Linear Models

In this chapter we will introduce linear regression models, their extension to generalized linear models and discuss their form for different types of response variables. Additionally, we will discuss how to perform model selection, introduce the most prominent information criteria and finally discuss how to evaluate models. All given details are only an introduction to this broad topic, serving as a foundation to what we will develop in the later chapters. Therefore the given theory is only a short excerpt and might be studied in more detail in McCullagh and Nelder (1989) or Agresti (2002). The mentioned references serve as a basis for our short introduction to linear models and the related topics.

2.1 Linear Regression

The main idea of a linear model is quite simple. We assume that the expected value of a variable Y, also known as response variable or dependent variable, can be written as a function of variables $x_0, x_1, \ldots, x_{p-1}$. These $x_1, \ldots, x_{p-1}$ are called covariates, explanatory variables, predictor variables or predictors. Additionally, the variable $x_0 = 1$ corresponds to an intercept (respectively the parameter β_0).

In ordinary linear regression we assume the response variables to be independently normal distributed with mean $\boldsymbol{x}_i^\intercal \boldsymbol{\beta}$ and variance σ^2. This means that we can model the expected value of the response variable by

$$\mathbb{E}[Y_i] = \sum_{j=0}^{p-1} \beta_j x_{ij}.$$

If we now assume that we have n observations of a response, which are modelled by $\boldsymbol{y} = (y_1, \ldots, y_n)^\intercal$ with all components being independent and that we know the values of the corresponding explanatory variables which are described by the so called design matrix $\boldsymbol{X}$, where each row contains all explanatory variables regarding the respective observation,

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{pmatrix},$$

we can write the linear regression model in matrix form

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}.$$

Here $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\intercal$ denotes a vector of independent and identically (iid) normally distributed unobservable error terms and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})^\intercal$ the vector of unknown parameters. We therefore want to find estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{p-1}$ of those parameters based on our observed sample. With these parameter estimates we can then estimate the expected value of Y_i, the so called fitted value $\hat{\mu}_i = \boldsymbol{x_i}^\intercal \hat{\boldsymbol{\beta}}$. The residuals are the difference between the variable y_i and the corresponding fitted value, i.e. $r_i = y_i - \hat{\mu}_i$.

The goal is now to find the best estimates for the parameters $\beta_0, \ldots, \beta_{p-1}$, which we get by the least square estimator $\hat{\boldsymbol{\beta}}$, which minimizes the sum of squared errors

$$\mathrm{SSE}(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X\beta})^\intercal(\boldsymbol{y} - \boldsymbol{X\beta}).$$

Under sufficient regularity assumptions we can calculate the least squares estimator explicitly by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal \boldsymbol{y}.$$

Since we assume normality of the response, the log-likelihood function is given by

$$\begin{aligned} \log f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i(\boldsymbol{\beta}))^2 \\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\mathrm{SSE}(\beta). \end{aligned}$$

This on the other hand implies that the least squares estimator $\hat{\boldsymbol{\beta}}$, which minimizes $\mathrm{SSE}(\boldsymbol{\beta})$, also maximizes the log-likelihood function with respect to $\boldsymbol{\beta}$. Therefore the least squares estimator is equivalent to the maximum likelihood estimator for $\boldsymbol{\beta}$. This is an especially nice property since when using generalized linear models we estimate the parameters by maximizing the respective log-likelihood function.

A value which is often discussed when dealing with linear regression models is the so called coefficient of determination R^2. R^2 denotes the proportion of the variance of the dependent variable, which can be explained by the covariates and is defined by

$$R^2 = \frac{\mathrm{SSR}(\hat{\boldsymbol{\beta}})}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}(\hat{\boldsymbol{\beta}})}{\mathrm{SST}}, \quad 0 \leq R^2 \leq 1.$$

This is based on the fact that the sum of squared errors $\mathrm{SSE}(\hat{\boldsymbol{\beta}})$ and the regression sum of squares $\mathrm{SSR}(\hat{\boldsymbol{\beta}})$ sum up to the total sum of squares SST, which is independent of the parameter estimates,

$$\begin{aligned}\mathrm{SSR}(\hat{\boldsymbol{\beta}}) + \mathrm{SSE}(\hat{\boldsymbol{\beta}}) &= \sum_{i=1}^{n} (\hat{\mu}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \bar{y})^2 = \mathrm{SST}.\end{aligned}$$

$R^2 = 1$ denotes perfect fit, implying $y_i = \hat{\mu}_i$ for all $i = 1, \dots, n$ and therefore $\mathrm{SSE}(\hat{\boldsymbol{\beta}}) = 0$. $R^2 = 0$ means that there is no linear dependency between the covariates and the response variable. Therefore $\boldsymbol{\beta} = \mathbf{0}$, which implies $\hat{\mu}_i = \bar{y}$ for all i and $\mathrm{SSR}(\hat{\boldsymbol{\beta}}) = 0$. For all other cases R^2 increases with every predictor added to the model. Therefore an often used modification is the so called adjusted R^2, denoted by R^2_{adj} and given by

$$R^2_{\mathrm{adj}} = 1 - \frac{\mathrm{SSE}(\hat{\boldsymbol{\beta}})/(n-p)}{\mathrm{SST}/(n-1)},$$

with $p - 1$ predictors in the model.

2.2 Generalized Linear Model

For several reasons it can be necessary to relax the linear regression assumptions and extend it to the so called generalized linear model. Again we face a vector $\boldsymbol{y} = (y_1, \dots, y_n)^\intercal$ of independent responses, with $\mathbb{E}(Y_i) = \mu_i$ and $var(Y_i) = a_i \phi V(\mu_i)$.

The product $a_i\phi$ is the so called dispersion, where ϕ is the dispersion parameter and a_i a known weight. Again we have a vector of explanatory variables to each y_i collected in the design matrix $\boldsymbol{X}$. The main difference to the linear regression setting is, that we do not assume the responses to be normally distributed, but allow for a much wider class of distributions. Therefore we connect the expected value of Y_i to a linear predictor $\eta_i = \boldsymbol{x_i^\intercal \beta}$ via a link function $h(\mu_i) = \eta_i$. Here the parameters $\boldsymbol{\beta}$ are estimated by the maximum likelihood estimator. However, since the resulting system of equations is not linear in general it is solved iteratively. There are several different methods available, with the Newton-Raphson method as prominent example. Further, the scaled deviance, a generalization of the sum of squared errors, is given by

$$\frac{1}{\phi}D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = -2\big(\ell(\hat{\boldsymbol{\mu}}|\boldsymbol{y}) - \ell(\boldsymbol{y}|\boldsymbol{y})\big),$$

where the second term denotes the log-likelihood of the saturated model. The deviance can be used to evaluate the goodness of fit, where large values indicate a lack of fit.

2.2.1 Logistic Regression

A special case of a generalized linear model is the so called logistic regression. In logistic regression we face a binary response variable, i.e. $Y_i \in \{0, 1\}$. We therefore set the probabilities

$$\mathbb{P}(Y_i = 0) = 1 - \pi_i \qquad \mathbb{P}(Y_i = 1) = \pi_i,$$

which can be seen as the probabilities for failure and success respectively. As nicely described by McCullagh and Nelder (1989), we have several different possibilities for the link function h. Two of the most prominent ones are the logit link and the probit link.

The logit link, is the logarithm of the odds of π,

$$h_{logit}(\pi) = \log\Big(\frac{\pi}{1-\pi}\Big).$$

We can therefore model the probability π by using the inverse function of the logit link, which corresponds to the cumulative distribution function of the logistic distribution

$$\pi = \frac{\exp(\boldsymbol{x}^\intercal \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}^\intercal \boldsymbol{\beta})}.$$

The second one, the probit link, which is the inverse of the cumulative distribution function of the standard normal distribution, is given by

$$h_{probit}(\pi) = \Phi^{-1}(\pi).$$

Therefore the probability can be modelled as

$$\pi = \Phi(\boldsymbol{x}^\intercal \boldsymbol{\beta}).$$

Both functions are symmetric, i.e. $h(\pi) = -h(1-\pi)$ and are in general very similar, while the logistic distribution has heavier tails compared to the normal distribution.

2.2.2 Log-linear Poisson Model

A second very prominent class of generalized linear models are the so called log-linear Poisson models. In this case we assume the dependent variable Y_i to be some form of counting, for example the number of goals scored in a game. It is common practice, that we then assume the variables Y_i to be Poisson distributed with intensity parameter λ_i and therefore $\mathbb{P}(Y_i = y|\lambda_i) = \frac{\lambda_i^y}{y!}\exp(-\lambda_i)$.

For the Poisson distribution the expected value is equal to the variance, $\mathbb{E}(Y_i) = \lambda_i = var(Y_i)$. Therefore the dispersion parameter ϕ is 1. The logarithm is the usual link function in Poisson regression, this is why it is also called log-linear Poisson model. We therefore result in a model of the form

$$\log(\lambda_i) = \eta_i = \boldsymbol{x_i}^\intercal \boldsymbol{\beta}.$$

2.2.3 Multinomial Response Models

In comparison to the logistic regression setting where we faced a binary response variable it is also possible to face a categorical response variable, which might take values in c different categories. There are several familiar examples for categorical response variables, like blood type (0, A, B, AB) or measurements of agreement (totally disagree, disagree, neutral, agree, totally agree) or physical and mental well-being.

As described by McCullagh and Nelder (1989) we can distinguish between three major types of scales underlying the categorical variables, nominal, ordinal and interval scale. Using a nominal scale implies that the categories do not show a structure and they are exchangeable. From the mentioned examples this would relate to the blood type. In an ordinal scale we assume the categories to be ordered, for example as a measurement of agreement, however we can not discuss the distance between categories. Using an interval scale, in which categories are ordered and we attach a score or numerical label to the categories, allows for a discussion of the distance between those categories. These scores can be for example the category mean or median. A special case of all mentioned, is the case $c = 2$ where we result in a binary measurement and therefore in logistic regression.

Since the main focus of the following chapters will be drawn to ordinal response variables which are therefore discussed in more detail, we will only shortly describe a model for nominal response variables at this point. We therefore assume a sample of independent multinomial responses $y_1, \dots, y_n$, where $y_i = (y_{i1}, \dots, y_{ic})$ and $\sum_{j=1}^{c} y_{ij}$ is fixed for each i. Let $\pi_i = (\pi_{i1}, \dots, \pi_{ic})$ denote the corresponding vector of categorical probabilities. In comparison to the binary logistic case, where we model the logarithm of the odds linearly, i.e. $\log \frac{\pi_{i1}}{\pi_{i2}} = \eta_i$, we can choose out of c possible reference categories in the multinomial case. However, it is common practice to choose the first category as reference category. Note, that by the structure of a nominal scale this category can be arbitrarily exchanged. We therefore result in a model of the form

$$\log \frac{\pi_{ij}}{\pi_{i1}} = \eta_{ij} = \boldsymbol{x_i}^\intercal \boldsymbol{\beta_j}, \quad j = 2, \dots, c.$$

In this model, the parameter vector $\boldsymbol{\beta_j}$ is dependent on the category, to model the effects of $\boldsymbol{x_i}$ on π_{ij}. Clearly, we restrict the sum of the probabilities to be one and therefore get the following expressions for the categorical probabilities

$$\begin{aligned}\pi_{i1} &= \frac{1}{\sum_{k=1}^{c}\exp(\eta_{ik})}\\ \pi_{ij} &= \frac{\exp(\eta_{ij})}{\sum_{k=1}^{c}\exp(\eta_{ik})}, \quad j = 2, \ldots, c.\end{aligned}$$

The described multinomial response model treats all categorical data as nominal. Therefore if we face an ordinal response variable, we would loose some information contained in the data, namely the ordering. As described by Agresti (2010), many advantages can be obtained by treating an ordered response variable as ordinal instead of nominal. They might apply in settings where the standard nominal models have too many parameters, extend the variety of possible models with simpler interpretations and can use measures similar to those used for quantitative variables, as correlations or slopes.

2.3 Model Selection

In model selection we have two essential tasks, the first one is the choice of the model type and the second one is to decide which subset of the available predictor variables is the best choice to compromise between a satisfying predictive performance and the simplicity of the model. As Agresti (2002) formulates it nicely: "The model should be complex enough to fit the data well. On the other hand, it should be simple to interpret, smoothing rather than overfitting the data."

Two of the most prominent model selection criteria are the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). It holds for both, that we want to minimize the respective value. Both base on the same idea, namely to measure the goodness of fit by the log-likelihood and to punish the number of predictors included in the model. The AIC was introduced by Akaike (1974),

$$\text{AIC} = 2[-\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) + p],$$

where $\hat{\boldsymbol{\theta}}$ denotes the MLE of the parameters $\boldsymbol{\theta}$ estimated in the model and p denotes the number of variables included.

Schwarz (1978) introduced the BIC, which has a very similar appearance as the AIC,

$$\text{BIC} = -2\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) + p \log n.$$

The main difference is the penalization of the number of parameters which is $\log n$ instead of 2. This on the other hand means that if $n > \exp(2) = 7.38$ the optimal model chosen by the BIC would not be more complex, compared to the one chosen by the AIC.

Hurvich and Tsai (1989) introduced a bias corrected form of the AIC, the so called corrected Akaike information criterion AICc, which is useful if facing a small sample size or a large fraction of fitted parameters in relation to the sample size,

$$\text{AICc} = -2\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) + 2p + \frac{2p(p+1)}{n-p-1} = \text{AIC} + \frac{2p(p+1)}{n-p-1}.$$

It is obvious that for $n \to \infty$, the AICc converges to the AIC.

When fitting models, in which the response variables are categorical we have several different possibilities to assess the goodness of fit. The first are Pearson χ^2 and deviance tests. The Pearson χ^2 test statistic is given by

$$\chi^2 = \sum_{j=1}^{c} \frac{O_j - E_j}{E_j},$$

where c denotes the number of response categories, O_j the number of observations in category j and E_j the expected number of observations in category j. The deviance goodness of fit test statistic is based on the deviance given in the previous section. For categorical data it is of the form

$$D = 2\sum_{j=1}^{c} O_j \log \frac{O_j}{E_j},$$

both test statistics are compared with the χ^2-distribution with $(c-1)$ degrees of freedom.

When predicting categorical variables, we actually predict the probability for each category. For comparability reasons we are going to choose the category with the highest predicted probability as predicted category. If we want to assess how good we perform when predicting categories we have several possibilities which essentially only distinguish by the choice of the training and the test sets. To evaluate the performance, we assess how often we predicted the correct category and are going to use three different methods, namely in sample, jackknife and data splitting. When assessing the in sample accuracy of a model, we use all available observations as training set and as test set. This has the disadvantage that the observation we are testing for was also part of the model fit and therefore had influence on the estimates. In contrast to jackknife, where we use all observations expect one to train the model and then assess the performance of the one left out. This is repeated for all observations separately.

When we are talking about data splitting, we mean that for a given probability p and n observations, we generate n Bernoulli random variables with success probability p. Each Bernoulli variable indicates whether the respective observation is in the training or in the test set. With the resulting training set we fit the model and then assess the accuracy of the test set. This procedure is repeated a presetted number of times N and is finally averaged to give the respective overall accuracy. The main difference to cross-validation as presented in Section 5.3 is that in cross-validation we build the folds once and then assess the accuracy for each fold, while in data splitting we only assess the accuracy for one test set in each loop.

3 Cumulative Logit Models

The goal of the following chapter is to find a model structure that is able to use the ordinal structure of the response variable while keeping already known methods available. We therefore, according to Agresti (2010), first define cumulative logits and a cumulative logit model.

Let Y be a response variable with c outcome categories, while $\pi_1, \dots, \pi_c$ describe the respective probabilities. The **cumulative logits** are defined as

$$\begin{aligned} \operatorname{logit}[\mathbb{P}(Y \leq j)] &= \log\left[\frac{\mathbb{P}(Y \leq j)}{1 - \mathbb{P}(Y \leq j)}\right] \\ &= \log\left[\frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_c}\right], \quad j = 1, \dots, c-1. \end{aligned} \tag{3.1}$$

These logits generalise the ordinary binary logit of the response outcome split into two results, $(Y \leq j)$ and $(Y > j)$. Each cumulative logit uses all c response categories. The cumulative logits are not defined for $j = c$ because $\mathbb{P}(Y \leq c) = 1$ and we would therefore divide by zero. Ordinal models use the $(c-1)$ logits to fit a single model. This approach may result in easier to interpret models than fitting separate models. There are some other logits, which should be mentioned because they allow for different assumptions on the structure of the ordinal response. For more details refer to Agresti (2002) or Agresti (2010).

- The **adjacent-categories logits**

$$\log\left(\frac{\pi_j}{\pi_{j+1}}\right), \quad j = 1, \dots, c-1.$$

- The **baseline-category logits**

$$\log\left(\frac{\pi_j}{\pi_c}\right), \quad j = 1, \dots, c-1.$$

- The **continuation-ratio logits**

$$\log\left(\frac{\pi_j}{\pi_{j+1}+\cdots+\pi_c}\right), \quad j=1,\ldots,c-1.$$

Suppose there are n observations, let y_i denote the outcome category and let $\boldsymbol{x}_i$ denote the vector of the corresponding explanatory variables of observation i. Our **cumulative logit model** then has the form

$$\operatorname{logit}[\mathbb{P}(Y_i \leq j)] = \alpha_j + \boldsymbol{x}_i^\intercal \boldsymbol{\beta} = \alpha_j + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots\,, \tag{3.2}$$

where $\boldsymbol{\beta}$ is a vector of parameters describing the effects of the explanatory variables. The equivalent model expression for the cumulative probabilities is

$$\mathbb{P}(Y_i \leq j) = \frac{\exp(\alpha_j + \boldsymbol{x}_i^\intercal \boldsymbol{\beta})}{1+\exp(\alpha_j + \boldsymbol{x}_i^\intercal \boldsymbol{\beta})}, \quad j=1,\ldots,c-1.$$

As a result the category probabilities are,

$$\mathbb{P}(Y_i = j) = \frac{\exp(\alpha_j + \boldsymbol{x}_i^\intercal \boldsymbol{\beta})}{1+\exp(\alpha_j + \boldsymbol{x}_i^\intercal \boldsymbol{\beta})} - \frac{\exp(\alpha_{j-1} + \boldsymbol{x}_i^\intercal \boldsymbol{\beta})}{1+\exp(\alpha_{j-1} + \boldsymbol{x}_i^\intercal \boldsymbol{\beta})}$$

with $\alpha_0 = -\infty$ and $\alpha_c = \infty$. Thompson and Baker (1981) called this link function for the category probabilities **composite link function**. To simplify notation and unless we need to refer to specific subjects or to specific values of the explanatory variables, we replace $\mathbb{P}(Y_i \leq j|\boldsymbol{x}_i)$ by $\mathbb{P}(Y \leq j)$. Keeping in mind that in the model this is a conditional probability at each fixed value of the explanatory variables.

Proposition 3.0.1. *The α_j are increasing in j.*

Proof. If $j<k$, then by ordering $\mathbb{P}(Y \leq j) \leq \mathbb{P}(Y \leq k)$ holds. Therefore,

$$\frac{\exp(\alpha_j + \boldsymbol{x}^\intercal \boldsymbol{\beta})}{1+\exp(\alpha_j + \boldsymbol{x}^\intercal \boldsymbol{\beta})} \leq \frac{\exp(\alpha_k + \boldsymbol{x}^\intercal \boldsymbol{\beta})}{1+\exp(\alpha_k + \boldsymbol{x}^\intercal \boldsymbol{\beta})},$$

after doing some simple calculations this results in $\exp(\alpha_j + \boldsymbol{x}^\intercal \boldsymbol{\beta}) \leq \exp(\alpha_k + \boldsymbol{x}^\intercal \boldsymbol{\beta})$. By monotonicity and positivity of the exponential function this directly implies $\alpha_j \leq \alpha_k$.

The α_j are strictly increasing in j, if $\pi_j \neq 0 \quad \forall j \in \{1,\ldots,c\}$. □

3.1 Single Continuous Predictor

Lets first consider the model described in (3.2), with a single continuous predictor x

$$\text{logit}[\mathbb{P}(Y \leq j)] = \alpha_j + \beta x, \quad j = 1, \ldots, c-1.$$

Figure 3.1 shows the model for $c = 4$ outcome categories of Y . For fixed j, the response curve is an ordinary logistic regression curve for a binary response with outcomes $(Y \leq j)$ and $(Y > j)$. The common effect β for the three cumulative logits implies that the three curves for the cumulative probabilities for $j = 1, 2, 3$ have the same shape. At any fixed x value, the curves have the same ordering as the cumulative probabilities, the one for $\mathbb{P}(Y \leq 1)$ being lowest. If $\beta = 0$, the graph of $\mathbb{P}(Y \leq j)$ as a function of x is a horizontal line for each j. Then, Y is statistically independent of x.

Theorem 3.1.1. For $j < k$, the curve for $\mathbb{P}(Y \leq k)$ is the curve for $\mathbb{P}(Y \leq j)$ shifted by $(\alpha_k - \alpha_j)/\beta$ units in the x direction,

$$\mathbb{P}[Y \leq k|X = x] = \mathbb{P}[Y \leq j|X = x + (\alpha_k - \alpha_j)/\beta].$$

Proof. We know that

$$\mathbb{P}[Y \leq k|X = x] = \frac{\exp(\alpha_k + \beta x)}{1 + \exp(\alpha_k + \beta x)}.$$

By

$$\begin{aligned}\exp(\alpha_k + \beta x) &= \exp\left(\alpha_k + \beta x + \beta\frac{\alpha_j - \alpha_j}{\beta}\right) = \exp\left(\beta\frac{\alpha_k}{\beta} + \beta x + \alpha_j - \beta\frac{\alpha_j}{\beta}\right) \\ &= \exp\left(\alpha_j + \beta(x + \frac{\alpha_k - \alpha_j}{\beta})\right),\end{aligned}$$

it holds that

$$\mathbb{P}[Y \leq k|X = x] = \frac{\exp\left(\alpha_j + \beta(x + \frac{\alpha_k - \alpha_j}{\beta})\right)}{1 + \exp\left(\alpha_j + \beta(x + \frac{\alpha_k - \alpha_j}{\beta})\right)} = \mathbb{P}[Y \leq j|X = x + (\alpha_k - \alpha_j)/\beta].$$

□

Figure 3.2 shows the corresponding curves for the category probabilities. For both, Figure 3.1 and Figure 3.2, β was chosen to be positive. If β would be smaller than 0, the curves in Figure 3.1 would descend and the labels in Figure 3.2 would reverse order.

