Masterarbeit

Application of Regression Analysis for Throughput Prediction in the Order Picking Process

eingereicht an der

Montanuniversität Leoben

erstellt am

Lehrstuhl für Informationstechnologie

Vorgelegt von:

Julia Lahovnik
m01235133

Betreuer/Gutachter:

Univ.-Prof. Dipl.-Ing. Dr.techn. Peter Auer
Dipl.-Ing. Stephan Spat

Leoben, am 17. September 2018

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich diese Arbeit selbständig verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt und mich auch sonst keiner unerlaubten Hilfsmittel bedient habe.

Affidavit

I declare in lieu of oath, that I wrote this thesis and performed the associated research myself, using only literature cited in this volume.

Leoben,am _____ _____

Datum Unterschrift

Abstract

The throughput is an important indicator for the performance of logistics systems, and a good estimation of the throughput allows a more precise planning of resources. In this Master's Thesis, regression analysis is applied to predict the throughput for the order picking process in a warehouse for beverages. At first, an introduction to mathematical modelling is given and different types of regression analysis as well as methods to evaluate and compare different models are presented. Then the initial situation and the available data of the application are presented and regression analysis is used to analyse the influence of the various input parameters on the performance of the picking process. Based on the result of this analysis, individual linear models for the operators are created which are used to predict their future performance. Finally, a procedure is described to estimate the throughput of the picking process by using these linear models. The evaluation of this procedure shows that it is possible to achieve a small reduction of personnel costs and to earlier inform the employees about changes in their working time.

Contents

1 Introduction

1.1 About warehouse optimisation

In today's highly competitive world, a good logistics performance is crucial for a company's success. Customers expect high quality, low costs and promptness of logistics services. Warehouses are a core element in the material flow between producers and customers. Thus, their effective and efficient operation has a high impact on quality, costs, time and flexibility which are the most important objectives in any logistics system.
Some of the characteristic challenges that need to be mastered to ensure an efficient operation of a warehouse are:

- Incoming deliveries are often not regular and cannot be planned in advance.
- The high variety in the assortment of goods requires a lot of different transport, handling and storage facilities.
- The throughput of some articles is subject to high fluctuations.
- Customers order small quantities that have to be assembled and summarised quickly.
- A lot of orders have to be finished and their sequence has to be optimised to ideally use the existing capacities.
- The system parameters are constantly changing.

Because of the high complexity of modern storage and distribution systems, computer-aided management systems are needed to handle warehouse operations. Those systems provide functionalities helping to monitor, control and optimise warehouse operation processes.
An ideal operation of a storage and distribution system is achieved if all customer orders can be fulfilled completely and on time. This has to be done under changing conditions, using as little time and resources as possible [5].
Because of the challenges listed above, it is not always easy to run the system in a way that its resources are used ideally. Mathematical prediction methods can help to better understand a system and therefore to make it more predictable. This enables the warehouse managers to set the right steps under changing conditions and as a consequence, to prevent over- or under-capacity.

1.2 About this Master's Thesis

In this Master's Thesis mathematical methods, namely regression analysis, will be applied to predict the throughput for the order picking process in a warehouse for beverages. In a warehouse for food-products the throughput of a process step can vary a lot due to different influencing factors. The main goal of this Master's Thesis is to identify these influencing factors for the order picking process step and to create a model that predicts the throughput for a certain configuration of these influencing factors, i.e. the throughput for a certain time under certain conditions.
This model shall afterwards be implemented in the software solution redPILOT which is a web application that helps to optimise logistics operations. It is based on warehouse process modelling and is designed to help managers and planners to fulfil their tasks more efficiently so that they can keep the system running smoothly.
Currently the throughput of a process step, which is the basis for the allocation of resources, is set statically based on experience. It is not updated if external conditions are changing. The aim is to adapt this throughput based on the configuration of some influencing factors.
To analyse the relationship between the throughput and different influencing factors, different types of data will be used. Firstly, data about performance, time, operators and weather conditions which is available in the redPILOT database is analysed. Additionally, data from the customer's warehouse management system is evaluated. This data contains order information that is not yet available in the redPILOt database, e.g., the planned time of an order which is a measure for the complexity of an order. For the analysis this data is provided in Excel spreadsheets but later it shall be directly transferred from the WMS to the redPILOT database.
In this Master's Thesis, at first a general introduction about mathematical models is given and their different types as well as methods for their evaluation are described. Afterwards, regression analysis, which will later be used to analyse the data and to predict the throughput, is treated in more detail. Different types of regression models are presented and methods for the estimation of the model parameters as well as for the evaluation of the model are described.
Finally, the methods described in the previous chapters are applied in the use case described above. At first plots are made to get a first overview of the data and simple regression models are created to identify the most relevant influencing factors. Different linear models to predict the actual time for certain orders using the identified influencing factors as input parameters are created, compared and the most appropriate one among them is chosen. This model is then used to predict the throughput for the order picking process step. Finally, an evaluation of the prediction of the throughput based on this model is done.

2 Mathematical Modelling

2.1 What is Mathematical Modelling?

According to [4], generally speaking, the aim of creating a model is to represent a complex real system or process. A model describes the most important aspects of a certain object in a certain context. The purpose of creating a model can either be to only describe the system or to also predict its future behaviour which is usually the case for mathematical models.
Mathematical methods play an important role as a universal language to formulate models and as a tool for the evaluation of different models. Thus, mathematical modelling can be seen as a process consisting of understanding a problem and creating a model, as well as calculating, interpreting and validating the results. This process is repeated until the defined goals are achieved.
The first step, the analysis of the problem, is necessary to precisely define the question that shall be answered. The whole problem can be divided into smaller sub-problems that are easier to solve. Furthermore, simplifying assumptions can be made, e.g., parameters that have only a very small influence on the output of the system can be neglected.
To find an appropriate mathematical model, first, the relevant system and model parameters have to be defined. The system parameters describe properties of the system and are predetermined whereas model parameters describe the properties of the model and have to be defined. To give information about the state of a system, state parameters are needed. All the state parameters together define the current state of a system. Some of the state parameters can also be unknown. Next, it has to be analysed which relationships exist between the system parameters, the state variables and the unknown parameters.
The created model then should be analysed. It should be made sure that the problem is properly represented. Very often it cannot be proved that the output of the model is correct but its plausibility can be checked using numerical calculations. For example, special cases where the solution of the problem is known can be analysed. Furthermore the sensitivity of the model with respect to noise in the input data should be evaluated.
After the calculation and simulation of the results of the model, it has to be validated and interpreted. At first it has to be checked whether the formulated mathematical problem has been solved. Then the mathematical results can be transferred to the real world application and it can be analysed if they correspond to the observations. After this process it should be clear if the model is appropriate or if it has to be adapted before it can be used. In the second

case it is necessary to repeat some or all of the steps described above.

2.2 Types of Mathematical Models

Mathematical models can be classified according to different criteria. Regarding mathematical structures, it can be distinguished between dynamic models where at least some state parameters are time-dependent and static models that describe the time-independent structure of systems or optimisation problems. Another differentiation that can be made is between discrete and continuous models as well as between deterministic models where the same input parameters always result in the same output and stochastic models where the results cannot be predetermined. In real world applications it often occurs that some state parameters are discrete and some continuous and that parts of the model are static and other parts dynamic. So, the differentiation between those types is often not very clear.
Another way to classify models is the approach of solving the problem. Such approaches can for example be linear or non-linear equation systems, differential equations, finite state machines, etc. A specific type are optimisation problems where an objective function is maximised or minimised.
Furthermore, models can be grouped by the characteristics of the phenomenon they describe, e.g., growth processes, transport processes or vibrations and waves.
There are different goals for a modelling process which is another possibility to distinguish between different types of models. One possible goal is to facilitate decision making by providing a prediction for the consequences of different decisions. Generally, creating predictions of the future behaviour of a system is a common goal of a modelling process. Another goal can be the optimisation of a system according to certain criteria. Other possible goals are, e.g., simulation, pattern recognition or verification of a hypothesis.
Models can also be divided into white-box, grey-box and black-box models. White-box models are derived from known principles and are fully specified. In contrast, grey-box models are based on plausible assumptions of the relations between the parameters of the observed system. Black-box models do not describe the inner structure of the system but only model relationships between its input and output [4].
The goal of the model that shall be found in this Master's Thesis is to predict the output of a system. Some of the input parameters are this discrete, others continuous. The output variable is continuous. As the output cannot be definitely predicted from the values of the input parameters, we need a stochastic model. The main purpose is to predict the output given the values of a set of input parameters but it would be desirable to also find relations between the parameters.

2.3 Accuracy of Mathematical Models

To select the best of several given models, measures to determine the quality of a model are necessary. For the different model types there are different measures to assess the accuracy of the model but there are also some basic concepts to measure the quality of a fit that can be used for different model types.
For regression models (i.e. the output variable is continuous) the most used quality measure is the mean squared error (MSE). If y_i is the true response value for the ith observation and $\hat{f}(\mathbf{x}_i)$ is the prediction for this observation, then the MSE is given by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(\mathbf{x}_i))^2. \tag{2.1}$$

It quantifies how much the predicted response values differ from the real response values. If the MSE is small, the prediction is good. Usually the MSE is computed using the same data that has been used to create the model (training data). But in fact, we are more interested on how well our model performs for new data. To get this information, we have to apply our model on some test data where the values of the response variables are known but have not been used to train the model. This test MSE can differ significantly from the training MSE. For example, if the model is very closely adjusted to the training data, it is very often not general enough to correctly predict the response values of new data. So, the training MSE is very low because the model perfectly fits the training data but the test MSE is high because the test data slightly differs from the training data so that the model does not fit anymore. This phenomenon is called overfitting. If test data is available, we can easily avoid overfitting by computing the test MSE instead of the training MSE [6]. Methods that can be used if not enough data is available are presented in Chapter 2.4.

The MSE on the test data $(\tilde{x}_1, \tilde{y}_1), ..., (\tilde{x}_n, \tilde{y}_n)$, which is given by

$$\frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i - \hat{f}(\tilde{x}_i))^2, \tag{2.2}$$

is also called generalisation error. If the complexity of the model increases, the generalisation error decreases first as the training error, but starts increasing at a certain point while the training error continues to decrease.
If we have the training data vector D $(x_1, y_1; x_2, y_2; ...; x_n, y_n)$ and the function f_D that minimises the error on the training data, the expected net error on the test data $E_D((\tilde{y}_i -$

$f_D(\tilde{x}_i))^2)$ can be written as a sum of three terms:

$$\begin{aligned} E_D((\tilde{y}_i - f_D(\tilde{x}_i))^2) =& E_{\tilde{y}_i}((\tilde{y}_i - E_{\tilde{y}_i}(\tilde{y}_i|\tilde{x}_i))^2) + \\ & E_D((E_{\tilde{y}_i}(\tilde{y}_i|\tilde{x}_i) - E_D(f_D(\tilde{x}_i)))^2) + \\ & E_D((E_D(f_D(\tilde{x}_i)) - f_D(\tilde{x}_i))^2) \end{aligned} \tag{2.3}$$

where $E_D(f_D(\tilde{x}_i)))$ is the compact notation for $E_D(f_D(\tilde{x}_i))|\tilde{x}_i)$, i.e. we are averaging over all possible training vectors D keeping $\tilde{x}_i$ fixed. The first term of this sum is the unavoidable error which occurs due to noise in the data. The second term is called bias. As this part of the error is mainly caused by the restrictive choice of F, the bias decreases for larger, more complex F. The third term, the variance, measures the extent to which $f_D(\bar{x}_i)$ as a function of D varies from $E_D(f_D(\bar{x}_i))$, its average over D. This term goes to zero for $n \to \infty$ for most traditional statistical schemes. If we work with finite, fixed n, the error usually becomes larger for more complex F. [2].
Thus, as the bias decreases for more complex F and the variance increases for more complex F, low bias and low variance are contradictory goals. For creating a good model we have to find a good balance between bias and variance which is called bias-variance trade-off.

2.4 Model Evaluation

After creating a model, we want to know how well it will predict the response values for new data. As already mentioned above, a low error on the training data does not guarantee that the model is a good predictor for new data. If enough data is available, we can simply split the data into a training dataset and a test dataset and compare the predicted values and the true response values for the test dataset. But often an appropriate test dataset is not available, so, we have to find other ways to evaluate the quality of the model. In this chapter, methods to assess the quality of the predictor and to choose the best among several models are presented.

2.4.1 Cross-validation

Cross-validation is a technique where the dataset is divided into several subsets and one of the subsets is used as a test dataset and the other subsets are used as training data. In k-fold cross-validation the dataset D is randomly split into k mutually exclusive subsets D_t (called folds). The predictor is created, tested k times and each time a different fold is used for the testing. Each time $t \in 1, 2, ..., k$ the training dataset is $D \backslash D_t$ and D_t is used for testing. In complete cross-validation the folds are created using all $\binom{m}{m/k}$ possibilities for choosing m/k instances out of m. A special case of complete cross-validation is leave-one-out where the quantity of folds is equal to the quantity of instances in the whole dataset. Usually, complete cross-validation is too expensive. So, only a single split of the data into the folds is used.

In stratified cross-validation the folds contain approximately the same proportions of specific response values as the original dataset. This type of cross-validation is used for classification to ensure that each of the folds contains approximately the same amount of examples for the different classes [7].

2.4.2 Akaike's Information Criterion and Bayesian Information Criterion

The Akaike's Information Criterion (AIC) combines the Log-likelihood of a given parameter vector $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)'$ and a penalty for complex models. It is given by

$$AIC = -2l(\hat{\theta}) + 2p \tag{2.4}$$

where $l(\hat{\theta})$ is the maximum value of the Log-likelihood function and $2p$ a penalty term for the number of parameters to avoid overfitting. When several models are compared, the model with the smallest value of the AIC is chosen. A variation of the AIC that includes a different penalty term is the corrected AIC. It is given by

$$AIC_{corr} = -2l(\hat{\theta}) + \frac{2n(p+1)}{n-p-2} \tag{2.5}$$

where n is the sample size.
The Bayesian Information Criterion (BIC) has a very similar form as the AIC. It is defined as

$$BIC = -2l(\hat{\theta}) + \log(n)p. \tag{2.6}$$

Thus, only the penalty term is different. In the BIC more complex models receive a much higher penalty and are therefore less likely to be selected [3].

3 Regression Analysis

Regression is one of the most common techniques to analyse empirical problems in economy, social and life sciences. There are several different types of regression models. In this paper, different types of linear regression will be presented.
We look for the influence of one ore several so-called independent variables on the so-called dependent variable. Usually, the model does not exactly represent the data and the relationship between the input and output parameters can only be approximated [3].

3.1 Linear Regression Models

Linear regression is a simple but very useful approach for predicting a quantitative response. According to [6], it can be used to answer questions such as:

- Is there a relationship between certain parameters?
- How strong is this relationship?
- Which of the input parameters have an effect on the dependent variable?
- How accurately can the effect of each input variable on the dependent variable be estimated?
- How accurate are the predictions for the dependent variable?
- Is the relationship linear?
- Are there interaction effects between the input variables?

3.1.1 The Classical Linear Model

As it is given in [3], we are interested in the influence of some independent variables $x_1, ..., x_k$ on a dependent variable y. The relationship between $x_1, ..., x_k$ and y is modelled using a function $f(x_1, ..., x_k)$ and is overlaid by some random noise ε. Using additive noise we get

$$y = f(x_1, ..., x_n) + \varepsilon. \tag{3.1}$$

Our goal is to estimate the unknown function f.
In linear models the following assumptions are made:

1. f is a linear combination of the co-variables, i.e.

$$f(x_1, ..., x_k) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k. \tag{3.2}$$

   The parameters $\beta_0, \beta_1, ...\beta_k$ are unknown and have to be estimated. Using the vectors $\mathbf{x} = (1, x_1, ..., x_k)'$ and $\boldsymbol{\beta} = (\beta_0, ..., \beta_k)'$ we can write

$$f(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta} \tag{3.3}$$

   We can also model non-linear relationships by transforming the input variable. We can define a new variable $z_j = g(x_j)$, e.g. $z_j = \log(x_j)$, and use this new variable instead of x_j in equation 3.2.

2. Additivity of the noise: This assumption seems to be restrictive but is at least approximately fulfilled in applications. An alternative to additive noise is multiplicative noise which is for example used in exponential models. Most models with multiplicative error terms can easily be written as models with additive noise using variable transformation (e.g., by taking the logarithm).

To estimate the unknown parameters $\boldsymbol{\beta}$ the observations y_i and $\mathbf{x_i} = (1, x_{i1}, ..., x_{ik})', i = 1, ..., n$ are used. For each observation we get the equation

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \varepsilon_i = \mathbf{x_i}'\boldsymbol{\beta} + \varepsilon_i. \tag{3.4}$$

Using the vectors $\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ and $\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$ and the design matrix

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}$$

we can write the n equations from 3.4 as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{3.5}$$

We assume that $\mathbf{X}$ is full rank, i.e. its columns are linearly independent. This condition is necessary to get a unique estimate for $\boldsymbol{\beta}$.
For the vector $\boldsymbol{\varepsilon}$ the following assumptions are made:

1. The average noise is zero, i.e. $E(\boldsymbol{\varepsilon}) = \mathbf{0}$.

2. The variance of the noise is constant for all observations, i.e. $\text{Var}(\varepsilon_i) = \sigma^2$. If this

condition is fulfilled, the errors are homoscedastic. If the residuals are heteroscedastic, this can be a sign that the model is not appropriate. Furthermore, we assume that the errors are uncorrelated, i.e. $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$. So, we get the covariance matrix $\text{Cov}(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\mathbf{I}$. This assumption is often violated for time series and longitudinal data because not all explaining variables can be included in the model since they are either not observable or not recorded.

3. The co-variables are stochastic so that all assumptions can be seen as conditioned by the design matrix. The assumptions on the errors must hold conditioned by $\mathbf{X}$: $E(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}$.

4. The noise is (approximatively) normally distributed, i.e. $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. If we assume normally distributed noise, we talk about a classical linear normal regression.

The co-variables only influence the expectation value of $\mathbf{y}$. The variance σ^2, respectively the covariance matrix $\sigma^2\mathbf{I}$, is independent from the co-variables.
Based on the estimation $\hat{\boldsymbol{\beta}}$ an estimate for the expected value $E(y_i)$ for y_i is given by

$$\widehat{E(y_i)} = \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_k x_{ik} = \mathbf{x}_i'\hat{\boldsymbol{\beta}}. \tag{3.6}$$

The difference between the true value y_i and the estimated value $\hat{y}_i$, the residual $\hat{\varepsilon}_i$, is given by

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}. \tag{3.7}$$

If we summarise the residuals to a vector, we get

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}. \tag{3.8}$$

The partial residual for the co-variable x_j is defined as

$$\hat{x}_{j,i} = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}} + \hat{\beta}_j x_{ij} = \hat{\varepsilon}_i + \hat{\beta}_j x_{ij}. \tag{3.9}$$

In this residual the influence of all co-variables except for x_j is removed.
As already mentioned above, we can also model non-linear relationships using linear models. One method to do so is variable transformation. In this case we use the regression model

$$y_i = \beta_0 + \beta_1 g_1(x_i) + ... + \beta_k g_k(x_i) + \varepsilon_i \tag{3.10}$$

where the g_j can be any function. It has to be defined before the estimation of the parameters. This gives us the linear model

$$y_i = \beta_0 + \beta_1 z_{i1} + ... + \beta_k z_{ik} + \varepsilon_i \tag{3.11}$$

where $z_{ij} = g_j(x_i) - \bar{g}$ with

$$\bar{g} = \frac{1}{k} \sum_{j=1}^{k} g_j(x_i). \tag{3.12}$$

The subtraction of $\bar{f}$ centres $\hat{\beta}_j z_{ij}$ around zero.
Another method is the use of a polynomial model:

$$y_i = \beta_0 + \beta_1 z_i + \beta_2 z_i^2 + ... + \beta_l z^l + ... + \varepsilon_i \tag{3.13}$$

By defining the new variables $x_{i1} = z_i, x_{i2} = z_i^2, ..., x_{il} = z_i^l$ we get the regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_l x_{il} + ... + \varepsilon_i. \tag{3.14}$$

To include categorical co-variables in the model we can use the so-called dummy coding. To model the effect of a categorical variable with c levels $c-1$ dummy variables which are defined by

$$x_{i1} = \begin{cases} 1 \ x_i = 1 \\ 0 \text{ otherwise} \end{cases} \quad ... \quad x_{i,c-1} = \begin{cases} 1 \ x_i = c-1 \\ 0 \text{ otherwise} \end{cases} \tag{3.15}$$

for $i = 1, ..., n$ are included in the regression model:

$$y_i = \beta_0 + \beta_{i1} x_{i1} + ... + \beta_{i,c-1} x_{i,c-1} + ... + \varepsilon_i \tag{3.16}$$

One category of x is used as a reference category and is not represented by one of the dummy variables in the model. The estimates for the other categories are compared to this reference category.
If the effect of one co-variable depends on the value of at least one other co-variable, an interaction between co-variables occurs. For the input variables x and z and the dependent variable y we can create the regression model

$$y_i = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz + \varepsilon. \tag{3.17}$$

The terms $\beta_1 x$ and $\beta_2 z$ are called main effects and the term $\beta_3 xz$ is the interaction between x and z. If $\beta_3 = 0$, there is no interaction and the effect of one co-variable does not depend on the value of the other one.

3.1.2 Estimation of the parameters

The most popular method to estimate the regression coefficients β is the least squares method. As described in [3], when using this method, the sum of the squared difference between the

estimate and the actual value

$$LS(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 = \sum_{i=1}^{n}\varepsilon_i^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \tag{3.18}$$

is minimised with respect to $\boldsymbol{\beta}$.
To find the minimum of $LS(\boldsymbol{\beta})$ we first rewrite 3.18:

$$LS(\boldsymbol{\beta}) = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \tag{3.19}$$

Then we differentiate with respect to β:

$$\frac{\partial LS(\beta)}{\partial \beta} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \tag{3.20}$$

Differentiating a second time gives $2\mathbf{X}'\mathbf{X}$. As the design matrix $\mathbf{X}$ is linearly independent, $\mathbf{X}'\mathbf{X}$ is positive definite and we can get the minimum by setting 3.20 to 0. So, the least squares estimate $\hat{\boldsymbol{\beta}}$ is the solution of

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}. \tag{3.21}$$

As $\mathbf{X}'\mathbf{X}$ is positive definite, it is also invertible and we get the LS estimate

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{3.22}$$

Another method to estimate the coefficients is the maximum likelihood method. For normally distributed noise $\boldsymbol{\varepsilon}$ we get $y \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ and the Likelihood is

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)}^{n/2} \exp(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})). \tag{3.23}$$

The Log-likelihood is

$$l(\beta, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{3.24}$$

If we maximize the Log-likelihood with respect to $\boldsymbol{\beta}$, we can neglect the first two summands because they do not depend on $\boldsymbol{\beta}$. Maximising $-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ is the same as minimising $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. Thus, the maximum likelihood estimate is equal to the least squares estimate.
Based on the estimate $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ for $\boldsymbol{\beta}$ we can estimate the expected value for $\mathbf{y}$ as

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}. \tag{3.25}$$

$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is called the prediction matrix or hat-matrix. It has the following properties:

- It is symmetric.
- It is idempotent.
- rank($\mathbf{H}$) = tr($\mathbf{H}$) = p where p is the number of coefficients
- $\frac{1}{n} \leq h_{ii} \leq \frac{1}{r}$ where r is the number of rows in $\mathbf{X}$ with identical x_i.
- The matrix $\mathbf{I} - \mathbf{H}$ is also symmetric and idempotent with rank($\mathbf{I} - \mathbf{H}$) = $n - p$.

The prediction matrix is used in the definition of the standardised and studentised residuals. These residuals are needed because the residuals themselves are neither homoscedastic nor uncorrelated. The residuals are usually used to check if the model assumptions are valid but heteroscedastic residuals do not necessarily mean that the noise is also heteroscedastic. We get the standardised residuals, which are homoscedastic if the model assumptions are correct, by dividing the residuals by the estimated standard deviation:

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} \tag{3.26}$$

The studentised residuals are defined by

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}} = r_i(\frac{n - p - 1}{n - p - r_i^2})^{1/2} \tag{3.27}$$

where $\hat{\sigma}_{(i)}$ is an estimation for $\hat{\sigma}$ that is not based on the ith observation.
The variance σ^2 can be estimated using the maximum likelihood method. We have already defined the Likelihood $L(\beta, \sigma^2)$ and the Log-likelihood $l(\boldsymbol{\beta}, \sigma^2)$ for the linear model. By partially differentiating 3.24 with respect to σ^2 and setting to zero we get

$$\frac{\partial l(\boldsymbol{\beta}, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0. \tag{3.28}$$

If we replace $\boldsymbol{\beta}$ by the estimate $\hat{\boldsymbol{\beta}}$, this gives

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = 0 \tag{3.29}$$

and therefore

$$\hat{\sigma}^2_{ML} = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n}. \tag{3.30}$$

The expected value of the squared sum of residuals is

$$E(\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}) = (n - p) * \sigma^2. \tag{3.31}$$

So,

$$E(\hat{\sigma}^2_{ML}) = \frac{n - p}{n}\sigma^2 \tag{3.32}$$

which means that the ML estimate for σ^2 is not unbiased and therefore rarely used.
An unbiased estimator for σ^2 is

$$E(\hat{\sigma}^2) = \frac{1}{n-p}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}. \tag{3.33}$$

It is called the Restricted Maximum Likelihood (REML) estimator and it can be shown that 3.33 maximises the marginal likelihood

$$L(\sigma^2) = \int L(\beta, \sigma^2) d\beta. \tag{3.34}$$

The least squares estimates have some important geometric properties:

- The estimated values $\hat{\mathbf{y}}$ and the residuals $\hat{\boldsymbol{\varepsilon}}$ are orthogonal, i.e. $\hat{\mathbf{y}}'\hat{\boldsymbol{\varepsilon}} = 0$.
- The columns of $\mathbf{X}$ and the residuals $\hat{\boldsymbol{\varepsilon}}$ are orthogonal as well, i.e. $\mathbf{X}'\hat{\boldsymbol{\varepsilon}} = 0$.
- On average the residuals are zero, i.e.