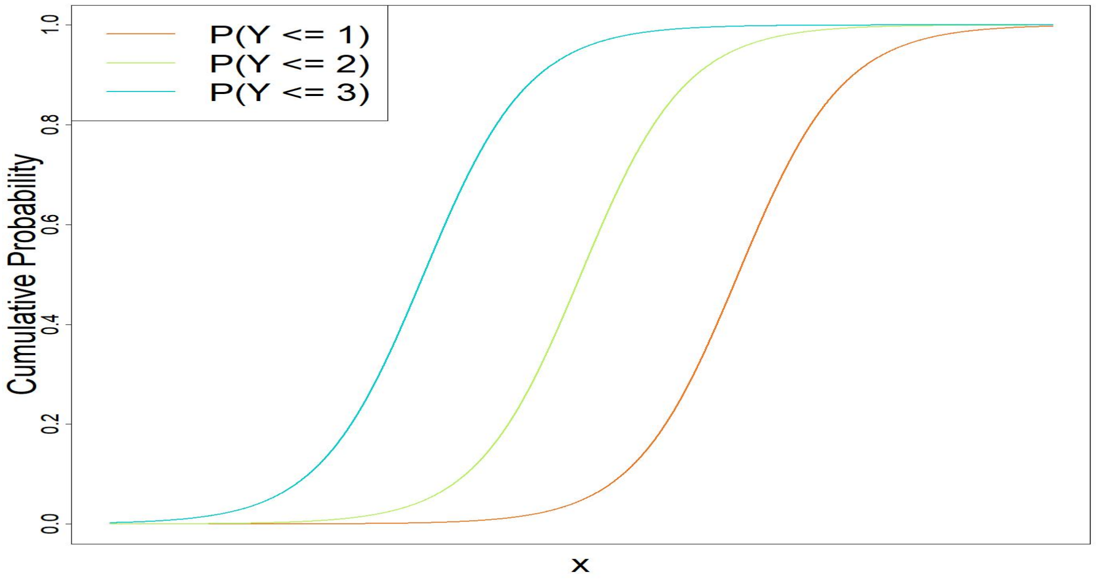

Figure 3.1: Cumulative Probabilities for a single continous predictor

3.2 Alternative Parametrization

Often, the cumulative logit model as shown in (3.2) is expressed with a negative sign in the parametrization, i.e.

$$\operatorname{logit}[\mathbb{P}(Y_i \leq j)] = \alpha_j - \boldsymbol{x}_i^\intercal \boldsymbol{\beta}, \quad j = 1, \ldots, c-1. \tag{3.3}$$

This parametrization assures a more natural interpretation, keeping the usual directory meaning. An increase in the explanatory variables, increases the probability that Y falls at the high end of the categorical spectrum. Some software use the parametrization as shown in (3.2) and some the one shown in (3.3). Therefore it is necessary to keep this in mind when actively using software to be aware of the interpretation. If not mentioned otherwise, we are going to use the form

$$\operatorname{logit}[\mathbb{P}(Y_i \leq j)] = \alpha_j + \boldsymbol{x}_i^\intercal \boldsymbol{\beta}, \quad j = 1, \ldots, c-1.$$

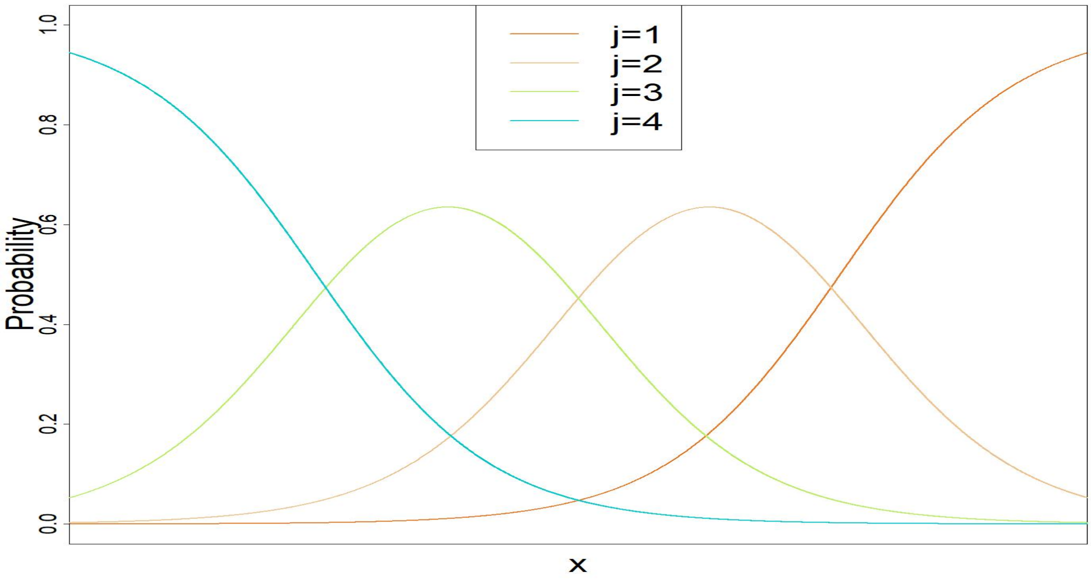

Figure 3.2: Category Probabilities

3.3 Proportional Odds

The model described in (3.2) satisfies

$$\begin{aligned}\operatorname{logit}[\mathbb{P}(Y \leq j|\boldsymbol{x}_1)] - \operatorname{logit}[\mathbb{P}(Y \leq j|\boldsymbol{x}_2)] &= \log\left[\frac{\mathbb{P}(Y \leq j|\boldsymbol{x}_1)/\mathbb{P}(Y > j|\boldsymbol{x}_1)}{\mathbb{P}(Y \leq j|\boldsymbol{x}_2)/\mathbb{P}(Y > j|\boldsymbol{x}_2)}\right] \\ &= (\boldsymbol{x}_1 - \boldsymbol{x}_2)^\intercal \boldsymbol{\beta}.\end{aligned}$$

This means, that the odds of $(Y \leq j)$ at $\boldsymbol{x} = \boldsymbol{x}_1$ are the odds at $\boldsymbol{x} = \boldsymbol{x}_2$ multiplied by $\exp((\boldsymbol{x}_1 - \boldsymbol{x}_2)^\intercal \boldsymbol{\beta})$. The log cumulative odds are proportional to the difference between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. This property is independent of the choice of j. McCullagh (1980) therefore called the model described in (3.2), **proportional odds model**. However, according to Agresti (2010) we will refer to this model as the **proportional odds version of the cumulative logit model**.

> ”The term ordered logit is vague, because there are also other types of logit models for ordinal data [..]. The term proportional odds is also vague, because these other logit models for ordinal data can also have a proportional odds structure.” (Agresti, 2010)

3.4 Latent Variable Motivation

In this section we want to motivate why it is legit to use a common effect $\boldsymbol{\beta}$ for the different cumulative logits. Suppose a latent, continuous variable Y^* is underlying Y and varies around a location parameter η, for example a mean, for which $\eta(\boldsymbol{x}) = \boldsymbol{x}^\intercal\boldsymbol{\beta}$. For the cdf of Y^*, then

$$\mathbb{P}(Y^* \leq y^*|\boldsymbol{x}) = G(y^* - \eta(\boldsymbol{x})).$$

Furthermore suppose $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_c = \infty$ are thresholds on the continuous scale, such that the ordinal variable Y falls in category j being equivalent to the latent variable Y^* taking values in the j-th interval,

$$Y = j \iff \alpha_{j-1} < Y^* \leq \alpha_j.$$

This is visualized in Figure 3.3. For this choice of the latent variable and its thresholds, it holds that

$$\mathbb{P}(Y \leq j|\boldsymbol{x}) = \mathbb{P}(Y^* \leq \alpha_j|\boldsymbol{x}) = G(\alpha_j - \boldsymbol{x}^\intercal\boldsymbol{\beta}).$$

Therefore we obtain the link function G^{-1}, applying to $\mathbb{P}(Y \leq j|\boldsymbol{x})$ to get a linear predictor,

$$G^{-1}[\mathbb{P}(Y \leq j|\boldsymbol{x})] = \alpha_j - \boldsymbol{x}^\intercal\boldsymbol{\beta}.$$

The proportional odds version of the cumulative logit model, with alternative parametrization is obtained if G is the cdf of the standard logistic distribution $G(x) = \frac{\exp(x)}{1+\exp(x)}$, then G^{-1} is the logit link function and therefore implies the model. With the cdf choosen above, we rather result in a model with linear predictor $\alpha_j - \boldsymbol{x}^\intercal\boldsymbol{\beta}$. In practice, it does not make a difference, as long as one is aware of which sign is used and how to interpret the model correctly. Another obvious choice would be that G is the cdf of the standard normal distribution, which would directly lead to the probit model (Anderson and Philips, 1981).

The latent variable motivation clarifies why the choice of the response categories does not make an impact on $\boldsymbol{\beta}$, regardless of how the thresholds $\{\alpha_j\}$ on the continuous scale where chosen. This property makes it possible to compare different studies on the same topic, with differently chosen responses. As an example imagine two studies on pain after a surgery, one with three response categories (pain, exhausted, well) and one with five (great pain, pain, exhausted, well, very well).

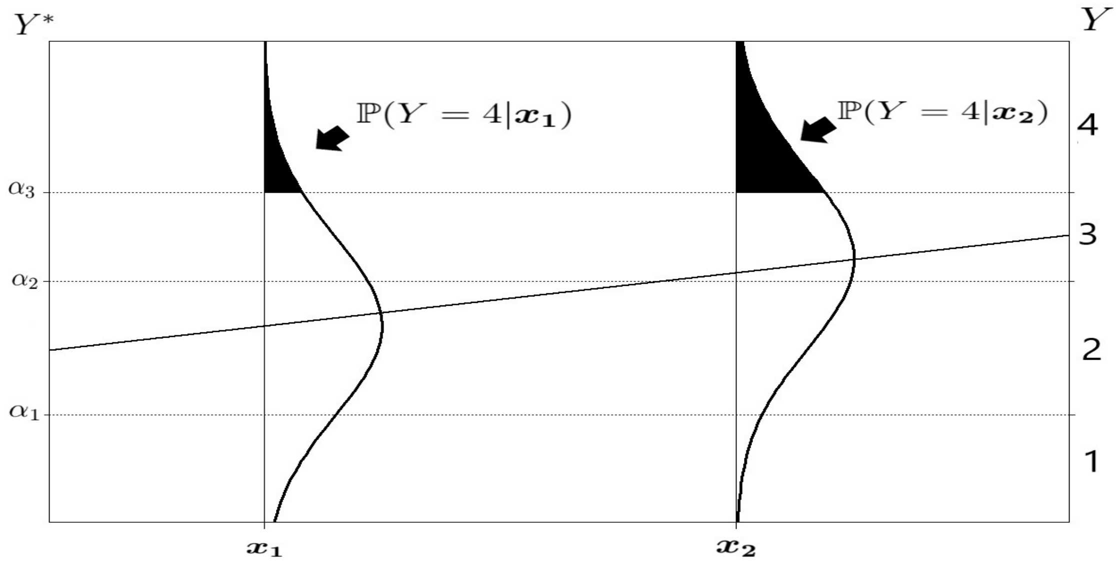

Figure 3.3: Latent variable motivation for 4 categories and an underlying linear regression model for the latent variable (cp. Agresti, 2002)

3.5 Other Cumulative Links

As already mentioned, the type of link functions depends on the choice of the underlying distribution function G. The cumulative logit model results for the choice of the standard logistic distribution, with the logit link function. Agresti (2002) also describes the probit link and the log-log link.

Using the probit link function, gives the so called **cumulative probit model**. It results from the choice of the standard normal cdf Φ for G, being appropriate if the latent variable Y^* is assumed to be normally distributed. In this model, $\boldsymbol{\beta}$ has the interpretation that an one unit increase in x_k corresponds to a β_k increase in $\mathbb{E}(Y^*)$ when keeping all other explanatory variables fixed at the same value. If the error term ϵ is not necessarily in standardized form with variance 1, an one unit increase in x_k corresponds to a β_k standard deviation increase in the expected value of Y^*.

If the latent variable Y^* is assumed to follow a Gumbel distribution, with cdf

$$G(y) = \exp\{-\exp[-(y-a)/b]\},$$

the resulting link function is the so called **complementary log-log link**, since the log-log link applies to the complement of the cumulative probability. This implies a model of the form

$$\log\{-\log[1-\mathbb{P}(Y \leq j)]\} = \alpha_j + \boldsymbol{x}^\intercal \boldsymbol{\beta}.$$

To clarify notation and to have a model to refer to, if we do not want to specify the link function, respectively the assumed distribution of the underlying latent variable, we define a more general form, the so called **cumulative link model**. Let h be an arbitrary link function. We then define the class of cumulative link models by

$$h[\mathbb{P}(Y \leq j)] = \alpha_j + \boldsymbol{x}^\intercal \boldsymbol{\beta}, \quad j = 1, \ldots, c-1. \tag{3.4}$$

This model again links the cumulative probabilities to a linear predictor and the effects of $\boldsymbol{x}$ are the same for each cumulative probability.

3.5.1 Dispersion Effects

As defined by Casella and Berger (2002), a cdf F_{X_1} is called **stochastically greater** than a cdf F_{X_2}, if $F_{X_1}(t) \leq F_{X_2}(t), \quad \forall t$ and $F_{X_1}(t) < F_{X_2}(t)$ for some t. Since the cumulative link models have the same effect $\boldsymbol{\beta}$ for all cumulative probabilities we have a stochastic ordering. This means that for given $\boldsymbol{x}_1$ and given $\boldsymbol{x}_2$ either $\mathbb{P}(Y \leq j|\boldsymbol{x}_1) \leq \mathbb{P}(Y \leq j|\boldsymbol{x}_2)$ for all j or $\mathbb{P}(Y \leq j|\boldsymbol{x}_2) \leq \mathbb{P}(Y \leq j|\boldsymbol{x}_1)$ for all j.

If dispersion changes substantially for different predictor values, this might result in a poor model fit. Agresti (2002) gives the following explanation:

> "Perhaps responses tend to concentrate around the same location but more dispersion occurs at $\boldsymbol{x}_1$ than at $\boldsymbol{x}_2$. Then perhaps $\mathbb{P}(Y \leq j|\boldsymbol{x}_1) > \mathbb{P}(Y \leq j|\boldsymbol{x}_2)$ for small j but $\mathbb{P}(Y \leq j|\boldsymbol{x}_1) < \mathbb{P}(Y \leq j|\boldsymbol{x}_2)$ for large j. In other words, at $\boldsymbol{x}_1$ the responses concentrate more at the extreme categories than at $\boldsymbol{x}_2$."

A **cumulative link model including dispersion effects** is

$$h[\mathbb{P}(Y \leq j|\boldsymbol{x})] = \frac{\alpha_j + \boldsymbol{x}^\intercal \boldsymbol{\beta}}{\exp(\boldsymbol{x}^\intercal \boldsymbol{\gamma})}.$$

As descripted in Section 3.2, it is again possible to change the sign in the linear predictor. The parameter $\boldsymbol{\gamma}$ describes the dispersion's dependence on $\boldsymbol{x}$. The ordinary cumulative link model (3.4) results if $\boldsymbol{\gamma} = \boldsymbol{0}$. The main difference between the ordinary cumulative link model and the cumulative link model including dispersion effects, is that the estimation in the second one becomes more complex, since it is not linear in the parameters anymore.

3.6 Thresholds

So far, we have not discussed restrictions on the thresholds $\boldsymbol{\alpha}$, except for the monotony described in Proposition 3.0.1. We call these unrestricted but ordered thresholds, **flexible thresholds**. Christensen (2018) describes the possibility of **structured thresholds**. Structured thresholds allow for restrictions on the appearance of the thresholds. We model this by a linear function $g(\boldsymbol{\theta}) = J^\intercal \boldsymbol{\theta} = \boldsymbol{\alpha}$, where $\boldsymbol{\theta}$ is the vector of parameters describing the thresholds. For example, when restricting the thresholds to be equidistant, we only need to estimate two parameters, namely the location of the first threshold and the distance between adjacent ones. For $c = 5$ we therefore result in a matrix of the form

$$J_{equidistant} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{pmatrix}.$$

Analogously it is possible to force the thresholds to be symmetric, which would result in a matrix of the form

$$J_{symmetric} = \begin{pmatrix} 0 & -1 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix},$$

again for $c = 5$. For the symmetry restriction it is possible to find more than one appearance of the matrix J. Most available software dealing with cumulative link models, have the option to restrict on the values and appearance of the thresholds. This might be useful in practice to ensure a better interpretability of the corresponding results.

4 Maximum Likelihood Estimation

In order to be able to estimate the parameters of the cumulative link model, we will derive the maximum likelihood equations as described in Agresti (2010) and we will see a regularized Newton-Raphson algorithm with step halving, which is the method of choice in the package `ordinal`, made available by Christensen (2019), used in R (R Core Team, 2019). Certainly, there are different methods available to maximize the likelihood function, since the Newton-Raphson algorithm is the one we are going to use in the application when dealing with unpenalized likelihood functions, we restrict on this algorithm. However, when trying to optimize penalized likelihood functions, which is needed when applying regularization techniques, we will use a different, more complex algorithm. This alternative algorithm is described in Section 5.3.1. In addition we will shortly discuss the possibility of infinite estimates.

4.1 Equations

In the following we will treat cumulative link models as multivariate generalized linear models and therefore assume a multinomial distribution. h describes the link function, which applies to a vector of means $(\pi_1(\boldsymbol{x}_i), \ldots, \pi_c(\boldsymbol{x}_i))$. For $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, c\}$ we define

$$y_{ij} = \begin{cases} 1 & \text{if} \quad y_i = j \\ 0 & \text{otherwise.} \end{cases}$$

Let $G = h^{-1}$ be the inverse link function and assume independent observations, then the likelihood-function is

$$\begin{aligned}\prod_{i=1}^{n}\prod_{j=1}^{c}\pi_j(\boldsymbol{x}_i)^{y_{ij}} &= \prod_{i=1}^{n}\prod_{j=1}^{c}[\mathbb{P}(Y_i \leq j|\boldsymbol{x}_i) - \mathbb{P}(Y_i \leq j-1|\boldsymbol{x}_i)]^{y_{ij}}\\ &= \prod_{i=1}^{n}\prod_{j=1}^{c}[G(\alpha_j + \boldsymbol{x}_i^\intercal\boldsymbol{\beta}) - G(\alpha_{j-1} + \boldsymbol{x}_i^\intercal\boldsymbol{\beta})]^{y_{ij}}\\ &= L(\boldsymbol{\alpha},\boldsymbol{\beta}).\end{aligned}$$

Therefore the log-likelihood function is

$$\log(L(\boldsymbol{\alpha},\boldsymbol{\beta})) = \ell(\boldsymbol{\alpha},\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{j=1}^{c} y_{ij}\log\left[G(\alpha_j + \boldsymbol{x}_i^\intercal\boldsymbol{\beta}) - G(\alpha_{j-1} + \boldsymbol{x}_i^\intercal\boldsymbol{\beta})\right].$$

Let now g denote the derivative of G, the density function of the corresponding cdf. Then, we result in the following equations

$$\frac{\partial\ell}{\partial\beta_k} = \sum_{i=1}^{n}\sum_{j=1}^{c} y_{ij}x_{ik}\frac{g(\alpha_j + \boldsymbol{x}_i^\intercal\boldsymbol{\beta}) - g(\alpha_{j-1} + \boldsymbol{x}_i^\intercal\boldsymbol{\beta})}{G(\alpha_j + \boldsymbol{x}_i^\intercal\boldsymbol{\beta}) - G(\alpha_{j-1} + \boldsymbol{x}_i^\intercal\boldsymbol{\beta})}$$

and

$$\frac{\partial\ell}{\partial\alpha_k} = \sum_{i=1}^{n}\sum_{j=1}^{c} y_{ij}\frac{\delta_{jk}g(\alpha_j + \boldsymbol{x}_i^\intercal\boldsymbol{\beta}) - \delta_{j-1,k}g(\alpha_{j-1} + \boldsymbol{x}_i^\intercal\boldsymbol{\beta})}{G(\alpha_j + \boldsymbol{x}_i^\intercal\boldsymbol{\beta}) - G(\alpha_{j-1} + \boldsymbol{x}_i^\intercal\boldsymbol{\beta})},$$

where δ_{jk} is the Kronecker delta, for which $\delta_{jk} = 1$ if $j = k$ and zero else. For notational purpose we will set $z_{ij} = \alpha_j + \boldsymbol{x}_i^\intercal\boldsymbol{\beta}$, then the second derivatives are

$$\begin{aligned}\frac{\partial^2\ell}{\partial\beta_k\partial\beta_l} &= \sum_{i=1}^{n}\sum_{j=1}^{c} y_{ij}x_{ik}x_{il}\left\{\frac{[G(z_{ij}) - G(z_{i,j-1})][g(z_{i,j-1})z_{i,j-1} - g(z_{ij})z_{ij}]}{[G(z_{ij}) - G(z_{i,j-1})]^2}\right.\\ &\left.-\frac{[g(z_{i,j-1}) - g(z_{ij})]^2}{[G(z_{ij}) - G(z_{i,j-1})]^2}\right\}\\ \frac{\partial^2\ell}{\partial\beta_k\partial\alpha_l} &= \sum_{i=1}^{n}\sum_{j=1}^{c} y_{ij}x_{ik}\left\{\frac{[g(z_{i,j-1}) - g(z_{ij})][\delta_{jl}g(z_{ij}) - \delta_{j-1,l}g(z_{i,j-1})]}{[G(z_{ij}) - G(z_{i,j-1})]^2}\right.\\ &\left.-\frac{[G(z_{ij}) - G(z_{i,j-1})][\delta_{jl}g(z_{ij})z_{ij} - \delta_{j-1,l}g(z_{i,j-1})z_{i,j-1}]}{[G(z_{ij}) - G(z_{i,j-1})]^2}\right\}\end{aligned}$$

$$\frac{\partial^2 \ell}{\partial \alpha_k \partial \alpha_l} = \sum_{i=1}^{n} \sum_{j=1}^{c} y_{ij} \left\{ \frac{[G(z_{ij}) - G(z_{i,j-1})][\delta_{j-1,k}\delta_{j-1,l} g(z_{i,j-1}) z_{i,j-1} - \delta_{jk}\delta_{jl} g(z_{ij}) z_{ij}]}{[G(z_{ij}) - G(z_{i,j-1})]^2} \right.$$
$$\left. - \frac{[g(z_{ij})\delta_{jk} - g(z_{i,j-1})\delta_{j-1,k}][g(z_{ij})\delta_{jl} - g(z_{i,j-1})\delta_{j-1,l}]}{[G(z_{ij}) - G(z_{i,j-1})]^2} \right\}.$$

Burridge (1981) showed that the log-likelihood function of the cumulative link model is concave and therefore there exists a unique global optimum.