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = 0 \quad \text{respectively} \quad \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i = 0 \tag{3.35}$$

- The average of the estimated values $\hat{y}_i$ is equal to the average of the observed values y_i.
- The regression hyperplane goes through the centroid of the data, i.e.

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + ... + \hat{\beta}_k \bar{x}_k \tag{3.36}$$

  where $\bar{x}_j$ is the mean of all the values given for the variable x_j.

From these properties we can derive a formula for the variance analysis:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}\hat{\varepsilon}_i^2 \tag{3.37}$$

By dividing by n respectively $(n-1)$ we get

$$s_y^2 = s_{\hat{\mathbf{y}}}^2 + s_{\hat{\boldsymbol{\varepsilon}}}^2 \tag{3.38}$$

which shows that the variance of the observed values $s_{\mathbf{y}}^2$ can be decomposed into the variance of the estimated values $s_{\hat{y}}^2$ and the variance of the residuals $s_{\hat{\varepsilon}}^2$.
This variance analysis formula is used in the definition of the coefficient of determination, also called coefficient of variation or coefficient of correlation, which gives information about the

quality of the fit. It is defined by

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{3.39}$$

If R^2 is close to one, the sum of the squared residuals is small and the model fits the data very well. An R^2 of zero means that that $\sum(\hat{y}_i - \bar{y})^2 = 0$ which means that the estimate for y_i is always equal to the average $\bar{y}$ and therefore independent of the explaining variables.
In general the coefficient of determination becomes larger when more input variables are included. Therefore, the coefficient of determination should only be used if all models have the same dependent variable y and the same number of regression coefficients.

3.1.3 Hypothesis tests and confidence intervals

It is given by [3] that we assume that $\varepsilon_i \sim N(0, \sigma^2)$, which makes the construction of exact tests and confidence intervals much simpler, but the tests and confidence intervals are also robust to small deviations from the normal distribution. For large data sets they also remain valid if the noise is not normally distributed.
The most common statistical hypotheses are:

- Test on the significance of an input variable:

$$H_0 : \beta_j = 0, \quad H_1 : \beta_j \neq 0 \tag{3.40}$$

- Test of a sub-vector $\boldsymbol{\beta}_1 = (\beta_1, ..., \beta_r)'$:

$$H_0 : \boldsymbol{\beta}_1 = 0, \quad H_1 : \boldsymbol{\beta}_1 \neq 0 \tag{3.41}$$

- Test on equality:

$$H_0 : \beta_j - \beta_r = 0, \quad H_1 : \beta_j - \beta_r \neq 0 \tag{3.42}$$

Those three tests are special cases of the test for general linear hypotheses

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d}, \quad H_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{d} \tag{3.43}$$

where $\mathbf{C}$ is a $r \times p$ matrix with rank$(\mathbf{C}) = r \leq p$. That means that for H_0 r linearly independent conditions are valid.
To find an appropriate test for the general problem (3.43), we follow the subsequent procedure:

1. Calculate the sum of squared residuals $SSE = \hat{\varepsilon}'\hat{\varepsilon}$ for the full model
2. Calculate the sum of squared residuals $SSE_{H_0} = \hat{\varepsilon}'_{H_0}\hat{\varepsilon}_{H_0}$ for the model if the null hypothesis is true, i.e. the restriction $\mathbf{C}\boldsymbol{\beta} = \mathbf{d}$ is fulfilled.

3. Evaluate the relative difference between the sum of squared residuals in the restricted model and the full model:

$$\frac{\Delta SSE}{SSE} = \frac{SSE_{H_0} - SSE}{SSE} \tag{3.44}$$

As the fit for the restricted model can be at the most as good as the full model, the difference $SSE_{H_0} - SSE$ is always greater than or equal to zero. The smaller the difference is, the more similar are the results for the two models and the higher is the probability that the null hypothesis is not rejected. The actual test statistic is given by

$$F = \frac{\frac{1}{r}\Delta SSE}{\frac{1}{n-p}SSE} = \frac{n-p}{r}\frac{\Delta SSE}{SSE} \tag{3.45}$$

where r is the number of restrictions, i.e. the number of rows in C. For a defined level of significance α the null hypothesis is rejected if the test statistic is bigger than the (1-α) quantile of the corresponding F-distribution:

$$F > F_{r,n-p}(1-\alpha). \tag{3.46}$$

For the specific test problems described above the test statistics are listed below:

- Test on the significance of an input variable (t-test):

$$H_0 : \beta_j = 0, \quad H_1 : \beta_j \neq 0; \quad j = 1, ..., p \tag{3.47}$$

  In this case F is given by

$$F = \frac{\hat{\beta}_j^{\,2}}{\widehat{\mathsf{Var}}(\hat{\beta}_j)} \sim F_{1,n-p}. \tag{3.48}$$

  We can also use the t-statistic

$$t = \frac{\hat{\beta}_j}{se_j} \tag{3.49}$$

  where $se_j = \widehat{\mathsf{Var}}(\hat{\beta}_j)^{1/2}$ is the estimated standard deviation of $\hat{\beta}_j$. The null hypothesis is rejected if

$$|t| > t_{1-\alpha/2}(n-p). \tag{3.50}$$

  For the more general hypothesis

$$H_0 : \beta_j = d_j, \quad H_1 : \beta_j \neq d_j; \quad j = 1, ..., k \tag{3.51}$$

  the modified test statistic

$$t = \frac{\beta_j - d_j}{se_j} \tag{3.52}$$

  can be used.

- For the test of a sub-vector $\boldsymbol{\beta}_1 = (\beta_1, ..., \beta_r)'$

$$H_0 : \boldsymbol{\beta}_1 = 0, \quad H_1 : \boldsymbol{\beta}_1 \neq 0, \tag{3.53}$$

the test-statistic is given by

$$F = \frac{1}{r}\hat{\boldsymbol{\beta}}_1{}'\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}_1)^{-1} \sim F_{r,n-p}. \tag{3.54}$$

The estimated co-variance matrix of the sub-vector $\hat{\boldsymbol{\beta}}_1$ consists of the corresponding elements of the co-variance matrix $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$.

- Test of the hypothesis "No relationship":

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0, \tag{3.55}$$

i.e. no co-variable has an impact on the output. If H_0 is valid, the LS-estimation only consists of the estimation $\beta_0 = \bar{y}$. SSE_{H_0} is given by

$$SSE_{H_0} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{3.56}$$

and the test statistic F is

$$F = \frac{n-p}{k}\frac{R^2}{1-R^2}. \tag{3.57}$$

Based on these tests we can construct confidence intervals for one single parameter β_j, $j = 0, ..., k$ respectively confidence ellipsoids for a sub-vector $\boldsymbol{\beta}_1$ of $\boldsymbol{\beta}$. To find a confidence interval for β_j presuming normal distribution, we use the test statistic $t = (\hat{\beta}_j - d_j)/se_j$ for the test of the hypothesis $H_0 : \beta_j = d_j$. The null hypothesis is rejected if $|t| > t_{n-p}(1-\alpha/2)$. The probability to reject H_0 if it is actually true is α. Thus, for H_0

$$P(|t| > t_{n-p}(1-\alpha/2)) = \alpha. \tag{3.58}$$

The probability that H_0 is not rejected is given by

$$P(|t| < t_{n-p}(1-\alpha/2)) = 1-\alpha. \tag{3.59}$$

This is equal to

$$P(\hat{\beta}_j - t_{n-p}(1-\alpha/2) * se_j < \beta_j < \hat{\beta}_j + t_{n-p}(1-\alpha/2) * se_j) = 1-\alpha \tag{3.60}$$

and we get the interval

$$[\hat{\beta}_j - t_{n-p}(1-\alpha/2) * se_j, \hat{\beta}_j + t_{n-p}(1-\alpha/2) * se_j]. \tag{3.61}$$

The confidence interval for a sub-vector $\boldsymbol{\beta}_1$ can be constructed the same way and is given by

$$\left\{\boldsymbol{\beta}_1 : \frac{1}{r}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1)'\widehat{\mathsf{Cov}}(\hat{\boldsymbol{\beta}}_1)^{-1}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) \leq F_{r,n-p}(1-\alpha)\right\}. \tag{3.62}$$

A confidence interval for the expected value $\mu_0 = E(y_0)$ of a future observation y_0 at the point x_0 is given by

$$\mathbf{x}_0'\hat{\boldsymbol{\beta}} \pm t_{n-p}(1-\alpha/2)\hat{\sigma}(\mathbf{x}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)^{1/2}. \tag{3.63}$$

If we are interested in finding an interval which has a high likelihood to contain the future observation y_0, we use a prediction interval which is in general much larger than the confidence interval. For a future observation y_0 at x_0 and a confidence level of $1-\alpha$, it is given by

$$\mathbf{x}_0'\hat{\boldsymbol{\beta}} \pm t_{n-p}(1-\alpha/2)\hat{\sigma}(1+\mathbf{x}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)^{1/2}. \tag{3.64}$$

3.1.4 The general linear model

The classical linear model that has been treated so far is a special case of the general linear model. We replace

$$\mathsf{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I} \tag{3.65}$$

by

$$\mathsf{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{W} \tag{3.66}$$

where $\mathbf{W}$ is a positive definite matrix. If the noise is heteroscedastic but uncorrelated, we get

$$\mathbf{W} = \mathsf{diag}(w_1, ..., w_n). \tag{3.67}$$

The variance of the noise is $\mathsf{Var}(\varepsilon_i) = \sigma_i^2 = \sigma^2 w_i$. If the LS-estimate $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is used for the general linear model, we get

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} \quad \mathsf{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \tag{3.68}$$

This means that the LS-estimate is still unbiased but the covariance matrix does not correspond to the covariance matrix for the classical linear model which implies that the variances for the regression coefficients and therefore also the tests and confidence intervals are incorrect.
There are different methods to get better estimates for the general linear model. One of them is the weighted least squares method where the dependent variables, the design matrix and the noise are transformed so that they conform to a classical linear model. For the use of this method the matrix $\mathbf{W}$ has to be known in advance.
For uncorrelated heteroscedastic errors, i.e. $\mathsf{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{W} = \sigma^2\mathsf{diag}(w_1, ..., w_n)$ we can multiply the noise ε_i by $1/\sqrt{w_i}$ and get the noise $\varepsilon_i^* = \varepsilon_i/\sqrt{w_i}$ which has the constant variances $\mathsf{Var}(\varepsilon_i^*) = \sigma^2$. The dependent variable and the co-variables have to be changed

accordingly: $y_i^* = y_i/\sqrt{w_i}, x_{i0}^* = 1/\sqrt{w_i}, x_{i1}^* = x_{i1}/\sqrt{w_i}, ..., x_{ik}^* = x_{ik}/\sqrt{w_i}$. This gives us the classical linear model

$$y_i^* = \beta_0 x_{i0}^* + \beta_1 x_{i1}^* + ... + \beta_k x_{ik}^* + \varepsilon_i^* \tag{3.69}$$

with homoscedastic errors ε_i^*. This corresponds to a left multiplication of the model equation $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ by the matrix $\mathbf{W}^{1/2} = \text{diag}(1/\sqrt{w_1}, ..., 1/\sqrt{w_n})$. The estimate for β is given by

$$\hat{\beta} = (\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{-1}\mathbf{y}. \tag{3.70}$$

It can be shown that this estimate maximises the weighted sum of least squares

$$\sum_{i=1}^{n} \frac{1}{w_i}(y_i - \mathbf{x_i}'\boldsymbol{\beta})^2. \tag{3.71}$$

Observations having a higher variance receive a lower weight than those with low variance. The expected value $E(\hat{\boldsymbol{\beta}})$ is $\boldsymbol{\beta}$ and therefore the LS estimate is unbiased. The REML-estimate for σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n-p}\hat{\boldsymbol{\varepsilon}}'\mathbf{W}^{-1}\hat{\boldsymbol{\varepsilon}} \tag{3.72}$$

and is also unbiased.
This method can also be used for an arbitrary covariance matrix $\sigma^2\mathbf{W}$. The matrix $\mathbf{W}^{1/2}$ is not unique but can be computed using the spectral decomposition

$$\mathbf{W} = \mathbf{P}\text{diag}(\lambda_1, ..., \lambda_n)\mathbf{P}'. \tag{3.73}$$

An application of the weighted least squares method is for grouped data. If several vectors of co-variables x_i are the same, we summarise them to one vector and note the number of repetitions n_i of the observation x_i and the average $\bar{y}_i$ of the values of the dependent variable. The covariance for the noise is then given by $\text{Cov}(\varepsilon) = \sigma^2\text{diag}(1/n_1, ..., 1/n_G)$ where G is the number of groups [3].

3.1.5 Selection of the model and the input parameters

One of the most important questions to be answered when a regression model is created is which of the input parameters should be included. It should be avoided that irrelevant variables are included in the model because this makes the model unnecessarily complex and increases its variance. As described in Chapter 2.3, a good trade-off between a low variance and a low bias has to be found to receive a model that is sufficiently complex to represent the data but simple enough to have a low variance and to give a good generalisation.
In Chapter 3.1.2, the coefficient of determination was described as a measure for the quality of a fit. It is given by [3] that it always increases when more co-variables are included and therefore it is not an appropriate tool for the comparison of different models. The adjusted

coefficient of determination includes a correction for the number of parameters so that it does not automatically increase when a new co-variable is included. It is given by

$$\bar{R}^2 = 1 - \frac{n-1}{n-p}(1-R^2) \tag{3.74}$$

where p is the number parameters in the model (including the intercept) and n the number of observations.
Another measure for the selection of a model is Mallows's C_p. It is defined as

$$C_p = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_{iM})^2}{\hat{\sigma}^2} - n + 2|M| \tag{3.75}$$

where $\hat{y}_{iM}$ is the estimation for y_i when certain parameters are used, $\hat{\sigma}^2$ is the estimated variance for the model including all variables and M is the number of included variables (i.e $M = p - 1$).
The methods described in Chapter 2.4 can also be used for the selection of a regression model. E.g. we can use the Akaike's Information criterion which is given by

$$AIC = -2 * l(\hat{\beta}_M, \hat{\sigma}^2_M) + 2|M + 1| \tag{3.76}$$

where $l(\hat{\beta}_M, \hat{\sigma}^2_M)$ is the maximum value of the Log-likelihood for a model including M variables, i.e. the Log-likelihood if the ML estimates $\hat{\beta}_M$ and $\hat{\sigma}^2_M$ are used. According to [3], for a linear model with normally distributed noise the AIC we get

$$\begin{aligned} -2 * l(\hat{\beta}_M, \hat{\sigma}^2_M) &= n\log(\hat{\sigma}^2_M) + \frac{1}{\hat{\sigma}^2_M}(\mathbf{y} - \mathbf{X}_M\hat{\boldsymbol{\beta}}_M)'(\mathbf{y} - \mathbf{X}_M\hat{\boldsymbol{\beta}}_M) \\ &= n\log(\hat{\sigma}^2_M) + \frac{1}{\hat{\sigma}^2_M} * \hat{\boldsymbol{\varepsilon}}'_M\hat{\boldsymbol{\varepsilon}}_M \\ &= n\log(\hat{\sigma}^2_M) + \frac{n\hat{\sigma}^2_M}{\hat{\sigma}^2_M} \\ &= n\log(\hat{\sigma}^2_M) + n. \end{aligned} \tag{3.77}$$

We receive

$$AIC = n * \log(\hat{\sigma}^2_M) + n + 2|M + 1|. \tag{3.78}$$

For $\hat{\sigma}^2$ the ML-estimate $\hat{\sigma}^2 = 1/n\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ is used where $\hat{\boldsymbol{\varepsilon}}$ is the expected value of the residuals. The BIC can be calculated using the following formula:

$$BIC = n * \log(\hat{\sigma}^2) + \log(n)|M|. \tag{3.79}$$

3.1.6 Model analysis

According to [3], after the first estimate the model has to be evaluated concerning its ability to represent the data. The first step of the evaluation is to check if the model assumptions are correct. In general, for a linear regression model the following assumptions are made:

- Homoscedasticity: To check if the assumption of homoscedastic variances is correct, residual plots or test for heteroscedasticity such as the Breusch-Pagan test can be used.
- Uncorrelated noise: Correlated noise can be detected by plotting the noise over the time and by using statistical tests such as the Durbin-Watson test.
- Linearity: Non-linear relationships can be detected by plotting the standardised respectively studentised residuals over the estimated values. Furthermore the partial residuals can be used.
- Normal distribution: To check for normal distribution Q-Q-plots can be used. In this plots the empirical quantiles are plotted over the theoretical quantiles of the distribution. If the data corresponds to the distribution, the points should be close to a line with a slope of 45 °.

If the model assumptions are not correct, it is possible that the model does not represent the data but this is not necessarily the case. Even if some of the model assumptions are violated, the representation of the data can still be reasonably good.
The next step is to check for collinearity of the co-variables. Highly correlated co-variables cause inaccurate estimations with high variance and should thus be avoided. If we look at the formula for the variance of β_j

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{1 - R_j^2 \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}, \tag{3.80}$$

where R_j is the coefficient of determination of a regression for x_j over all other input parameters, we can see that for a high correlation of x_j with the other explaining variables (measured by R_j^2) the variance increases. If R_j^2 is close to one, the variance becomes very high and the estimate for β_j is very inaccurate.
To check for collinearity the variance inflation factor can be used:

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2} \tag{3.81}$$

It indicates by which factor the variance of $\hat{\beta}_j$ is increased due to the linear dependency. The higher the correlation between x_j and the other co-variables, the higher R_j^2 and VIF_j. The problem of collinearity can be solved for example by leaving out some of the co-variables or by summarising the affected variables to one variable.

Another important step of the model analysis is the analysis of outliers. Outliers are observations that do not fit to the model and therefore have high residuals. They can strongly influence the estimates and the deductions made and can be detected by looking at the so-called "leave one out" residuals which are given by

$$\hat{\varepsilon}_{(i)} = y_i - \hat{y}_{(i)} = y_i - \mathbf{x}_i'(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}\mathbf{X}_{(i)}'\mathbf{y}_{(i)} \tag{3.82}$$

where $\mathbf{X}_{(i)}$ is the design matrix without the ith row. Those residuals are based on estimates where the ith observation is not considered. By standardising we get the studentised residuals

$$r_i^* = \frac{\hat{\varepsilon}_{(i)}}{\hat{\sigma}_{(i)}(1 + \mathbf{x}_i'(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}\mathbf{X}_{(i)}'\mathbf{x}_i)^{1/2}}. \tag{3.83}$$

In this equation $\hat{\sigma}_{(i)}$ is the variance estimated without using the ith observation. For a correctly specified model, the studentised residuals follow a t-distribution with $n - p - 1$ degrees of freedom (n is the number of observations, p the number of parameters). For a given significance level α we can compare the $\alpha/2$-quantile respectively the $1 - \alpha/2$-quantile of that t-distribution with the value of the residual r_i^*. If it is smaller than the $\alpha/2$-quantile or larger than the $1 - \alpha/2$-quantile, the observation can be considered as an outlier.
Outliers should be closely observed because they can be a hint for errors in the data or deliver information that has not been known yet. To reduce the influence of outliers, so-called robust methods can be used (see Chapter 3.1.7).
Next to outliers, also observations that have a high impact on $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{y}}$ should be identified. Often these observations are outliers that strongly influence the estimated parameters. To find especially influential observations, the leverages or the Cook's distances can be computed.
The leverages are the diagonal elements h_{ii} of the prediction matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and have values between $1/n$ and one. A leverage close to one means that the variance is very small and the regression line is certainly close to that point, no matter what values the other observations have. Thus, this observation has a high impact. Ideally, the leverages should be equally distributed.
If $\hat{\mathbf{y}}_{(i)}$ is an estimate based on all observations except for the ith one and p is the number of parameters, then the Cook's distance is given by

$$D_i = \frac{(\hat{\mathbf{y}}_{(i)} - \hat{\mathbf{y}})'(\hat{\mathbf{y}}_{(i)} - \hat{\mathbf{y}})}{p * \hat{\sigma}^2}. \tag{3.84}$$

Observations with $D_i > 0.5$ can be considered as noticeable and if $D_i > 1$ the observation should necessarily be inspected.

3.1.7 Robust Regression

The main disadvantage of the Least Squares method is that it is very vulnerable to outliers. As the errors are squared, high residuals caused by outliers can have a strong impact on the estimates. Therefore, estimators that are robust to outliers have been developed.
According to [11] there are three different types of outliers:

- Vertical outliers which have a very high error term but are within the "normal" range in the space of the explanatory variables;
- Good leverage points which are outliers in the space of the explanatory variables but are close to the estimated regression line;
- Bad leverage points which are far away from both, from the regression line and from most other points in the space of the explanatory variables;

Riani et al. [9] divide robust regression methods into three classes:

1. Hard (0,1) Trimming: In the Least Trimmed Squares (LTS) method the sum of squares of the residuals of h observations is minimised. Thus, the amount of trimming is defined by the trimming parameter h.
2. Adaptive Hard Trimming: In this case the value of h is determined by the data. One starts with a very robust fit using a low h and then adds more and more observations until all of them are included. This is called Forward Search. During the search different parameters, such as measures of fit, are monitored and afterwards the most appropriate model is chosen.
3. Soft Trimming (downweighting): In Soft Trimming Methods observations that are far away from the centre of the distribution receive less weight. Some examples for soft trimming, notably M-estimators and derived methods, will be presented below.

One possibility to reduce the influence of outliers is to minimise the absolute values of the residuals:

$$\hat{\beta}_{L_1} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} |r_i(\beta)| \tag{3.85}$$

This method, called median regression, protects against vertical outliers but not against bad leverage points. M-estimators are a generalisation of median regression. They also consider other functions than the absolute value:

$$\hat{\beta}_{L_1} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} \rho \left\{ \frac{r_i(\beta)}{\sigma} \right\} \tag{3.86}$$

where $\rho()$ is a loss function which is even, non-decreasing for positive values and less increasing than the square function. M-estimators are implemented using an iteratively reweighted Least

Squares algorithm. We assume that σ is known and define weights $\omega_i = \rho(r_i/\sigma)/r_i^2$. Equation 3.86 can be rewritten as

$$\hat{\beta}_{L_1} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} \omega_i r_i^2(\beta). \tag{3.87}$$

However, the weights are a function of β and therefore unknown. To start the LS algorithm, the weights can be computed using an initial estimate $\hat{\beta}$ for β. This algorithm converges to the global minimum of 3.86 only for monotone M-estimators which are not robust against bad leverage points [11] .

A popular M-estimator is the so-called Huber estimator. Its objective and weight functions are given by

$$\rho(r) = \begin{cases} 0.5r^2 & |r| \leq t \\ |r|t - 0.5r^2 & |r| \geq t \end{cases} \tag{3.88}$$

and

$$w(r) = \begin{cases} 1 & |r| \leq t \\ t/|r| & |r| \geq t \end{cases} \tag{3.89}$$

where t is a tuning constant to achieve the desired efficiency [1].

It is given by [11] that higher robustness can be achieved by minimising a measure of dispersion that is less sensitive to extreme values than the variance. This class of estimators is called S-estimators. For the ordinary least squares method the goal is to minimise the variance $\hat{\sigma}^2$ of the residuals, which can be written as $1/n \sum_{i=1}^{n} (r_i/\hat{\sigma})^2 = 1$. To increase the robustness, the square function can be replaced by another loss function that is less sensitive to large residuals. The estimation problems consists of finding the smallest robust scale of the residuals. The robust dispersion $\hat{\sigma}^S$ satisfies

$$\frac{1}{n} \sum_{i=1}^{n} \rho \left\{ \frac{r_i(\beta)}{\hat{\sigma}^S} \right\} = b \tag{3.90}$$

where $b = E\{\rho(Z)\}$ with $Z \sim N(0,1)$. The S-estimator is the value of β that minimises $\hat{\sigma}^S$:

$$\hat{\sigma}^S = \underset{\beta}{\text{argmin}}\ \hat{\sigma}^S\{r_1(\beta), ..., r_n(\beta)\} \tag{3.91}$$

3.2 Generalised Linear Models

Linear models are especially suitable for regression analysis where the dependent variable is continuous and can be modelled approximately with a normal distribution. Besides, it has to be possible to describe the expected value of the dependent variable as a linear combination of - possibly transformed - co-variables. Generalised linear models also assume a linear relationship between the input variables and the dependent variable but the dependent variable does not necessarily have to be continuous or have normally distributed values. They have the following common properties [3]:

1. The expected value $\mu = E(y|x)$ of the dependent variable y is connected with the linear predictor $\eta = \mathbf{x}'\boldsymbol{\beta}$ through a response function h respectively a link function $g = h^{-1}$:

$$\mu = h(\eta) \quad \text{respectively} \quad \eta = g(\mu) \tag{3.92}$$

2. The distributions of the dependent variables (e.g. normal distribution, Poisson distribution) can be written in the form of a single-parameter exponential family. The density of a single-parameter exponential family for the dependent variable y is given by

$$f(y|\theta) = \exp\left(\frac{y\theta - b(\theta)}{\phi}\omega + c(y, \phi, \omega)\right) \tag{3.93}$$

The parameter θ is called canonical parameter. The function $b(\theta)$ has to fulfil the condition that $f(y|\theta)$ can be normalized and the that $b'(\theta)$ and $b''(\theta)$ exist. ϕ is a dispersion parameter and ω a known value (usually a weight).

3.2.1 Binary Regression

If we have the data $(y_i, x_{i1}, ..., x_{ik}), i = 1, ..., n$ for a binary dependent variable y that is coded with 0 and 1 and the co-variables $x_1, ..., x_k$, we are looking for

$$\pi_i = P(y_i = 1|x_{i1}, ..., x_{ik}) = E(y_i|x_{i1}, ..., x_{ik}). \tag{3.94}$$

I.e., we want to determine the conditional probability for $y_i = 1$ given the values of the co-variables $x_{i1}, ..., x_{ik}$. In binary regression models the probability π_i is linked with the linear predictor η_i through a relationship of the form

$$\pi_i = h(\eta_i) = h(\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}). \tag{3.95}$$

The function h is a strictly monotically increasing distribution function so that always $h(\eta) \in [0, 1]$ and the relationship can be written in the form of

$$\eta_i = g(\pi_i) \tag{3.96}$$

by using the inverse function $g = h^{-1}$. h is the response function and $g = h^{-1}$ is the link function. Different functions can be used to model the relationship:

- Logit model:

$$\pi = \frac{\exp(\eta)}{1 + \exp(\eta)} \Leftrightarrow \log\frac{\pi}{1 - \pi} = \eta \tag{3.97}$$

- Probit model:

$$\pi = \Phi(\eta) \Leftrightarrow \Phi^{-1}(\pi) = \eta \tag{3.98}$$

where Φ is the distribution function of the standardised normal distribution

- Complementary log-log model:

$$\pi = 1 - \exp(-\exp(\eta)) \Leftrightarrow \log(-\log(1-\pi)) = \eta \tag{3.99}$$

[3]

3.2.2 Regression for Count Data

Count data occurs when the number of certain events during a time period or the frequencies in contingency tables are analysed. The dependent variables y_i have values from $0, 1, 2, ...$ and are independent for given co-variables $x_{i1}, ..., x_{ik}$. If all events appear very often, an approximation with a normal distribution can be sufficient. But usually discrete distributions considering the special characteristics of count data are the most appropriate. The most common approach is too use the Poisson distribution [3]:

$$P_\lambda(k) = \frac{\lambda^k}{k!} e^{-\lambda} \tag{3.100}$$

There are different types of Poisson regression models:

- Log-linear Poisson model: $y_i|\mathbf{x_i} \sim P_{\lambda_i}(k)$ with

$$\lambda_i = \exp(\mathbf{x_i}'\boldsymbol{\beta}) \quad \text{respectively} \quad \log\lambda_i = \mathbf{x_i}'\boldsymbol{\beta} \tag{3.101}$$

- Linear Poisson model:

$$\lambda_i = \eta_i = \mathbf{x_i}'\boldsymbol{\beta} \tag{3.102}$$

- Model with an overdispersion parameter:

$$E(y_i|\mathbf{x_i}) = \lambda_i = \exp(\mathbf{x_i}'\boldsymbol{\beta}),\ Var(y_i|\mathbf{x_i}) = \phi\lambda_i \tag{3.103}$$

where ϕ is the overdispersion parameter .

3.2.3 Models for Positive Continuous Dependent Variables

The classical linear model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i,\ E(\varepsilon_i) = 0,\ \text{Var}(\varepsilon_i) = \sigma^2 \tag{3.104}$$

is appropriate especially in cases where the errors ε_i are at least approximately normally distributed. In that case, the dependent variables y_i are independent for a given input vector $\mathbf{x}_i$

and normally distributed with

$$y_i|\mathbf{x_i} \sim N(\mu_i, \sigma^2),\ \mu_i = E(y_i|\mathbf{x}_i) = \mathbf{x}_i'\boldsymbol{\beta}. \tag{3.105}$$

In many applications the dependent variables cannot have negative values and their distribution is right skewed. On possibility to use the methodology of linear models in that case is to take the logarithm of the dependent variable and to use a linear model for $\bar{y} = \log(y)$, i.e.:

$$\bar{y}_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \quad \text{respectively} \quad \bar{y}_i|\mathbf{x}_i \sim N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2). \tag{3.106}$$

The original variable y is log-normally distributed with

$$E(y_i) = \exp(\mathbf{x}_i'\boldsymbol{\beta} + \sigma^2/2),\ \mathrm{Var}(y_i) = \exp(2\mathbf{x}_i'\boldsymbol{\beta} + \sigma^2)(\exp(\sigma^2) - 1). \tag{3.107}$$

We can insert the LS estimate $\hat{\boldsymbol{\beta}}$ and the estimated variance $\hat{\sigma}^2$ for the linear model. When transforming back using the exponential function, there can be considerable distortions at the estimation $\hat{\mu}_i = \exp(\hat{\eta}_i) = \exp(\mathbf{x}_i'\boldsymbol{\beta} + \sigma^2/2)$.
Therefore it could be better to assume a gamma distribution with the expected value $E(y_i) = \mu_i$ and the scale parameter ν for the dependent variables $y_i|\mathbf{x}_i$. The variance $\mathrm{Var}(y_i) = \sigma_i^2$ is

$$\mathrm{Var}(y_i) = \mu_i^2/\nu \tag{3.108}$$

For the non-negative, gamma-distributed dependent variables $E(y_i) = \mu_i > 0$ is valid. Therefore, instead of a direct linear approach a multiplicative exponential model

$$y_i = \exp(\eta_i) = \exp(\mathbf{x}_i'\boldsymbol{\beta}) = \exp(\beta_0)\exp(\beta_1 x_{i1}) \cdot ... \cdot \exp(\beta_k x_{ik}) \tag{3.109}$$

with the response function $h(\eta) = \exp(\eta)$ is used [3].

3.3 Mixed Models

Mixed models include random effects or coefficients in the predictor $\eta = \mathbf{x}'\beta$. Therefore they are also called random effect models. Important applications of mixed models are the analysis of longitudinal or cluster data. Longitudinal data is data that is collected if repeated observations of individuals or objects are made. Cluster data is gained if several objects from one primary unit (cluster) are selected and the values of the corresponding variables are collected. For each individual respectively for each cluster we have $i = 1, ..., m$ repeated observations

$$(y_{i1}, ..., y_{ij}, ..., y_{in}, \mathbf{x}_{i1}, ..., \mathbf{x}_{ij}, ..., \mathbf{x}_{in}). \tag{3.110}$$

For longitudinal data y_{ij} is an observation of the individual i at time t_j, whereas for cluster data y_{ij} is the value of the dependent variable for the jth object of the cluster i. To model the effects that are specific for the individuals or clusters, the linear predictor $\eta_{ij} = \mathbf{x}'_{ij}\beta$ for the observation y_{ij} is extended to $\eta_{ij} = \mathbf{x}'_{ij}\beta + \mathbf{u}'_{ij}\boldsymbol{\gamma}_i$. $\mathbf{u}_{ij}$ is a sub-vector of the covariables and $\boldsymbol{\gamma}_i$ the vector for the random effects. An advantage of this type of model is that correlations due to repeated observations for one individual or cluster are considered [3].

3.3.1 Mixed Linear Models for Longitudinal Data and Cluster Data

For longitudinal data

$$(y_{ij}, \mathbf{x}_{ij}), \quad i = 1, ..., m, \quad j = 1, ..., n_i \tag{3.111}$$

where m is the number of individuals or clusters and n_i the number of observations for individual/cluster i with observation at the times $t_{i1} < ... < t_{ij} < .. < t_{in_i}$, we receive a model of the form

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{u}'_{ij}\boldsymbol{\gamma}_i + \varepsilon_{ij}. \tag{3.112}$$

In this equation x_{ij} is a vector of covariates where one component is 1 to include a constant β_0. The vector u_{ij} also includes a 1 and can additionally contain components that are already included in x_{ij}. The vector $\boldsymbol{\gamma}_i$ stands for the specific deviation of one individual from the coefficients $\boldsymbol{\beta}$ that are estimated for the whole population. In matrix notation a mixed linear model (MLM) for longitudinal and cluster data is given by

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{U}_i\boldsymbol{\lambda}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, ..., m \tag{3.113}$$

where y_i is the n_i-dimensional vector of the dependent variables for the individual respectively the cluster i, m is the number of individuals/clusters, $\mathbf{X}_i$ and $\mathbf{U}_i$ are $(n_i \times p)$- and $(n_i \times q)$-dimensional design matrices for known covariates, $\boldsymbol{\beta}$ is the p-dimensional vector of the fixed effects, λ_i is a q-dimensional vector for the random effects and $\boldsymbol{\varepsilon}_i$ is an n_i-dimensional error vector. The following assumptions about the distributions are made:

$$\boldsymbol{\lambda}_i \sim N(\mathbf{0}, \mathbf{D}) \quad \boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \Sigma_i) \tag{3.114}$$

$\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_m$, and $\boldsymbol{\varepsilon}_1, ..., \boldsymbol{\varepsilon}_m$ are independent. Thus, the effects that are specific for the individuals respectively the clusters are considered as random values that are independent and normally distributed. The covariance matrix for the random effects is

$$\mathbf{D} = \begin{pmatrix} \tau_0^2 & \tau_{01} & \dots & \tau_{0q} \\ \tau_{10} & \tau_1^2 & \dots & \tau_{1q} \\ \vdots & \vdots & & \vdots \\ \tau_{q0} & \tau_{q1} & \dots & \tau_q \end{pmatrix}. \tag{3.115}$$

The covariances τ_{kl} represent dependencies between the random effects, the values of the diagonal elements τ_{kk} how far the individual effects the spread around the global effects [3].

3.3.2 The general linear mixed model

It is described in [3] that if we summarise all observations, the model can be written in a more compact form. We define the vectors

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_i \\ \vdots \\ \mathbf{y}_m \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon}_i \\ \vdots \\ \boldsymbol{\epsilon}_m \end{pmatrix}, \quad \boldsymbol{\gamma} = \begin{pmatrix} \boldsymbol{\gamma}_1 \\ \vdots \\ \boldsymbol{\gamma}_i \\ \vdots \\ \boldsymbol{\gamma}_m \end{pmatrix} \tag{3.116}$$

of all dependent variables, error terms and random effects. The design matrices are given by

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_i \\ \vdots \\ \mathbf{X}_m \end{pmatrix}, \quad \mathbf{U} = \text{diag}(\mathbf{U}_1, ..., \mathbf{U}_i, ..., \mathbf{U}_m). \tag{3.117}$$

The linear mixed model is given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{3.118}$$

with

$$\begin{pmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\varepsilon} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{pmatrix} \right) \tag{3.119}$$

where the covariance matrices are block-diagonal:

$$\mathbf{R} = \text{diag}(\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_i, ..., \boldsymbol{\Sigma}_m), \ \mathbf{G} = \text{diag}(\mathbf{D}_1, ..., \mathbf{D}_i, ..., \mathbf{D}_m). \tag{3.120}$$

3.3.3 Estimation of the Parameters

A common estimator for the parameters of the linear mixed model is the Best Linear Unbiased Predictor (BLUP) described in [10]. The estimates are linear because they are linear functions of the data, unbiased because the average value of the estimate is equal to the average value of the quantity being estimated and best because they have the minimum squared error among linear unbiased predictors. It is assumed that the variance-covariance structure is known except for a constant σ^2.

The BLUP estimates are defined as the solutions of the following simultaneous equations:

$$\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}'\mathbf{R}^{-1}\mathbf{U}\hat{\boldsymbol{\gamma}} = \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \tag{3.121}$$

$$\mathbf{U}'\mathbf{R}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} + (\mathbf{U}'\mathbf{R}^{-1}\mathbf{U} + \mathbf{G}^{-1})\hat{\boldsymbol{\gamma}} = \mathbf{U}'\mathbf{R}^{-1}\mathbf{y}. \tag{3.122}$$

If $\mathbf{X}$ is full rank, the covariance matrix of estimation errors is

$$E\left(\begin{pmatrix}\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\\ \hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\end{pmatrix}\begin{pmatrix}\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\\ \hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\end{pmatrix}'\right) = \begin{pmatrix}\mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{U}\\ \mathbf{U}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{U}'\mathbf{R}^{-1}\mathbf{U}+\mathbf{G}^{-1}\end{pmatrix}^{-1}\sigma^2 \tag{3.123}$$

3.3.4 Hypothesis tests

Usually, primarily test for the fixed effects are made. The test for one component β_j of $\boldsymbol{\beta}$

$$H_0 = \beta_j = d_j \quad H_1 : \beta_j \neq d_j \tag{3.124}$$

can - assuming an approximately normal distribution for $\hat{\boldsymbol{\beta}}$ - be made using a confidence interval for $\hat{\beta}_j$. If d_j is not in the confidence interval H_0 is rejected. Equivalently the test statistics

$$t_j = \frac{\hat{\beta}_j - d_j}{\hat{\sigma}_j} \tag{3.125}$$

where $\hat{\sigma}_j$ is the square root of the jth diagonal element of the covariance matrix for $\hat{\boldsymbol{\beta}}$ can be used. For big samples a normal distribution of t_j is assumed and H_0 is rejected if $|t_j| > z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ is a quantile of the standard normal distribution.
More general hypotheses of the form

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d} \quad H_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{d} \tag{3.126}$$

can be tested using the Wald statistics

$$w = (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})'(\mathbf{C}\mathbf{V}\mathbf{C}')^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}) \tag{3.127}$$

where $\mathbf{V}$ is the covariance matrix of $\boldsymbol{\beta}$ and therefore $\mathbf{C}\mathbf{V}\mathbf{C}'$ the covariance matrix of $\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}$. It measures the weighted difference between $\mathbf{C}\hat{\boldsymbol{\beta}}$ and $\mathbf{d}$. If H_0 is true, it should be small [3].

3.4 Regression Analysis in R

R is an open source project that has been developed specially for statistical computing. It provides an environment with a lot of built-in functions for interacting with data as well as a language for programming. It also includes several functions for regression analysis. The basic function for regression analysis is the `lm` function which can be used to estimate the

coefficients of a linear model containing one or several input parameters and one output parameter. The function uses the least squares method to find the best fit to the data. To get some information about the fit, the summary function can be used. Figure 3.1 shows the output of this function for a simple linear model with one dependent variable and one predictor. In

```
Call:
lm(formula = actual_time_min ~ revised_planned_min, data = data_op)

Residuals:
    Min      1Q  Median      3Q     Max
-6.6457 -2.1199 -0.9868  0.4002 31.5890

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         -5.7146     1.3347  -4.281 2.56e-05 ***
revised_planned_min   1.1430     0.1243   9.193  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.371 on 276 degrees of freedom
Multiple R-squared:  0.2344,    Adjusted R-squared:  0.2317
F-statistic: 84.52 on 1 and 276 DF,  p-value: < 2.2e-16
```

Figure 3.1: Example output of the summary function

the first line the formula that has been used to call the function is repeated. Next, information about the residuals, i.e. the difference between the actual values and the estimated values, is given. Under the heading `Coefficients` the coefficients, (column `Estimate`), their standard error (column `Std. Error`), t-values and the corresponding p-values for the two-sided t-test (column `Pr(>t)`) are shown. The `t` values are the estimated coefficients divided by their estimated standard deviation and the null hypothesis for the test is $\beta_j = 0$ (see Chapter 3.1.3). After the p-value a symbol indicating the the significance level is displayed, e.g. for a p-value below 0.01 one star ($\star$) is shown, three stars ($\star\star\star$) indicate a very small p-value. Furthermore the residual standard error, which is a measure providing information about the residuals, is given. It is defined as the square root of the sum of the squared residuals divided by $n-p$ where n is the number of observations and p the number of input parameters. Another measure for the quality of a fit that is displayed is the coefficient of determination (`Multiple R squared`, see Chapter 3.1.2) respectively the adjusted R^2 (`Adjusted R squared`) which also takes into account the number of predictors used in the model (see Chapter 3.1.5) [8].
R also provides a function that - instead of performing least squares approximation - uses a robust estimator, namely the Huber M estimator, to estimated the coefficients of the linear model.

4 Application

4.1 Description of the Use Case

We are examining a warehouse for beverages where the order picking is done manually, i.e employees go or drive through the storage area to collect the different articles of an order. An order can consist of only one article or of several different articles. Of each article one or several units have to be picked. The work is done in two shifts which are planned by shift managers. To support the shift managers with the allocation of the operators, the redPILOT software is used.
As already mentioned in the introduction, the methods described in the previous chapters shall be used to predict the throughput, i.e. the units processed during a certain timespan, of the order picking process step for a certain time in the future, given certain conditions. At first, different input parameters are analysed to find the most relevant among them. Using these input parameters, a model shall be created to estimate the throughput.

4.1.1 Initial Situation and Goals

The estimated throughput for a process step is an important input factor for the planning process. Based on this estimation the number of operators is determined and the available operators are allocated to the different shifts. A better estimation of the throughput can help to improve the shift planning and to allocate the operators for the system and the time they are needed.
Currently the throughput of one operator is set statically and based on experience by a planner and is not updated if operational conditions are changing. Depending on the expected orders for different weeks, the number of operators needed is estimated based on this static throughput. In this Master's thesis it will be evaluated if there are input parameters that have an impact on the throughput of the order picking process step. The information about the influencing factors shall then be used to predict and automatically configure the throughput for a certain time in the future.
Based on discussion with planners and experts in warehouse optimisation, the following possible influencing factors have been identified:

- **Operators:** As the picking is done manually, the individual performances and the motivation of the operators have a big impact on the throughput.

- **Time:** The throughput can vary over the year, on different days of the week and at different hours of the day.
- **Weather:** The weather conditions could have an influence on the throughput.
- **Order structure:** The composition of the orders (e.g., types of goods, number of articles per order, ...) is also influencing the throughput.

4.1.2 Available Data

The data that is analysed comes from different sources: The redPILOT database provides information about the performance during a specific timespan, the operators, the weather conditions and about the orders that need to be finalised at a certain time. The performance and order information are available from June 2016 until now, the information about weather conditions beginning with June 2017. This data is stored in an SQL database. The values of the following input parameters are available:

- Performance of a certain operator at a certain time and day: The performance, i.e. the number of packing units processed, is given for intervals of one minute which are given by a timestamp for the start and end time. From this timestamp the month, the hour of the day, the day of the week and other time parameters can be deduced.
- Information on the operators: A specific operator is identified by a unique id. Operators belong to operator groups (e.g. full-time personnel, part-time employees, contract workers) and have competences for different process steps. Those competences indicate for which process step an operator should preferably be allocated. The higher the competence of an operator for a specific process step is, the higher is the likelihood that he will be allocated for this process step. Operators can also have additional skills such as first-aider or fire protection officer and they can decide if they are ready to do paid overtime or not. This information could be a measure for the motivation of an operator.
- Orders: In the redPILOT database only information about order quantities are available, which means that we do not know anything about the articles contained in the order or the complexity of the order. There are two different types of quantities: the total order quantity and the quantity of orders that have to be finalised. Both quantities are given for intervals of one hour. The total order quantity gives the total number of orders that are registered in the system at that hour of the day whereas the quantity of orders that have to be finalised indicates the number of orders that have to be finished within the given interval. Both quantities are given as a number of packing units.
- Weather conditions: The minimal and maximal temperature on a certain day as well as the noon temperature and the weather conditions (sun, rain, fog, etc.) are stored in the database.

Table 4.1 gives a list all the analysed parameters from the redPILOT database.

Table 4.1: Descriptions of the items in the redPilot database

Variable	Description	Type	unit	Availability
Performance	Number of packing units	Performance information	Packing Unit	Starting with June 2016
Month	Month	Time information	-	Starting with June 2016
Hour	Hour of the day	Time information	-	Starting with June 2016
Day	Day of the week	Time information	-	Starting with June 2016
Factor overtime premium	Information about overtime premium	Time information	-	Starting with June 2016
Distance to holiday	Days until next holiday	Time information	-	Starting with June 2016
Operator id	Internal id for an operator	Operator information	-	Starting with June 2016
Personal Number	Personal number of the operator	Operator information	-	Starting with June 2016
Operator group id	Group the operator belongs to	Operator information	-	Starting with June 2016
Competence KOM	Competence for process step order picking	Operator information	-	Starting with June 2016
Number other competences	Competences for other process steps	Operator information	-	Starting with June 2016
Number skills	Additional skills	Operator information	-	Starting with June 2016
Paid overtime	Willingness for overtime hours	Operator information	Yes/No	Starting with June 2016
Order quantity total	All current orders	Order quantity information	Packing units	Starting with June 2016
Order quantity finalise	Orders to be finalised	Order quantity information	Packing units	Starting with June 2016
Min Temperature	Minimal temperature on that day	Weather information	°C	Starting with June 2017
Noon Temperature	Temperature at noon	Weather information	°C	Starting with June 2017
Max Temperature	Maximal temperature on that day	Weather information	°C	Starting with June 2017
Weather	Weather conditions on that day	Weather information	-	Starting with June 2017

Additionally, data from the warehouse management system that gives information about the orders that are realised during the order picking can be delivered. The following tables will be used:

- Comparison of planned times and actual times: In this table the planned time for an order, the operator who fulfils the order, the timestamps of when the operator started respectively finished the order as well as the planned and actual packing units the orders consist of are available. The planned time gives information about the complexity of an order. It is the sum of the estimated times for all the steps needed to fulfil the order. Therefore, the difference between the planned time and the time the operator needs can give information about the performance of an operator. Very performant operators usually fulfil the orders in less than the plan time whereas less performant operators have a higher likelihood to need more time than planned.

- Order information: Usually one order consists of several articles that have to be picked. This table gives information about the articles included in an order that has to be fulfilled by an operator.

This data is not stored for a longer period of time, so, it is only available from the end of January (comparison of planned and actual times) respectively the beginning of February (order information) until now. It is delivered in Excel spreadsheets and then loaded into an SQL tables.
Table 4.2 summarises the relevant variables from the customer's WMS system.
Figure 4.1 shows an entity-relationship diagram diagram of the raw data from both, the redPILOT database and the data from the warehouse management system. Each operator belongs to an operator group. An operator group consists of several operators whereas one operator can only belong to one operator group. Therefore there is an 1:n relationship between

Table 4.2: Descriptions of the items of the additional data from the WMS

Variable	Description	Type	unit	Availability
User number	Personal Number of the operator	Operator information	-	Starting with end of January (gap of 11 days)
Company	information if internal or external operator	Operator information	-	Starting with end of January (gap of 11 days)
GroupRef	Reference to a customer order	Order structure information	-	Starting with end of January (gap of 11 days)
Units planned	Number of planned units for an order	Order quantity information	Packing units	Starting with end of January (gap of 11 days)
Actual units	Number of actually picked units for an order	Order quantity information	Packing units	Starting with end of January (gap of 11 days)
Planned time	Planned time for an order	Order structure information	Minutes	Starting with end of January (gap of 11 days)
Revised planned time	Updated planned time for an order	Order structure information	Minutes	Starting with end of January (gap of 11 days)
Actual time	Time actually needed to fulfil an order	Time information	Minutes	Starting with end of January (gap of 11 days)
Timestamp start	Time when the order was started	Time information	Time	Starting with end of January (gap of 11 days)
Timestamp end	Time when the order was finished	Time information	Time	Starting with end of January (gap of 11 days)
Article Number	ID of an article	Order structure information	-	Starting with beginning of February (gap of 17 days)
Number distinct articles	Number of different articles in one order	Order structure information	-	Starting with beginning of February (gap of 17 days)

operators and operator groups. One operator has many performance records which are given for intervals of one minute. The relationship between the operators and the performances is 1:n because one performance record only belongs to one specific operator. One record in the operator table gives information about one specific operator. It contains his or her id, personnel number, competences, skills, willingness for paid overtime and the id of the operator group he or she belongs to. One performance record consists of two timestamps, the start of the interval and the end of the interval, the operator id and the number of packing units processed. The orders are given for intervals of one hour which are defined by two timestamps. At latest at the time of the timestamp_end the orders to finalise have to be finished. The weather information is given for days which are indicated by a date. From the timestamps in all the tables time information like day of week, month or hour can be derived which can be used to connect performances, orders and weather. Furthermore there are two tables that give information about overtime premium and holidays. One record in the planned times table coming from the WMS consists of an id (groupRef), two timestamps (start and end time of the order), planned and revised planned time and planned and actual units. Furthermore the personnel number of the operator who executed the order is given which can be used to connect this table to the operator information. One operator executes several orders and therefore this is an 1:n relationship. Between the planned times and the order information there is also an 1:n relationship, i.e one order with a specific planned time consists of several articles which are given in the order information table. One row corresponds to one item that is picked and it is connected to the planned times by the groupRef value. The timestamp_start in this table indicates when the operator has received the information that he has to pick this article, the timestamp_end gives information when it has been confirmed that the article has been picked. The planned times and the order information can also be connected to the performances and the order quantities using the timestamp.