4.2 Infinite Parameter Estimates

McCullagh (1980) argued that the probability of a unique maximum tends to one, if the sample size increases. Therefore, a sufficiently large n guarantees a well-defined maximum. However, for finite n, we might face infinite parameter estimates. Some indicators, that this might be the case are relatively small sample sizes, highly unbalanced data or a large number of model parameters. If the estimate $\hat{\beta}_k = \infty$ occurs, this means that the log-likelihood function continues to increase as β_k increases, the same holds for the reversed direction. From binary logistic regression models, it is known that an estimate is infinite or does not exist if the space of predictor variables is separable by a hyperplane into the ones with $y = 0$ and $y = 1$. For a cumulative logit model this is true, if the separation is possible for all $(c - 1)$ resulting binary responses.

When facing infinite parameter estimates, one possible solution is to use a simpler model in which all parameters and interactions which have an infinite estimate are removed from the model. However, if the simpler model fits poorly this has to be taken into account when interpreting the model. Agresti (2010) describes how to still use models with an infinite estimate. Regardless, if we face this situation in practice, we are going to use methods to remove or improve the estimate of the corresponding parameter. For numerical reasons most software is not able to detect infinite estimates, therefore unusually large estimates with huge standard errors are indicators that we might face a parameter which has an infinite estimate.

4.3 Regularized Newton-Raphson Algorithm

In the following we will see a **regularized Newton-Raphson algorithm with step-halving** using analytical expressions for the gradient and Hessian of the negative log-likelihood function as described in Christensen (2018). Christensen argues the choice of this algorithm with the following statement.

> ”Due to computationally cheap and efficient evaluation of the analytical derivatives, the relative well-behaved log-likelihood function [..] and the speedy convergence of the Newton-Raphson algorithm, the estimation of CLMs is virtually instant on a modern computer even with complicated models on large datasets. This also facilitates simulation studies. More important than speed is perhaps that the algorithm is reliable and accurate.”

4.3.1 The Algorithm

The regularized Newton-Raphson algorithm is an iterative algorithm, whose output is a sequence of estimates $\boldsymbol{\psi}^{(1)}, \dots, \boldsymbol{\psi}^{(i)}$. Given the i-th estimate, we get the $(i+1)$-th estimate by

$$\boldsymbol{\psi}^{(i+1)} = \boldsymbol{\psi}^{(i)} - c_1 \boldsymbol{h}^{(i)}$$

where

$$\boldsymbol{h}^{(i)} = \tilde{\boldsymbol{H}}(\boldsymbol{\psi}^{(i)}; \boldsymbol{y})^{-1} \boldsymbol{g}(\boldsymbol{\psi}^{(i)}; \boldsymbol{y})$$

with

$$\tilde{\boldsymbol{H}}(\boldsymbol{\psi}^{(i)}; \boldsymbol{y}) = \boldsymbol{H}(\boldsymbol{\psi}^{(i)}; \boldsymbol{y}) + c_2(c_3 + \min(\boldsymbol{e}^{(i)}))\boldsymbol{I}.$$

Here $\boldsymbol{g}(\boldsymbol{\psi}^{(i)}; \boldsymbol{y})$ is the gradient of the negative log-likelihood function evaluated at the current estimates and $\boldsymbol{H}(\boldsymbol{\psi}^{(i)}; \boldsymbol{y})$ the respective Hessian matrix. $\boldsymbol{e}^{(i)}$ denotes the vector of eigenvalues of the Hessian. c_2 and c_3 are scalar parameters which control the regularization in the Newton-Raphson algorithm. This regularization only takes place if the Hessian is not positive definite, therefore $c_2 \in \{0, 1\}$. Due to numerical reasons $c_2 = 1$ if $\min(\boldsymbol{e}^{(i)}) < \tau$, where τ is an appropriate tolerance. In our case $c_3 = 1$ and therefore simplifies the described algorithm. In the general case, c_3 is arbitrary, however positive. Finally c_1 is the scalar parameter which controls the step-halving. Step-halving takes place, when the full step $\boldsymbol{h}^{(i)}$ results in a decrease in the likelihood function. In this case c_1 is repeatedly halved $c_1 = \frac{1}{2}, \frac{1}{4}, \ldots$ until the step is small enough to increase the likelihood function or until the maximum number of step-halvings is reached and the algorithm is forced to stop.

4.3.2 Convergence Properties

The algorithm described in Section 4.3.1 is used in the package `ordinal`, made available by Christensen (2019). It has two convergence criteria, an absolute and a relative criterion. While the absolute criterion requests

$$\max |\boldsymbol{g}(\boldsymbol{\psi}^{(i)}; \boldsymbol{y})| < \tau_1,$$

the relative criterion asks for

$$\max |\boldsymbol{h}^{(i)}| < \tau_2,$$

where τ_1 and τ_2 are set to $\tau_1 = \tau_2 = 10^{-6}$. The algorithm used in the package, attempts to satisfy the absolute convergence criterion first.

The goal, to find a well-defined optimum, is achieved, if the gradient with respect to the parameters is small and the correpsonding Hessian matrix is positive definite. It is not uncommon in practice, that the likelihood function is almost flat in one or more directions and we therefore result in identifiability problems. The so called condition number of the Hessian, which is the ratio of the largest and the smallest eigenvalue, is a possible method to measure the empirical identifiability. The function `clm` reports this condition number after stopping the algorithm. According to Christensen (2018) a condition number less than 10^4 strongly indicates, that a well-defined optimum has been reached.

5 Regularization

To determine a model, we have to fulfil several tasks, two of them are variable selection and parameter estimation. One, well known, variable selection procedure is stepwise selection. In stepwise selection, we start with a given model and check if adding, removing or replacing a predictor would improve some kind of previously chosen criterion. Common choices for this criterion are the Akaike Information Criterion, the Bayesian Information Criterion or a corrected version of the Akaike Information Criterion if one wants to adjust for a high number of possible predictors in relation to a small sample size. Examples for parameter estimation are Ordinary Least Square Estimation or Maximum Likelihood Estimation. The goal of the regularization methods described in the following is to avoid overfitting or to handle infinite parameter estimates, by adjusting in variable selection and in parameter estimation. An overfitted model is an model that corresponds too closely to the available sample and therefore might fail to predict observations, which are not included in the sample, satisfactorily. Figure 5.1 shows a visual representation of an overfitted model against a balanced model. As Bickel et al. (2006) formulate loosely, regularization is the class of methods needed to modify maximum likelihood to give reasonable answers in unstable situations. We will focus on popular and well known regularization techniques, even though there are several other methods available. We are going to discuss ridge regression, first introduced by Hoerl (1962) and well described and discussed in Hoerl and Kennard (1970). In ridge regression we add a quadratic penalty term to the maximization problem, which is therefore also referred to as L^2-regularization. A very similar approach is the so called lasso, introduced by Tibshirani (1996). The main difference to ridge regression is the form of the penalty term, which is a sum of absolute values in the lasso case. It is therefore also referred to as L^1-regularization. A combination of both was introduced by Zou and Hastie (2005), as the so called

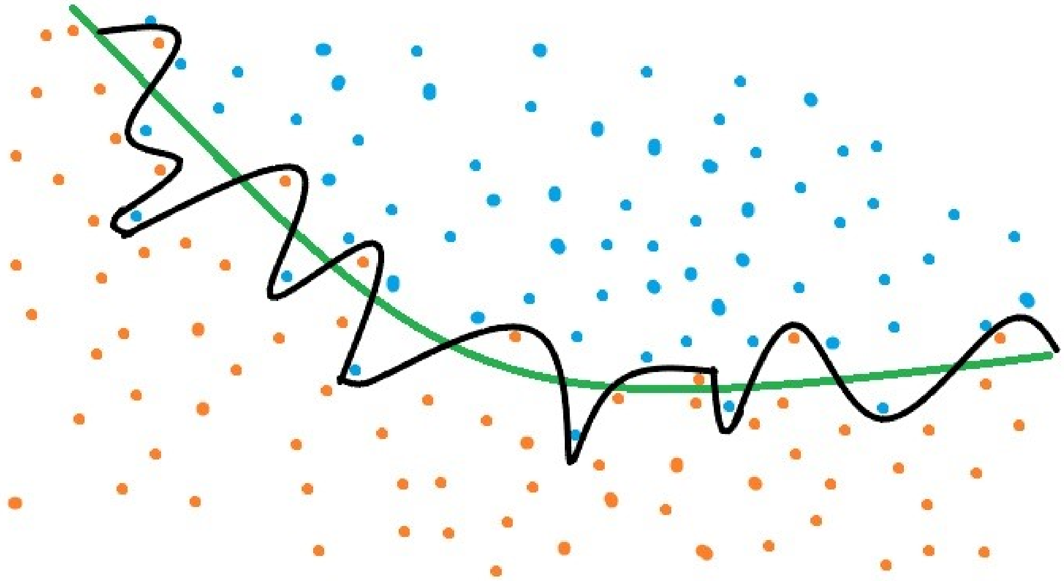

Figure 5.1: Graphical representation of overfitting; the green line represents a balanced model and the black line an overfitted model

elastic net penalty. The elastic net penalty is simply a convex combination of the L^1 and the L^2 penalty terms. While ridge regression does not perform variable selection, lasso and the elastic net penalty are able to reduce several parameters to zero and therefore increase the interpretability of the model.

5.1 Collinearity and Motivation

When we face a linear model including many correlated variables, their parameters might become poorly determined with high variance. A large negative parameter estimate on one variable can be cancelled by an equally large positive parameter estimate on its correlated twin. One possible solution is to constraint the size of the parameter estimates, which is the basic idea behind ridge regression. When talking about strongly correlated variables, we have to distinguish between two different types of multi-collinearity, firstly **exact multi-collinearity**, in which two or more explanatory variables are linearly dependent. This however implies that the design matrix $\boldsymbol{X}$ does not have full rank, as well as $\boldsymbol{X}^\intercal \boldsymbol{X}$. Therefore, $\boldsymbol{X}^\intercal \boldsymbol{X}$ is not invertible. If the covariates are only approximately linearly dependent, we face **approximate multi-collinearity**. In the situation of approximate multi-collinearity the design matrix $\boldsymbol{X}$ and $\boldsymbol{X}^\intercal \boldsymbol{X}$ do have full rank, due to the almost linear dependency the estimates might become very unstable. This is caused by that fact that $\det(\boldsymbol{X}^\intercal \boldsymbol{X})$ reaches a value near zero and therefore at least one eigenvalue λ_i becomes very small. Again the so called condition number, as mentioned in 4.3.2, the ratio between the largest and the smallest eigenvalue, is a possible measure of multi-collinearity, with a high value indicating the presence of approximate multi-collinearity. There exist some other methods to detect multi-collinerity. The first to mention is the so called **Farrar-Glauber test**, introduced by Farrar and Glauber (1967). It is actually a composition of three hypothesis tests, a chi-square test on the presence of multi-collinearity, i.e. checking if the null hypothesis that $\boldsymbol{X}$ is orthogonal has to be rejected, a F-test for the location of multi-collinearity and a t-test for the pattern. It is to mention, that several authors, like Kumar (1975) or Wichers (1975), criticized the Farrar-Glauber test in their work. The second method is the so called **variance inflation factor (VIF)**

$$VIF_j = \frac{1}{1 - R_j^2}, \quad j = 1, \ldots, p.$$

Here R_j^2 is the coefficient of determination of x_j being treated as dependent variable and the remaining x_i with $i \neq j$ as predictors. A large value of VIF, highlights the possibility of multi-collinearity. Alin (2010) claims that the threshold value between small and large is usually taken to be 10. With a VIF of 10, the corresponding R_j^2 would be 0.9, indicating that a huge amount of variability of x_j can be explained by the other predictors.

If the sample size is large enough, it might be possible to determine the parameters accurately by maximum likelihood. In practice however, the sample size often is not big enough to result in reliable estimates, sometimes not even unique. It therefore might be beneficial to use a penalized version of the log-likelihood function.

Tibshirani (1996) introduces the lasso assuming that the predictor variables are standardized and Zou and Hastie (2005) additionally assume that the response variable is centred. However, since we are going to discuss ordinal regression models, we only assume the predictor variables to be standardized. This means that we transform location and scale, simply by using the mean and standard deviation. Standardizing the predictors has several advantages. Firstly, rescaling all variables to equal size and unit standard deviation makes them more comparable. As pointed out by Hastie, Tibshirani, and Friedman (2009), standardizing the predictors also makes the penalty term more meaningful. In addition we observed an improved numerical stability when using standardized predictors. So from now on all predictors are assumed to be standardized, i.e.

$$\sum_{i=1}^{n} x_{ij} = 0 \quad \text{and} \quad \sum_{i=1}^{n} x_{ij}^2 = 1, \quad j = 1, \ldots, p$$

5.2 Regularization Methods

The main idea of the following regularization methods is, to add a penalization term to the optimization problem, either as a constraint or equivalently, directly in the functional. All presented methods where introduced using least squares in an ordinary regression setting and can therefore be written as **penalized least squares**

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta}[(\boldsymbol{y} - \boldsymbol{X\beta})^{\intercal}(\boldsymbol{y} - \boldsymbol{X\beta}) + P(\lambda, \boldsymbol{\beta})]. \tag{5.1}$$

Here $P(\lambda, \boldsymbol{\beta})$ denotes the penalty term and $\lambda \geq 0$ a tuning parameter, controlling the influence of the penalty term. If $\lambda = 0$ we would get the ordinary least squares estimate.

However, when using generalized linear models we might face some difficulties using least squares estimates. Tibshirani (1996) suggests in his work introducing the lasso, to use the log-likelihood when applying it to generalized linear models. Therefore we will rewrite (5.1), replacing the least squares by the log-likelihood function. We result in the optimization of the **penalized likelihood**

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta}[-\ell(\boldsymbol{\theta}) + P(\lambda, \boldsymbol{\beta})]. \tag{5.2}$$

$\boldsymbol{\theta}$ denotes the vector of unknown parameters, possibly including intercept and thresholds. However, the penalty term only depends on the coefficients of the covariates.

5.2.1 Ridge Regression

As already mentioned, Hoerl and Kennard (1970) introduced ridge regression for ordinary least squares regression, we will therefore first introduce it analogously and then extend it to our purposes. In **ridge regression** the length of the parameters is restricted, therefore the optimization problem is of the form

$$\hat{\boldsymbol{\beta}}_{ridge} = \arg\min_{\beta}[(\boldsymbol{y} - \boldsymbol{X\beta})^{\intercal}(\boldsymbol{y} - \boldsymbol{X\beta})], \quad \text{s.t.} \sum_{j=1}^{p} \beta_j^2 \leq t, \quad t \geq 0. \tag{5.3}$$

This formulation makes the constraint on the length of the parameters explicit. An equivalent formulation, in the form of a penalized least squares is then

$$\hat{\boldsymbol{\beta}}_{ridge} = \arg\min_{\beta}[(\boldsymbol{y} - \boldsymbol{X\beta})^{\intercal}(\boldsymbol{y} - \boldsymbol{X\beta}) + \lambda \sum_{j=1}^{p} \beta_j^2]. \tag{5.4}$$

It is clear to see that t and λ are closely related. There exists a t such that for a specific value of λ the estimates resulting as a solution of (5.3) and (5.4) are equal. In terms of an ordinary regression problem we can calculate the solution explicitly by

$$\hat{\beta}_{ridge} = (\boldsymbol{X}^{\intercal}\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^{\intercal}\boldsymbol{y},$$

where $\boldsymbol{I}$ is the identity matrix. By adding the identity matrix to $\boldsymbol{X}^{\intercal}\boldsymbol{X}$, we assure that the matrix is invertible. Finally, to extend it to generalized linear models we simply specify the penalty term in (5.2), corresponding to the ridge penalty or L^2-penalty $P(\lambda, \boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} \beta_j^2$.

As mentioned by Hesterberg, Choi, Meier, and Fraley (2008), using a ridge penalty includes all predictors with typically smaller coefficients in comparison to the unconstrained version. As λ increases, the coefficients approach zero but do not equal zero and therefore ridge regression do not perform variable selection. This lack of variable selection is one major drawback of ridge regression and is one of the reasons, why similar approaches which do perform variable selection, like lasso and elastic net penalty, were developed. In practice the variables are scaled, such that the penalty is invariant to the scale of the original data and the intercept is not penalized. Flexeder (2010) mentions the major advantages of ridge regression, it reduces the variance of the estimates, possibly improves prediction accuracy and is less sensitive to changes in data.

5.2.2 Lasso

When Tibshirani (1996) introduced the **least absolute shrinkage and selection operator**, or short lasso, the main idea was to improve prediction accuracy on the one hand and interpretability on the other hand. Prediction accuracy is improved by shrinking or setting some coefficients to zero. Loosely said, we trade in a little bias to reduce the variance in the predictions. As we saw, this is what ridge regression does. However, to improve interpretability we want less predictors included in the final model. This is one of the main advantages of lasso regularization, it sets the parameter estimates of covariates with low or no influence on the response to zero. Therefore lasso is a useful variable selection procedure. The main difference to ridge regression is the form of the penalty term, where the L^2-penalty is replaced by a L^1-penalty. Again for motivational reasons we will start introducing lasso using least squares and will then simply extend it to the more general case. The optimization

$$\hat{\boldsymbol{\beta}}_{lasso} = \arg\min_{\beta}[(\boldsymbol{y} - \boldsymbol{X\beta})^{\intercal}(\boldsymbol{y} - \boldsymbol{X\beta})], \quad \text{s.t.} \sum_{j=1}^{p}|\beta_j| \leq t, \quad t \geq 0,$$

is again equivalent to

$$\hat{\boldsymbol{\beta}}_{lasso} = \arg\min_{\beta}[(\boldsymbol{y} - \boldsymbol{X\beta})^{\intercal}(\boldsymbol{y} - \boldsymbol{X\beta}) + \lambda\sum_{j=1}^{p}|\beta_j|].$$

The one to one relationship between t and λ is again obvious. If $t > \sum_{j=1}^{p}\hat{\beta}_j^{OLS}$, where $\hat{\beta}_j^{OLS}$ denote the estimates by unpenalized ordinary least squares, we result in the same estimates. Therefore if we want to force a shrinkage of the coefficients and perform variable selection, $t < \sum_{j=1}^{p}\hat{\beta}_j^{OLS}$ is a necessary restriction. To finally extend lasso to our purposes, we set the penalty term in (5.2) to the L^1-penalty or lasso penalty $P(\lambda, \boldsymbol{\beta}) = \lambda\sum_{j=1}^{p}|\beta_j|$.

An extension of the lasso can be found in Zou (2006), the so called adaptive lasso, in which we add extra weight in the penalty term to adjust for possible unfair penalization. Unfair in the sense that each coefficient is equally penalized. We therefore result in

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta}[-\ell(\boldsymbol{\theta}) + \lambda\sum_{j=1}^{p}\omega_j|\beta_j|],$$

where the ω_j denote known weights. For the linear regression setting Zou (2006), propose to use $\omega_j = 1/|\hat{\beta}_j^{OLS}|^{\nu}$, with $\nu > 0$. However we might face missing implementation in R for several models, especially regarding ordinal responses.

Even though lasso possesses the variable selection property, it has some drawbacks. As stated by Tibshirani (1996), ridge regression dominates lasso when we face a large number of small effects. In addition, Zou and Hastie (2005) point out that for the usual $n > p$ case, with highly correlated predictors, the predictive performance of lasso is dominated by ridge regression. A further drawback stated in this work is, that the lasso does not have a grouping property, meaning that if there is a group of highly correlated variables, lasso tends to select arbitrarily not taking into account which one is included.

5.2.3 The L^q Penalty

In Figure 5.2 we see the comparison of lasso and ridge regression if there are only two parameters, β_1 and β_2. The left part of the Figure shows the lasso, while the square is the constraint $|\beta_1| + |\beta_2| \leq t$ and the right part shows ridge regression with the constraint $\beta_1^2 + \beta_2^2 \leq t$. The residual sum of squares are ellipses centered at the ordinary estimate. Both methods determine the point where the ellipses first touch the constraint region. If this point is at the corner of the lasso restriction area one parameter equals zero, unlike the disk which does not posses corners. As Hastie et al. (2009) state, in the higher dimensional case the square becomes a rhomboid and therefore has many corners and other opportunities for the estimated parameters to be zero.

Lasso and ridge regression can be naturally extended to the **L^q-penalty term**

$$\hat{\boldsymbol{\beta}}_{L^q} = \arg\min_{\beta}[(\boldsymbol{y} - \boldsymbol{X\beta})^{\intercal}(\boldsymbol{y} - \boldsymbol{X\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|^q]. \tag{5.5}$$

Figure 5.3 shows the contours of the constraint regions for different values of q in the two dimensional case. We will shortly discuss properties of the parameter q, as described by Hastie et al. (2009). For $q = 0$ we count the number of non-zero parameters and therefore this method corresponds to variable subset selection.