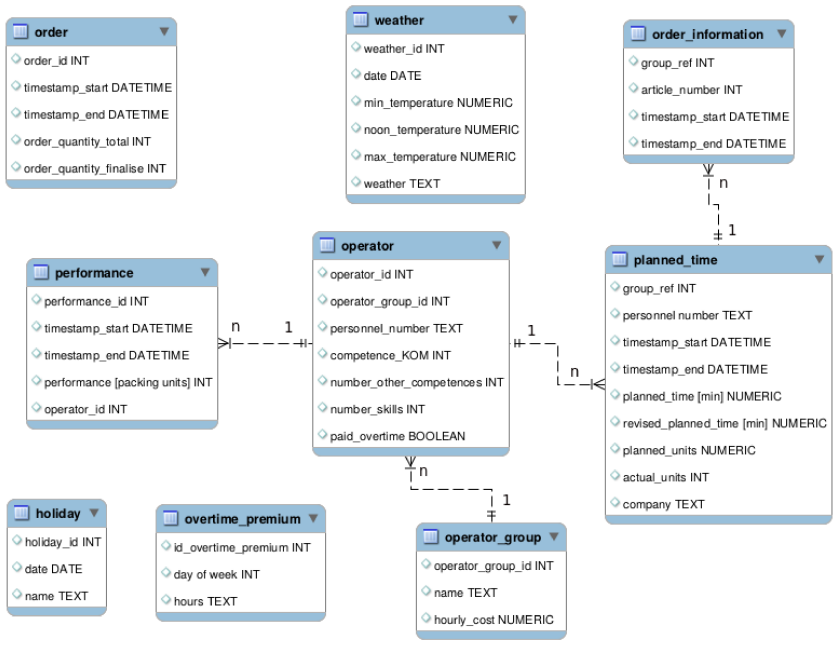

Figure 4.1: ER diagram of the raw data

4.2 Data Analysis

4.2.1 Aggregation of the Data

To prepare the data for the analysis which is done in R, it first has to be processed and aggregated in SQL. Different tables have to be joined and calculations and String operations have to be performed to retrieve information from the entries of the tables (e.g. to get the hour of the day and the day of the week from a timestamp). The data from the redPILOT database has been aggregated to the following tables:

- **Time data:** This table contains the overall performance of one hour at a certain date. The month, day and day of the week are extracted from the date. It also provides information to which shift the hour belongs, if overtime premium is paid and what the distance to the next holiday is. The dataset contains information from June 2016 to December 2017 and has around 8000 entries.

- **Performance operators:** This table contains the performance of a specific operator during one specific hour on a certain day. Informations about the operator (operator group, competences, skills, paid overtime) are included. The dataset contains information from June 2016 to December 2017 and has around 93300 entries.

- **Performance weather:** This table contains the performance on a specific day and the weather conditions on that day. The three different temperature values (see Chapter 4.1.2) are aggregated to an average temperature. The dataset contains information from June 2017 to December 2017 and has around 160 entries.

- **Performance orders:** This table contains the performance during a certain hour on a certain day and the number of orders for that time. The dataset contains information from June 2016 to December 2017 and has around 2300 entries.

- **Planned times:** This table contains data that gives information about the planned and actual times for a specific order. Each rows stands for an order that is executed by a certain operator. The entries include the personnel number of the operator, information if he or she comes from a temporary employment agency, a reference to a customer order, the original planned times and quantities, revised planned times (if the originally planned order cannot be fulfilled and therefore the calculated time to fulfil the order changes, the planned times are updated), actual quantities and actual start and end time of the order. The actual fulfilment times as well as the absolute and relative differences between planned and actual times have been calculated. It includes data from 20th of January 2018 to the 8th February 2018 and the 19th of February 2018 to the 28th of April 2018 and contains about 105000 observations.

- **Order information:** This table contains detailed information about the orders. One row corresponds to one article that has to be picked. It includes the article number and description, a reference to a customer order, the user number (personnel number of the operator) as well as the planned and actual number of items that have to be picked. Furthermore, the planned and actual times of the orders and their absolute and relative differences are given. This dataset includes is given from the 2nd of February 2018 to the 9th of February 2019 and the 26th of February 2018 to the 28th of April 2018 and includes around 564000 entries.

- **Distinct articles** This table provides the same information as the planned times table and additionally includes the number of distinct articles an order contains. The data is given for the same period as the order information data and contains about 76000 entries.

The gap where no data for the planned times respectively the order information is available occurs because the data is not stored automatically for a longer timespan and therefore has to be stored manually which has not been done during that time.
In Table 4.3 all tables of aggregated data, the included variables and the the availability of the data are listed.
Figure 4.2 shows an ER-diagram of the aggregated data. The data from the redPILOT database has been aggregated to 4 different tables (Time data, Performance operators, Performance weather and Performance orders) which are all connect via time attributes (e.g. date, hour of the day). The data from the warehouse management system has been aggregated in three different tables (Planned times, Order information, Distinct articles). Those three tables are all connected via the groupRef value. The relationship between the planned times and the

Table 4.3: Tables of Aggregated data

Data	Variables	Available period	Number rows
Time data	Performance per hour, Month, Day of week, Hour, Factor overtime premium, Distance to holiday	June 2016 - December 2017	8000
Performance operators	Operator id, Operator group id, Competence KOM, Number other competences, Number skills, Paid overtime, Performance per hour and operator	June 2016 - December 2017	93300
Performance weather	Min temperature, Noon temperature, Max temperature, Weather, Performance per day	June 2016 - December 2017	93300
Performance orders	Order quantity total, Order quantity finalise, Performance per hour	June 2017 - December 2017	160
Planned times	Operator id, personnel number, Company, GroupRef, Planned units, Planned time, Actual Units, Revised planned time, Actual time, Absolute/Relative difference planned time - actual time	20th of January - 8th February, 19th of February - 28th of April	105000
Order information	GroupRef, User number, Article number, Planned time, Revised Planned time, Actual Time, Absolute/Relative difference planned time - actual time	2nd of February - 9th February, 26th of February - 28th of April	564000
Distinct articles	Operator id, personnel number, Company, GroupRef, Planned units, Planned time, Actual Units, Revised planned time, Actual time, Absolute/Relative difference planned time - actual time, Number distinct articles	2nd of February - 9th February, 26th of February - 28th of April	76000

distinct articles table is a 1 to 1 relationship, i.e. one record in the planned times table belongs to one record in the distinct articles table. The relationship to the order information table is 1 to n, i.e. there are several records in the order information table which belong to one record in the planned times and distinct articles table.

4.2.2 First Tests on redPILOT data

To get an overview of the data, plots are very helpful. They can indicate which input parameters have a significant influence on the dependent variable and how the values are distributed. Besides, regression analysis is used to evaluate the influence of the different input variables on the performance.

4.2.2.1 Weather Data

First, we have a look on the data about the weather conditions. It could be possible that under certain weather conditions, for example very hot weather, working is harder and therefore the performance is lower. We will examine the influence of the average temperature and the weather conditions on the performance for one day. The Q-Q-Plot (Quantile-Quantile-Plot) shows if the values are approximately normally distributed. If the points are close to the 45° line as it is the case in for the daily performance (see Figure 4.3), a normal distribution can

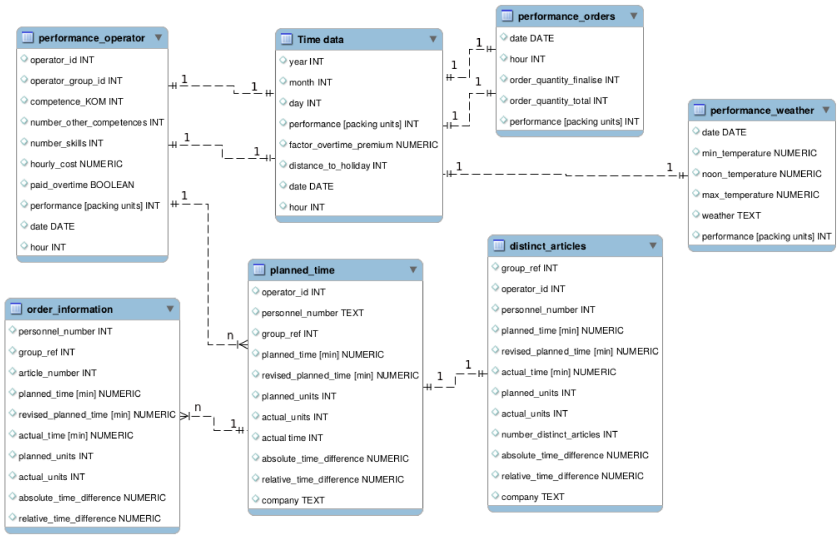

Figure 4.2: ER diagram of the aggregated data

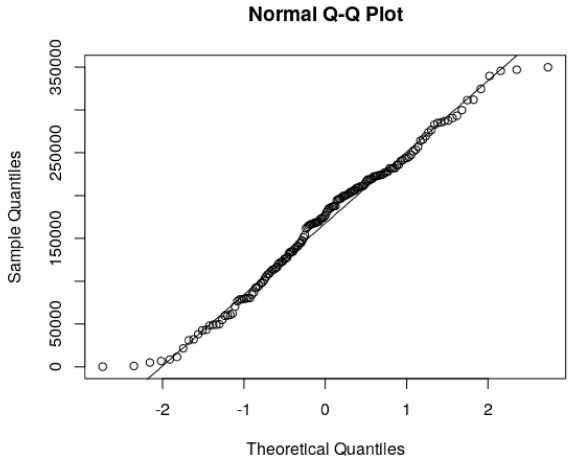

Figure 4.3: Q-Q-Plot Daily Performance

be assumed. Figure 4.4 shows the performance for days with a certain average temperature. The values seem to be randomly distributed, so, there seems to be no relationship between the temperature and the performance. The plot next to it is a boxplot for the performance of days with certain weather conditions. The box gives the range within which 50% of the values can be found, the bold line marks the mean value. Apparently for the weather conditions "snow" and "wind" only very few values are available and the performance on cloudy and rainy days seems to be lower. If we plot only the values for the summer (15^{th} of June until 15^{th} of September) to eliminate the influence of seasonal differences in the order quantities, we still cannot see any patterns for the temperature and the lower performance for cloudy days is more distinct. We can try to created a linear regression model with the daily performance as the dependent variable and the average temperature and the weather conditions as input parameters. For the weather conditions we have several categories which can be modelled using

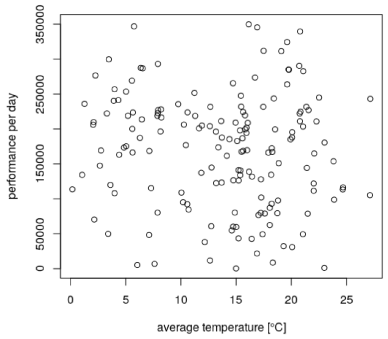

Figure 4.4: Performance over average temperature

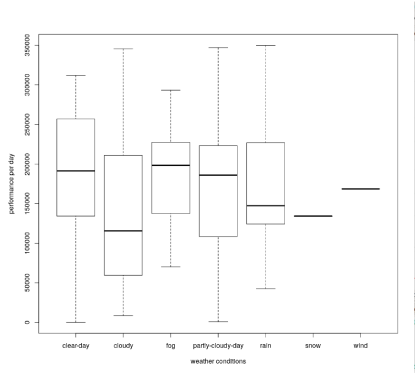

Figure 4.5: Performance over weather conditions

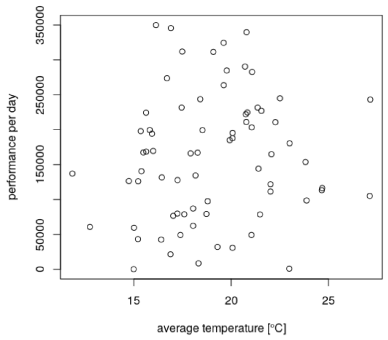

Figure 4.6: Performance over average temperature summer

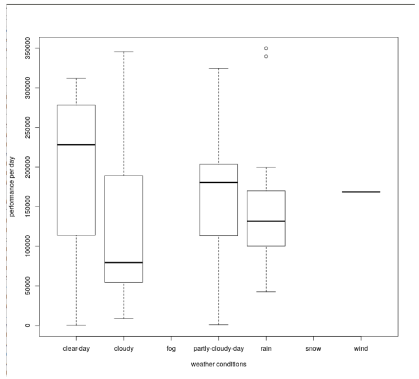

Figure 4.7: Performance over weather conditions summer

dummy coding. We use sunny weather conditions as a reference category to which the other weather conditions are compared. As expected after looking on the plots, a linear model is not appropriate to describe the relationship between the weather conditions and the performance. For most coefficients the t-values are very low and the significance level for which the null hypothesis (no significant influence of the input variable) can be rejected is high. Only for the intercept and the dummy variable for the category "cloudy weather" the null hypothesis can be rejected at a relatively low significance level (0.035) which means that the performance on cloudy days is significantly lower than on sunny days. We have only observed the total performance per day, so a lower performance does not necessarily mean that the operators are working less. As we are analysing the performance in a storage for beverages, there are usually more orders on sunny days. Therefore probably more operators are working and the overall performance increases. This impression seems to be confirmed by looking at the plots for the performance per operator per day (Figures 4.8 and 4.9). In these plots no tendencies for a higher performance under certain weather conditions is visible. If we again create a linear model for the performance over the average temperature and the weather conditions (using dummy coding with sunny as a reference category), we get high p-values for all weather

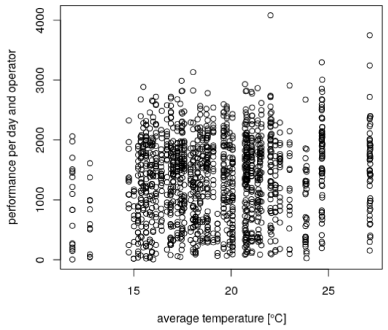

Figure 4.8: Performance per operator over average temperature summer

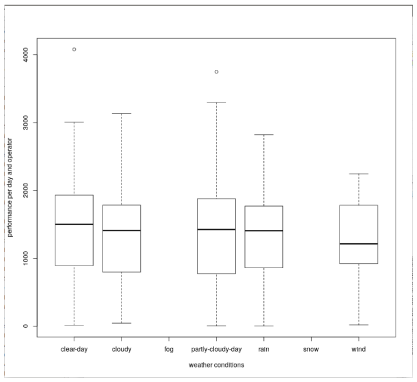

Figure 4.9: Performance per operator over weather conditions summer

conditions except for the category partly-cloudy which has a p-value of 0.035. The value of this coefficient is -98.9. For the coefficient of the average temperature we get a very low p-value and the coefficient of the parameter is 25. This means that the average performance per operator increases by 25 if the temperature rises by 1°C. This is not very much given that the average performance per operator and day is about 1360 and that the standard deviation of the performance values is 689. Thus, the influence is very small. Besides, the residual standard error has a value of 685 and is thus high and does not differ much from the standard deviation of the performance values. The adjusted R^2 is very small (0.01). This indicates that the variance of the performance cannot really be explained by changing weather conditions.

4.2.2.2 Operator data

Next, we analyse the data about the operators working at the order picking process step. As described in Chapter 4.1.2, different information about the operators is available. We know the id of an operator, the id of the operator group and its hourly wage rate, the competences, the number of skills and the willingness for paid overtime. For the competences two different values are available: The value of the competences for the process step order picking, which can be between 0 and 9, and the number of competences the operator has. By looking at the Q-Q-plot (Figure 4.10), we can see that for a certain range the values of the performances per hour of the operators are normally distributed. The first two plots (Figure 4.11 and 4.12) are boxplots of the performance for certain operators, identified by their id, and certain operator groups, also identified by an id. It can be seen that there are some differences in the performances of different operators and operator groups. As both variables are categorical, the coefficients calculated for the dummy variables by the linear model function in R represent the average performance of an operator respectively an operator group. Looking at the output of the summary for the linear model, we can see that the performances for many operators differ significantly from the performance of the "reference operator" whereas only for three operator groups the performance is significantly different from the one of the reference group.

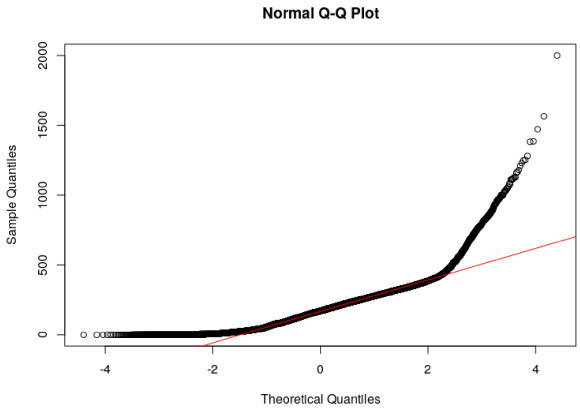

Figure 4.10: Q-Q-Plot Performance per operator

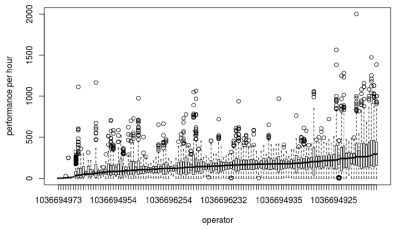

Figure 4.11: Performance over operator id

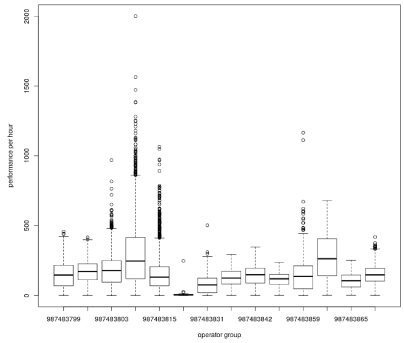

Figure 4.12: Performance over operator groups

The adjusted coefficient of determination (adjusted R^2) which gives information how well the variance in the dependent variable is explained by the model is 0.17 for the model including the operator id. This implies that the operator has a significant influence on the performance. For the operator group the adjusted R^2 is about 0.06. Thus, the difference between the operator groups does not seem to be that significant.

Next, we compare the performance for operators with (Y) and without (N) paid overtime (Figure 4.13). The median of the performances per hour for operators with and without paid overtime seems to be similar but for operators without paid overtime we can see much more outliers above the 50% range that is given by the box. This can be explained by the fact that only two of the operators in our dataset have paid overtime and we therefore have much more observations for operators without paid overtime. Therefore, we cannot draw a valid conclusion concerning the influence of this parameter.

The next plot (Figure 4.14) shows a comparison of the performances for the different levels of hourly costs, i.e. the personnel cost per hour for a specific operator group. Those operators who belong to an operator group where the hourly cost is 0 are apprentices who receive a fixed salary that does not depend on their working time or their performance. For the different

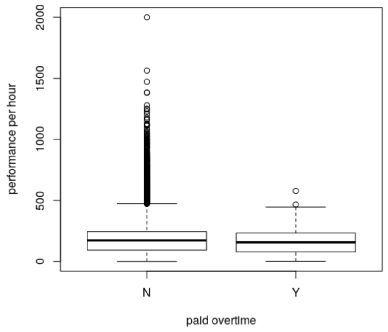

Figure 4.13: Performance over paid overtime

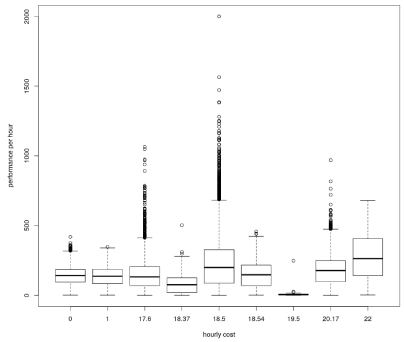

Figure 4.14: Performance over hourly cost

values of hourly cost, differences in the performance can be seen. These differences for the hourly costs correspond to the differences between the operator groups. E.g., the performance values for the category 18.5 seem to be higher. These costs belong to the group "987483859" for which the boxplot also indicates a higher performance. If we again create a linear model where different levels of hourly costs are modelled with dummy coding to compare the groups, we receive p-values below 0.05 for 6 of the groups but the adjusted R^2 is only 0.04 which is even lower than for the operator groups which means that the influence of the hourly costs on the performance does not seem to be big.

Finally, we have a look on the plots for the competences and skills. Figure 4.15 shows a boxplot over the categories of the competences for the process step order picking. It could be expected that operators with a higher competence have higher performances but the boxplot does not confirm this assumption. We can see some small differences in the median values of the performance values for different values of competences but for most of the categories there a quite a lot of outliers. Furthermore, the plot does not really support the hypothesis that operators with a higher competence value have higher performances. In contrast, the median for competence 0 is very similar to the median for competence 9 and we have a lot more high outliers for competence 0. Another assumption would be that operators with competences for several process steps and some additional skills have more experience and therefore a higher performance but this hypothesis is not supported if we look at the plots of the data. In Figure 4.16, which shows a boxplot of the performances over the number of other competences, we can see that the median values are relatively similar for all the values of the number of other competences and that we have a lot of outliers. For 6 other competences the performance seems to be a bit higher and for 7 other competences it seems to be a bit lower than for the other values but for these values there is only one operator who has this number of other competences and therefore we cannot conclude that operators with six other competences generally have a higher performance. As we can see that the median values are very similar and that there are a lot of outliers for all values of the number of other

competences, we furthermore cannot conclude that a specific number of competences leads to higher performance. Figure 4.17 shows a boxplot for the performance over the number of skills. We can see that there are only three different values for the number of skills: 0, 1 and 2. For one and two skill(s) the performance values seem to be relatively homogeneous whereas for no skills there are a lot of outliers. This can be explained by the fact that there are a lot more operators who have no skills. There are 118 operators with no skills, two operators with one skill and eight operators with two skills. Therefore we have a lot more observations for zero skills which leads to a higher variety of values and it is not really possible to draw a conclusion about the impact of the number of skills on the performance.

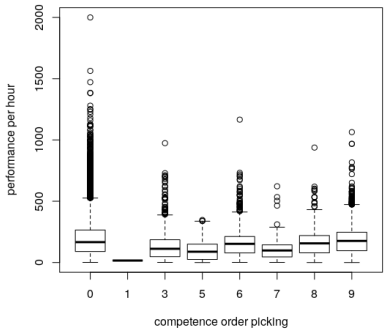

Figure 4.15: Performance over competence process step order picking

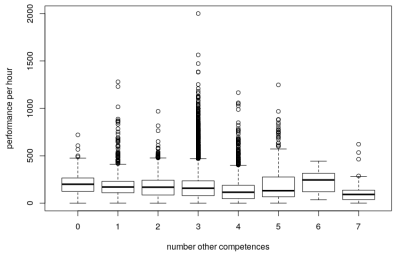

Figure 4.16: Performance over number of other competences

4.2.2.3 Order data

The next input parameters that are analysed are the two variables for order quantities: the total order quantity and the order quantity to be finalised. The dependent variable is the performance per hour.

In Figure 4.18 we can see that the different performance values are relatively equally distributed

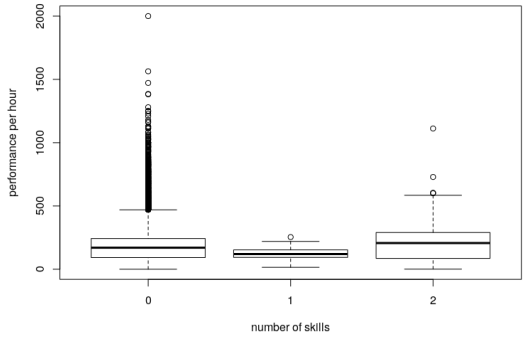

Figure 4.17: Performance over number of skills

over the different values for the total order quantity. If we create a linear model for the performance over the total order quantity, we receive a coefficient of 0.0066 for the order quantity which means that the performance only increases very little if the order quantity increases. Furthermore the adjusted R^2 is only 0.02 which means that very little of the variance in the performance is explained by the variance of the total order quantity.
Figure 4.19 shows the performance over the number of orders to finalise. It can be seen that this number is very often zero or very small and that we have a wide range of performance values for these small order quantities. The many small values occur because the quantities of orders that have to be finalised are not spread evenly over the hours of the day. Often, there is one hour at the end of the day where the quantity of orders that have to be finished is very high and during the rest of the day the numbers are small. This does not mean that all those orders are actually done during that hour, it can rather be seen as a deadline. This makes this input parameter not very valuable for creating a model. As it can be expected, creating a linear model does not reveal any correlation. The coefficient for the order quantity to finalise is very small and has a very high p-value (0.87).