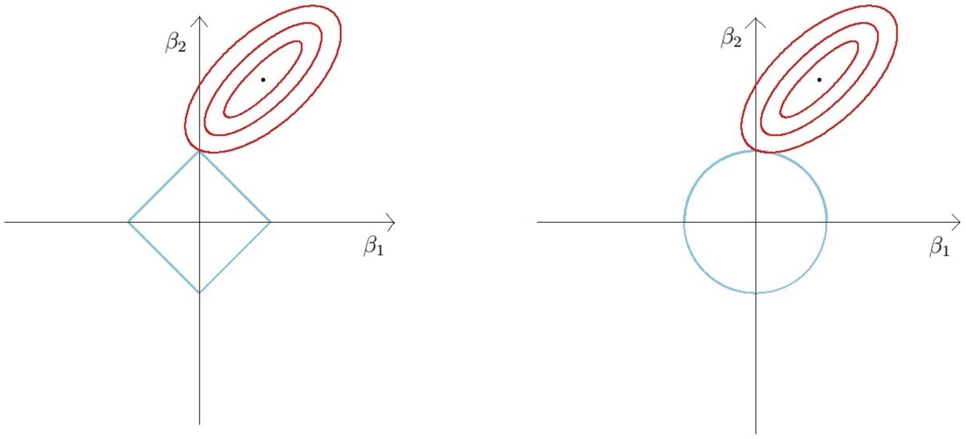

Figure 5.2: Estimation for lasso (left) and ridge regression (right) under the respective constraints in the two dimensional space (cp. Hastie et al., 2009)

Lasso results for $q = 1$ and ridge regression for $q = 2$, therefore values of $q \in (1, 2)$ somehow suggest a compromise between lasso and ridge regression. However, the value of $q = 1$ is a limit for two very important properties. Firstly for all q less than one, the constraint region is not convex any more, making the optimization problem more complex. On the other hand for all values of q being larger than one, we do not have the ability to set parameters exactly to zero. To some extend these properties might have been the motivation to introduce the elastic net penalty, described in the next section.

5.2.4 Elastic Net Penalty

As previously mentioned both, lasso and ridge regression, do have several drawbacks and the compromise (5.5) for $q \in (1, 2)$ looses the variable selection property. Therefore Zou and Hastie (2005) introduced the so called **elastic net penalty**. The elastic net penalty is a convex combination of both, ridge regression and lasso.

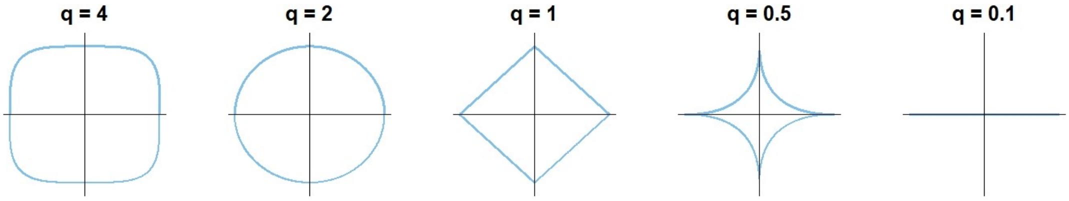

Figure 5.3: Contours of constraint regions for given values of q

The penalty term in (5.2), with fixed α is given by

$$P(\lambda, \boldsymbol{\beta}) = \lambda\Big[\alpha \sum_{j=1}^{p} \beta_j^2 + (1-\alpha) \sum_{j=1}^{p} |\beta_j|\Big]$$

or equivalently

$$P(\lambda, \boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} \Big(\alpha\beta_j^2 + (1-\alpha)|\beta_j|\Big). \tag{5.6}$$

Therefore, the parameter estimates are then given as the solution of

$$\hat{\boldsymbol{\beta}}_{elasticnet} = \arg\min_{\beta}[-\ell(\boldsymbol{\theta}) + \lambda \sum_{j=1}^{p}(\alpha\beta_j^2 + (1-\alpha)|\beta_j|).]$$

Note that in contrary to ridge regression and lasso, the elastic net penalty possesses two tuning parameters. For $\alpha = 1$ we result in ridge regression, for $\alpha = 0$ we get the lasso penalty. The elastic net penalty term can therefore been seen as a generalization of both. It is obvious that $\alpha \in [0, 1]$, otherwise the penalty term in the optimization would perform other than intended. λ should be, as already mentioned, chosen to be bigger zero. A discussion of how to choose the parameters is found in Section 5.3.

As stated by Hastie et al. (2009), the elastic net combines the advantages of lasso and ridge regression, namely the variable selection property and the grouping property. Figure 5.4 shows a comparison between the constraint regions for the L^q penalty with $q = 1.2$ and the elastic net penalty with $\alpha = 0.2$. Even though they

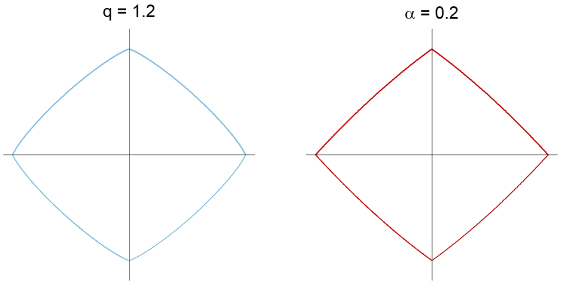

Figure 5.4: Constraint regions of the L^q penalty term with $q = 1.2$ and the elastic net penalty with $\alpha = 0.2$

look very similar, the elastic net possesses sharp corners, while the L^q penalty does not. As already mentioned in the previous section this means that the elastic net penalty possesses the variable selection property while the L^q penalty term does not.

Zou and Hastie (2005) discuss the grouping property of the elastic net penalty in further details. Essentially the grouping property means, that the parameter estimates of highly correlated variables tend to be equal, with a negative sign for negatively correlated ones. One disadvantage of the elastic net might be the computational costs, due to the tuning parameters. This is especially noteworthy in high dimensional settings, however can be overcome for example by presetting a few different values for α and only tuning λ on these predefined values.

5.3 Estimation of Tuning Parameters

In the penalized likelihood model it is crucial to choose the appropriate tuning parameter, such that the performance of the fitted model is optimized with respect to some previously set criteria. Firstly we have to distinguish between two problems, namely finding the tuning parameter λ for ridge regression, lasso and elastic net penalty and on the other hand determining the tuning parameter α in the elastic net penalty term. Basically, we can compare two tuning procedures, which

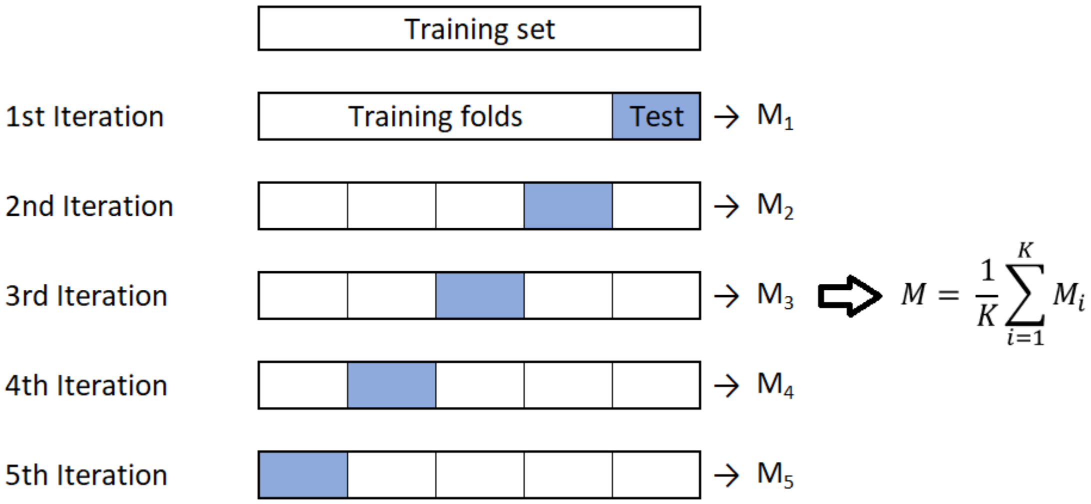

Figure 5.5: Five-fold cross-validation. In each iteration one of the five folds was used to determine the measure M_i, while the other four folds where used to fit the model

both do not differ dramatically from ordinary model selection procedures. One option is to select the model which is best with respect to a specific criterion, like the AIC or BIC. The other option is to use cross-validation. Since many authors, like Hastie et al. (2009) and Zou and Hastie (2005) in their introducing paper choose cross-validation to determine the tuning parameters and the R-Package in use, `ordinalNet`, provided by Wurm, Rathouz, and Hanlon (2020) also uses cross-validation, we will follow their choice and use cross-validation when determining the tuning parameters.

The main idea of cross-validation is rather simple, namely to use a part of the available data to fit the model and to use the remaining part to test it. While there are numerous different types of cross-validation, most of them can be described by the so called **K-fold cross-validation**. Figure 5.5 gives a visual representation of five-fold cross-validation. In K-fold cross-validation we split the available data set into K parts of (approximately) equal size. For the k-th part, where $k = 1, \ldots, K$, we fit the model to the remaining $k - 1$ parts and assess the prediction error of the fitted model, or another measure of fit, on the k-th part. We repeat this procedure for all k and then combine the K prediction error estimates. In practice, when trying to determine the tuning parameter λ, it is common to pre-set a sequence of λ-values and to perform cross-validation for each value separately. Finally one chooses the λ-value with the best performance.

In the elastic-net penalty case we have to estimate two tuning parameters and therefore have to cross-validate on a two dimensional surface. This might result in high computational costs, especially in high dimensional settings. As stated by Zou and Hastie (2005), one practical solution might be early stopping, meaning that one simply forces the algorithm to stop when a specific, chosen number of variables in the model is reached. Another possibility is, as already mentioned in the previous section, to reduce the possible number of α-values.

Independently of how one chooses to handle the determination of the tuning parameters, an essential question is how to choose the number of folds, K. As mentioned by Hastie et al. (2009), common choices are $K = 5, 10, n$. In the case $K = n$, where n is the number of observations, we talk about **leave-one-out cross-validation**. For $K = 5, 10$ we talk about **five-fold cross-validation** and **ten-fold cross-validation**, respectively. When choosing five-fold cross-validation, this implies that we use 80% of the data as training set and 20% as test set, while changing to 90%/10% in ten-fold cross-validation. Since all training sets are very similar when using leave-one-out cross-validation, the estimator might show high variance even though it should be approximately unbiased for the true prediction error. In addition the computational costs are much higher in comparison to the situation where we build less folds. Figure 5.6 shows a hypothetical learning curve and is used to visualize the impact of the number of folds. Note that, since this is a hypothetical learning curve, the stated numbers are not valid in general. In the case of 200 observations, five-fold cross-validation would result in a training set of 160 observations, implying not much bias, since it has almost the same performance as if one uses the whole data set for training. However, if we face a situation where only 40 observations could be used as training set we would result in a much higher bias. Summarizing, five-fold and ten-fold cross-validation would overestimate the true prediction error if the learning curve shows a substantial slope at the training set size. Leave-one-out cross-validation on the other hand reduces the bias while possibly increasing the variance. Hastie et al. (2009) recommend five-fold and ten-

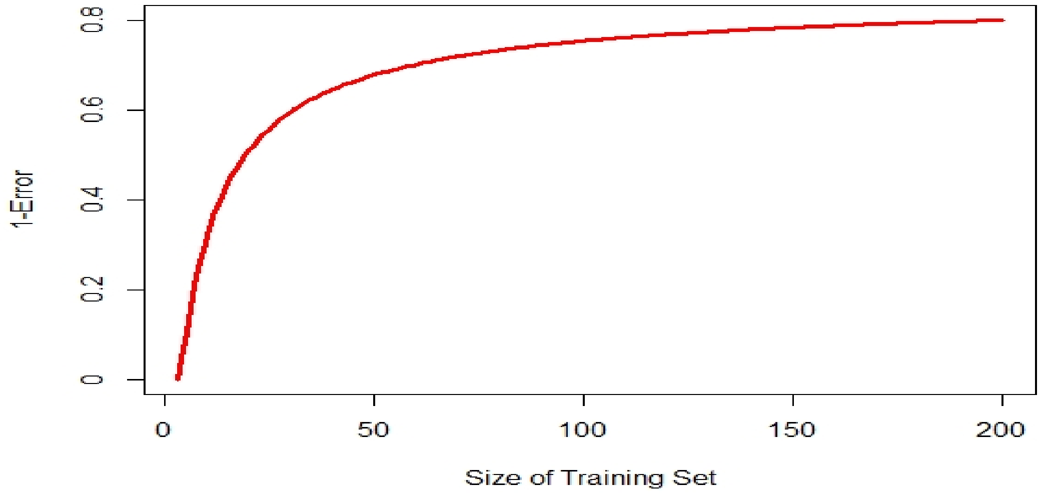

Figure 5.6: A hypothetical learning curve, the plot of 1-Error against the size of the training set. The total data set consists of 200 observations (cp. Hastie et al., 2009)

fold cross-validation as a good compromise. However, in practice it might be useful to compare the results of both with the results of leave-one-out cross-validation.

5.3.1 Optimization

The algorithm described in Section 4.3, works perfectly for unregularized likelihoods. However, when using the elastic net penalty, including lasso and ridge regression, we get in need of an adjusted optimization algorithm. Again, even though there might be numerous different algorithms design for that purpose we will only discuss the algorithm used in the R-package `ordinalNet`, since it is the one we are going to use in the application. Wurm, Rathouz, and Hanlon (2017) describe the algorithm used for a more general class of models, of which our models are a subset. We are only going to shortly discuss the basic idea and refer the reader interested in more details to the work published by Wurm et al. (2017) basing on ideas by Friedman, Hastie, and Tibshirani (2010) and Friedman, Hastie, Höfling, and Tibshirani (2007).

The algorithm described is an iterative one, with an inner and an outer loop. In the outer loop we construct a quadratic approximation of the log-likelihood, i.e. a Taylor expansion around the current parameter estimates $\hat{\boldsymbol{\beta}}^{(r)}$. The inner loop then computes the next estimate $\hat{\boldsymbol{\beta}}^{(r+1)}$ by optimizing the penalized quadratic approximation using coordinate descent. This is necessary since the optimization step does not have a closed form when using the elastic net penalty. Using a coordinate descent procedure, informally means that we cycle through the coefficient estimates, update each one with the marginally best value and iterate this cycle until a convergence criterion is met.

Since we are interested in finding the solutions for different λ-values, it makes sense to determine the λ_{max}-value, which is the value where the parameter estimate of the first covariate is non-zero. This is simply done by determining the threshold λ-value for each penalized coefficient where its estimate becomes non-zero and then setting λ_{max} to the maximum of all those thresholds. Friedman et al. (2010) propose the strategy, after finding λ_{max}, to set $\lambda_{min} = 0.01 * \lambda_{max}$ and to construct a decreasing sequence of λ-values from λ_{max} to λ_{min} on the log-scale. To improve the efficiency of the algorithm a technique called warm starts is used. This means that the starting value for each λ-value is the parameter estimate resulting from the previous one. Unfortunately, there is no numerical solution provided of how to estimate the tuning parameter α. Therefore the very pragmatic way chosen is to simply presetting different α-values and compare the respective performances.

6 Goodness of Fit

In the following chapter we will have a look at methods of how to evaluate regression models for categorical response variables with respect to goodness of fit. All three following test statistics are based on estimated probabilities, rather than on a particular model. The strategy is to group the observations by a score and to derive the test statistics only depending on the ordinal response variable and the respective probability estimates. All tests base on the Pearson χ^2 test and the approach by Hosmer and Lemeshow (1980) for binary data, which follow the idea to compare observed and expected frequencies. We will first derive the test statistics for the proportional odds version of the cumulative logit model, will then compare the test statistics and shortly discuss advantages and drawbacks regarding the null distributions and the power of the tests. Finally we will have a brief look at the extension to the adjacent-category and the continuation-ratio models.

6.1 Test Statistics

Fagerland and Hosmer (2013) derived a test statistic based on the Hosmer-Lemeshow test for binary logistic regression for the proportional odds version of the cumulative logit model and compared it to a goodness of fit test proposed by Lipsitz, Fitzmaurice, and Molenberghs (1996) and a modification of the Pearson χ^2 and deviance statistics for ordinal models published by Pulkstenis and Robinson (2004). Fagerland and Hosmer (2016) extended their work to the adjacent-category and the continuation-ratio models and compared the already mentioned test statistics also for the two alternative models. Due to the different structure of the tests and models, it is important to note that the distributions of the test statistics might vary between them.

6.1.1 Lipsitz Test

First we derive the goodness of fit test for ordinal regression models, including the proportional odds version of the cumulative logit model, proposed by Lipsitz et al. (1996). We will therefore refer to it as the **Lipsitz test**. Let $\pi_{ij} = \mathbb{P}(Y_i = j|\boldsymbol{x}_i)$ for $j = 1, \dots, c$ and $i = 1, \dots, n$, and equivalently let $\hat{\pi}_{ij}$ denote the estimated probabilities calculated from a fitted ordinal regression model. We then assign a score

$$s_i = 1\hat{\pi}_{i1} + 2\hat{\pi}_{i2} + \cdots + c\hat{\pi}_{ic}, \quad i = 1, \dots, n \tag{6.1}$$

to each observation. In the next step we group the observations into g groups of (approximately) equal size based on the assigned score, where the first group should contain the n/g lowest score observations and the last group the n/g highest score observations. Based on this grouping we create $g - 1$ indicator variables I_k, such that

$$I_{ik} = \begin{cases} 1 & \text{if observation } i \text{ is in group } k \\ 0 & \text{otherwise.} \end{cases}$$

for $i = 1, \dots, n$ and $k = 1, \dots, g - 1$. With these variables we fit a new ordinal regression model of the same type as the previous model, additionally including the indicators,

$$h[\mathbb{P}(Y \leq j|\boldsymbol{x})] = \alpha_j + \boldsymbol{x}^\intercal\boldsymbol{\beta} + \gamma_1 I_1 + \gamma_2 I_2 + \cdots + \gamma_{g-1} I_{g-1}, \quad j = 1, \dots, c - 1.$$

If the previously fitted model is the correct model, we would result in $\gamma_1 = \cdots = \gamma_{g-1} = 0$. To obtain a goodness of fit test we compare the value of the likelihood ratio test statistic $-2(\ell_1 - \ell_0)$ with the quantiles of the χ^2-distribution with $g - 1$ degrees of freedom. Here ℓ_1 and ℓ_0 denote the two log-likelihoods of the fitted model without indicators and with indicators, respectively.

Lipsitz et al. (1996) suggested that the number of groups should fulfill $6 \leq g \leq n/5c$ or to follow the simpler rule, provided by Hosmer and Lemeshow (1980) which suggest to form 10 groups of equal size. In addition Lipsitz et al. (1996) state, that using a Taylor series expansion, it can be shown that the test statistic under the null hypothesis, that $\gamma_1 = \cdots = \gamma_{g-1} = 0$, is a linear combination of $(O_{kj} - E_{kj})$.

Where

$$O_{kj} = \sum_{i=1}^{n} I_{ik} y_{ij}, \tag{6.2}$$

is the observed number of elements in group k with response j, with

$$y_{ij} = \begin{cases} 1 & \text{if} \quad y_i = j \\ 0 & \text{otherwise,} \end{cases}$$

and

$$E_{kj} = \sum_{i=1}^{n} I_{ik} \hat{\pi}_{ij}, \tag{6.3}$$

is the estimated number of elements in group k with response j.

6.1.2 Pulkstenis and Robinson Test

The idea of Pulkstenis and Robinson (2004) is to modify the Pearson chisquare and deviance test, in which we construct a contingency table of estimated and observed frequencies. The columns consist of all levels of the response variable and the rows consist of the possible covariate patterns. If all covariates are categorical, we can easily construct the contingency table and assess the goodness of fit using the Pearson chisquare and deviance test. However, this approach fails if continuous covariates are part of the model, because the number of possible covariate patterns makes the contingency table sparse. Pulkstenis and Robinson (2004) describe the following procedure to derive the test statistics: Starting with assigning a score, as defined in (6.1), to each observation, we then specify all covariate patterns determined by the categorical covariates and remove all unobserved ones. We can then sort within each covariate pattern and split the group into two subgroups, one smaller or equal to the median of the estimated scores within this group and one larger than the median. The modified test statistics are then given by

$$\chi^{*2} = \sum_{l=1}^{2} \sum_{k=1}^{K} \sum_{j=1}^{c} \frac{(O_{lkj} - E_{lkj})^2}{E_{lkj}},$$
$$D^{*2} = 2 \sum_{l=1}^{2} \sum_{k=1}^{K} \sum_{j=1}^{c} O_{lkj} log \frac{O_{lkj}}{E_{lkj}},$$

where c is the number of response categories, K denotes the number of different observed categorical covariate patterns and l relates to the two score-based subgroups. Both test statistics are compared with the quantiles of the χ^2-distribution with $(2K-1)(c-1)-p_{cat}-1$ degrees of freedom, with p_{cat} denoting the number of categorical covariates. For example a model with 4 dichotomous covariates and one factor with 4 levels, represented by three design variables, would result in $p_{cat}=4+3=7$. Pulkstenis and Robinson (2004) point out that sample size is an important consideration when performing their tests. They suggest that about 80% of expected cell counts should exceed 5, or at least to be very cautious when dealing with results of sparse data. As a pragmatic solution they suggest to simply combine rows with small sample sizes.

6.1.3 Fagerland and Hosmer Test

The test statistic proposed by Fagerland and Hosmer (2013) is based on an approach first suggested by Hosmer and Lemeshow (1980) for binary logistic regression and later adapted to multinomial logistic regression by Fagerland, Hosmer, and Bofin (2008). In the binary setting, the observations are grouped with respect to the estimated success probability. Again the number of groups g, can be arbitrary. However, 10 groups seem to be a reasonable choice. Then a contingency table containing the observed and estimated frequencies, within each group, can be constructed. The Pearson χ^2 is then the corresponding test statistic and the reference distribution is the χ^2-distribution with $g-2$ degrees of freedom. In the multinomial setting, we analogously group the observations with respect to the complement of the estimated probability of the reference response category. Again we can then construct a contingency table containing the estimated and observed frequencies for each group and each response category. The test statistic is again the Pearson χ^2 statistic and is compared with the quantiles of the χ^2-distribution with $(g-2)(c-1)$ degrees of freedom. For the proportional odds version of the cumulative logit model Fagerland and Hosmer (2013) choose a very similar approach. First they calculate the estimated probabilities $\hat{\pi}_{ij}$ and compute the ordinal scores as defined in (6.1). We again split the observations into g groups of (approximately) equal size, with the same ordering as described for the Lipsitz test. Let again O_{kj} and E_{kj}, as defined in (6.2) and (6.3), denote the observed

	$Y=1$		$Y=2$		$\cdots$	$Y=c$		
Group	Obs.	Est.	Obs.	Est.	$\cdots$	Obs.	Est.	Sum
1	O_{11}	E_{11}	O_{12}	E_{12}		O_{1c}	E_{1c}	n/g
2	O_{21}	E_{21}	O_{22}	E_{22}	$\cdots$	O_{2c}	E_{2c}	n/g
$\vdots$	$\vdots$	$\vdots$				$\vdots$	$\vdots$	$\vdots$
g	O_{g1}	E_{g1}	O_{g2}	E_{g2}	$\cdots$	O_{gc}	E_{gc}	n/g

Table 6.1: Observed and estimated frequencies sorted and summed into g groups

and estimated number of observations in each group for each response level. The resulting table is shown in Table 6.1. The ordinal test statistic, which we will refer to as the **Fagerland and Hosmer test statistic**, is then the Pearson χ^2 statistic given by

$$C_g = \sum_{k=1}^{g} \sum_{j=1}^{c} \frac{(O_{kj} - E_{kj})^2}{E_{kj}}. \tag{6.4}$$

Fagerland and Hosmer (2013) posit that the degrees of freedom are $(g-2)(c-1)+(c-2)$. They also proof, that sorting using a score of the form $s_i = \hat{\pi}_{i1}$ is equivalent to the ordinal score we used so far. This is noteworthy, since this is the score used in the multinomial setting.