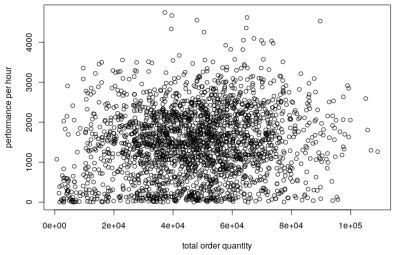

Figure 4.18: Performance over total order quantity

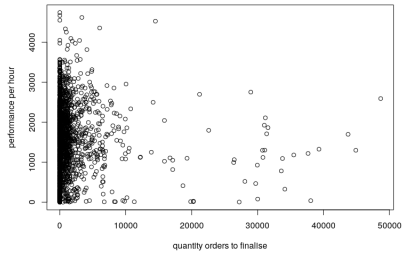

Figure 4.19: Performance over order quantity to finalize

4.2.2.4 Time data

In this section we analyse the influence of the time when the order is executed on the performance. This time we analyse the data for the whole year to also identify possible seasonal fluctuations. The performance is aggregated per operator and per hour. In the Figures 4.20 and 4.21 we can see that the performance of the operators is lower for certain hours of the day whereas the performance seems to stay constant over the year. The next figures show that the performance is almost equal from Monday to Saturday (day 2-6) but higher on Sunday (day 1) and that it is higher when the factor for overtime premium is 1.75 which is the case for Sunday evening. Thus, this plot gives the same information as a combination of Figure 4.20 and Figure 4.22. As it can be seen in Figure 4.24, the distance to the next holiday does not seem to have any influence. The performance values for the different numbers of days are relatively evenly distributed.
If we create a linear model with dummy variables for the different hours and use the aver-

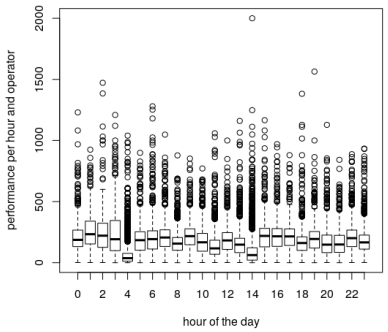

Figure 4.20: Performance over hour

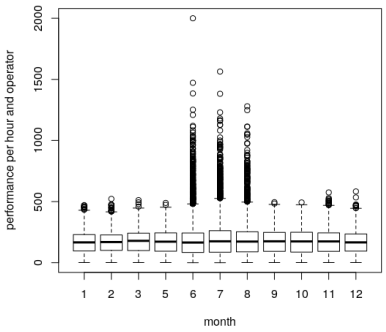

Figure 4.21: Performance over month

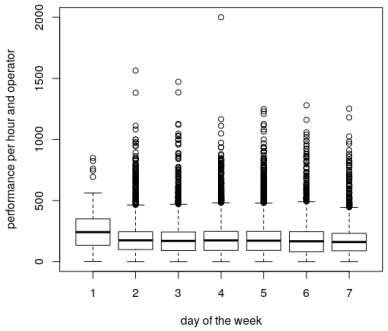

Figure 4.22: Performance over day of week

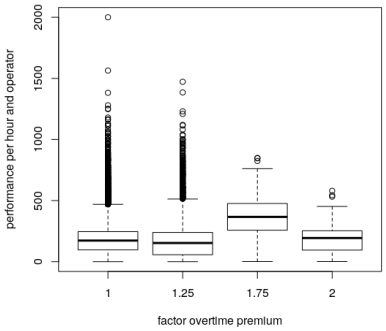

Figure 4.23: Performance over factor overtime premium

age performance as a reference value, we get high p-values (greater than 0.25) for all hours. This means that in none of the hours the performance differs significantly from the average performance. If we do the same for the day of the week, we get similar results: For all the coefficients we get p-values greater than 0.45 which means that the performance on none of the days is significantly different from the average performance. Concerning the factor of overtime premium, it can be observed that for the factors 1.25 and 1.75 the performance differs significantly from the performance at the time without any overtime premium whereas the performance for the time with factor 2 is not significantly different. The coefficient for the factor 1.25 is -11.9 which means that the average performance is a bit lower whereas for the factor 1.75 the coefficient is 190 and thus the average performance is a lot higher. The p-values for those coefficients are very small (10^{-16}) which means that the difference seems to be significant. For the factor 2 the coefficient is 6.2 and the p-value is 0.38 which means that the performances are very similar.

4.2.2.5 Conclusion

From the analysis done so far we can conclude that the performance of the individual operators seems to be significantly different. Besides, the performance seems to vary for the different

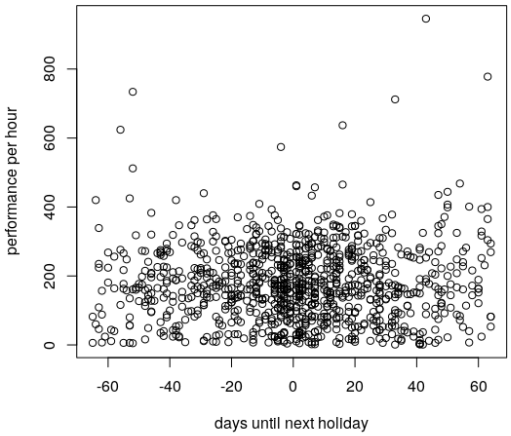

Figure 4.24: Performance over distance to holiday

hours of the day. It also seems to be higher on Sunday. Thus, we have possibly found some factors that influence the throughput but the information is not sufficient to predict the future throughput for a certain configuration. Besides, there are many factors which are not considered yet. We do not know if some operators have to do "harder" tasks and therefore have a lower performance or if the reason for the lower performance is that they a working slower. So, we need some information on the difficulty of the orders to be able to make better judgements on the influencing factors.

4.2.3 Analysis of Planned Times

4.2.3.1 Analysis of Whole Dataset

The planned time for an order is a guideline during which amount of time the operator should be able to accomplish an order. The salary of the operators partly depends on the fulfilment of this planned times. Thus, the fulfilment of the planned times can be a good measure for the performance of an operator. For this first analysis the data from the 20th of January to the 8th of February and from the 19th of February to the 17th of March will be used which is the data that is available so far. This dataset contains around 52000 observations.
At first we will analyse the relationship between the planned and the actual times, i.e. if the actual times are usually higher, lower or approximately equal as the planned times and how big the differences are. Figure 4.25 and Figure 4.26 show the frequencies of certain values for the absolute and relative differences between the planned times and the actual times for specific orders (in minutes). We can see that most of the values are positive which means that the time the operator needed was lower than the planned time. Only for a few orders the actual time was higher than the planned time. We also can see that there are some outliers where the actual time was a lot higher than the planned time. Furthermore we observe that in the

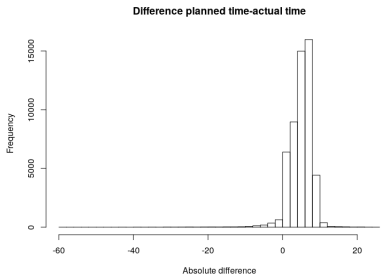

Figure 4.25: Histogram absolute differences

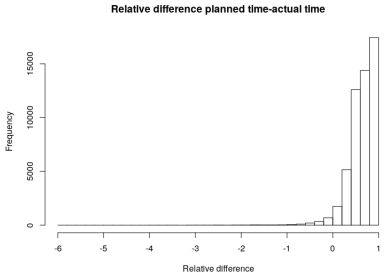

Figure 4.26: Histogram relative differences

histogram for the relative differences there are a lot of values between 0.8 and 1. If we have a closer look at the data, we find that many of the relative differences are exactly 1 and that the corresponding actual times are zero whereas the planned time is bigger than zero. The reason for these zero values is that there are some orders where the operator does not have to move and therefore can confirm the order the same time as he starts it. The values for the planned times of this orders, however, are not zero which causes a relative difference of 100%. As these values could distort the results of the analysis, we do not use them for creating the models which reduces the amount of data by around 10%. If we look at the histograms

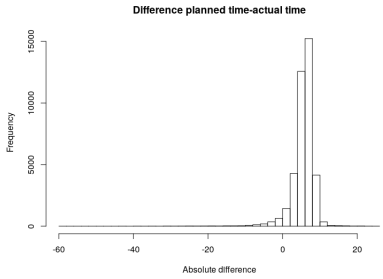

Figure 4.27: Histogram absolute differences without zero values

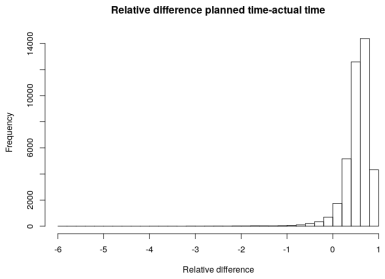

Figure 4.28: Histogram relative differences without zero values

for the absolute and relative differences for the data without the zero values for the actual time (Figures 4.27 and 4.28), we can see that there are now much fewer values between 0.8 and 1 for the relative difference and that the distribution of the absolute differences has also changed. Most relative differences now lie between 0.4 and 0.8.

Figure 4.29 shows the values of the actual times over the planned times for a sample of 500 observations. We can see that the relationship between the planned and the actual times seems to be relatively linear with some outliers. The red line is a regression line fitted to the data using least squares approximation. The coefficient for the planned time is approximately 0.863 which means that if the planned time increases by 1, the actual time increases by on average 0.863. The p-values for both coefficients (intercept and slope) are very small which means that there seems to be a significant between the planned and the actual time. The adjusted R^2 is 0.321 and therefore, only around 32.1% of the deviation in the actual time is

explained by the planned time. Thus, there must be other influencing factors.

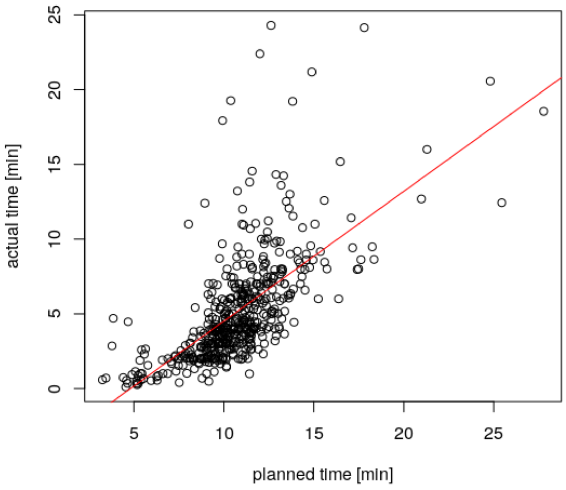

Figure 4.29: Actual times over planned times

One possible influencing factor could be the individual performances of the operators who are executing the orders. To see if there are any differences between the different operators, we have a look at the boxplot for the absolute and relative differences between the planned and actual times of an order for each operator. Figures 4.30 and 4.31 show the distribution of the absolute and relative differences for 15 randomly selected operators. Both of these plots indicate that there are significant differences between the performances of the different operators.

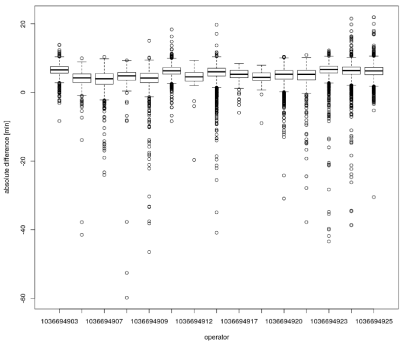

Figure 4.30: Boxplot absolute differences for individual operators

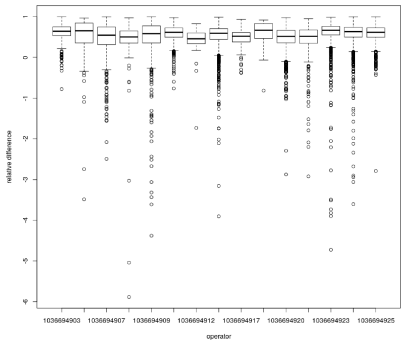

Figure 4.31: Boxplot relative differences for individual operators

To confirm this impression, we use the `lm` function for the formula `difference (absolute and relative)` $\sim$ `id` which computes the average deviation of the performance of all operators from the first operator which is taken as a reference. Many coefficients have low p-values which indicates that the values for the differences are significantly different from the reference value. Thus, it is sensible to create a different model for each operator.

4.2.3.2 Separate Models for Different Operators

For some operators we have only a few observations which makes it difficult to derive general rules from the values of the data of these operators. Therefore we will only create models for operators where we have a least 100 observations. Instead of the original 63 operators we only use the data of 54 operators.
If we plot the actual times over the planned times separately for each operator, we can see that the relationship between the input (planned times) and the output (actual time) seems to be much more linear than if we look at the data for all operators at once. Figure 4.32 shows plots for the actual times over the revised planned times (i.e. the planned times for the orders the operators actually fulfilled which can have changed on a very short-term basis) for 4 different operators. In this figure there are still some outliers but the data points are in general much closer to a virtual regression line.

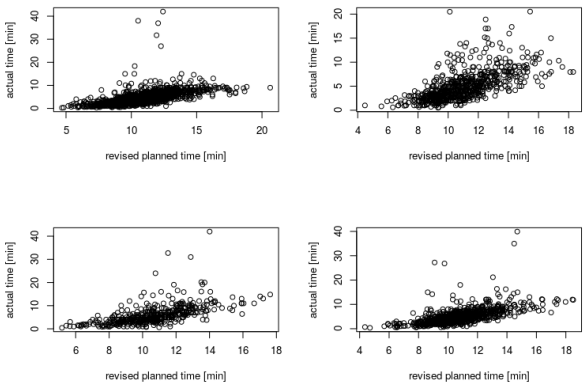

Figure 4.32: Actual times over planned times for 4 different operators

We can again use the linear model function with the formula `actual time` $\sim$ `planned time` to create the different models for all the operators. Figure 4.33 shows the intercepts and slopes of the different models created. For some operators the slope is greater than 1 whereas for other it is smaller than 1. Thus, for some operators the time they actually need - on average - increases by more than 1 when the planned time increases by 1 and for others the increase of the actual time is smaller than the increase of the planned time. This again shows that there are significant differences between the operators. The mean squared error on the training data for the "combined" model consisting of the single models for all operators is 9.215.
In Figure 4.33 the coefficients of the linear models are sorted with respect to the slopes and we can see that there seems to be an anticorrelation between the slopes and the intercepts of the models for the different operators, i.e. the higher the slope of the model, the smaller the intercept. Figure 4.34 and Figure 4.35 show the distribution of the values of the planned times and the actual times. We can see that most of the values of the planned times lie between 5 and 15 minutes and most of the values of the actual times lie between 0 and 10 minutes. In Figure 4.36 and Figure 4.37 which show the distribution of the planned respectively the

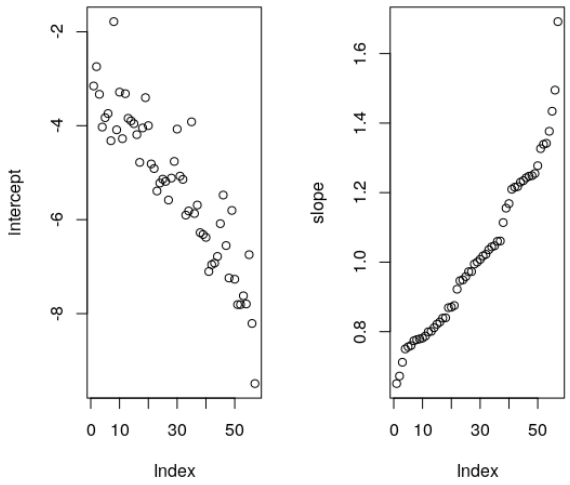

Figure 4.33: Intercepts and slopes of the different linear models

actual times of the orders done by four different operators we can see that the distribution of the planned times of all four operators is relatively similar to the distribution of all orders whereas the distribution of the actual times varies among the four operators. However, the range within which most values of the actual time can be found is similar for all four operators. As in the distribution of the actual times of all the orders, most of the values lie between 0 and 10 minutes. The observations we can make for those four operators are also true for most of the other operators. This means that for all the operators most of the data points lie in the same area of the graph but their distribution within this area is different. Therefore, the regressions lines for the different operators lie within a similar area of the graphs but their slopes are different. Lines with higher slopes intercept with the y-axis of the coordinate system at a lower position which leads to the anticorrelation.

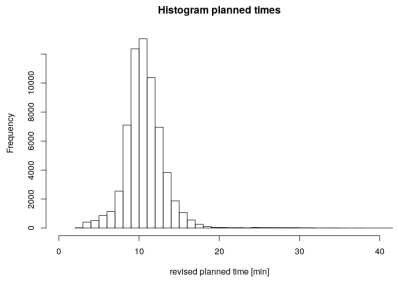

Figure 4.34: Histogram of the planned times

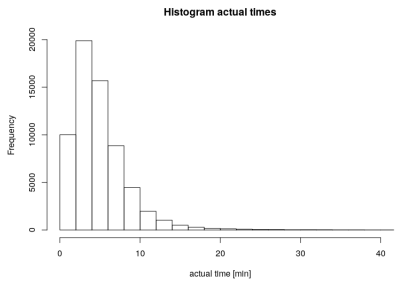

Figure 4.35: Histogram of the actual times

If we look at the plots in Figure 4.32, we can see that for all the 4 different operators shown there are some outliers, i.e. data points that lie far from most other points. The least squares estimation used by the `lm` function is very vulnerable to those outliers because the errors are

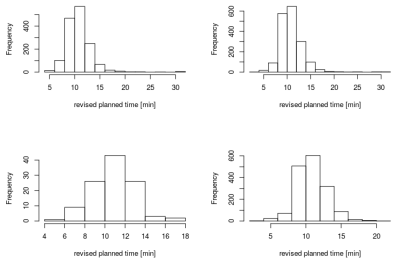

Figure 4.36: Histogram of the planned times

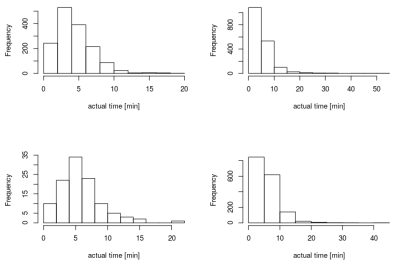

Figure 4.37: Histogram of the actual times

squared. To weaken the influence of those outliers, robust regression methods (see Chapter 3.1.7) can be used. As before, we use the data for operators who have at least 100 observations. For each operator we fit a linear model using the `rlm` function in R which uses the Huber M-estimator to fit the linear model. If we combine the different models to one model as we did before and compute the MSE on the training data, we get a value of 8.422. This is slightly higher than the one for the LS linear model. The reason is that the robust estimation gives less weight to the outliers and therefore lies further apart from these points which leads to high error terms for them. Figure 4.38 shows the differences between the coefficients for the linear models fitted by LS regression (black) and the models fitted using robust regression (red) ordered by the slopes of the linear models. The points of the corresponding models for one operator are always above each other. We can see that for some operators the differences between the coefficients using the different methods are very small whereas for other operators the points lie further apart. In general, the slopes for the models fitted by robust regression are lower than those for the models fitted by LS regression which is again caused by the lower weights of outliers.

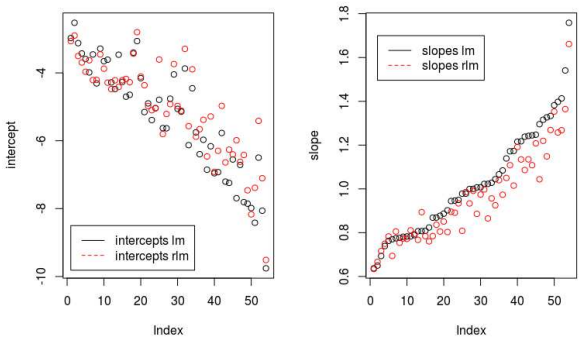

Figure 4.38: Comparison intercepts and slopes lm and rlm

Another way to deal with outliers is to exclude those data points from the data that is used to create the model. To identify outliers we can choose two different approaches: We can eliminate points which are far away from the regression line by looking at the studentised

residuals or we can eliminate those points that have a high impact on the estimation of the coefficients, i.e. those points that "pull" the line in a certain direction. Usually those points are also outliers. The impact can be measured by computing the leverage of the data points with respect to the linear model.
We exclude all the data points where the absolute value of studentised residual is larger than 4 times the mean value of the absolute values of studentised residuals. How many data points we exclude, depends on their distribution. For most models the number of points that are eliminated lies between 5 and 20. When using the leverages, we proceed in a similar way: we exclude all data points where the leverage is more than 4 times larger than the mean of the leverages. The two methods eliminate approximately the same number of observations but the data points excluded are in general not the same. For example, in the model for the first operator using the leverages, the 8^{th} observation (planned: 29.88, actual: 16) is excluded whereas using the studentised residuals, it is kept. For this pair of values the planned time is relatively high and therefore on the x-axis it lies relatively far away from the other points and has thus a high impact on the estimation of the parameters. For the 6^{th} data point (planned: 18.16, actual: 18) we have the reverse case: It is kept using the leverages and excluded using the studentised residuals. For this point the planned time lies in the normal range but the difference between the planned and actual time is smaller than for most other values and therefore the point lies far away from the regression line.

4.2.3.3 Comparison of the Models

We can now compare the coefficients we get for the models by using the different methods. Figure 4.39 shows the estimated values for some of the coefficients using the 4 different methods described. The first four columns contain the values for the slopes, the other four columns the values for the intercepts of the linear models. We can see that - similarly to the models fitted by robust regression (second column) - the slopes for the models fitted using the data without the data points with high studentised residuals (forth column) we get lower slopes than for the models fitted using all the data. In contrast, the coefficients for the models fitted for the data set where observations with high leverage are excluded are sometimes higher, sometimes lower than for the models where the whole data set was used. The MSEs on the training data for the two new models are 9.35 (without points with high residuals) and 9.26 (without points with high leverage). Comparing the MSEs on the training data does not give meaningful information about the quality of the models because we do not know how well they work on new data. Therefore we apply the models on a test data set to compare the qualities of the predictions. In a new dataset we are likely to find data for operators for which we have not created a model. To make predictions for these operators, we can use the models that have been created using the same methods but not separately for each operator but for all operators at once. As a test dataset we use the observations from the 19^{th} of March until the 24^{th} of March. This dataset contains about 8000 observations. For two of the operators

	slopes	slopes_robust	slopes_w_outl_lev	slopes_w_outl_studres	intercepts	intercepts_robust	intercepts_w_outl_lev	intercepts_w_outl_studres
1	0.80664702645381	0.76719621427922	0.836888277111414	0.753250370542063	-4.47518460911964	-4.21746487081308	-4.8036853298262	-4.0505462335025
2	1.04413622588777	0.924992294846773	1.04413622588777	0.977047325664082	-4.45334412180258	-3.89422246607113	-4.45334412180258	-4.18279602327486
3	0.877438298457831	0.805314782456312	0.877438298457831	0.811062090036966	-3.06173137461549	-2.79994332502764	-3.06173137461549	-2.78843843553633
4	1.0069513570841	0.885989915966653	1.06417575942013	0.883059856091091	-4.04506398156824	-3.73958626855481	-4.64238106986602	-3.71279184006532
5	1.023047749996646	0.864938804637795	0.996306417943754	0.899764872307995	-3.87001069749226	-3.29207265424439	-3.64430949469148	-3.37349597004018
6	0.739015782244542	0.749315890795051	0.859873016807171	0.741793551960632	-3.42774678399614	-3.69402319205367	-4.68592351509043	-3.58370300847155
7	0.944660143530286	0.895540802347465	0.961904935019321	0.946339871896594	-4.8989987454585	-4.97745083779682	-5.0682207280077	-5.29842735000251
8	1.171787680904864	1.108190978481020	1.229344366275190	1.105257888926140	-6.85349886973242	-6.46176514698591	-7.49003779300386	-6.36888570880595
9	1.173145496170940	1.015038220573430	1.135016412796060	1.000783322511489	-6.16189771081550	-5.28428312354306	-5.77050179406759	-5.05052516830640
10	0.769931469038618	0.693878154785651	0.799319702415507	0.707091897684880	-3.98584804957137	-3.62785020737543	-4.30170147999293	-3.66316076477171
11	0.868622682531188	0.784643496228786	0.863946348197006	0.812366055961479	-4.64534502320998	-4.27398180660664	-4.5877273824757	-4.42738272595487
12	0.649843438510718	0.666641297438431	0.726001207463370	0.677994316399857	-2.52351634509029	-2.8988612166918	-3.32640527450313	-2.95164816987123
13	0.762416616674318	0.782889522096568	0.874685022170862	0.814535455239236	-3.59126242206216	-3.9660422707969	-4.76851452299274	-4.23200697691028
14	1.084302556061920	0.973446975937501	1.152388535886720	0.983307195496500	-6.39626782786571	-5.65869532964437	-7.12567140368509	-5.64391934677526
15	1.027934533478523	0.956294646616779	1.051913686590280	0.951943569414099	-6.12860102598378	-5.56773037008093	-6.39324014203372	-5.49642517244499
16	0.774460136930156	0.805376900250310	0.907673388996502	0.803920525002556	-3.45576272376217	-4.21135568733471	-4.85559583990153	-4.04054565343648
17	0.978254610850250	0.985544357741452	0.967354828050752	0.980973105047601	-5.62618286618758	-5.80075936528498	-5.51564303195046	-5.76842317063541
18	0.781519239490874	0.773593992717069	0.846794851177165	0.750348772264107	-3.28540411776426	-3.45617521968212	-3.98406211911197	-3.11827139629013
19	0.978181650946581	0.808439656967830	1.031069570421670	0.816719582799804	-4.79169242384658	-3.60589807404097	-5.34152584607142	-3.47046287774725
20	0.776668853965921	0.753924271239195	0.818279199207683	0.751416838150841	-4.30948693810351	-4.20667626156321	-4.75776945548293	-4.14713553132906
21	1.295414700985430	1.043883739760590	1.200210874829210	1.067168409517800	-7.69324309895603	-5.98473238525185	-6.69261890610765	-5.96917947263163
22	0.886818488218039	0.851034237071312	0.845683676911904	0.879547802192308	-4.15517803666369	-4.10900330733902	-3.74983868071128	-4.3117211515871
23	1.241994369597300	1.134875626094320	1.261773311814820	1.078161585661110	-7.20418515623882	-6.63809568498291	-7.43030055890375	-5.96296963156516
24	1.022687695072130	0.997492966925500	1.063597590822480	0.987872454525651	-5.11424511394003	-5.11324503371354	-5.55506072056516	-4.92782869261066

Figure 4.39: Estimated coefficients for the different methods

in this dataset we have not created a model and will therefore use the model for all operators. We get the following values of the MSE for the different model types:

- Linear model on the whole dataset: 8.04
- Robust linear regression: 7.83
- Linear model on dataset without points with high leverage: 8.30
- Linear model on dataset without points with high studentised residuals: 7.76

Thus, the predictions using the last model type are closest to the real values but the difference to the second type is very small.
All models have negative coefficients for the intercepts which leads to negative predictions for the actual time if the value of the planned time is very small. As the actual time cannot be negative, we can set these values to zero. If we do this and then compute the MSE for the corrected predictions, we get:

- Linear model on the whole dataset: 6.56
- Robust linear regression: 6.63
- Linear model on dataset without points with high leverage: 6.54
- Linear model on dataset without points with high studentised residuals: 6.60

Thus, we can significantly reduce the MSE and the values for all models are very similar now.