Theorem 6.1.1. Sorting the observations based on the ordinal score as defined in (6.1) is equivalent to sorting the observations using the multinomial score.

Proof. We know that $\pi_1 = \mathbb{P}(Y \leq 1|\boldsymbol{x})$ and for $j = 2, \ldots, c$,

$$\pi_j = \mathbb{P}(Y = j|\boldsymbol{x}) = \mathbb{P}(Y \leq j|\boldsymbol{x}) - \mathbb{P}(Y \leq j-1|\boldsymbol{x}),$$

with $\mathbb{P}(Y \leq c|\boldsymbol{x}) = 1$. Then the ordinal score is

$$\begin{aligned} OS &= \pi_1 + \cdots + c\pi_c \\ &= \mathbb{P}(Y \leq 1|\boldsymbol{x}) + 2 * [\mathbb{P}(Y \leq 2|\boldsymbol{x}) - \mathbb{P}(Y \leq 1|\boldsymbol{x})] + \cdots + c * [1 - \mathbb{P}(Y \leq c-1|\boldsymbol{x})] \\ &= -\mathbb{P}(Y \leq 1|\boldsymbol{x}) - \mathbb{P}(Y \leq 2|\boldsymbol{x}) - \cdots - \mathbb{P}(Y \leq c-1|\boldsymbol{x}) + c \\ &= c - \sum_{j=1}^{c-1} \mathbb{P}(Y \leq j|\boldsymbol{x}). \end{aligned}$$

The multinomial score is

$$MS = 1 - \pi_1 = 1 - \mathbb{P}(Y \leq 1|\boldsymbol{x}).$$

To show that the sorting is equivalent, we will now show that for two independent observations A and B

$$MS_A > MS_B \iff OS_A > OS_B.$$

Let $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$ denote the corresponding covariate vectors. First, we assume that $MS_A > MS_B$. This implies that

$$1 - \mathbb{P}(Y \leq 1|\boldsymbol{x}_A) > 1 - \mathbb{P}(Y \leq 1|\boldsymbol{x}_B),$$

and by the monotonicity of $\mathbb{P}(Y \leq j|\boldsymbol{x})$ with respect to $\alpha_j + \boldsymbol{x}^\intercal\boldsymbol{\beta}$ it holds that

$$\alpha_1 + \boldsymbol{x}_A^\intercal\boldsymbol{\beta} < \alpha_1 + \boldsymbol{x}_B^\intercal\boldsymbol{\beta}.$$

By Proposition 3.0.1 we already know that, under sufficient regularity assumptions, $\alpha_j < \alpha_{j+1}$. Hence,

$$\alpha_2 + \boldsymbol{x}_A^\intercal\boldsymbol{\beta} = \alpha_1 + \boldsymbol{x}_A^\intercal\boldsymbol{\beta} + (\alpha_2 - \alpha_1) < \alpha_1 + \boldsymbol{x}_B^\intercal\boldsymbol{\beta} + (\alpha_2 - \alpha_1) = \alpha_2 + \boldsymbol{x}_B^\intercal\boldsymbol{\beta}.$$

Again by the monotonicity this implies that $\mathbb{P}(Y \leq 2|\boldsymbol{x}_A) < \mathbb{P}(Y \leq 2|\boldsymbol{x}_B)$. Inductively, by the same argument, we can conclude that

$$\mathbb{P}(Y \leq j|\boldsymbol{x}_A) < \mathbb{P}(Y \leq j|\boldsymbol{x}_B), \quad \text{for } j = 1, \ldots, c-1$$

and therefore

$$\sum_{j=1}^{c-1} \mathbb{P}(Y \leq j|\boldsymbol{x}_A) < \sum_{j=1}^{c-1} \mathbb{P}(Y \leq j|\boldsymbol{x}_B),$$

which finally means that

$$OS_A = c - \sum_{j=1}^{c-1} \mathbb{P}(Y \leq j|\boldsymbol{x}_A) > c - \sum_{j=1}^{c-1} \mathbb{P}(Y \leq j|\boldsymbol{x}_B) = OS_B.$$

Now assume that $OS_A > OS_B$ and show that this implies $MS_A > MS_B$. By assuming $OS_A > OS_B$, we again get that

$$\sum_{j=1}^{c-1} \mathbb{P}(Y \leq j|\boldsymbol{x}_A) < \sum_{j=1}^{c-1} \mathbb{P}(Y \leq j|\boldsymbol{x}_B).$$

It follows that

$$\sum_{j=1}^{c-1}(\alpha_j + \boldsymbol{x}_A^\intercal\boldsymbol{\beta}) < \sum_{j=1}^{c-1}(\alpha_j + \boldsymbol{x}_B^\intercal\boldsymbol{\beta}),$$

immediately implying $\boldsymbol{x}_A^\intercal\boldsymbol{\beta} < \boldsymbol{x}_B^\intercal\boldsymbol{\beta}$ and therefore

$$\mathbb{P}(Y \leq 1|\boldsymbol{x}_A) < \mathbb{P}(Y \leq 1|\boldsymbol{x}_B).$$

This however means that

$$MS_A = 1 - \mathbb{P}(Y \leq 1|\boldsymbol{x}_A) > 1 - \mathbb{P}(Y \leq 1|\boldsymbol{x}_B) = MS_B.$$

□

6.2 Comparison of the Tests

Fagerland and Hosmer (2013) performed several simulations to check the null distribution of the test statistics for a response with three, four and five levels. Additionaly they assessed the possible effects of sample size, number of response levels and covariate distribution. In total they compared six test statistics, the Pulkstenis and Robinson tests, denoted by $PR(\chi^2)$ and $PR(D^2)$, the Lipsitz test where $g = \min(10, n/5c)$ and three Fagerland and Hosmer tests, C_8, C_{10} and C_{12}.

The results of the simulations for the three Fagerland and Hosmer tests did not differ much, indicating that the test do not strongly depend on the choice of the number of groups. In addition the Lipsitz test and the Pulkstenis and Robinson tests had slightly too high rejection rates in comparison with the nominal level, while the Fagerland and Hosmer test were slightly below the nominal level. More details and explicit numbers can be found in the web-based supporting materials (Fagerland and Hosmer, 2013). They found, that none of the tests had good power when assessing for a missing quadratic term or for a wrong functional form of a covariate. However, the power to detect a missing interaction term was high for all test. For detection of a violation of the proportional odds assumption, the Lipsitz test did not show satisfying results, whereas the Pulkstenis and Robinson tests and the Fagerland and Hosmer tests performed well. The same holds for the situation where we face a nominal response instead of an ordinal one, with especially good results for the Pulkstenis and Robinson tests. In summary the C_g tests were able to detect lack of fit in five out of six investigated situations, with an slightly decreasing power when increasing g. The Pulkstenis and Robinson tests were able to detect four cases and the Lipsitz test three. However, the Fagerland and Hosmer tests have not had the highest power in any of the five situations, when compared to the other tests.

Fagerland and Hosmer (2016) extended their work also to two other logits, namely the adjacent-categories and the continuation-ratio logits and again performed several simulations to check the null distribution and to assess the power of the tests. Since it is possible to estimate the probabilities for each response level, conditioned on a vector of covariates, the test statistics do not change in comparison to the proportional odds version of the cumulative logit model. The results of this simulation studies where quite similar to the ones mentioned beforehand.

The consistency of the two papers allows to make some general conclusions respectively recommendations. Due to the different lack of fits the test might detect and the different behaviours with respect to rejection rates and power, it is recommended to use all three types of tests to assess potential lack of fit. This might assure that together they have reasonable power when facing samples of moderate to large size. Fagerland and Hosmer (2013) caution that the tests have low power for small sample sizes. Therefore they suggest to choose a significance level of 10% and to be cautious when interpreting the results of the tests.

7 Application

In the following we will discuss previously derived results on a real data example. The available data contains performance information on team basis from two ice hockey seasons, which corresponds to 104 games in the final data set. All games were part of the "Erste Bank Eishockey Liga" and either the home or the away team was one specific Austrian professional ice hockey team. The information was provided by *InStat* (2020), a worldwide leader in sports performance analysis for professional leagues, clubs, players and media in football, ice hockey, basketball and futsal. For more details on the data itself, the extraction procedure, discussion about variables and an exploratory data analysis, refer to Friedl and Embacher (2020). To improve stability of the estimation procedures, to ensure consistency with the described methods and to guarantee comparability between the different methods we will use standardized predictor variables. The following methods can essentially been split in two different approaches. The first one uses expected number of goals to either predict the actual number of goals or to use this as a basis to predict the outcome of a game. The second approach uses ordinal regression models to directly predict the outcome.

7.1 Prediction of Scores using Poisson Regression

We are going to start the analysis by looking at the expected number of goals scored, either by the home team or by the away team independently. It seems somehow natural to model scored goals by a Poisson distribution, since we are counting the occurrence of an event, namely a goal being scored. Ice hockey is as complex as it is simple, if a team wants to win it has to score more goals than its opponent. The same principle is valid for many other sports, like football,

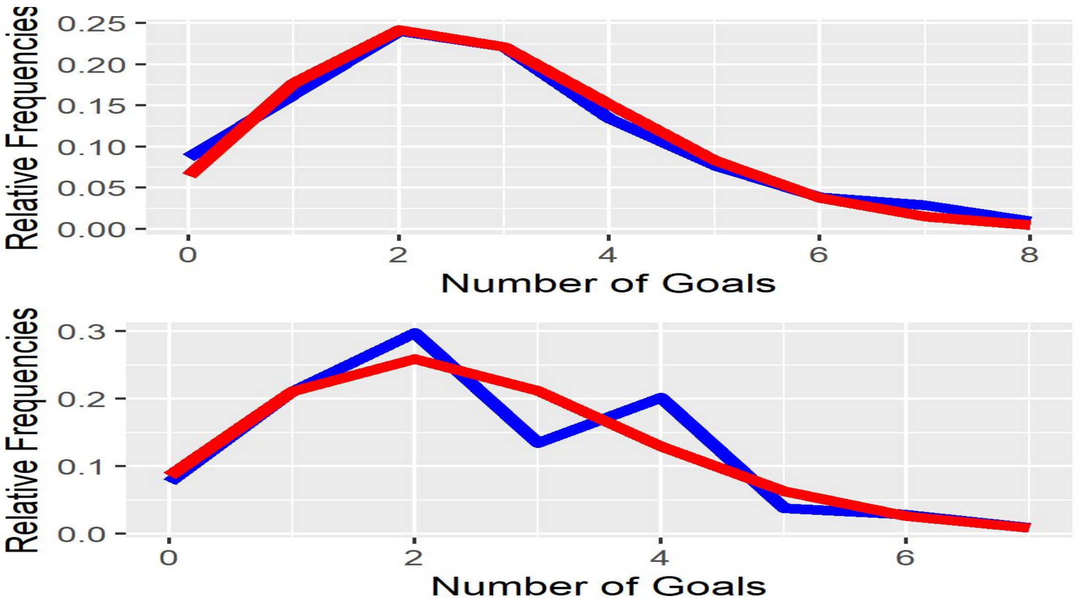

Figure 7.1: Comparison of observed frequencies (blue) and expected Poisson frequencies (red), for home (upper) and away team (lower)

handball or basketball. It is very intuitive that shooting on the goals is increasing your chance to actually score a goal and that having possession of the puck or winning a faceoff might as well increase your chance of scoring a goal and therefore increases the number of expected goals. In the following we are going to look at which variables do have significant influence on the expected number of goals, on basis of our data, and how good this information might be used to actually predict the outcome of a game.

7.1.1 Methods

Our first class of models will be log-linear models on the expected number of goals scored by the home team and the away team respectively. As described by Groll and Schauberger (2019) we will assume that the scored goals are Poisson distributed, where $X \sim Poi(\lambda)$ denotes the goals scored by the home team and $Y \sim Poi(\mu)$ the goals scored by the away team. A comparison of the observed relative frequencies and the theoretical probabilities, when estimating the parameter of the Poisson distribution as the mean of the scored goals, for both home and away team, is shown in Figure 7.1. There we can see that the goals scored by the home team almost perfectly follow a Poisson distribution, while the goals scored by the away team somehow struggle to. This however might smooth out with an increasing number of games in the sample.

In most models, the Poisson distributions are considered independent and we therefore result in a simple joint distribution

$$\mathbb{P}(X = x, Y = y) = \frac{\lambda^x \exp(-\lambda)}{x!} \frac{\mu^y \exp(-\mu)}{y!}. \tag{7.1}$$

We now perform some further simplifications to increase the sample size. First we assume, as already mentioned, that the goals of the home team and the goals of the away team are independent. Additionally we suppose that for both teams the same covariates are of significance and we therefore state that

$$\begin{aligned} \lambda &= \exp(\beta_0 + \delta + \boldsymbol{x}^\intercal \boldsymbol{\beta}_H) \\ \mu &= \exp(\beta_0 + \boldsymbol{x}^\intercal \boldsymbol{\beta}_A), \end{aligned} \tag{7.2}$$

where β_0 denotes the intercept, δ is assumed to be a global home effect and $\boldsymbol{\beta}_H$ and $\boldsymbol{\beta}_A$ the respective parameters of the covariates. The two parameter vectors are simply connected via $\boldsymbol{\beta}_H = \boldsymbol{\beta}_A + \boldsymbol{\beta}_\delta$. $\boldsymbol{\beta}_\delta$ is necessary to adjust for possible different effects for the home and the away team. From a modelling point of view $\boldsymbol{\beta}_\delta$ can be seen as the parameter of the interaction terms between the factor 'home' and the covariates. Having the model fitted, we get two parameter estimates $\hat{\lambda}_i$ and $\hat{\mu}_i$ for the i-th game, which represent the expected values of the two independent Poisson distributions. To now predict the outcome of the game, we have several options.

The first and very simple way is to use the fitted values of the two Poisson distributions and determine the result via the difference of the intensity estimates. Using this approach, one has to think about, how to specify a draw. It might be very rare that $\hat{\lambda}_i - \hat{\mu}_i = 0$, therefore defining an interval in which we determine a draw is reasonable. We say that if $\hat{\lambda}_i - \hat{\mu}_i \in (-0.5, 0.5)$ the predicted result is a draw.

The next, very similar approach is to use the estimated distribution of both scores to calculate the probabilities of an ordinal outcome, like home win, draw and away win. For example, the probability of a draw is given by

$$\mathbb{P}(Draw) = \mathbb{P}(X - Y = 0).$$

These probabilities can be derived using a Skellam distribution, which is a discrete probability distribution, resulting as the difference between two independent Poisson distributions. The Skellam distribution has mean $\lambda - \mu$ and variance $\lambda + \mu$. Even though it was initially introduced as the difference between two independent Poisson distributions, Karlis and Ntzoufras (2008) proved that it can also be derived as the difference of distributions which have a specific trivariate latent variable structure. Since we assume the Poisson distributions to be independent we do not go into further detail here. To now predict the outcome of a game using the Skellam distribution we simply set the predicted outcome to the one with the highest predicted probability.

The third possibility is to use random samples, representing the final score, (X, Y) from the respective distributions. As mentioned by Groll and Schauberger (2019) due to the high variability of this approach, a large number of replications should be considered. However, this approach might be very useful when simulating tournaments. Often it is necessary to determine the final standing of a group stage or a tournament itself, including scored and received goals to correctly determine the following knock-out stage or the winner, respectively.

7.1.2 Results

We now will have a look at the results of our real data example. Since we assume independence between home and away goals, we result in 208 observations out of 104 games. For each of these 208 observations we have 122 possible covariates and the factor 'home', which indicates whether the respective team played home or away. In addition, we allow for all interactions between the explanatory variables and this factor, to adjust for possible different effects for the home and the away team, modelled by $\boldsymbol{\beta}_\delta$. We will fit a model of the form (7.2). For model fitting we use a stepwise variable selection procedure with the corrected Akaike Information Criterion as decision base. As mentioned, we will use standardized covariates, reflecting zero mean and unit variance.

Variable	Coefficient	Std. Error	p-value	Mean	Std. Dev.
Intercept	0.8854	0.0459	0.0000		
ShotsOG60	0.2909	0.0575	0.0000	29.61	7.48
FaceoffsNZ	0.1765	0.0430	0.0000	8.44	2.78
Possession3P	-0.1223	0.0501	0.0145	538.10	89.07
ShotsAES	-0.1286	0.0582	0.0273	42.67	10.49
PassesACkey	0.1162	0.0427	0.0065	3.57	2.27
Takeaways3P	0.1352	0.0466	0.0037	30.88	4.54
Giveaways3P	-0.1342	0.0489	0.0060	20.22	5.27
ChallengesWNZ	0.1124	0.0425	0.0082	9.34	4.29
PossessionDZ	0.0917	0.0463	0.0477	612.91	69.92

Table 7.1: Parameter estimates, standard errors and p-values for the Poisson model using standardized covariates from 208 games, with the mean used for centering and the standard deviation for scaling

In the first step we allow δ to be non zero, i.e. we include a home effect in the model. However, when fitting this model we see that no interaction term between the factor home and the predictors in the model is significant, i.e. we can not reject a hypothesis like $\boldsymbol{\beta}_\delta = \mathbf{0}$. Even more surprising is, that the home effect is strongly insignificant (p-value: 0.58029). It therefore seems reasonable to drop the home effect out of the model and to set $\delta = 0$. We only choose covariates to be included in the model if the respective p-value is not exceeding 0.05. Table 7.1 shows the coefficients in the fitted model described by (7.2) and δ being set equal to zero. Hence we get the same model for both teams, only depending on the respective performance. The deviance of the fitted model is 168.24 on 198 degrees of freedom. In the Appendix, in Figure A.1, one can find the diagnostic plots for the fitted log-linear model. As an example we will use the coefficients on one specific game. The expected number of goals is given by

$$\mathbb{E}[\text{Goals}] = \exp\left(\beta_0 + \sum_{j=1}^{9} \beta_j \frac{x_j - c_j}{s_j}\right), \tag{7.3}$$

	In Sample	Jackknife	Data Splitting
Goals Home	22.11%	23.08%	22.88%
Goals Away	28.85%	32.69%	29.91%
Goals	25.48%	25.00%	25.14%
$\hat{\lambda}_i - \hat{\mu}_i$	55.77%	53.85%	54.00%
Skellam Distribution	60.58%	57.69%	58.11%
Random Sampling	60.58%	57.69%	58.30%

Table 7.2: Accuracy of the model with parameter estimates as shown in Table 7.1 using in sample accuracy, jackknife and data splitting

where c_j denotes the observed mean used for centering and s_j the observed standard deviation used for scaling, both found in Table 7.1. From the home team's point of view in our example game we observed 36 shots on goal, 9 won faceoffs in the neutral zone, 575 seconds of possession in the third period, 40 shots at even strength, 6 accurate key passes, 31 takeaways in the third period, 37 giveaways in the third period, 6 won challenges in the neutral zone and 746 seconds of possession in the defensive zone. Plugging this in Model (7.3) results in 2.56 estimated goals, while we observed 2 scored goals.