4.2.3.4 Analysis of other influencing factors

We now know that the operators have a significant influence on the relationship between the planned times and the actual times and we can analyse if there are other influencing factors. As for the data from the redPILOT database, we can examine if there are time variables that have an influence on the performance.
First, we have a look at the hour of the day. We make boxplots for the different operators for the relative differences between the planned time and the actual time at the different hours of the day. Figure 4.40 shows plots for four different operators which indicate that there are some differences for the different hours. To see if there are really significant differences between the

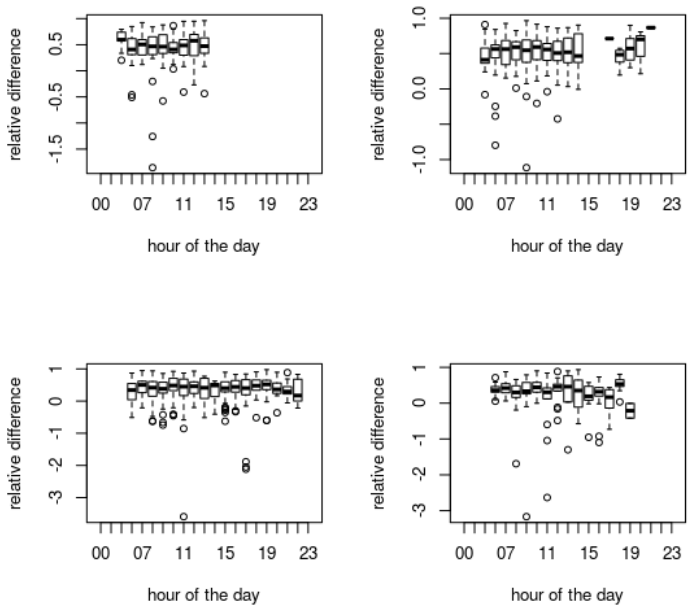

Figure 4.40: Relative differences planned time - actual time over hour of the day

hours of the day, we can include dummy variables for the different hours in our linear models. We take one hour as reference and test if the other hours differ significantly from it. If we use the linear model function for the formula `actual time` $\sim$ `planned time + hour`, we get models where the coefficient for each hour gives information how much the intercept for this hour differs from the intercept of the reference hour. For all operators these coefficients have high p-values and the adjusted R^2 only increases very little or even decreases in comparison with the models that only includes the planned times (formula: `actual time` $\sim$ `planned time`). We can also include an additional interaction term `planned time*hour` to see if the coefficient for the planned time depends on the hour of the day. Similarly as before, we get mostly high p-values. Thus, we can conclude that the influence of the hour of the day does not seem to be significant.
Next, we can analyse the differences in performance for different days of the week. We proceed

as we did before and first have a look a the boxplots of the differences between planned and actual times over the day of the week for the different operators. In Figure 4.41, which shows

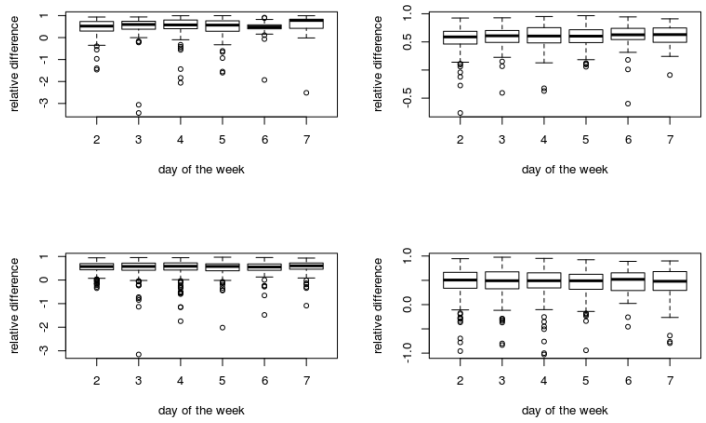

Figure 4.41: Relative differences planned time - actual time over day of week

the plots for four operators, we can see that the differences for the different days of the week seem to be very small. This impression is confirmed by the summaries for the linear models created with the linear model function for the formula `actual time` $\sim$ `planned time + day of week`. Almost all p-values are large which means that the variable does not have a significant influence.

4.2.3.5 Conclusion

From the results of the analyses performed above, we can conclude that the performance of the operators are significantly different. The differences between the planned times and the actual times that are needed to execute an order differ depending on the operator who is responsible for the execution. We thus know that, to make a good prediction for the throughput of the order picking process step, we have to consider the differences between the individual operators. Depending on which operators are working, it will be higher or lower. In contrast, the hour of the day and the day of the week do not seem to have a significant influence on the performance.

4.2.4 Analysis of Order Information Data

4.2.4.1 Analysis of Articles

In addition to the data analysed in the previous chapter, more detailed information about the orders is available, for example we know of which articles the order consists. We can analyse if there is any difference between the different articles concerning the difference between the actual time and the planned time, i.e. if there are some articles for which it is more likely that

the operator will not fulfil the order within the planned time if they are included in the order. For our analysis we use a dataset where the information about the planned and actual times has been combined with the order information. Each row contains the information about one article (article number and description), the picking times for that specific article as well as the planned and actual times of the order to which the article belongs and the operator that is executing this order. The dataset contains the data from the 2^{nd} of February until the 9^{th} of February and from the 26^{th} of February until the 17^{th} of March which includes about 235000 observations. As we did for the analysis of the planned times data before, we will exclude rows where the actual time is zero which reduces the dataset by around 10000 entries.
There are more than 1800 different articles and for some of them we have only very few observations. Similarly as we did for the operators, we will exclude those articles from our analysis because from those few observations we cannot make any general conclusions. We only analyse articles for which we have at least 30 observations which leads to a dataset with around 220000 rows containing information for 1338 different articles.
As there are so many different articles, we cannot use a boxplot anymore to get a first idea of the differences between them. But as we already did for the operators, we can create a linear model with dummy coding for the relative difference between planned and actual times over the article number. The output of the summary of that model gives us information if there is a significant difference between the first article which is taken as a reference category and the rest of the articles. We can see that some of the coefficients do have low p-values and thus this article is significantly different from the reference article but many of the articles have high p-values and are therefore similar to the reference. If we use the average relative difference as a reference instead of one specific article, we get relatively high p-values (above 0.05) for all the articles which means that for none of the articles the average relative difference between planned time and actual time seems to be significantly different from the overall average relative difference. Although the variations of the relative differences between planned and actual times do not seem to be that significant, we can still try to make some kind of categorisation of the articles. It does not make sense to treat each article separately but we can try to find out if there are some articles where the actual times are frequently higher than the planned times and others where the order is often fulfilled a lot faster than scheduled. To get an overview of the values of the relative differences between planned and actual time, we can have a look at the histogram of the values. We can see that most of the values are positive, i.e. the actual time is shorter than the planned time and that most of the values lie below 0.8. So, we can define three categories: 1: "relative difference < 0"; 2: "$0 <=$ relative difference < 0.8"; 3: "relative difference $>= 0.8$"; To find out if there are any articles that fall into a certain category more frequently than the others, we first assign one of the three categories to each row and then we analyse how many observations for a specific article fall into a certain category. As we have a different number of observations for each article, we analyse the relative frequencies for the categories, i.e. how many percent of the orders

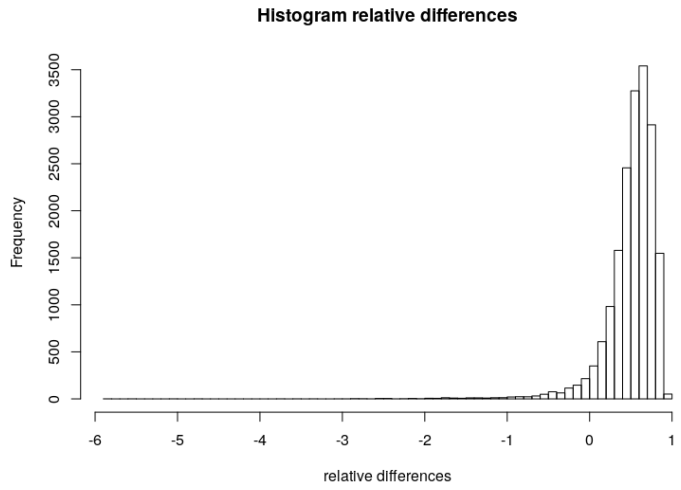

Figure 4.42: Relative differences planned time - actual time

containing the article fall into a specific category.
At first we have a look at the histogram for the relative frequencies of the first category (see Figure 4.43). This histogram shows for how many articles the relative frequency of orders with category 1 lies within a certain range. We can see that for most of the articles less than 20% of the observations have a higher planned time than actual time. Therefore we will consider articles having a relative frequency for category 1 which is higher than 0.2 as "slow" articles.

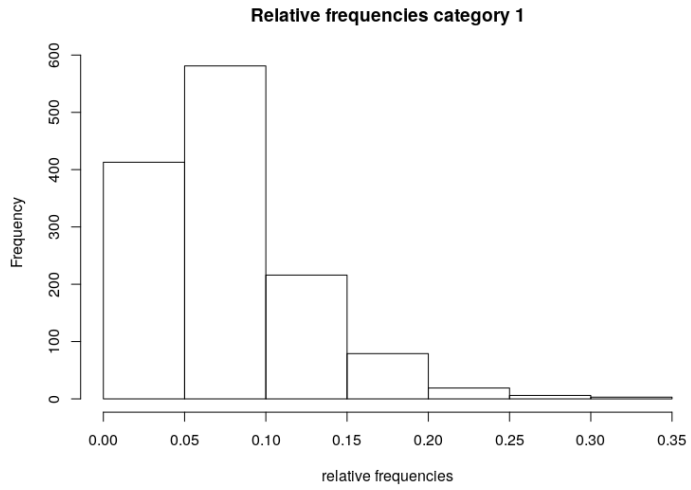

Figure 4.43: Relative frequencies category 1

Next, we can do the same for category 3. In Figure 4.44 we can observe that usually less than 10% of the relative differences between planned and actual time are higher than 0.8. So, we will categorise articles where the relative frequency is more than 0.1 as "fast". All articles that are neither categorised as "slow" nor as "fast" are considered as "normal". Figure 4.45 shows a boxplot for the relative differences between planned and actual times for the different

categories of articles. It seems as if there are significant differences between the observations for articles of the different categories.

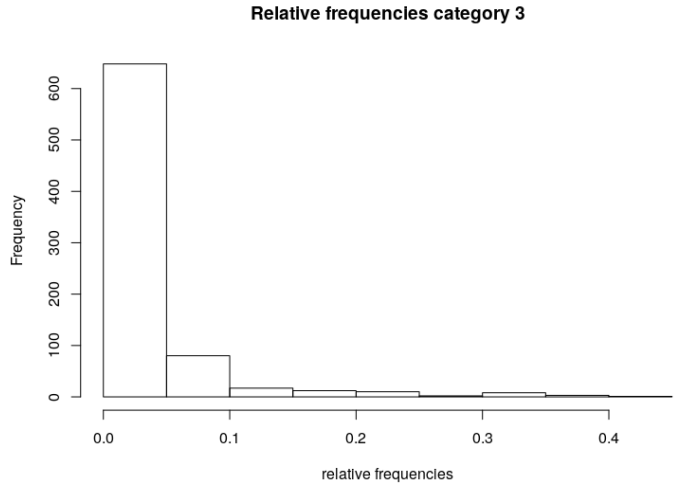

Figure 4.44: Relative frequencies category 3

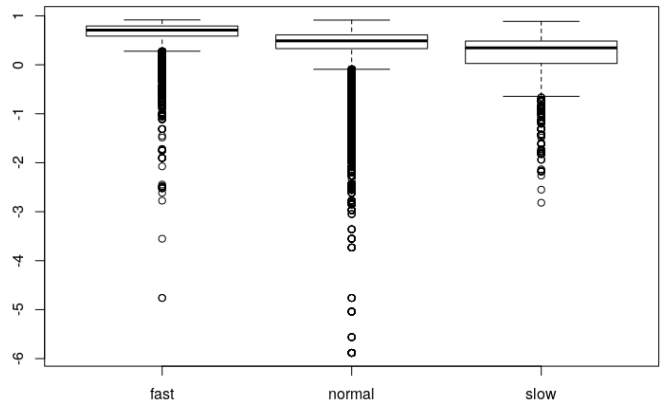

Figure 4.45: Boxplot relative differences over category article

For each order we can now count how many "slow" and "fast" articles it contains and analyse if those two numbers have an influence on the relationship between the planned and the actual time. We can create a linear model for the formula `actual time` $\sim$ `planned time + number slow + number fast` on the full dataset (without observations where the actual time is zero) to see if the number of "slow" respectively "fast" articles has an impact on the actual time. All the coefficients have low p-values, the residual standard error is 3.29 which is slightly lower than for the linear model only including the planned times (3.33) and the adjusted R^2 is 0.2897 which is slightly higher than for the model without those coefficients (0.2723). That indicates that including the number of "fast" and "slow" articles can slightly

improve the prediction.

4.2.4.2 Number of Articles per Order

Apart from the articles themselves, another factor that has an influence on the relationship between the planned times and the actual times could be the number of articles per order. If an order contains more articles, the planned time itself increases because it is the sum of all necessary steps to fulfil an order. Additionally, if more different articles have to be picked, the operator is also more likely to encounter some problems and therefore not to be able to fulfil the order within the planned time. It is thus sensible to include the number of different articles per order into our model. We can again compare the model including the new coefficient for the number of different articles to the model where the planned time is the only predictor using the data for all operators without entries where the actual time is zero. The p-value for the additional coefficient is small, the residual standard error is 3.004 and the adjusted R^2 is 0.4079. Thus, the quality of the prediction can be clearly improved.
If we additionally add the two coefficients for the categories of the articles (formula for the linear model: `actual time` $\sim$ `planned time + number distinct articles + number slow + number fast`), we can achieve a small further improvement: The residual standard error is 2.984 and the adjusted R^2 is 0.4159.

4.2.4.3 Separate Model for Each Operator

As we have seen in chapter 4.2.3, we can improve the quality of the prediction by creating a separate model for each operator. This time we will add the coefficients for the numbers of distinct articles and articles categorised as "slow" respectively "fast". Again we will only create models for operators where at least 100 observations are available and remove entries where the actual time is zero. For each operator we can create three different models to analyse the impact of the different input parameters:

- One model where the planned time is the only predictor (formula: `actual time` $\sim$ `planned time`)
- One model additionally including the number of distinct articles per order (formula: `actual time` $\sim$ `planned time + number distinct articles`)
- One model including the number of distinct articles as well as the number of "slow" respectively "fast" articles (formula:`actual time` $\sim$ `planned time + number distinct articles + number slow + number fast`)

If we look on the output of the summary function for the models created using the linear model function in R, we can see that the coefficient for the number of distinct articles always has a low p-value. Thus, this parameter has a significant influence on the actual time. For the two

coefficients for the number of "slow" and "fast" articles an order contains we sometimes get high p-values. This means that it cannot be shown that this input parameter influences the output. This can either mean that there is no relationship between the output variable and this parameter or that we do not have enough data to evaluate the influence (i.e. the number of articles of that category is mostly zero). Figure 4.46 shows a comparison of the values of the adjusted R^2 and the residual standard error for the three different models for some of the operators. We can see that the adjusted R^2 strongly increases and the RSE strongly decreases when the coefficient for the number of different articles is added. When the coefficients for the numbers of "slow" and "fast" articles are added, the values of the adjusted R^2 slightly increase and those of the RSE slightly decrease. This confirms the impression we received by looking at the models for the whole dataset at once.

	only planned times R^2	+number distinct articles R^2	+number slow/fast R^2	only planned times RSE	+number distinct articles RSE	+number slow/fast RSE
1036694903	0.4054084	0.5693042	0.5976258	1.817573	1.546920	1.495194
1036694907	0.3509873	0.5431012	0.5501448	3.563897	2.990257	2.967119
1036694909	0.3279457	0.5019781	0.5082011	3.715404	3.198365	3.178320
1036694911	0.4839835	0.6869683	0.7040047	1.737572	1.353334	1.315992
1036694913	0.2893348	0.4457064	0.4821353	3.087222	2.726499	2.635382
1036694920	0.4056749	0.6154514	0.6277366	2.902121	2.334421	2.296829
1036694921	0.2297784	0.3607805	0.3590597	4.547600	4.142852	4.148425
1036694923	0.1756186	0.3037906	0.3084710	3.543742	3.256626	3.245661
1036694924	0.2921128	0.3665413	0.3847326	3.352198	3.171077	3.125213
1036694925	0.3309024	0.5627084	0.5933985	2.121034	1.714701	1.653436
1036694926	0.4451607	0.6360605	0.6511993	1.938358	1.569875	1.536877
1036694927	0.3028480	0.4700627	0.4713673	3.601116	3.139681	3.135814
1036694928	0.3904925	0.5402311	0.5653517	2.894142	2.513623	2.443990
1036694929	0.3076736	0.4798773	0.4957107	2.437969	2.113127	2.080715
1036694931	0.3823836	0.6300154	0.6660349	2.225660	1.722627	1.636628
1036694932	0.2758687	0.3518329	0.3617955	4.333337	4.099749	4.068120
1036694933	0.4579307	0.7261888	0.7561778	1.798032	1.277896	1.205887
1036694936	0.2203057	0.2783721	0.2794102	4.803721	4.621387	4.618061
1036694938	0.3092882	0.5301528	0.5373766	2.707747	2.233254	2.216019
1036694947	0.2319597	0.3022816	0.3285079	3.902332	3.719395	3.648822

Figure 4.46: Comparison of the adjusted R^2 and RSE for linear models including different parameters

As we did in Chapter 4.2.3.3, we can use the data from the 19th of March until the 24th as test data to compare the quality of the different models. For the first model the mean squared error is 8.57, for the second it is 5.49 and for the third 5.31. Again, we receive prediction below zero for small planned times. If we set those values to zero, the MSEs are 6.4, 5.1 and 5.01.

4.2.4.4 Conclusion

From the analyses performed before we can gather that the structure of the orders has a significant influence on the difference between the planned time and the actual time. Especially the number of distinct articles has a strong impact. By including this parameter into our model, we can clearly improve the quality of the prediction. There are also some differences between the articles themselves. We have discovered that for some articles the actual time is more often higher than the planned time than for most other articles. In contrast, if some other specific articles are included in an order, the operator is more likely to finish the order in a lot less time than planned. If we add the number of these articles contained in an order to our model, it seems that we can achieve a small further improvement. However, this improvement is not as clear as for the number of distinct articles. Therefore, some further analysis is necessary.

4.3 Model for estimating the actual times

4.3.1 Model selection

Our previous analyses have shown that there are some differences in the performance of different operators and thus, it makes sense to create a separate model for each operator. Furthermore, we have seen that the number of articles per order has a significant influence on the fulfilment of the planned time. The higher the number of articles per order, the higher is the actual time is compared to the planned time. Therefore it seems reasonable to include this input parameter into our model. Besides, there might also be a small effect that is caused by the articles that are contained in an order. We measure this effect by counting the number of "slow" and "fast" articles as described in Chapter 4.2.4.1. It is not clear if these two parameters have a significant influence on the output variable (the actual time) and should be added to the model. To decide if we should include those two variables, we can use cross-validation.
For the selection of the model we can now use a dataset containing observations from the 2nd of February to the 8th of February and from the 26th of February to the 21st of April , which means we have information about around 65700 orders. To assure that each of the folds used in the cross-validation process contains a representative sample of the data for training as well as for testing, we shuffle the data. We compare six different models:

1. A linear model fit by least squares approximation only for the actual time over the planned time (formula: `lm(actual time ~ planned time)`)

2. A linear model fit by least squares approximation for the actual time over the planned time and the number of distinct articles (formula: `lm(actual time ~ planned time + number distinct articles)`)

3. A linear model fit by least squares approximation for the actual time over the planned

time, the number of distinct articles and the number of "slow" and "fast" articles (formula: `lm(actual time ~ planned time + number distinct articles + number slow + number fast)`)

4. A linear model fit by robust regression (Huber M estimator) only for the actual time over the planned time (formula: `rlm(actual time ~ planned time)`)

5. A linear model fit by robust regression (Huber M estimator) for the actual time over the planned time and the number of distinct articles (formula: `rlm(actual time ~ planned time + number distinct articles)`)

6. A linear model fit by robust regression (Huber M estimator) for the actual time over the planned time, the number of distinct articles and the number of "slow" and "fast" articles (formula: `rlm(actual time ~ planned time + number distinct articles + number slow + number fast)`)

During the training phase we eliminate observations where the actual time is zero to avoid distortions. In the part of the data that is used for testing we of course keep these values because if we apply the model to make predictions on new data, we do not know which of the observations will have an actual time of zero. Furthermore, in each fold we only create models for operators where we have at least 100 observations. To predict the actual time for operators in the test data where we do not have a specific model, we use the model that is created using the same formula on the data for all operators in that fold. If the predicted value for the actual time is below zero, we set it to zero. To select the most appropriate model from our candidates, we compare their mean square errors. As it is recommended by Kohavi [7], we use ten-fold cross-validation for selecting the model. This number of folds provides a good balance between the bias and the variance of the results of the cross-validation.
To select the best model, we compare the averages of the mean squared errors derived from the predictions on the test data in each of the folds. We get the following results:

1. `lm(actual time ~ planned time)`: mean(MSE) = 7.25

2. `lm(actual time ~ planned time + number distinct articles)`: mean(MSE) = 5.78

3. `lm(actual time ~ planned time + number distinct articles + number slow + number fast)`: mean(MSE) = 5.71

4. `rlm(actual time ~ planned time)`: mean(MSE) = 7.38

5. `rlm(actual time ~ planned time + number distinct articles)`: mean(MSE) = 5.91

6. `rlm(actual time` $\sim$ `planned time + number distinct articles + number slow + number fast)`; mean(MSE) $= 5.82$

We can see that the third model has the lowest average MSE but the results are similar for some of the models. The MSE for the second model, which is simpler than the third one, is just slightly higher whereas the MSE for the simplest model, the one only including the planned time, is much higher. The difference between the results for the second and the third model are so small that we cannot really say that the second model is the better one. We will therefore apply the principle of Ockham's razor and choose the more simple one of the two.
We will thus create a linear model for each operator using the revised planned time and the number of distinct articles in the order as input parameters. Some of the operators are employed at a temporary employment agency. We do not have any influence on those operators because the plans for them are made by their employer. Therefore we will not create models for those operators. As in the cross-validation process, we will only create separate models for operators where we have at least 100 observations. Additionally we will create a model using the same input parameters for the whole dataset which can be used to make predictions for operators where we have not enough observations for a separate model. For the estimation of the coefficients for the two variables we will use least squares approximation.