Due to the independence of the expected goals, we are able to discuss the model performance with respect to the home goals, away goals or goals in general on the one hand and with respect of the match outcome on the other hand. As mentioned, there are multiple possibilities to actually predict the outcome of the game on 3-way basis (Home win, Draw, Away win). We will compare the results for all three mentioned methods using in sample accuracy, jackknife and data splitting where we use 80% as training data and 20% as test data. In addition we will use the same three test methods on the expected goals, where we simply round the prediction to the nearest integer to assure comparability. One result, presented in Table 7.2 is not really surprising, using the Skellam distributions shows similar results as drawing random samples with a large number of replications. This is due to the fact that the Skellam distribution is resulting as the difference of two independent Poisson distributions and in the other case we draw random samples out of two independent Poisson distributions and build their difference. Table 7.2

Home											Away										
	Predicted											Predicted									
Observed	0	1	2	3	4	5	6	7	8	Total	Observed	0	1	2	3	4	5	6	7	8	Total
0		3	4	2						9	0		2	6							8
1		5	9	3						17	1		3	15	3	1					22
2		3	10	8	3	1				25	2		3	16	12						31
3		1	13	5	3	1				23	3			5	7	2					14
4			2	7	3	2				14	4			8	8	4	1				21
5			1	5	1		1			8	5				3	1					4
6			1	2	1					4	6				2	1					3
7					1	1			1	3	7							1			1
8					1					1	8										0
Total	0	12	40	32	13	5	1	0	1	104	Total	0	8	50	35	9	1	1	0	0	104

Table 7.3: Predicted and observed in sample frequencies for the goals scored by the home team and goals scored by the away team

also shows the resulting numbers comparing all methods. There we see that we approximately predict 25% of the scored goals correctly, with a slightly better performance for the goals scored by the away team. Table 7.3 shows the contingency tables for both, where we can observe that in both cases we fail to predict games with no goals and tend to underestimate the scored goals in high scoring games. When trying to predict the threeway outcome of game i using $\hat{\lambda}_i - \hat{\mu}_i$ we can see in Table 7.2, that we are able to predict more than half of the games correctly. In Table 7.4 we see the contingency table of both, the method using $\hat{\lambda}_i - \hat{\mu}_i$ and using the highest predicted probability by the Skellam distribution as predicted outcome. This immediately shows a major drawback of the latter. It is unable to predict draws. This however might relate to the fact, that only 19 out of the 104 games in the sample where draws. Figure 7.2 shows the boxplots with the resulting predicted probabilities. The same holds, as previously mentioned, when drawing random samples with a large number of replications. Even though this is a massive drawback when trying to predict the correct outcome, it does not prevent us from using the predicted probabilities as further insight. When drawing random samples on game basis we result in the same predictions as using the Skellam distribution. However, if we draw the random samples on tournament basis, we can simulate tournament trees, the scored and received goals and standings in the table at different timepoints. Also the Skellam distribution might be useful when

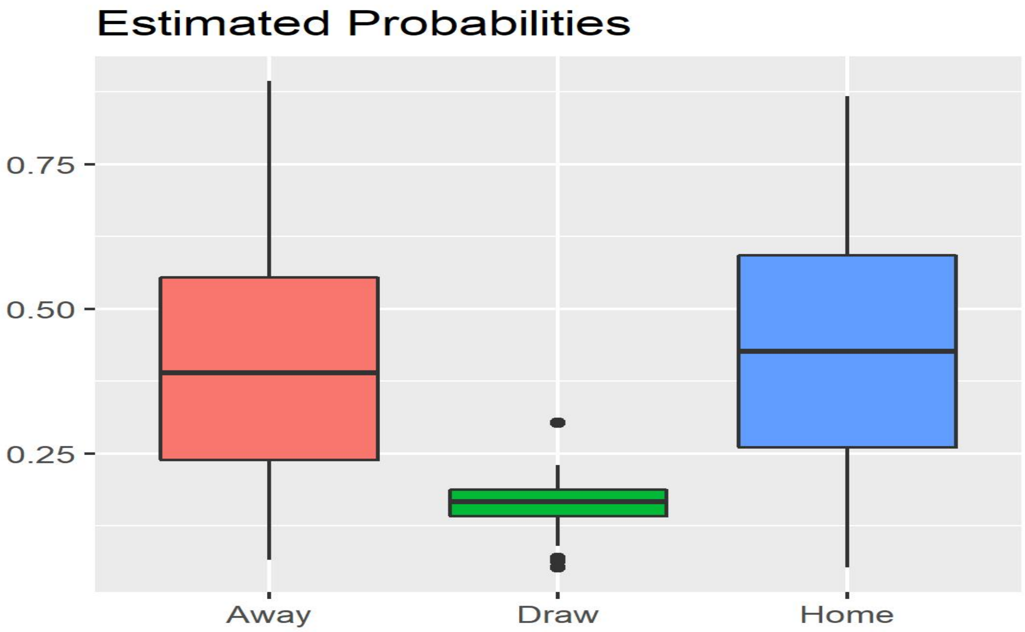

Figure 7.2: Boxplots of the predicted probabilities using the Skellam Distribution

not discussing game results after 60 minutes, but the full time result. It is very common to play an overtime if an ice hockey game has not seen a winner after 60 minutes.

7.1.3 Possible Extensions and Further Development

The assumptions made in the previous sections are very strong. There are several extensions possible to improve the model. As proposed by Lee (1997), it is possible to determine ability parameters, which can be separated into offensive and defensive parameters. In Maher (1982), we see an even more specified approach, in which we additionally assume that the offensive and defensive abilities are differently determined, when playing home or away. This would result in intensity parameters

$$\lambda = \exp(\beta_0 + \tau_H - \gamma_A)$$
$$\mu = \exp(\beta_0 + \tau_A - \gamma_H).$$

$\hat{\lambda}_i - \hat{\mu}_i$

	Predicted			
Observed	Away win	Draw	Home win	Total
Away win	23	5	6	34
Draw	4	7	8	19
Home win	8	15	28	51
Total	35	27	42	104

Skellam Distribution

	Predicted			
Observed	Away win	Draw	Home win	Total
Away win	26	0	8	34
Draw	9	0	10	19
Home win	14	0	37	51
Total	49	0	55	104

Table 7.4: Predicted and observed in sample frequencies for $\hat{\lambda}_i - \hat{\mu}_i$ and the Skellam Distribution

Here τ_H denotes the attacking ability of the home team, γ_A the defensive ability of the away team, while τ_A is the attacking ability of the away team and γ_H the home team's defensive ability. It is clearly not necessary to account for a home effect.

Secondly, assuming independence between scored goals of two teams, which actually compete against each other does not seem natural. Dixon and Coles (1997) found empirically that the assumption of independence is reasonable for football matches, except for low scoring games. They therefore added an dependence term for these low scoring games in the joint distribution (7.1). Karlis and Ntzoufras (2003) discuss the effects of using a bivariate Poisson distribution, which allows a dependence term by definition. McHale and Scarf (2007) use copulas to generate bivariate Poisson distributions with flexible dependence structure.

7.2 Prediction of Match Outcomes using Ordinal Regression

In the following we will apply the developed ordinal regression methods on our real data example. Starting with a discussion of how to choose the response categories for our data situation and the advantages of using a 4-point scale, we will then see the problems arising when fitting an unregularized proportional odds version of the cumulative logit model as described in Chapter 3. We will use these problems as a motivation to see why it is necessary to use some form of regularization. Explicitly we are discussing the model fit and tuning procedure for ridge regression and the least absolute shrinkage and selection operator (lasso) and compare their respective properties with regard of our data situation. Additionally we will see different models using the elastic net penalty term, discuss the tuning procedure, compare the results to ridge regression and lasso and shortly discuss if using other logits, building ratios of our data or evaluating the teams performances separately would improve the model performance. All regularization methods can be found in Chapter 5.

We will then discuss the performance of the different models fitted, again using in sample accuracy, jackknife and data splitting. Since our models do not predict an actual category but probabilities for each category, we will simply use the one with the highest predicted probability as predicted category. It is to mention that predicted probabilities give a more differentiated discussion basis when evaluating team performance, hence in practice it might be more reasonable to discuss the probabilities for the respective outcomes. As an example, imagine the situation where the fitted model predicts probabilities for 4 categories and results in values like $(0.23, 0.24, 0.27, 0.26)$. However, to evaluate the accuracy of our predictions we are going to discuss it with respect to the categories. Finally, we will compare the results from using a log-linear Poisson model, as derived in Section 7.1, to the ones obtained by ordinal regression.

7.2.1 Choice of Categories

When discussing ordinal regression models, one crucial point is how the response categories are chosen. In some situations the categories are already determined by the observed responses, i.e. in a study where every participant can choose between five possible answers like totally agree, agree, don't know, do not agree and totally disagree, the categories are already set. In our case, however, we have a more flexible situation. As we use the on-ice performance after 60 minutes it seems somehow natural to discuss the result after 60 minutes. This means that we allow for draws, even though there would be an overtime following and if necessary a shootout to determine a winner of the respective game. In Section 7.1.2 when using the Skellam distribution, we faced some difficulties in actually predicting draws. The same problem occurred with most of the fitted regularized ordinal regression models. This is the reason why we decided to replace all draws with the final result, i.e. the standing after overtime and shootout.

Schauberger, Groll, and Tutz (2018) found that using a 5-point scale instead of a 3-point scale showed improved performance when predicting match outcomes for the German football Bundesliga. As they included the possibility of draws in their model, a 5-point scale means that the away win and the home win are both divided into high and close wins. The main idea, is that using a 5-point scale instead of a 3-point scale possibly provides more information on the dominance of the respective performance and therefore might increase the explanatory power of the on-ice covariates. Since we exclude the possibility of draws we use a 4-point scale as defined in (7.4). The categories are chosen in this manner to take care of the possibility of an empty net goal scored in the end of the game. Empty net goals are usually scored in games in which the score difference is one and the team trailing by one pulls the goalie to replace him by a skater. Which is common practice and often actually results in an additional goal for the team leading. We therefore let our response variable Y be out of four categories and set

$$Y_i = \begin{cases} 1 & \text{if the away team wins with 3 or more goals difference,} \\ 2 & \text{if the away team wins after OT or with 1 or 2 goals difference,} \\ 3 & \text{if the home team wins after OT or with 1 or 2 goals difference,} \\ 4 & \text{if the home team wins with 3 or more goals difference.} \end{cases} \tag{7.4}$$

We are going to refer to category 1 and category 4 as high wins and to category 2 and category 3 as close wins.

We again use the same data set as before. However, since we use the outcomes of the respective games as response, we only result in 104 observations. For each observation we have 253 possible covariates. These are two times the 122 possible covariates in Section 7.1.2 plus 9 information that can not be identified with one of the two competing teams, for example the total challenges in a game. The observed goal differences after 60 minutes are shown in Figure 7.3, where the 19 games with a goal difference of zero, went to overtime. When classifying the observed games as described in (7.4), we result in the following numbers

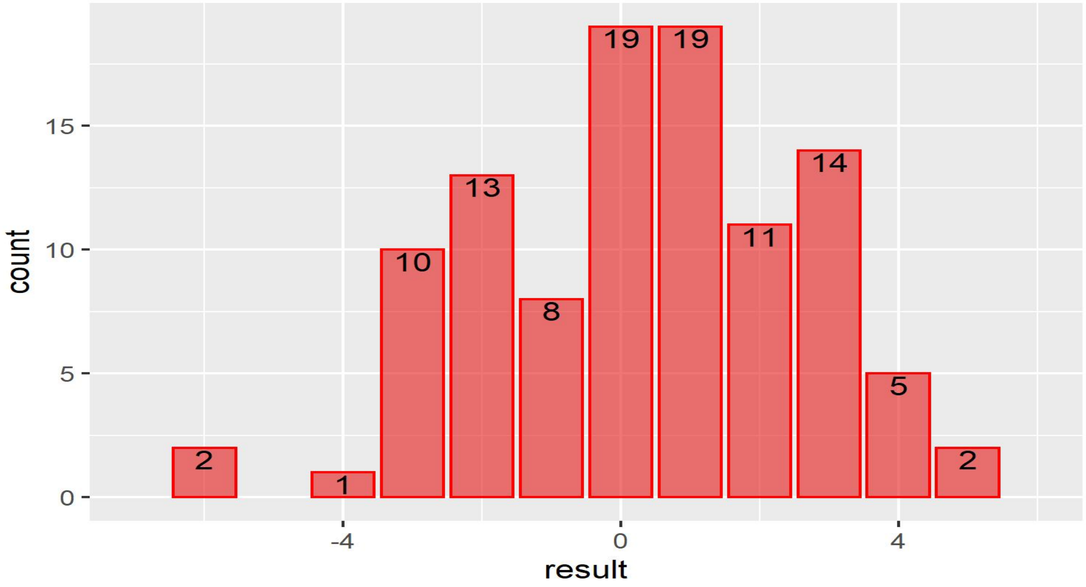

Figure 7.3: Observed goal differences after 60 minutes of play

Category	1	2	3	4	Total
Number of games	13	34	36	21	104

In comparison to the setting when using a log-linear Poisson-model, we do not specifically need to add, respectively test for the need of a homeeffect. This is due to the fact that including a homeeffect would effectively only mean that we shift the respective thresholds or vice versa the flexible thresholds already include the homeeffect. Additionally, since we are going to find a model for the categorical probabilities, for which we use information from both teams, the parameter estimates should automatically adjust for the home advantage.

7.2.2 Unregularized Cumulative Logit Model

First we start with an unregularized proportional odds version of the cumulative logit model as described in Chapter 3. To fit the model we are going to use the function `clm` out of the R-Package `ordinal` provided by Christensen (2019). We then use a stepwise selection procedure, choosing the AIC as criterion. When simply running the stepwise procedure, in form of the R-function `stepAIC`, we

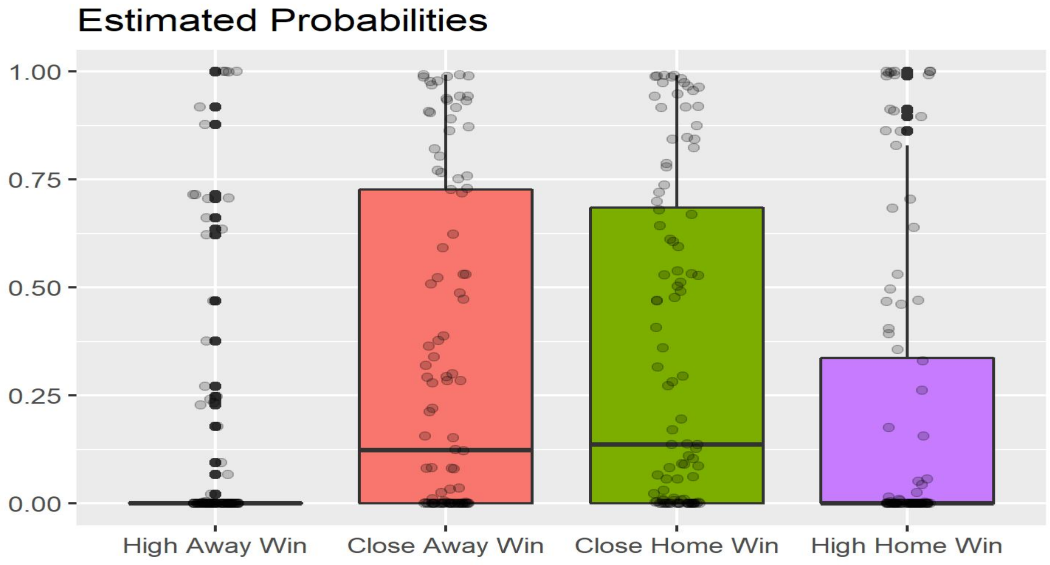

Figure 7.4: Boxplots of the predicted probabilities for an unregularized model

observe numerical instabilities as NAs for the estimated standard errors, z-values and p-values or very large parameter estimates and a condition number larger than 10^8, while Christensen (2018) suggests a condition number less than 10^4. We therefore, in the truest sense, take one step back and simply use the model fit right before facing these problems. Then we exclude all covariates with p-values which exceed 0.05, which results in the model presented in Table A.2 found in the Appendix. The mentioned function fits a model using the alternative parametrization described in Section 3.2.

It is also noticeable that the estimates with the highest standard errors, are the ones for which a strongly correlated covariate is also included in the model. For example, the variables `PassesACdumpout_h` and `Passesdumpout_h`, both have relatively large standard errors and a Pearson correlation coefficient of 0.954. The same holds for other pairs. When checking the variance inflation factor, as defined in Section 5.1, we observe several covariates exceeding the threshold of 10 by far, and some being slightly below this threshold. This is a strong indication that we are facing some form of multi-collinearity. Figure 7.4 shows a box plot of predicted probabilities for the four categories with a grouping at zero and one,

	log-likelihood	misclassification rate
five-folds	2.229673	3.082796
ten-folds	1.986049	7.961686
leave-one-out	2.178668	3.624904

Table 7.5: Tuned ridge regression λ-parameters for out of sample log-likelihood and misclassification rate

which is clearly displaying that we face over-fitting. All these problems together are suggesting that we are in need of some form of regularization. We are therefore going to have a look at the same type of model being fitted with a penalty term.

7.2.3 Ridge Regression and Lasso

The first model we are going to discuss is ridge regression, i.e. including a L^2-penalty in the penalized likelihood. As described in Section 5.2.1, ridge regression does not perform variable selection. Therefore we will result in a model which includes all 253 covariates. Even though, we intend to reduce the predictors included in the model to increase the interpretability, we still fit the model using ridge regression, simply to compare the results and the respective performances.

To tune the parameter λ we use the function `ordinalNetTune` included in the package `ordinalNet`, provided by Wurm et al. (2020). As tuning criteria we choose out of sample log-likelihood and the out of sample misclassification rate, both on five-folds, ten-folds and 104-folds (leave-one-out) and each on 200 different λ-values. We therefore result in six tuned λ-values, shown in Table 7.5.

The second regularized model we are going to discuss for our real data example is lasso, i.e. including a L^1-penalty in the penalized likelihood. As already discussed in Section 5.2.2, the major advantage of lasso is that it performs variable selection. Again we will use the function `ordinalNetTune` and tune the parameter λ on five-folds, ten-folds and 104-folds, each on 200 different λ-values being equidistant on the log-scale. Performing the tuning procedure results in the values presented in Table 7.6.

	log-likelihood	misclassification rate
five-folds	0.1075619	0.07601592
ten-folds	0.1324679	0.03976468
leave-one-out	0.1152946	0.102697

Table 7.6: Tuned lasso λ-parameters for out of sample log-likelihood and misclassification rate

In Table 7.5 and Table 7.6 we can already observe what will become even more obvious when tuning the λ parameters for the elastic net penalty. The estimates for λ using the misclassification rate are more unstable compared to the ones tuned by the log-likelihood. Since we want to choose one specific λ-value for each model, we will choose it out of the available values tuned by the out of sample log-likelihood and base our decision on the respective AICs and BICs. Consequently, we choose the value determined by ten-fold cross-validation for ridge regression ($\lambda = 1.986$) and the one determined by five-fold cross-validation for the lasso ($\lambda = 0.108$). As already mentioned, ridge regression does not perform variable selection, but shrinks the respective parameter estimates by the structure of the penalty term. Lasso does perform variable selection, i.e. sets several parameter estimates to zero. For the chosen lasso model we result in a total of 8 predictors and 3 thresholds included in the model.

Figure 7.5 shows a comparison of the coefficient paths for ridge regression and lasso. To not overload Figure 7.5 we have randomly chosen the path of 80 predictors out of the 253 available. While it seems, that we see all 80 path for ridge regression, the number seems to be less for the lasso. This is due to the fact, that some of the coefficients chosen were estimated to be zero for all shown λ-values.

Figure 7.6 shows the number of predictors included in the lasso model plotted against the tuning parameter λ. There we can see nicely, how increasing the tuning parameter λ, i.e. increasing the influence of the penalty term, reinforces the variable selection property. Even though we have two completely different models in terms of non zero coefficients, with the lasso including 8 predictors and

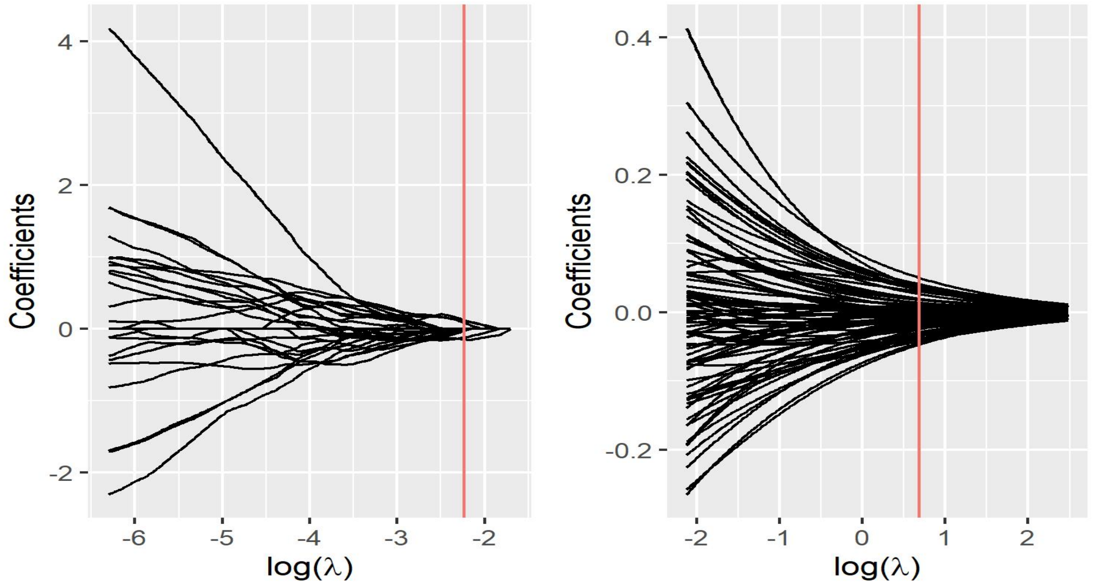

Figure 7.5: Coefficient paths for 80 randomly chosen predictors, with the lasso on the left and ridge regression on the right. The red line represents the tuned λ-values

ridge regression including all possible covariates, Table 7.7 shows a very similar behaviour. Both models are very unlikely to predict high wins. This inability might be explained by the fact that we choose the category with the highest predicted probability and that, at least for the two models, the probability of a high win is almost always less than the probability of a close win. Even though this might be true for a lot of games, there should be at least some games where the respective performance was so dominant that a high win is the most likely possibility.

7.2.4 Elastic Net

As we discussed in Section 5.2.4, the elastic net combines advantages of both, lasso and ridge regression, simply by using a convex combination of the penalty terms. In comparison to ridge regression and lasso, where we only need to tune one parameter, using the elastic net penalty asks for two parameters to be tuned, namely α and λ. Since these are two dependent parameters we are going to denote λ by $\lambda(\alpha)$ where it is needed.

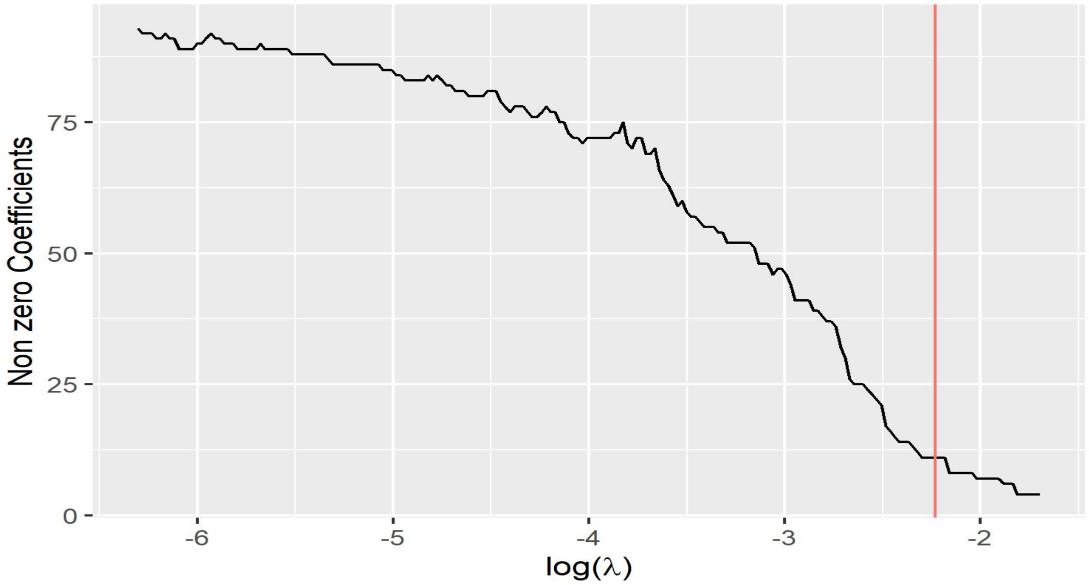

Figure 7.6: Number of non zero coefficients for the lasso, with the red line representing the tuned λ-value

Since computational costs are an issue when tuning the parameters for the elastic net penalty we restrict ourselves on 21 possible α-values with equidistant values between 0.05 and 0.95 and 0.01 and 0.99 as most extreme values. For each α-value we again tune the λ-values using the function `ordinalNetTune` on 200 possible values using five-fold, ten-fold and leave-one-out cross-validation, all using the out of sample log-likelihood and the out of sample misclassification rate as tuning criterion.