4.3.2 Model creation and evaluation

The dataset used for the creation of the models contains observations from the 2nd of February to the 8th of February and from the 26th of February to the 28th of April. It has about 76300 entries. Before we create our models, we split the available data into one dataset for training and one for testing. In total we have data for 55 days, the data for 28 days is used for training and the data for the remaining 27 will later be used for testing. In the training dataset we remove the observations for operators from the temporary employment agency, in the test dataset we keep them. This gives us a training dataset containing about 31600 entries and a test dataset containing around 37000 entries.
As we did before, we do not use observations where the actual time is zero for training. After removing the zero values, about 21000 observations for the actual model creation process are left. Among the 45 internal operators there are 34 where we have more than 100 observations and for who we thus create a separate model. In all those 34 models for one single operator as well as in the model for the whole dataset, the coefficient for the number of distinct articles has a p-value below (mostly far below) 0.05 which means that the significance level for which we can reject the null hypothesis that the input parameter does not have an impact on the output is very low. The p-values for the planned time are higher than 0.05 in five of the models which means that for those operators the influence of the planned time is not that significant. The residual standard errors lie between 1 and 6.5. Our estimates for the actual times are thus likely to differ from the real actual times by a few minutes. Most of the values

for the adjusted R^2 lie between 0.35 and 0.65 which means that between 35 and 65% of the variation in the output variable can be explained by the model.
One of the assumptions made for linear models is that the average noise is zero. In the 5^{th} column of Figure 4.48 the mean of the residuals is given. It is always very small which means that this assumption holds for all our models. Another assumption for the classical linear model is a normal distribution of the residuals. This can be verified by looking at quantile-quantile-plots. Figure 4.47 shows the Q-Q-plots for four different operators. We can see that the residuals are close to the line which indicates normally distributed values within a certain range but for higher values the points get further away from the line. This means that the assumption of normally distributed residuals only holds within a certain range. To check if the model assumption of homoscedasticity holds for the different linear models, we can apply the Breusch-Pagan (BP) test. In this test the null hypothesis that all error terms have the same variance is tested against the alternative hypothesis that this is not the case and we thus have heteroscedastic noise. For 12 of the models the p-value of this test is greater than 0.05 which means that we do not reject the null hypothesis at a significance level of 0.05 and therefore homoscedasticity can be assumed. In the other models we can find heteroscedastic noise. Furthermore we can apply the Durbin Watson (DW) test to check if the noise is uncorrelated. The null hypothesis of this test is that the autocorrelation of the noise is 0. At a significance level of 0.05 this hypothesis holds for 19 of our 34 models. In the other models the null hypothesis can be rejected at this significance level which means that there seems to be some correlation in the noise. We can also check if there is a collinearity between the two input parameters. As explained in Chapter 3.1.6, this can be done by calculating the variance inflation factor (VIF). The variance inflation factor for all our models is between 1.5 and 2.5 which means that there does not seem to be a serious collinearity.

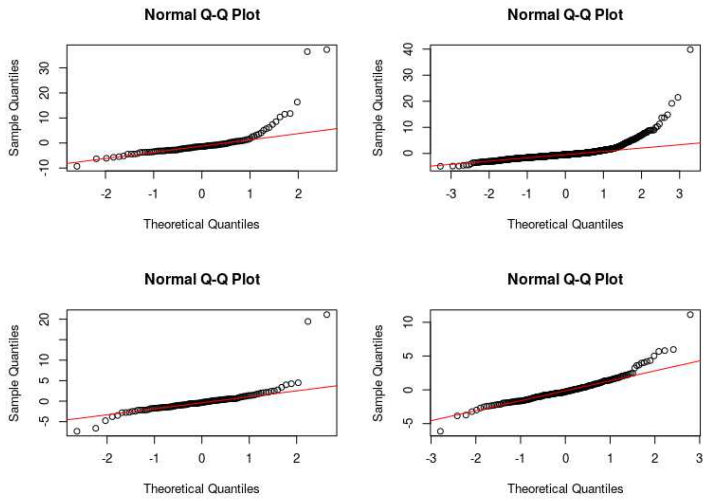

Figure 4.47: Q-Q-Plot of the residuals of the linear models for four operators

Figure 4.48 shows the values of the p-values, the RSE, the adjusted R^2, the mean value of the

residuals, the p-values of the BP-test and the DW-test as well as the variance inflation factor for some of the models.

	p-value planned min	p-value dist. articles	RSE	adjusted R^2	mean residuals	p-value BP-test	p-value DW-test	VIF
1036694903	1.138394e-24	4.245784e-66	1.262571	0.6165145	1.085426e-16	1.274379e-08	7.950752e-01	2.086898
1036694907	9.641030e-02	1.507680e-28	3.193103	0.4463768	8.978936e-18	5.436756e-02	6.168608e-05	2.216503
1036694909	8.379206e-01	1.544634e-12	5.453584	0.3600834	2.936979e-16	2.564917e-04	3.614809e-04	2.103650
1036694911	2.031501e-19	1.139888e-69	1.260745	0.6518215	-1.120504e-16	8.266347e-05	4.162515e-01	2.019706
1036694913	2.576252e-03	2.917860e-37	2.925581	0.3717396	-3.678004e-18	3.716854e-02	9.421698e-01	1.772605
1036694918	3.448302e-03	1.298923e-06	2.259355	0.5006149	-9.652799e-17	1.274650e-01	3.237194e-02	2.025407
1036694920	2.208423e-13	3.316218e-76	2.060712	0.6315894	-1.649505e-16	4.760026e-15	3.977491e-02	1.747923
1036694921	1.011237e-01	4.162219e-04	5.412305	0.2947191	4.914714e-17	4.245041e-01	9.944068e-01	2.238715
1036694923	2.697240e-03	3.940013e-28	2.705011	0.2744804	-8.633079e-17	9.683424e-01	1.694613e-03	1.864239
1036694924	1.257672e-01	2.940052e-22	3.335447	0.2651432	7.450741e-17	2.296900e-01	9.034414e-01	1.833705
1036694925	7.233639e-12	2.025363e-74	1.679116	0.5478201	5.745444e-17	8.259443e-06	3.267868e-02	1.806291
1036694926	5.158679e-22	1.016781e-96	1.212087	0.7327678	-7.448867e-19	2.800188e-07	1.538147e-01	1.968058
1036694927	1.076094e-06	4.224400e-65	2.334543	0.5299473	-4.009834e-17	7.426074e-07	2.920543e-01	1.849774
1036694928	1.488319e-05	2.499587e-30	2.462031	0.5029301	3.666493e-17	1.502148e-01	2.347946e-02	1.988622
1036694929	1.065829e-08	9.277735e-53	2.141290	0.5056061	6.426259e-17	1.718894e-05	9.372877e-01	1.842051
1036694931	2.903307e-03	2.109536e-18	1.722627	0.6300154	-2.547516e-17	1.282783e-02	3.828853e-02	1.724300
1036694932	2.701616e-16	2.821002e-20	3.354852	0.3913556	7.858347e-17	1.073838e-06	8.622389e-01	1.702497
1036694933	1.015382e-27	4.609342e-107	1.416383	0.6276747	-3.953490e-17	1.752456e-07	8.635132e-01	1.744418
1036694936	8.406823e-08	6.511046e-05	4.813331	0.2123435	-1.208048e-17	3.255002e-03	3.081342e-01	1.908186
1036694937	2.452585e-06	4.733293e-48	1.893769	0.5899966	2.198573e-17	9.765482e-16	2.795019e-01	1.858008
1036694938	1.494699e-03	2.799865e-61	2.360797	0.4474290	6.864182e-17	4.175744e-01	4.074881e-03	1.724366
1036694947	4.906958e-05	3.875824e-04	5.328584	0.1492649	3.803234e-16	8.565272e-01	1.297402e-01	1.952175

Figure 4.48: Evaluation parameters of some of the models for estimating the actual times

To see how far the predicted values for the actual time lie from the real values, we can apply the model to the test dataset. As we did for the model selection, we set predicted values that are below zero to zero. Figure 4.3.2 shows the difference between the real values for the actual times in the training dataset (left side) as well as the test dataset (right side) and the predicted values for the corresponding dataset after setting the negative predictions to zero for a sample of 500 observations. As it can be expected, the residuals on the training dataset are smaller. We can see that most of the residuals lie close to zero but, especially on the test dataset, we also have some outliers, in particular on the positive side which means that some of the real values are a lot higher than our predicted values. This can also be seen in the histogram for the residuals which is shown in Figure 4.50.

Figure 4.49: Differences between predicted and real values of the actual times

4.3.3 Improvement of the parameters

The dependent variable in our model, i.e. the actual time needed for an order can only be greater than or equal to zero. In the chapter before we have tried to predict this variable

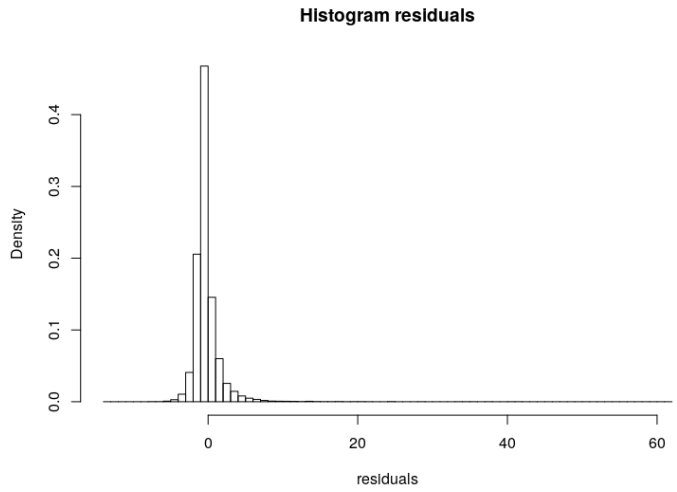

Figure 4.50: Histogram of the residuals

by creating a linear model with two predictors, the planned time and the number of distinct articles, for each operator. The least squares approximation which we have used to estimate the coefficients (intercept, coeff_1, coeff_2) has lead to negative estimates for the intercept for all models. As a consequence we receive negative predictions for the actual time if the planned time and the number of distinct articles are small. This happens in about 11.9% of the predictions in out test dataset. So far we have set those values to zero after the prediction. We can try to achieve a better model by including the restriction to allow only positive values for the actual time in our model. Instead of minimising

$$\sum_{i=1}^{n}(\text{actual\_time}_i - (\text{intercept} + \text{planned\_time}_i * \text{coeff}_1 + \text{distinct\_articles}_i * \text{coeff}_2))^2, \quad (4.1)$$

we minimise the cost function

$$\sum_{i=1}^{n}(\text{actual\_time}_i - \max(0, \text{intercept} + \text{planned\_time}_i * \text{coeff}_1 + \text{distinct\_articles}_i * \text{coeff}_2))^2 \quad (4.2)$$

where actual_time_i, planned_time_i and $\text{distinct\_articles}_i$ are the planned and the actual time respectively the number of distinct articles of a certain order.

This means that we already exclude negative values for predicted values of the actual time during the estimation of the parameters. We can take the coefficients we estimated before as starting points and try if we can improve the prediction of the actual times by varying them a little bit. We can calculate the cost function of the original estimated and the changes estimate to see which one is better. If the variation results in a reduction of the value of the cost function, we adapt the values of the coefficients. We try different intercepts in an interval of [original estimate - 3; original estimate +3] with a step size of 0.2 and different values for the coefficients of the slope of the planned time and the number of distinct articles in a range

of [original estimate - 0.3; original estimate + 0.3] with a step size of 0.01. As it can be seen in Figure 4.51 where the varied coefficients are given in the columns 2 to 4 and the original coefficients in the columns 5 to 7, the coefficients have changed for almost all operators which means that we could reduce the value of the cost function by our variations. By using this

	operator	intercepts	slopes_pt	slopes_da	intercepts_lm	slopes_pt_lm	slopes_da_lm
1	1036694903	-1.8146013	0.34090518	0.2115985	-1.61460133	0.32090518	0.2115985
2	1036694905	-3.1629655	0.61961384	0.4281838	-2.56296547	0.60961384	0.3781838
3	1036694907	-0.9188334	0.17915978	0.5631762	0.08116657	0.13915978	0.5131762
4	1036694908	-1.0499167	0.01523986	0.7534959	-1.04991669	0.01523986	0.7534959
5	1036694909	-1.2029199	0.10512822	0.7479209	-1.00291990	0.08512822	0.7479209
6	1036694911	-1.6851342	0.31068248	0.2760317	-1.48513420	0.30068248	0.2660317
7	1036694913	-1.4427218	0.26968856	0.3571724	-1.24272182	0.24968856	0.3571724
8	1036694918	-1.7122237	0.34624108	0.3346451	-1.51222368	0.34624108	0.3146451
9	1036694920	-2.5194498	0.40864068	0.3942494	-2.51944979	0.41864068	0.3842494
10	1036694921	-3.9283862	0.55577082	0.4251106	-3.32838616	0.49577082	0.4351106
11	1036694923	-1.3414786	0.27674534	0.2551841	-0.74147863	0.22674534	0.2551841
12	1036694924	-1.0605546	0.20439179	0.3265782	-0.46055462	0.15439179	0.3265782
13	1036694925	-1.6031333	0.31494625	0.2671882	-1.00313332	0.26494625	0.2671882
14	1036694926	-2.0660848	0.33204372	0.3228404	-2.06608484	0.33204372	0.3228404
15	1036694927	-1.7858988	0.29378806	0.3848952	-1.58589878	0.27378806	0.3848952
16	1036694928	-2.4191289	0.40063452	0.3559342	-1.81912889	0.34063452	0.3659342
17	1036694929	-1.9996203	0.32206332	0.3279578	-1.99962031	0.32206332	0.3279578
18	1036694931	-1.6612845	0.27848567	0.3991249	-1.46128451	0.26848567	0.3891249
19	1036694932	-5.1498127	0.71054813	0.3071722	-4.54981273	0.66054813	0.3071722
20	1036694933	-1.9564589	0.32452749	0.2456370	-1.95645889	0.32452749	0.2456370

Figure 4.51: Variation of the coefficients

procedure we can reduce the mean squared error on the training data from 4.54 to 4.42. If we use the new coefficients to predict the actual times of the test dataset, we can reduce the MSE from 5.46 to 5.39. Thus, we can conclude that by applying our optimisation procedure, we can improve our predictions a bit. In Figure 4.3.3 we can see the residuals for the training data (left side) and the test data (right side). We can see that they seem to have become a little bit smaller compared to the residuals for the original coefficients.

Figure 4.52: Differences between predicted and real values of the actual times

4.4 Throughput prediction

We now want to use the information we gained to make a prediction about the performance during a certain shift so that we, e.g., know if the allocated operators can fulfil the demand. The linear model described above can be used to estimate the actual time for a picking order given the planned time, the number of distinct articles and the operator executing the order.

This model shall now be used to estimate the necessary working time of specific operators for some given orders.
If we make a very short-term prediction, the real orders are already available. The prediction is made at the beginning of a shift to estimate the time that is needed to fulfil the given orders. It shall give the shift manager the information if the available operators can finish the orders within their scheduled working time or if more or less time than planned is needed. Based on this information the shift manager can ask the employees to stay longer than planned or to go home earlier. If the shift manager can make these decisions already at the beginning of the shift and not only one or two hours before the end of the shift, he can inform the operators earlier which helps to prevent dissatisfaction among them. Additionally, if employees go home earlier, personnel costs can be saved.
If we want to predict the throughput of a shift more time in advance, the real orders are not available yet. Therefore, reference data has to be used. For each week of the year a reference week, i.e. a week from the past which is similar to this week, is defined. Currently these reference weeks are used in redPILOT to get reference data for the number of units that have to be done. In the future also reference data for the order structure shall be available. As this reference data is real data from the past, it has the same format as the data used in the short-term forecast. Therefore the same procedure can be used to estimate the throughput.

4.4.1 Calculation of the Throughput

Our goal is to estimate how much time is needed to fulfil certain orders. The input data that is available is a list of orders that have to be processed which contains the planned time, the number of distinct articles and the number of packing units for each order. It is not known in advance which operator will do which orders. In general an operator who has finished one order will just receive the next order from the list. There is only some optimisation done by the Warehouse Management System which tries to avoid that the next article to be picked is very far away. Furthermore, we also do not know which or how many of the orders will be done by operators from the external employment agency. There is only a contract with the agency how many packing units the external operators have to process. If we subtract this number from the sum of the units for all orders in the list, we know how many units have to be done by the internal operators. To predict how much time the allocated (internal) operators will need to finish this amount of orders, we need an estimate for how many orders they can do within a certain time, i.e. the units per hour which they can process given a certain order structure. If we sum up the units per hour of all operators working during a certain shift, we get the throughput per hour for that process step, given certain conditions, namely the allocated operators and the structure of the orders.
As already mentioned above, we do not know in advance which operators will do which of the orders given in the list. We will therefore calculate the throughput for each operator using

all orders. We thus assume that each of the operators receives the same "mix" of orders. By using the model described in the previous chapter we can get an estimate for the time a specific operator will need for an order. To estimate the time he needs to fulfil several orders, we also have to consider the time between the orders. It is not realistic to assume that the operators will do one order after another. There will usually be some time between the orders. So, to get a realistic estimation of the throughput per hour, we also have to examine those order transition times.

4.4.1.1 Order transition times

To examine the transition times between the orders, we can use the same dataset we have used for training the linear models. As for the creation of models, we only use the data of operators where we have at least 100 observations.
The transition time is the time between the end of one order and the start of the next order an operator has done. We do not know if this time is a planned break, a waiting time or an unplanned interruption, e.g. because of technical problems. The majority of these transition times (about 86%) lies between 0 and 10 minutes. Usually, for each operator once per shift we can see a larger value (around 45 minutes) which corresponds to the planned break time. There are also a few outliers which are larger than 100 but these values only amount for 0.3% of all values.
For the estimation of the throughput we will use the average transition time for those operators where we have at least 100 observations and the overall average for those operators where we only have few observations. For most of the operators the average transition time lies between 4 and 7 minutes but we have three outliers where the average transition time is higher than 10 minutes. The overall average is about 5.5 minutes.

4.4.1.2 Estimation of the Throughput

As already described above, our input dataset is a list of orders containing the following information for each order:

- the planned time (minutes)
- the number of distinct articles
- the planned units (packing units)

Additionally we know how many packing units will be done by external operators and which internal operators will work at which time. Using our linear models for the different operators and the transition times, we can follow the subsequent procedure to calculate the throughput:

1. Estimate the actual times for all the orders for each operator. For operators with at least 100 observations in the training dataset their specific coefficients are used, for the

other operators the linear model generated on the whole dataset for all operators is used. These models contain an intercept, a coefficient for the input parameter "planned time" (coeff1) and a coefficient for the input parameter "distinct articles" (coeff2). If the predicted value for the actual time is smaller than zero, it is set to zero. For operator i and order j, we get the formula:

$$\text{actual\_time}_{ij} = \max(0, \text{intercept}_i + \text{coeff1}_i * \text{planned\_time}_j + \text{coeff2}_i * \text{number\_distinct\_articles}_j) \tag{4.3}$$

2. Sum up all the estimated actual times for all n orders:

$$\text{sum\_actual\_times}_i = \sum_{j=1}^{n} \text{actual\_time}_{ij} \tag{4.4}$$

3. Add the transition times. For operator i for whom the average transition time transition_time_i calculated based on the training data we get:

$$\text{sum\_times}_i = \text{sum\_actual\_times}_i + (n-1) * \text{transition\_time}_i \tag{4.5}$$

4. Divide the sum of the planned units for all n orders by the sum of the times and multiply by 60 to get the throughput per hour:

$$\text{sum\_units} = \sum_{j=1}^{n} \text{planned\_units}_j \tag{4.6}$$

$$\text{throughput}_i = \frac{\text{sum\_units}}{\text{sum\_times}_i} * 60 \tag{4.7}$$

5. To get the units the operator can do during the shift, we can multiply the throughput by the planned working time of the operator which can be the whole shift or only some hours:

$$\text{units\_working\_time}_i = \text{throughput}_i * \text{working\_time}_i \tag{4.8}$$

6. If we sum up the units for all operators, we can get the units that can be done during the whole shift:

$$\text{units\_shift} = \sum \text{units\_working\_time}_i \tag{4.9}$$

As already mentioned above, by using this procedure we make the assumption that employees have to do approximately the same "mix" of orders. In Figure 4.53 we can see a comparison of the average number of units contained in an order for a specific operator and the average throughput per hour of that operator. We can see that there are some operators who seem to be mainly doing orders with only a few units. By looking at the order information data we

can see that the articles in those orders are large heavy articles such as multipacks of mineral water or barrels of beer. In general, those operators who have a lower number of average units per orders also have lower throughputs than operators who are mainly doing orders with more units. If we assume that operators who have done mainly orders with only a few units in the

	operator	average_units	average_throughput
1	1036694903	37.4839494163424	307.276883188711
2	1036694905	2.54766734279919	37.3968238189766
3	1036694907	9.61208475019618	110.764814486917
4	1036694908	18.4352941176471	155.4867972037
5	1036694909	5.53073953073953	61.6114423569496
6	1036694911	37.996719160105	271.689787497046
7	1036694913	38.0100439422473	245.028385058857
8	1036694918	6.24024483550115	93.1097580752791
9	1036694920	35.4589963280294	205.914703646565
10	1036694921	35.341935483871	189.397525498479
11	1036694923	29.9066068515498	249.451421646371
12	1036694924	29.1340909090909	219.288266728982
13	1036694925	34.1757281553398	276.513400817545
14	1036694926	36.3579902302861	256.06357535693
15	1036694927	26.5788742182071	206.746095465801
16	1036694928	29.772	229.87237193892
17	1036694929	37.1408614668219	255.880961576851
18	1036694930	29.5725190839695	329.433274635872
19	1036694931	34.8165680473373	197.048941812404
20	1036694932	26.7796086508754	207.124862555544
21	1036694933	31.7011450381679	323.096251589559
22	1036694936	29.7063492063492	187.621980106232

Figure 4.53: Average units per order and throughput for different operators

past will also do mainly orders with only a few units in the future, by using the procedure described above to calculate the throughput, we will probably overestimate the throughput for these operators. In contrast, for other operators who do fewer orders with only a few units we will underestimate the throughput.

The aim of the throughput prediction is to estimate how much time the operators allocated for one shift need to finish the given orders. Thus, we are mainly interested in the throughput of all of them together and not so much in the throughput of the individual operators. If we underestimate the throughput for some of the allocated operators and overestimate it for others as it will probably be the case due to uneven distribution of units per order, the estimation of the throughput for all operators should still be reasonably good.

4.4.2 Evaluation

To evaluate if using the model described above to predict the throughput can bring any benefit to the company, we will now perform a qualitative and quantitative evaluation. As for a long-term prediction not enough reference data is available, we will only perform the evaluation for a short-term forecast. For this evaluation we will use the test dataset we have retrieved from the available data (see Chapter 4.3.2). This test dataset contains the observations from those 27 days we have not used for the training of the model.

4.4.2.1 Assumptions

To evaluate the model, we have to make some simplifying assumptions. Those assumptions mainly concern the measures taken as a consequence of the prediction as well as the cost savings or additional costs for correct or incorrect predictions. The result of the prediction will be presented to the shift manager and he or she can decide how to react. For our evaluation we assume that two different measures are taken:

- Tell one, several or all operators that they can leave one or several hours earlier
- Tell one, several or all operators that they should stay longer

In reality the shift manager could also make other decisions, such as to do some cleaning or maintenance work if orders are finished early, or to call additional operators if orders cannot be finished.
Furthermore we assume that the working time reduction or additional working time is spread evenly among the employees. If we discover for example that we need two hours of working time less than it was planned, we send two people home one hour earlier and not one person two hours earlier. If only one ore some and not all people can go home earlier, the decision who can go home earlier will be made by the shift manager. For his decision he will probably consider criteria such as overtime hours or personal duties of the employee. In our evaluation we will therefore determine the number of operators that can leave earlier in a way that even if the best operators (the operators with the highest throughput according to our prediction) leave, the work can still be done. Similarly, we will determine the number of operators that have to stay longer in a way that even if the weakest operators (the operators with the lowest throughput according to our prediction) stay, the orders can be fulfilled. Moreover, we assume that the costs per hour for all operators are the same and therefore concerning the cost savings or additional costs it does not make any difference which operators leave early or stay longer.

4.4.2.2 Creation and use of the prediction

To make our prediction we use the following input data:

- A list of orders containing the planned time, the number of distinct articles and the number of packing units for each order
- A schedule for the shift including the information which operators will work during which timespan (the whole shift or a part of it)

We use the procedure described in Chapter 4.4.1.2 to estimate the throughput. As an output of this procedure we get an estimation of how many units can be processed within the planned working time of the allocated operators given a certain mix of orders. Besides, we also have an estimation of the throughput per hour for each operator for that order structure. To decide

which measures shall be taken, we compare the possible throughput of all operators for the whole shift to the number of units that has to be done. This number is the sum of the planned units for all orders in the list minus the number of packing units that is planned for the operators of the temporary employment agency:

$$\text{units\_todo} = \text{sum\_units} - \text{planned\_units\_external} \tag{4.10}$$

We calculate the difference between the possible units, i.e. the units the operators can do during their scheduled working time (units_shift), and the units to be done:

$$\text{diff\_units} = \text{units\_shift} - \text{units\_todo} \tag{4.11}$$

Depending on the value of that difference, different measures are taken.