It is to mention that `ordinalNet` uses a different parametrization of the convex combination in the elastic net penalty term (5.6). While in our case α corresponds to the weight on the L^2-penalty and $(1-\alpha)$ to the weight on the L^1-penalty, the implemented parametrization changed the roles of the weights, i.e. $P(\lambda, \boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} \left((1-\alpha)\beta_j^2 + \alpha|\beta_j|\right)$. However, we are going to use the parametrization described in Section 5.2.4, when discussing results or referencing on models.

Lasso

	Predicted				
Observed	1	2	3	4	Total
1		10	3		13
2		16	18		34
3		7	29		36
4		4	17		21
Total	0	37	67	0	104

Ridge Regression

	Predicted				
Observed	1	2	3	4	Total
1		13			13
2		24	10		34
3		6	30		36
4		1	15	5	21
Total	0	44	55	5	104

Table 7.7: Observed and predicted in sample frequencies for the response categories using lasso (left) and ridge regression (right)

In Table A.1, found in the Appendix and visualized in Figure 7.7 we see a comparison of the tuned $\lambda(\alpha)$-values for both tuning criteria and the three mentioned cross-validations. As mentioned in the previous section, we can clearly observe, that using the out of sample log-likelihood as tuning criteria produces more stable results. This might relate to the small sample size and the therefore resulting very sensitive misclassification rate. For example using ten-fold cross-validation means that our test fold only consists of 10 games, which by chance can only consist of one specific category and therefore heavily influence the tuned λ parameter. We can observe these stability problems in several λ-values being close to zero, which implies that cross-validation identifies an nearly unpenalized model as the best choice. A small λ-value also implies a larger number of non zero coefficients, which at least for $\alpha > 0.5$ results in even more predictors included in the model compared to the unregularized version. All together we therefore decided to only use the values provided by the log-likelihood and again base our final decision on the respective AICs and BICs.

To compare some result we choose $\alpha = 0.2, 0.8$ and the corresponding $\lambda(0.2) = 0.1455$ and $\lambda(0.8) = 0.5376$. A somehow surprising outcome is that for all α values smaller than 0.6 we result in very similar results, with all models including almost the same predictors and very similar coefficients. This might relate to the increasing influence of the L^1-penalty term and the implied variable selection property. Additionally, these predictors are also part of the models including more predic-

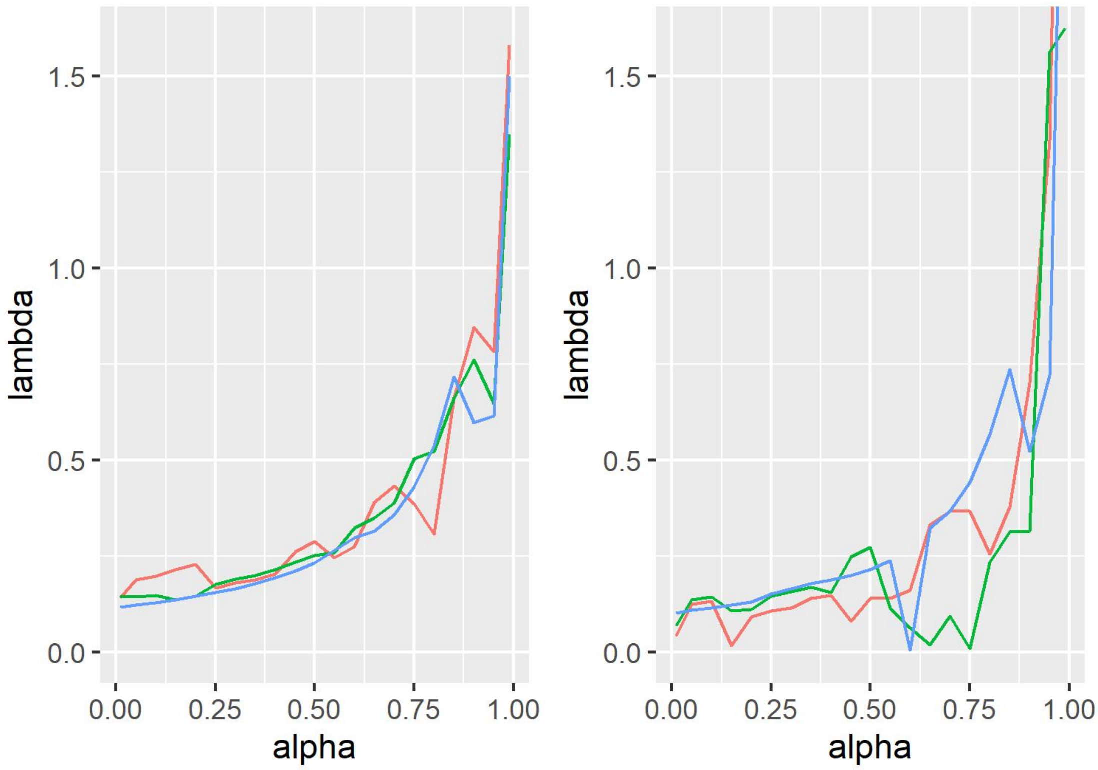

Figure 7.7: $\lambda(\alpha)$ for different α values, comparing out of sample log-likelihood (left) and out of sample misclassification rate (right) using five-fold (red), ten-fold (green) and leave-one-out (blue) cross-validation

tors, we therefore can identify them as the most influential. However, we would expect the number of non zero coefficients to steadily decrease with increasing weight on the L^1-penalty term and not to stagnate that early. From a mathematical point of view this might relate to the low number of observations compared to the high number of possible covariates and that most predictors do not or at least only have little influence on our dependent variable and are therefore set to be zero by the lasso part of the penalty term. From a sportive point of view, especially the latter might be a reasonable explanation. It is not uncommon to hear a player or coach to say something like "It's the little things that count". And there are a lot of little things, all having their possibly little influence on the result.

Figure 7.8 shows the coefficient paths for the two chosen models. Especially the

$\alpha = 0.2$

	Predicted				
Observed	1	2	3	4	Total
1		10	3		13
2		16	18		34
3		8	28		36
4		4	17		21
Total	0	38	66	0	104

$\alpha = 0.8$

	Predicted				
Observed	1	2	3	4	Total
1		9	4		13
2		15	19		34
3		7	29		36
4		4	17		21
Total	0	35	69	0	104

Table 7.8: Observed and predicted in sample frequencies for the response categories using elastic net with $\alpha = 0.2$ (left) and $\alpha = 0.8$ (right)

graph for $\alpha = 0.2$ looks very similar to the graph for the lasso in Figure 7.5. In Table 7.8 we see the observed and predicted in sample frequencies for both model, where we can observe that their predictive performance is very similar and also very close to the one by the lasso, shown in Table 7.7. Its again obvious that both models are unable to predict high wins and that there is some kind of home bias. We will compare all models and their performance in further detail in Section 7.2.6.

7.2.5 Possible Adjustments

In the following we are trying to solve some of the problems arising in the previous sections. We are therefore going to discuss if we can improve the models with slight changes on the structure of the data or by using other logits.

When comparing the covariates included in the lasso or elastic net models we can observe that most of the used information is from the home team. We therefore start by adjusting the data in such a way, that we force the model, when including a predictor from the home team to also include the respective predictor from the away team. We therefore create new variables, which are simply the ratio of the two respective variables and fit the model with these new variables, again in a standardized form. To tune the parameters and to fit the resulting models we are again using the same procedure as used in the previous sections. However, we are not going into further detail and simply compare the results for lasso and

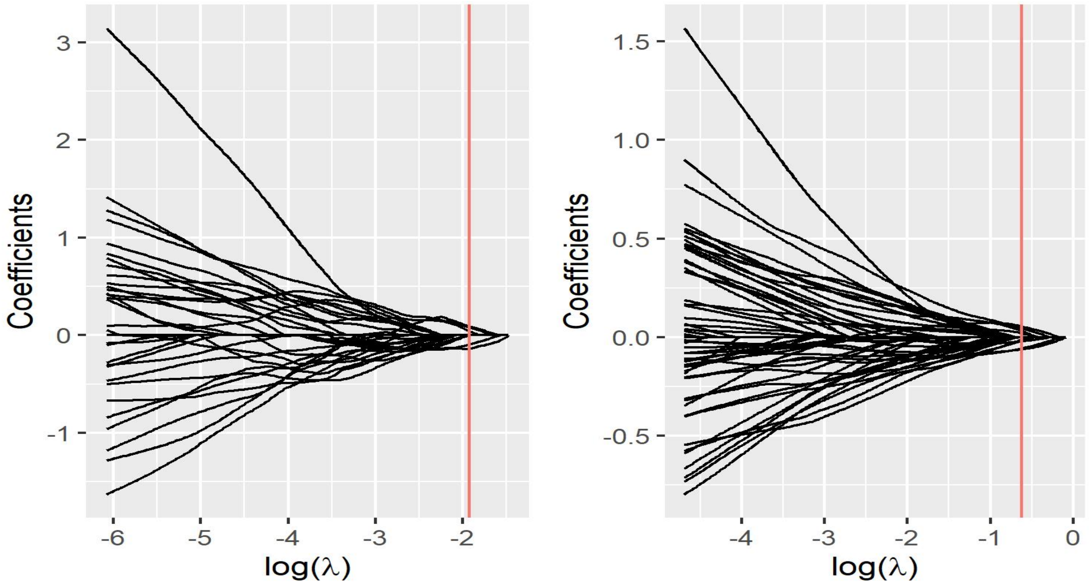

Figure 7.8: Coefficient paths for 80 randomly chosen predictors, with $\alpha = 0.2$ on the left and $\alpha = 0.8$ on the right. The red line represents the tuned λ-values

elastic net with $\alpha = 0.2$ and $\alpha = 0.8$. Since using lasso and using the elastic net penalty with $\alpha = 0.2$ results in very similar models, both including the same predictors with very close parameter estimates, we are only going to compare the two elastic net models.

It is noteworthy that using the ratio variables results in a significant higher number of predictors included in the model. For the lasso and the elastic net penalty with $\alpha = 0.2$ we include 23 predictors, which are 46 when splitting them to the initial variables. However, as we can observe in Table 7.9, using the ratios instead of the initial variables substantially reduces the home bias and also, at least for the home wins, enables the model to predict high wins correctly. Even though we include a high number of initial variables in the models we can not observe numerical instabilities or over-fitting. Since the performances of the two models are very close, with the one using the elastic net penalty with $\alpha = 0.8$ including a higher number of predictors (76 initial variables), we are only going to compare the other model, namely using the elastic net penalty with $\alpha = 0.2$ and $\lambda(0.2) = 0.0572$, in

$\alpha = 0.2$

	Predicted				
Observed	1	2	3	4	Total
1		13			13
2		24	10		34
3		7	27	2	36
4		1	11	9	21
Total		45	48	11	104

$\alpha = 0.8$

	Predicted				
Observed	1	2	3	4	Total
1		13			13
2		25	9		34
3		6	28	2	36
4		1	14	6	21
Total		45	51	8	104

Table 7.9: Observed and predicted in sample frequencies for the response categories using elastic net with $\alpha = 0.2$ (left) and $\alpha = 0.8$ (right) on the ratios of the paired covariates

Section 7.2.6.

As described in Section 7.2.2, we faced serious numerical issues when we fitted an unregularized version of the cumulative logit model. This might be due to the high number of possible explanatory variables compared to the sample size. We are therefore going to discuss the results if we adjust our data set in a very similar way as seen in Section 7.1, where we used a log-linear Poisson model to describe the number of goals for both team seperately. The main idea is that we do not describe the outcome of the game depending on the performances of both teams but use only one team's performance to predict the match outcome. For that purpose, we introduce the factor 'home', which indicates whether the respective team played home or away. In addition, we allow for all interactions between the explanatory variables and this factor, to adjust for possible different effects for the home and the away team. Additionally, we have to slightly modify the ordinal response variable in such a way, that ascending from 1 to 4 we model high and close losses and close and high wins. For instance, a high win from the home team which corresponds to a 4 in the original classification (7.4), would remain a 4 for the home team in the modified case but change to a 1 for the away team. The main benefit of this adjustment is that we double the available sample size by halving the number of possible covariates.

	Predicted				
Observed	1	2	3	4	Total
1	13	17	4	0	34
2	6	44	20	0	70
3	1	20	41	8	70
4	0	1	20	13	34
Total	20	82	85	21	208

Table 7.10: Observed and predicted in sample frequencies for the response categories using the teams performances separately

To fit the model with the modified data, we again use the same R-functions. However, we do not face numerical issues as we did without the modifications. This might be due to the new dimensions of the modified data and the resulting stabilisation of the estimates. Finally we use the model resulting from the stepwise selection procedure and exclude all covariates with p-values which exceed 0.1. We therefore result in the model presented in Table A.3 found in the Appendix. There we can observe that even though the thresholds are flexible they are almost symmetric around zero. Further there are no notably large parameter estimates or standard errors, as we observed when using both teams performances simultaneously. In Table 7.10 we can see, that we not only improve the numerical stability of the model fit but also result in a nicely fitting model. However, Table 7.10 also shows a major disadvantages of this approach, namely that the predictions are not symmetric, which means that we might face two different predictions for the same game. This is due to the fact, that we predict the outcome of the game separately for both teams, i.e. result in two predictions, which are not necessarily the same. In our sample we actually observed this difference in predictions for almost half of the games. Therefore we have to be cautious when interpreting the results. However, this is not only a disadvantage, since the mentioned approach allows to interpret the performances of the teams separately. As an example imagine a game which ended with a close win for the home team, while the model predicted a close win for the home team and a close win for the away team. This states that the performance of the home team was good enough to be a deserved winner. However also the away team showed a good performance and the outcome of the

	In Sample	Jackknife	Data Splitting
Teams separately	53.37%	44.23%	43.77%

Table 7.11: Model accuracy for the model evaluating the teams performances separately using in sample accuracy, jackknife and data splitting

game, from their point of view, might be seen as bad luck. Table 7.11 shows the predictive performance of the model.

In the following we are going to discuss if using different logits, as described in Chapter 3 or changing the assumption of the underlying distribution, i.e. using the probit link instead of the logit link, would improve our model performance. Since the package `ordinalNet` allows to specify the logits and the link function in the function call, we can simply adjust and compare the results. We again use the initial data set, not building the ratios. When tuning the parameter λ we again use cross-validation and the out of sample log-likelihood. Without discussing further details, we can observe that there is no significant improvement when applying any of the adjustments mentioned above. Unfortunately, in R there is no function or package available which combines structured thresholds and regularization methods for ordinal responses. Since the unregularized model has shown numerical instabilities and multi-collinearity, we are not going to discuss the possible effects of structured thresholds, even though they might provide a good solution to the home bias and the lack of fit for high wins. While forcing the thresholds to be symmetric should improve the symmetry of the fits, equidistant thresholds might be beneficial regarding high wins.

7.2.6 Discussion of Results

To now evaluate the predictive performance of several models, namely the described unregularized proportional odds version of the cumulative logit model, the models resulting from using ridge regression and lasso, the elastic net models with $\alpha = 0.2$ and $\alpha = 0.8$ and the model using the ratios of the initial variables and the elastic net penalty term with $\alpha = 0.2$, we compare their accuracy regarding in sample, jackknife and data splitting. All models were fitted using 4 categories

	In Sample	Jackknife	Data Splitting
Unregularized	88.46%	59.61%	58.26%
Ridge	56.73%	41.34%	39.98%
Lasso	43.26%	36.53%	33.50%
EN $\alpha = 0.2$	42.31%	37.50%	33.23%
EN $\alpha = 0.8$	42.31%	37.50%	32.61%
EN ratio $\alpha = 0.2$	57.69%	44.23%	41.73%

Table 7.12: Comparison of the model accuracy using in sample accuracy, jackknife and data splitting on a 4-point scale

	In Sample	Jackknife	Data Splitting
Unregularized	95.19%	87.50%	84.59%
Ridge	84.61%	61.50%	61.68%
Lasso	70.19%	60.57%	53.90%
EN $\alpha = 0.2$	68.27%	57.69%	53.44%
EN $\alpha = 0.8$	69.23%	55.78%	54.97%
EN ratio $\alpha = 0.2$	83.65%	76.92%	72.98%

Table 7.13: Comparison of the model accuracy using in sample accuracy, jackknife and data splitting on a 2-point scale

and can therefore be used to predict the outcome of a game on a 4-point scale. However, a 4-point scale can be easily combined to the 2-point scale, which is the reason why we are going to discuss their accuracy for both. The model in which we evaluate both teams separately is somehow not comparable to the others, since we might get two different predictions per game. However, Table 7.11 shows the accuracy for this model and it is noteworthy that this model compared to the ones given in Table 7.12 shows promising results. Although Fagerland and Hosmer (2013) and Pulkstenis and Robinson (2004) warn to be very cautious when using their goodness of fit test when facing a sample size as small as ours, we are shortly discussing the problems arising when using the implemented functions in the R-package `generalhoslem` provided by Jay (2019).

Table 7.12 displays the accuracy of the respective models on a 4-point scale. We can observe that the unregularized model seems to show the best results, especially regarding the in sample accuracy. However, as described previously this model shows strong signs of overfitting. Another interesting result is, that the models using lasso and the elastic net penalty term show very similar results. It is especially surprising that this is seemingly independent of the choice of α. While the model using the L^2-penalty might be impractical due to the fact that the model includes all possible covariates, the most promising results can be observed when manipulating the appearance of the possible covariates to be ratios. While this substantially increases the number of predictors in the model, it also solves the problem, that most models exclusively include predictors related to the home team, which might be the reason that we can observe a home bias.

To now discuss the results on a 2-point scale, which means that we only distinguish between home and away wins, we simply combine the predicted probabilities for high and close wins for both, home and away and use the one with the higher predicted probability as predicted outcome. Clearly, the home bias is again present. The respective accuracies are shown in Table 7.13. It is not surprising that we can draw very similar conclusions compared to the 4-point scale. Again the unregularized model shows the best results with the model using the ratios shows the most promising results. While predicting the correct winner in more than 70% of the games seems to be a reasonable result, slightly above 50% is somehow unsatisfying and not much of an improvement to simply guessing.

It is somehow intuitive to fit a binary logistic regression model when trying to predict the probabilities for a variable with two possible values. When fitting an unregularized version of this model, we can again observe the numerical instabilities we faced when fitting our ordinal model. When using the R-package `glmnet` provided by Friedman et al. (2010) and the corresponding functions to fit a model with the L^1-penalty, using the default measure in cross-validation, which uses the deviance for tuning, we result in a λ-value where the resulting regularized model sets all parameter estimates to be zero. This strongly corresponds to the results we observed when fitting the elastic net models in Section 7.2.4, where all models with $\alpha < 0.6$ included the same and small number of predictors. Again observing this property when trying to fit a regularized logistic regression model is a strong

	C_5	C_6	C_{10}
Unregularized	0.999	0.936	0.996
Ridge	0.001	0.001	0.014
Lasso	0.017	0.123	0.148
EN $\alpha = 0.2$	0.001	0.050	0.009
EN $\alpha = 0.8$	0.013	0.034	0.015
EN ratio $\alpha = 0.2$	0.039	0.309	0.398

Table 7.14: p-values for the Fagerland and Hosmer test, with 5, 6 and 10 groups

indication that most covariates have little influence on the response and are set to be zero when we penalize that they are included in the model.

A clear indication that we face some issues regarding our small sample size can be found in the fact that when we try to determine the number of groups for the goodness of fit tests by the formula suggested by Lipsitz et al. (1996), $6 \leq g \leq n/5c$. Since in our case $n = 104$ and $c = 4$ we result in $6 \leq g \leq 5.2$ which is a clear contradiction, due to the small sample size n. Additionally in the available package, the Lipsitz test is not implemented for regularized ordinal regression models. Furthermore, since the Pulkstenis and Robinson tests construct a contingency table using the categorical covariate patterns, while we have not observed any categorical covariates, these goodness of fit tests are not applicable on our data situation. All together, we remain with the Fagerland and Hosmer test statistic and compare their results where we set the number of groups g to 5, 6 and 10. Table 7.14 shows the respective p-values. We can again see that we face sample size problems, due to the fact that the results differ widely by just changing the number of groups, while Fagerland and Hosmer (2013) stated that their test statistic do not depend strongly on the number of groups. Additionally for almost all tests we get the warning that the Chi-square approximation might be incorrect because at least one expected cell frequency is less than 1. When having a look at the form of the test statistic given in (6.4), where we divide by E_{kj}, it becomes obvious why this might give missleading results. Even though the results of the tests performed might not be trustworthy, we can still observe results which somehow go in line with the previous ones. Although the p-values differ strongly between the tests choosing different numbers of groups it indicates that using the elastic net penalty and ridge regression might result in poor fit. Additionally the high p-values for the unregularized model might be reasoned with the issue that we face overfitting and therefore the tests naturally do not reject the null hypothesis that we fitted a correct model. Besides the unregularized model, the model using the ratios of the initial variables shows the most promising results.

Summarizing, we can conclude that using a regularized proportional odds version of the cumulative logit model, clearly improves the numerical stability and does not show any signs of multi-collinearity. On the other hand, the observed performances are not satisfying when trying to predict the outcome of a game. As already mentioned this might be due to the fact that in ice hockey, in sports in general, there are a lot of things that impact the result, while most of them have little statistical influence. As an example, it might have very low impact if there are two more faceoffs won, while a faceoff in the last minute on a powerplay in the offensive zone could be a crucial point in the game. Instead of treating all information separable, it seemed that comparing the respective team performances clearly improved the predictive performance of the models. This suits the fact, that the two teams actually compete against each other and building ratios takes this into account and also increases the number of covariates included in the model, while preserving the stability benefits we get from using a regularization method. Even though, the unregularized model shows multiple problems it might strongly benefit, as all other models do, from increasing the available sample size. This argument is reinforced by the fact, that the unregularized proportional odds version of the cumulative logit model in which we doubled the sample size and halved the possible covariates by evaluating the teams performances separately did not show any numerical issues and resulted in satisfying results.

7.3 Comparison of the two Approaches

We have seen two different approaches when trying to predict the outcome of a game. In the first, we use a log-linear Poisson model to predicted the number of goals scored by the home team and the away team independently. In the second one we used ordinal regression on 4 categories, incorporating information from both teams to predict the outcome of a game using the predicted probabilities of the respective categories. While using the first approach, where we doubled the available sample size by assuming independence, we did not observe any numerical issues and resulted in a model with 9 included covariates, a totally different situation appeared when trying to fit an ordinal regression model. Without penalizing the log-likelihood, we faced serious numerical issues and a variance inflation factor strongly indicating multi-collinearity. However, simply applying the regularization methods on the available information resulted in models showing unsatisfying results, due to the fact that they shrunk most parameter estimates to zero and included almost only home team information. Building ratios of the respective performance pairs, i.e. shots of the home team and shots of the away team or faceoffs won in the neutral zone by the home team and faceoffs won in the neutral zone by the away team, seemed to show promising results, while increasing the number of covariates included in the model. The ordinal regression model in which we doubled the sample size, analogously as in the log-linear Poisson model, also showed improved numerical stability and promising results. However, one has to be cautious when interpreting the results of this model.

When comparing the predictive performances of the two approaches, we have to distinguish between the one where we predict the outcome of a game using the difference of the two estimated Poisson intensities and all other approaches, which predict the probability of all outcomes. Ultimately, the rules of ice hockey and how they deal with draws after 60 minutes of play imply that predicting the correct winner of a game is the most crucial part, while all others provide additional information. Therefore, we discuss the predictive performance on a 2-point scale, with two possible outcomes, namely a home win or an away win. While predicting probabilities gives a more differentiated performance evaluation base, predicting the number of goals scored by the home and away team is also a reasonable approach. While using the Skellam distribution always allows for draws, simply because the difference of two Poisson variables can be zero, using the difference of the intensity estimates allows to predict on a two point scale. This can be achieved by not defining an interval in which we set the predicted result to a draw and to define that if the difference is exactly zero, we predict an home win (even though this might be very rare). As already mentioned and observable in

	In Sample	Jackknife	Data Splitting
$\lambda_i - \mu_i$	67.31%	63.46%	65.38%
EN ratio $\alpha = 0.2$	83.65%	74.03%	72.98%

Table 7.15: Comparison of the model accuracy using in sample accuracy, jackknife and data splitting on a 2-point scale

Table 7.13, the ordinal regression model which shows satisfying results and incorporates both teams performances simultaneously is the one using the ratios of the initial variables.

When comparing the in sample, jackknife and data splitting accuracy as shown in Table 7.15 we can observe that the regularized ordinal model performs better than the one using the estimated Poisson intensity parameters. However, the ordinal model uses more than twice non zero coefficients. When discussing predictive performances in sports, one should remember that there is always some kind of luck involved and often the outcome of the match is not the performance based fair outcome. In our case we can say that in approximately two thirds to three fourth of the games the winner was, from a statistical point of view, a deserved winner, while in the rest we saw a lucky winner. When using the results in practice, the best choice would be to combine the advantages of both approaches.

8 Conclusion

In this thesis we introduced the cumulative logit model which, by assuming an underlying latent variable, takes the ordinal structure of the response variable into account. The idea of penalizing the resulting maximum likelihood function was discussed in detail and several different possible penalty terms were compared. Regularizing is especially useful when facing multi-collinearity or a large number of possible predictors compared to a small sample size. Further, to assess goodness of fit we discussed modifications of existing goodness of fit tests on ordinal data. The theoretical concepts where then applied to a real data example in sports.

In this practical part we faced a sample of 104 ice hockey games, where we observed 253 possible covariates, with each covariate either describing the home team's performance, the away team's performance and some providing general information. The aim was to find a model which allows for predictions on the outcome of a game, without knowing how many goals were scored by the teams. We therefore started with a log-linear Poisson model, estimating the number of goals scored by both teams separately. By discussing the teams performances separately we doubled the available sample size while halving the possible covariates, which seemed to be sufficient to not face numerical issues and to result in a nicely fitting model. When modelling the expected number of goals we found a model, which does not distinguish between the home and the away team. This means that the number of goals scored is not influenced whether the respective team played at home or away. The resulting model showed satisfying results, however was unable to predict zero goals scored. This might be seen as an argument that not conceding a goal in an ice hockey game is a special achievement. On the other hand the model tended to underestimate the number of goals scored in high scoring games. With the estimated number of goals scored by both teams, we then have different possibilities to assess the outcome of a game, all having their advantages and disadvantages.

In the second part of the practical part we used the developed theoretical concepts to model the outcome of the game directly by ordinal regression. We saw that the unregularized version of the cumulative logit model had several issues indicating numerical problems, most of them relating to the high number of possible covariates compared to a relatively small sample size. We therefore fitted different regularized models, while the one using ridge regression is impracticable since it does not estimate parameters to be zero, the lasso and elastic net penalty showed unsatisfying results. They tended to estimate all parameters to be zero, with a very low number of exceptions, all being information from the home team. This might relate to the fact, that in a sport competitions there is a high number of factors all having a possibly small influence and are therefore estimated to be zero, if we penalize their inclusion in the model. We therefore transformed our data set, to provide reasonable estimates of the match outcome. In the first approach we built the ratios of the pairs of covariates to then fit a regularized model, which showed satisfying results. This indicates that building ratios of the respective teams performances considered the competitive nature of the game and therefore allowed for reasonable predictions with a larger number of predictors being included in the model. In the second approach we transformed the data in the same way as when fitting the log-linear Poisson model, which resulted in a nicely fitting unregularized model evaluating the teams performances separately. This reinforces the argument that we face sample size problems with the initial data set.

Conclusively, even though predicting the correct outcome of a sport competition is always challenging simply because there are numerous unmeasured influences and luck involved, modelling the expected number of goals scored by a log-linear Poisson model and using an ordinal regression model to predict the outcome of a game directly are promising approaches. While the number of goals seemed to not be directly influenced by the corresponding team playing at home or away, predicting the match outcome directly required to take the competition into account or to evaluate the teams separately.

9 References

Agresti, A. (2002). *Categorical Data Analysis.* New York: John Wiley and Sons. (2nd Edition)

Agresti, A. (2010). *Analysis of Ordinal Categorical Data.* New York: John Wiley and Sons. (2nd Edition)

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723.

Alin, A. (2010). Multicollinearity. *WIREs Computational Statistics*, *2*(3), 370-374.

Anderson, J., and Philips, P. (1981). Regression, discrimination and measurement models for ordered categorical variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *30*(1), 22-31.

Bickel, P., Li, B., Tsybakov, A., Geer, S., Yu, B., Valdés, T., ... Vaart, A. (2006). Regularization in statistics. *TEST*, *15*(2), 271-344.

Burridge, J. (1981). A note on maximum likelihood estimation for regression models using grouped data. *Journal of the Royal Statistical Society. Series B (Methodological)*, *43*(1), 41-45.

Casella, G., and Berger, R. L. (2002). *Statistical Inference.* Belmont, CA: Duxbury Press. (2nd Edition)

Christensen, R. H. B. (2018). *Cumulative link models for ordinal regression with the R package ordinal.* (submitted for publication in Journal of Statistical Software)

Christensen, R. H. B. (2019). *ordinal—Regression models for ordinal data.* (R package version 2019.12-10. https://CRAN.R-project.org/package=ordinal)

Dixon, M. J., and Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *46*(2), 265-280.

Fagerland, M. W., and Hosmer, D. W. (2013). A goodness-of-fit test for the proportional odds regression model. *Statistics in Medicine*, *32*(13), 2235-2249.

Fagerland, M. W., and Hosmer, D. W. (2016). Tests for goodness of fit in ordinal logistic regression models. *Journal of Statistical Computation and Simulation*, *86*(17), 3398-3418.

Fagerland, M. W., Hosmer, D. W., and Bofin, A. M. (2008). Multinomial goodness-of-fit tests for logistic regression models. *Statistics in Medicine*, *27*(21), 4238-4253.

Farrar, D. E., and Glauber, R. R. (1967). Multicollinearity in regression analysis: The problem revisited. *The Review of Economics and Statistics*, *49*(1), 92–107.

Flexeder, C. (2010). *Generalized lasso regularization for regression models* (Unpublished master's thesis). Ludwig-Maximilians-Universität München.

Friedl, H., and Embacher, S. (2020). *Convertion and Verification of PDF Data Using R and Excel* (Tech. Rep.). Institute of Statistics, Graz University of Technology.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, *1*(2), 302-332.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22.

Groll, A., and Schauberger, G. (2019). Prediction of soccer matches. In N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, and J. Teugels (Eds.), *Wiley statsref: Statistics reference online.* John Wiley and Sons.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer.

Hesterberg, T., Choi, N. H., Meier, L., and Fraley, C. (2008). Least angle and L1 penalized regression: A review. *Statistics Surveys*, *2*, 61–93.

Hoerl, A. E. (1962). Application of ridge analysis to regression problems. *Chemical Engineering Progress*, *58*(3), 54-59.

Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55-67.

Hosmer, D. W., and Lemeshow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics - Theory and Methods*, *9*(10), 1043-1069.

Hurvich, C. M., and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*(2), 297-307.

Instat. (2020). (`https://instatsport.com/hockey`)

Jay, M. (2019). generalhoslem: Goodness of fit tests for logistic regression models [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=generalhoslem` (R package version 1.3.4)

Karlis, D., and Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *52*(3), 381-393.

Karlis, D., and Ntzoufras, I. (2008). Bayesian modelling of football outcomes: Using the Skellam's distribution for the goal difference. *IMA Journal of Management Mathematics*, *20*(2), 133-145.

Kumar, T. K. (1975). Multicollinearity in regression analysis. *The Review of Economics and Statistics*, *57*(3), 365–366.

Lee, A. J. (1997). Modeling scores in the Premier League: Is Manchester United really the best? *CHANCE*, *10*(1), 15-19.

Lipsitz, S. R., Fitzmaurice, G. M., and Molenberghs, G. (1996). Goodness-of-fit tests for ordinal response regression models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *45*(2), 175-190.

Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, *36*(3), 109-118.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, *42*(2), 109-142.

McCullagh, P., and Nelder, J. (1989). *Generalized Linear Models.* Chapman & Hall. (2nd Edition)

McHale, I., and Scarf, P. (2007). Modelling soccer matches using bivariate discrete distributions with general dependence structure. *Statistica Neerlandica*, *61*(4), 432-445.

Pulkstenis, E., and Robinson, T. J. (2004). Goodness-of-fit tests for ordinal response regression models. *Statistics in Medicine*, *23*(6), 999-1014.

R Core Team. (2019). R: A Language and Environment for Statistical Computing [Computer software manual]. Vienna, Austria. Retrieved from `https://`

www.R-project.org/

Schauberger, G., Groll, A., and Tutz, G. (2018). Analysis of the importance of on-field covariates in the german bundesliga. *Journal of Applied Statistics*, *45*(9), 1561-1578.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461 – 464.

Thompson, R., and Baker, R. (1981). Composite link functions in generalized linear models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *30*(2), 125-131.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

Wichers, C. R. (1975). The detection of multicollinearity: A comment. *The Review of Economics and Statistics*, *57*(3), 366–368.

Wurm, M., Rathouz, P., and Hanlon, B. (2017). *Regularized ordinal regression and the ordinalnet R package.*

Wurm, M., Rathouz, P., and Hanlon, B. (2020). ordinalnet: Penalized ordinal regression [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=ordinalNet (R package version 2.9)

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418-1429.

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *67*(2), 301–320.

Appendix

α	5-fold (l)	10-fold (l)	104-fold (l)	5-fold (mc)	10-fold (mc)	104-fold (mc)
0.99	1.58	1.35	1.50	3.16	1.62	2.69
0.95	0.78	0.65	0.62	1.33	1.56	0.72
0.90	0.85	0.76	0.60	0.70	0.32	0.52
0.85	0.66	0.66	0.72	0.38	0.31	0.74
0.80	0.31	0.52	0.54	0.26	0.24	0.57
0.75	0.39	0.50	0.43	0.37	0.01	0.44
0.70	0.43	0.39	0.36	0.37	0.09	0.37
0.65	0.39	0.35	0.32	0.33	0.02	0.32
0.60	0.28	0.32	0.30	0.16	0.06	0.01
0.55	0.25	0.26	0.27	0.14	0.12	0.24
0.50	0.29	0.25	0.23	0.14	0.27	0.22
0.45	0.26	0.24	0.21	0.08	0.25	0.20
0.40	0.20	0.22	0.19	0.15	0.16	0.19
0.35	0.19	0.20	0.18	0.14	0.17	0.18
0.30	0.18	0.19	0.17	0.11	0.16	0.17
0.25	0.17	0.18	0.16	0.11	0.15	0.15
0.20	0.23	0.15	0.15	0.09	0.11	0.13
0.15	0.22	0.14	0.14	0.02	0.11	0.12
0.10	0.20	0.15	0.13	0.13	0.14	0.12
0.05	0.19	0.14	0.12	0.13	0.14	0.11
0.01	0.14	0.15	0.12	0.04	0.07	0.10

Table A.1: A comparison of the tuned $\lambda(\alpha)$-values for five-fold, ten-fold and leave-one-out cross-validation using the out of sample log-likelihood (l) and the out of sample misclassification rate (mc) as criterion

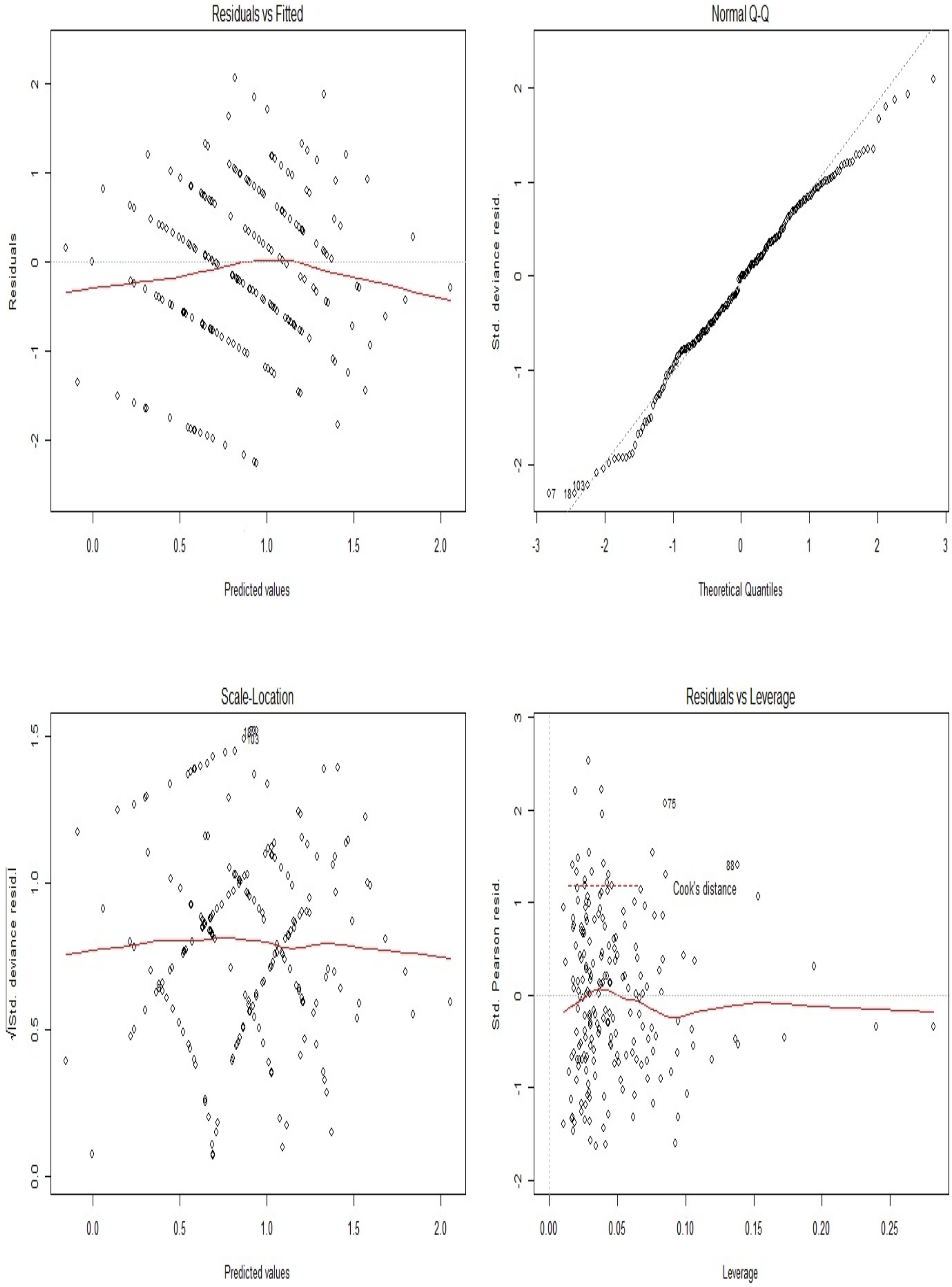

Figure A.1: Diagnostic plots of the fitted log-linear model

Variable	Coefficient	Std.Error	p-value	Mean	Std. Dev.
Entries3P_h	-2.4446	0.8696	0.0049	20.05	4.15
ShotsOG60_h	2.6639	0.9039	0.0032	31.09	7.03
ChallengesStickh_h	-8.2583	2.5225	0.0010	22.32	7.61
ChallengesWNZ_h	5.7984	1.7817	0.0011	9.13	4.11
ChallengesNZ_h	-5.5966	1.8487	0.0024	18.68	7.32
Passes3P_a	-6.6218	2.1805	0.0023	126.97	25.78
PassesACOZ_a	-11.3986	2.8639	0.0000	127.27	37.38
ChallengesWOZ_a	11.4467	3.1305	0.0002	22.72	8.84
Faceoffs3P_a	5.9911	1.6679	0.0003	9.83	2.76
EBGTcountatws_a	-2.8119	0.9465	0.0029	6.35	2.71
EBGTcountatws_h	6.0370	1.6031	0.0001	7.04	3.34
FaceoffsOZ_h	2.1918	0.7262	0.0025	12.69	4.44
Hits_h	-6.4928	1.8646	0.0004	14.41	5.89
Entries1P_a	-4.4425	1.2293	0.0003	18.69	4.38
GiveawaysDZ1P_h	2.8999	0.8457	0.0006	8.64	3.30
Passesdumpout_h	11.1974	3.2786	0.0006	21.14	10.01
ChallengesWStickh_h	3.3000	1.2264	0.0071	12.77	5.62
possessionEVOZavertime_h	3.1366	1.0356	0.0024	8.65	1.76
PassesAC60_a	7.7155	2.0803	0.0002	311.05	43.71
FaceoffsNZ_a	-5.7601	1.6970	0.0006	8.46	2.76
FaceoffsOZ_a	3.1428	0.9742	0.0012	11.12	4.11
ShotsOG60_a	-3.3573	0.9759	0.0005	28.12	7.64
possessionDZ_a	-2.6718	0.8401	0.0014	610.21	70.07
Hits_a	3.1035	0.9356	0.0009	14.07	6.03
PassesACdumpout_h	-9.2939	2.9621	0.0017	16.32	7.78
Giveaways1P_h	2.9107	0.9936	0.0033	21.69	5.11
PassesAC3P_a	9.2352	2.6627	0.0005	101.79	23.05
Shots2P_h	-3.4214	1.1900	0.0040	19.91	5.68
Passesdumpin_a	-2.5813	0.9311	0.0055	30.43	8.07
ChallengesW60_a	-6.9300	2.0526	0.0007	63.79	19.34
Giveaways3P_a	2.1864	0.7651	0.0042	19.94	4.91
ChallengesWownslot_a	1.9025	0.7365	0.0097	3.03	2.16
FaceoffsDZ_h	-2.3119	0.8587	0.0070	9.61	3.66

Table A.2: Parameter estimates, standard errors and p-values for the proportional odds version of the cumulative logit model using standardized covariates from 104 games, with the mean used for centering and the standard deviation for scaling. The thresholds are $\alpha_1 = -14.94$, $\alpha_2 = -0.81$ and $\alpha_3 = 11.61$

Variable	Coefficient	Std.Error	p-value	Mean	Std. Dev.
home	-0.1094	0.3166	0.7296		
ChallengesWNZ	0.5249	0.1557	0.0007	9.34	4.28
ChallengesStickh	0.4921	0.2215	0.0263	21.53	7.85
Pen	-0.4245	0.1624	0.0089	4.29	1.79
PassesACdumpout	0.6253	0.1772	0.0004	17.10	7.59
PassesAC3P	-0.6268	0.1713	0.0002	105.28	23.49
PassesOTB	0.7346	0.1715	0.0000	107.49	19.84
PP	0.6961	0.2579	0.0069	3.53	1.48
FaceoffsPP	-1.1275	0.3068	0.0002	4.65	2.74
ShotsOGSH	0.3566	0.1675	0.0332	0.61	0.99
Hits	-0.5376	0.1513	0.0003	14.24	5.95
Giveaways3P	-0.7788	0.1973	0.0000	20.21	5.27
GiveawaysDZ3P	0.8543	0.2163	0.0000	8.34	3.54
ShotsOG60	0.5181	0.1710	0.0024	29.61	7.47
Possession1P	0.5637	0.1653	0.0006	533.71	83.00
ChallengesWoppslot	-0.3530	0.1512	0.0195	1.17	1.12
Defensiveactions	-0.3921	0.1773	0.0270	519.66	45.75
PPshotspermin	0.3440	0.1662	0.0385	1.70	0.61
PassesACPP	1.0220	0.4423	0.0208	78.29	38.78
PPmininOZ	-0.7669	0.4111	0.0621	225.59	114.25
homeyes:ChallengesStick	-0.6901	0.2848	0.0154		
homeyes:FaceoffsPP	0.8059	0.3120	0.0097		

Table A.3: Parameter estimates, standard errors and p-values for the proportional odds version of the cumulative logit model using the teams performances separately, with the mean used for centering and the standard deviation for scaling. The thresholds are $\alpha_1 = -2.503$, $\alpha_2 = 0.007$ and $\alpha_3 = 2.507$