- Positive difference:
  - If the difference is smaller than the highest one of the estimated throughputs per hour for the different operators, no action is taken.
  - If the difference is bigger than the highest one of the estimated throughputs per hour for the different operators, one operator is sent home one hour earlier.
  - If the difference is bigger than the sum of the two highest throughputs, two operators are sent home one hour earlier.
  - ...
  - If the difference is bigger than the sum of all estimated throughputs, all operators are sent home one hour earlier, i.e. the shift ends one hour earlier.
  - If the difference is bigger than the sum of all estimated throughputs plus the highest one of the estimated throughputs, one operator is sent home two hours earlier and all the other operators are sent home one hour earlier.
  - ...
- Negative difference:
  - If the absolute value of the difference is smaller than the lowest one of the estimated throughputs per hour, one operator stays one hour longer.
  - If the absolute value of the difference is higher than the lowest one of the estimated throughputs per hour but lower than the sum of the two lowest estimated throughputs, two operators stay one hour longer
  - ...

- If the absolute value of the difference is smaller than the sum of all estimated throughputs but bigger than the sum of all estimated throughputs except for the highest, all operators have to stay longer and the shift ends one hour later.
- If the absolute value of the difference is bigger than the sum of all estimated throughputs but smaller than the sum of all estimated throughputs plus the highest one, one operators has to stay two hours longer and the other operators have to stay one hour longer.
- ...

4.4.2.3 Check for the correctness of the prediction

To check if our prediction was correct and the measures taken were good, we can use the real working time and the sum of actually processed units for the operators given in our test dataset. As in the determination of the measures to be taken, we also assume that the operator(s) with the highest throughput leave earlier and the one/those with the lowest throughput stay longer. To validate the prediction, the following steps have to be done:

- Calculate the real throughputs of the operators by dividing the sum of the units they have done by their working time. For operator i:

$$\text{actual\_throughput}_i = \frac{\sum_{j=1}^{n_i} \text{actual\_units}_{ij}}{\text{working\_time}_i} \tag{4.12}$$

  It might be the case that an operator who was planned for a certain shift did not actually work during that shift and therefore we cannot calculate his actual throughput for that shift. In that case, we take the average throughput of all the working days of that operator in the test dataset.

- Find the best/weakest operators
- Reduce/augment their working time according to the measures described before
- Multiply the actual throughputs for the different operators by the changed working times:

$$\text{units\_working\_time\_new}_i = \text{actual\_throughput}_i * \text{working\_time\_new}_i \tag{4.13}$$

- Sum up the units for all the operators:

$$\text{units\_shift\_new} = \sum \text{working\_time\_new}_i \tag{4.14}$$

We can than compare the possible units after our changes (units_shift_new) according to the real throughputs of the operators to the units that have to be done. If we have reduced the

working time for one or several and the number of possible units is still higher than the number of units to be done, our decision was correct and we can save personnel costs. In contrast, if the number of possible units is now lower than the number of units to be done the decision was incorrect. To check if the decision to let operators stay longer was correct, we have to calculate the number of possible units for the originally scheduled working times (which - due to short-term changes done after receiving the actual orders - will usually not be equal to the actual working time we can see in the test dataset) of the different operators. If the number of the possible units for the original schedule was lower than the number of units to be done and the number of units that can be done after the changes is higher than the units that have to be done, the decision to let some operators stay longer was correct. However, if the number of units that could be done according to the original plan is higher than the number of units to be done, it was not necessary to let anybody stay longer and our decision was incorrect. The decision is also incorrect if the number of units that can be done after the changes is smaller than the number of units that have to be done.

4.4.2.4 Evaluation of Costs

Two different cost factors will be considered for the evaluation of the throughput prediction: personnel costs and operator satisfaction. For a warehouse where the order picking is done manually, personnel costs are one of the biggest cost factors. Therefore, this is one of the areas where the highest cost savings are possible. Furthermore, the operators have a high impact on the performance of the whole system and therefore it is important that they are satisfied and motivated. If we can already inform them about changes in their working time at the beginning of the shift, we can make them more content, if we give them wrong information they will get more dissatisfied.
For the evaluation of personnel costs we assume that even if we have told them that they can go home earlier at the beginning of the shift, we can still withdraw that decision and tell them to stay for the whole working time to avoid that orders are not finished. However, if we tell them at the beginning of the shift that they have to stay longer, they will actually stay longer and we therefore have additional personnel costs.

4.4.2.5 Results of the evaluation

For our evaluation we make predictions for 43 different shifts, we take the measures described, check for their correctness as described above and calculate the changes of operator satisfaction and personnel costs. For the calculation of the personnel costs, we assume a cost per hour of 22.5 euros.
Table 4.5 shows the results of the evaluation. The first column indicates the start time of the shift, i.e. the date and if it was a morning or afternoon shift. The second column shows the sum of the planned working times in hours for all operators of the shift. In the third column

Table 4.4: Potential costs

	Go home earlier	**Stay longer**
Correct prediction	Personnel costs: -cost per hour*number hours reduced Operators happy: +1 for all operators Operators angry: 0	Personnel costs: neutral Operators happy: +1 for all operators Operators angry: 0
Incorrect prediction	Personnel costs: neutral Operators happy: +1 per correct information Operators angry: +1 per incorrect information	Personnel costs: +cost per hour*number hours added Operators happy: +1 per correct information Operators angry: +1 per incorrect information

the sum of the working times after the measures taken based on the prediction are given. The column "Ideal WT" gives the ideal working time calculated base on the actual throughputs. The next three columns show the units that have to be done and the possible units based on the actual throughputs for the planned working time and the changed working time. The measure that has been taken after the prediction is given in the seventh column, followed by a column that gives information if it was correct. The next two columns show how many operators we have given a correct information about their working time (operators happy) and how many we have informed incorrectly (operators angry). In the last column the change of personnel costs is given.

For three shifts the estimated throughput (i.e. the number of possible units) is lower than the units that have to be done and we therefore decide to let operators stay longer. If we calculate the possible units based on the actual throughputs of the operators, we can see that this decision was correct because the orders could not have been finished within the scheduled working time. We can also see that the number of units that can be done after our changes of the working times of the operators is still lower than the number of units that have to be done which means that we did not add enough hours and therefore the measure was not entirely correct. In one of those three shifts the number of units that have to be done is not a lot higher than the number of possible units for the changed working times and we therefore have informed some operators correctly about their working time. For four of them this information was that the working time stays the same and for two of them that they have to stay one hour longer. However, five other operators also had to stay longer although we did not give them that information at the beginning of the shift. In the other two shifts the number of units that has to be done is a lot higher than the possible units after the changes and we therefore gave all operators a wrong information.

In one shift we did not take any action. For all the other shifts the estimated throughput is higher than the units that have to be done and we therefore have reduced the working times of the operators. In most of the cases (28 shifts) this decision is basically right, i.e. the number of units that can be done according to the actual throughputs is also higher than the units that have to be done. But we have reduced the working times too much and therefore the number of possible units after the changes is lower than the number of units that have to be done. In 16 of those shifts all of the operators actually have to stay longer than we have told them at the beginning of the shift. Thus, we give the operators a wrong information, the operator satisfaction decreases and we cannot save personnel costs. In the other shifts the number of possible units after the changes is only a bit smaller than the units that have to be done which means that we gave a correct information to some and a wrong information to other employees. In those shifts we can also reduce personnel costs because some operators go home earlier. In total we reduce the costs by 1147.5€. We give wrong information to operators 408 times and correct information 57 times.

Table 4.5: Results evaluation

Shift start	WT plan h	WT new h	Ideal WT	Units todo	Possible units plan	Possible units new	Measure	Correct	Operators happy	Operators angry	Personnel costs
2018-03-08 06:00:00	64	64	81	17280	13701	13701	none	no measure	0	0	0
2018-03-08 15:00:00	88	53	62	14781	21363	12738	go home earlier	no	2	9	-45
2018-04-07 05:00:00	132.75	80.75	109.75	23149	28367	16868	go home earlier	no	0	15	0
2018-04-20 05:00:00	81	42	47	11704	19721	10198	go home earlier	no	6	5	-135
2018-03-03 05:00:00	96.75	69.75	82.75	17600	20932	14708	go home earlier	no	0	11	0
2018-02-27 06:00:00	108	67	99	18483	20566	12557	go home earlier	no	0	14	0
2018-02-27 15:00:00	80	41	55	11725	17278	8651	go home earlier	no	0	10	0
2018-04-05 05:00:00	122	98	126	26380	25802	20249	go home earlier	no	0	14	0
2018-04-05 15:00:00	72	68	77	18932	17761	16456	go home earlier	no	0	8	0
2018-03-20 05:00:00	115	64	86	17661	23816	12961	go home earlier	no	0	14	0
2018-03-20 15:00:00	90	55	64	15148	21235	12643	go home earlier	no	1	9	-22.5
2018-04-09 05:00:00	103	89	108	24742	23879	20455	go home earlier	no	0	12	0
2018-04-09 15:00:00	96.75	85.75	92.75	21275	22169	19637	go home earlier	no	4	7	-90
2018-04-17 05:00:00	112	91	116	25406	24792	19734	go home earlier	no	0	13	0
2018-04-17 15:00:00	96.75	66.75	68.75	16771	23874	16140	go home earlier	no	9	2	-202.5
2018-04-24 05:00:00	129	110	131	31743	31577	26586	go home earlier	no	0	15	0
2018-04-24 15:00:00	105.75	81.75	93.75	20224	23081	17843	go home earlier	no	0	12	0
2018-02-03 05:00:00	123.75	94.75	118.75	24131	25155	19091	go home earlier	no	0	14	0
2018-03-16 06:00:00	74	35	42	10530	19453	8889	go home earlier	no	4	7	-90
2018-03-02 06:00:00	58	44	71	12572	10231	7565	go home earlier	no	0	9	0
2018-02-02 06:00:00	61	47	57	12895	14076	10467	go home earlier	no	0	9	0
2018-03-07 06:00:00	48	39	47	10218	10776	8543	go home earlier	no	0	6	0
2018-03-07 15:00:00	88	90	95	25436	23425	23730	stay longer	no	6	5	0
2018-03-05 05:00:00	69	65	83	16910	14149	13057	go home earlier	no	0	8	0
2018-03-05 15:00:00	96.75	91.75	103.75	23534	22100	20655	go home earlier	no	0	11	0
2018-03-10 05:00:00	90	71	79	16403	18909	14729	go home earlier	no	2	8	-45
2018-04-03 05:00:00	112	50	61	13805	25152	11076	go home earlier	no	2	11	-45
2018-04-03 15:00:00	90	48	52	12091	21287	11221	go home earlier	no	6	4	-135
2018-03-17 05:00:00	117	74	90	20013	26214	16180	go home earlier	no	0	13	0
2018-03-28 05:00:00	78	63	72	17271	18865	15023	go home earlier	no	0	9	0
2018-03-28 15:00:00	90	94	118	26725	20591	21191	stay longer	no	0	10	0
2018-03-19 05:00:00	106	83	103	21221	22183	17038	go home earlier	no	0	13	0
2018-03-19 15:00:00	90	53	65	15742	22324	12866	go home earlier	no	0	10	0
2018-02-28 06:00:00	100	64	89	16580	18934	11761	go home earlier	no	0	13	0
2018-02-28 15:00:00	64	58	81	16642	13140	11737	go home earlier	no	0	8	0
2018-03-01 06:00:00	78	72	96	19566	15954	14365	go home earlier	no	0	10	0
2018-03-01 15:00:00	62.75	47.75	60.75	12468	12952	9706	go home earlier	no	0	8	0
2018-02-26 05:00:00	110	77	112	19585	19436	13123	go home earlier	no	0	13	0
2018-02-26 15:00:00	81	50	63	14414	18577	11276	go home earlier	no	0	9	0
2018-04-25 05:00:00	105	97	105	27973	28504	25871	go home earlier	no	4	8	-90
2018-04-25 15:00:00	99	78	86	20587	23869	18594	go home earlier	no	3	8	-67.5
2018-04-19 05:00:00	96	103	124	28240	21790	23058	stay longer	no	0	11	0
2018-04-19 15:00:00	96.75	81.75	84.75	23268	26949	22437	go home earlier	no	8	3	-180
Sum									57	408	-1147.5

As we always reduced the working time of the operators by too many hours or did not add enough hours, we can try if we can achieve better results by using a more conservative prediction. An easy approach would be to multiply the estimated amount of possible units by a certain factor $0 < f <= 1$ and then use the same procedure as described before to determine the measures. We can try different values for f and compare the results. Table 4.6 shows a comparison of the number correct and incorrect decisions as well as the sums of personnel

costs and operator satisfaction for values of f between 0.8 and 1. We can see that the number of incorrect decisions decreases with the value of f. The reduction of personnel cost first increases as the factor decreases, but for $f = 0.8$ it is lower than for $f = 0.85$. In contrast, the operator satisfaction is the highest for $f = 0.8$ because for this value of f we have the highest number of correct decisions. Furthermore we can see that the lower the factor is, the more often we increase the working times. For $f = 0.8$ we already have some shifts where we unnecessarily increase the working times and therefore have additional personnel costs. If we further decrease the value of f, the number of incorrect decisions will start increasing again because we will unnecessarily increase the number of working hours.
In our table we can see that the operator satisfaction is the highest for $f = 0.8$ whereas the

Table 4.6: Comparison of different estimates for the possible units

Factor	Measure go home earlier	Measure stay longer	Number correct decisions	Number incorrect decisions	Personnel costs	Operators happy	Operators angry
1	39	3	0	42	-1147.5€	57	408
0.95	37	4	2	39	-3082.5€	134	320
0.9	33	10	10	33	-6120€	236	237
0.85	29	12	18	23	-7605€	290	157
0.8	25	18	23	20	-6637.5€	367	106

reduction of personnel costs is the highest for $f = 0.85$. It thus seems that the two goals to increase operator satisfaction and to reduce personnel costs are contradictory. If we make more conservative predictions, we are less likely to give wrong information but we also reduce the working times by fewer hours and can therefore not always exploit the cost reduction potential. If we want to achieve the highest possible number of "happy" operators for our test dataset, we have to chose a factor of 0.8. To get the highest possible reduction of personnel costs (-7762.5€), we have to choose $f = 0.84$. A factor of $f = 0.83$ seems to provide a good balance between operators satisfaction and reduction of personnel costs. We give correct information to operators 340 times and incorrect information 133 times and we can reduce the personnel costs by 7357.5€.
We can use some new test data that has not been used yet to test if this factor is really appropriate and which cost reductions and changes of operator satisfaction we can achieve. This new test dataset contains data for the time between the 1st of June and the 15th of June which means that we have 15 shifts for which we can evaluate the prediction. Table 4.7 shows the results of this evaluation. We can see that we decide to let operators stay longer four times which is basically always correct but twice we did not increase the working time enough. In the other shifts we sent people home some time earlier. Six times the measure is fully correct and the units to be done can be fulfilled within the changed working time. Once it was correct to reduce the working time but we reduced it too much and four times we decided that operators can go home earlier although it was actually necessary that they stay longer or stay the planned time. In total we could achieve a reduction of personnel costs of 2520€. We gave employees the right information 114 times and wrong information 58 times. We can thus conclude that we can achieve a small reduction of personnel costs and also an increase of operator satisfaction by applying the procedure described to predict the possible units for a

shift and by multiplying this quantity by a factor of 0.83.

Table 4.7: Results evaluation new test data

Shift start	WT plan h	WT new h	Ideal WT	Units todo	Possible units plan	Possible units new	Measure	Correct	Operators happy	Operators angry	Personnel costs
2018-06-01 05:00:00	80	77	87	16923	15784	14951	go home earlier	no	1	10	-22.5
2018-06-02 05:00:00	96.75	84.75	100.75	20240	19663	17172	go home earlier	no	0	11	0
2018-06-04 05:00:00	114	128	130	30325	26615	29749	stay longer	no	11	2	0
2018-06-04 15:00:00	96.75	71.75	65.75	13741	20520	14992	go home earlier	yes	11	0	-562.5
2018-06-05 05:00:00	113	130	123	26961	24910	28349	stay longer	yes	13	0	0
2018-06-05 15:00:00	87.75	69.75	67.75	12810	16877	13178	go home earlier	yes	10	0	-405
2018-06-11 05:00:00	91	88	95	22169	21367	20439	go home earlier	no	4	7	-90
2018-06-11 15:00:00	96.75	114.75	99.75	26653	25880	30467	stay longer	yes	11	0	0
2018-06-12 05:00:00	93	85	91	19399	19761	17801	go home earlier	no	5	6	-112.5
2018-06-12 15:00:00	105.75	103.75	100.75	23420	24999	24400	go home earlier	yes	12	0	-45
2018-06-13 05:00:00	97	85	97	18811	18877	16555	go home earlier	no	0	12	0
2018-06-13 15:00:00	105.75	104.75	100.75	25314	26876	26489	go home earlier	yes	12	0	-22.5
2018-06-14 05:00:00	93	99	109	22154	18937	19895	stay longer	no	1	10	0
2018-06-14 15:00:00	105.75	85.75	81.75	19372	25321	20199	go home earlier	yes	12	0	-450
2018-06-15 05:00:00	81	45	42	9471	18688	10228	go home earlier	yes	11	0	-810
Sum									114	58	-2520

We can compare these results to the results we get if we use a static value for the throughput of one operator to estimate the the necessary time to finish the orders as it is currently done in the redPILOT application. This throughput is configured as 240 units per hour. To get the appropriate working time for a certain shift we simple divide the number of units that have to be done by the configured throughput. If the estimated necessary working time is lower than the planned working time, we send some random operators home earlier, if it is higher, we let some random operators stay longer. The calculation of the units that can actually be done before and after the changes is done as before. In Table 4.8 we can see that the number of possible units after our changes is almost always lower than the number of units that have to be done. It thus seems that on average the actual performance of the operators is lower than the configured throughput. By using this kind of prediction with a static throughput of 240 we give correct information to the operators 75 times and we can reduce the personnel costs by 1035€. Both numbers are lower than those achieved by using the linear models to estimated the throughputs.

Table 4.8: Results evaluation with static throughput new test data

Shift start	WT plan h	WT new h	Ideal WT	Units todo	Possible units plan	Possible units new	Measure	Correct	Operators happy	Operators angry	Personnel costs
2018-06-01 05:00:00	80	71	87	16923	15784	13981	go home earlier	no	0	11	0
2018-06-02 05:00:00	96.75	85	100.75	20240	19663	17172	go home earlier	no	0	11	0
2018-06-04 05:00:00	114	127	130	30325	26615	29657	stay longer	no	10	3	0
2018-06-04 15:00:00	96.75	58	65.75	13741	20520	12144	go home earlier	no	3	8	-67.5
2018-06-05 05:00:00	113	113	123	26961	24910	24910	none	no measure	0	0	0
2018-06-05 15:00:00	87.75	54	67.75	12810	16877	10592	go home earlier	no	0	10	0
2018-06-11 05:00:00	91	93	95	22169	21367	21734	stay longer	no	9	2	0
2018-06-11 15:00:00	96.75	112	99.75	26653	25880	30060	stay longer	yes	11	0	0
2018-06-12 05:00:00	93	81	91	19399	19761	17176	go home earlier	no	1	10	-22.5
2018-06-12 15:00:00	105.75	98	100.75	23420	24999	23019	go home earlier	no	9	3	-202.5
2018-06-13 05:00:00	97	79	97	18811	18877	15457	go home earlier	no	0	12	0
2018-06-13 15:00:00	105.75	106	100.75	25314	26876	26970	stay longer	no	11	1	22.5
2018-06-14 05:00:00	93	93	109	22154	18937	18937	none	no measure	0	0	0
2018-06-14 15:00:00	105.75	81	81.75	19372	25321	19404	go home earlier	yes	12	0	-562.5
2018-06-15 05:00:00	81	40	42	9471	18688	9278	go home earlier	no	9	2	-202.5
Sum									75	73	-1035

4.4.2.6 Conclusion of the evaluation

In this chapter it has been presented how the linear models created before can be used to predict the throughput of a process step on a short-term basis given a certain list of orders and a certain allocation of operators. We have seen that by using the described approach

to estimate the throughput we often receive a too optimistic prediction and therefore take measures that seem to be wrong according to the results of our evaluation. We have therefore introduced a factor f by which we multiply the estimate for the units that can be done during the shift. Using a test dataset we have analysed different values for f and have found out that a factor of 0.83 seems to provide a good balance between the goal to exploit the potential for personnel cost reductions and the goal to give correct information to the employees. We have tested this factor using a new test dataset and have seen that we can improve operator satisfaction and reduce personnel costs by using the procedure described above to estimate the throughputs of the operators and by multiplying the estimated units for one shift by $f = 0.83$. Furthermore we have seen that we can achieve better results than those we get if we use the currently configured static throughput to make the prediction.
We can thus conclude that the use of our linear models in combination with the described method to get an estimation for the throughput can bring some benefits in terms of reducing personnel costs and increasing operator satisfaction. The disadvantage of the procedure is that it assumes that the real orders are already available which is only true for a very short-term prediction, i.e. a prediction for the same day or maybe the next day. An approach how to make a prediction longer in advance is presented in the next chapter. Furthermore it has to be highlighted that we have made a lot of assumptions for the evaluation. E.g. we have only taken two different measures (let operators stay longer or send them home earlier), we have always spread the hours reduced or added evenly among the operators and we have assumed that the costs for all employees are the same. In reality, based on the results of the prediction, the shift manager could also decide to take other measures, such as to do cleaning or maintenance work. Furthermore he or she could decide to send one person home two hours earlier instead of two persons one hour and there are different groups of employees for which the cost per hour is not the same.
Therefore, to really make a valid statement about the benefits of the use of the prediction, the procedure has to be used in practice and reviewed by shift managers and planners. Those people can then say if it helps them to make correct decisions and to ideally use the available resources.

4.4.3 Long-term prediction

The throughput prediction is especially useful if we can make it some time, i.e. one or several weeks, in advance during the planning phase. In this phase the schedule for a shift is created which means that operators are allocated and the duration of the shift is fixed. A good prediction of the throughput can help to optimise the use of resources and to avoid short-term changes. E.g., the planner can replace full-time employees by part-time employees if less time is needed or request temporary helpers if the units cannot be finished within the shift by the previously allocated operators. Thus, if we know longer in advance what the throughput in

a shift will be, much higher cost saving can be achieved than with a prediction done on a short-term basis.
The procedure described and evaluated above can only be used if the real orders are already available. This information is only available for the same day. So, if we want to make a prediction for a longer time frame, we have to replace this data. Therefore we store reference data that can in the future be used to use the throughput prediction as it is described above. For each week of the year a reference week is defined. This reference week is a week in the past that was similar than this week. E.g., for the Easter week of this year the Easter week of the last year can be used. As the data containing the planned times and the order information has not been collected in the past, it is only available for this year and we thus do not have reference data for all weeks. We therefore cannot fully evaluate the long-term prediction but the procedure will be shown in an example.

4.4.3.1 Example long-term prediction

One of the days contained in the test dataset used in the evaluation for the short-term prediction is the 5th of March. For the week from the 4th of March to the 10th of March and the week from the 11th of March to the 17th of March the same reference week is used. The data for this reference week is not available but in our training dataset we can find the data for the 12th of March. As the two weeks have the same reference weeks, they can be seen as similar and we can therefore use the data for the 12th of March as a reference data for the 5th of March. Of course, in reality this is not possible because the 12th of March is after the 5th of March but for this example, we will pretend that the data for the 12th is data from the last year.
The 12th and the 5th of March are Mondays which means there are two shifts. We analyse the morning shift which lasts from 6 a.m. to 2 p.m.. The input for the throughput prediction are the order list from the reference day, i.e. the orders processed in the morning shift of the 12th of March, and the schedule for the morning shift of the 5th of March, i.e. the operators and their planned working times. Using this input data we can apply the procedure described in Chapter 4.4.1.2 to estimate the throughputs per hour of the different operators and subsequently the units that can be processed within the shift. According to this estimation, 18438 units can be done by the operators within their planned working times. In total, 29716 units have to be done within the shift and 6000 units shall be done by operators of a temporary employment agency. Therefore, 23716 units have to be done by the internal employees which is more than they can do according to our estimation. This means that we have to increase the number of working hours of some operators. According to the procedure described in Chapter 4.4.2.2, we have to increase the working time by 20 hours in total, which means that we increase the working time of four operators by three hours and the working time of the other four operators by two hours.
Now we can use the actual data for the morning shift of the 5th of March to check if the

estimation and the measures taken were correct. At first we can analyse if the reference data and the real data are similar and if it was therefore appropriate to use the reference data for the prediction. In the reference dataset the order list contains 983 orders with a total of 29716 units to be done whereas in the actual dataset the list contains only 765 orders with 22975 units. On both days 6000 units are planned for the employees of the temporary employment agency which means that 23716 respectively 16975 units have to be done by the internal operators. Thus, on our reference day more units had to be done during the morning shift than on the day for which we make the prediction. Also relevant for the accuracy of our estimation is the structure of the orders. Figure 4.54 and Figure 4.55 show histograms for the planned

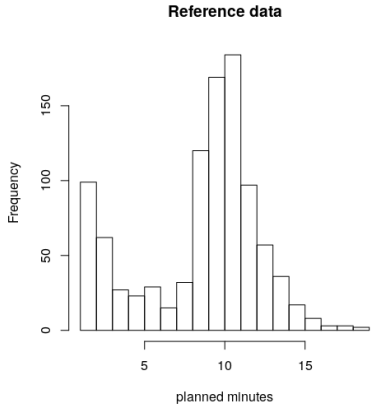

Figure 4.54: Histogram planned minutes reference data

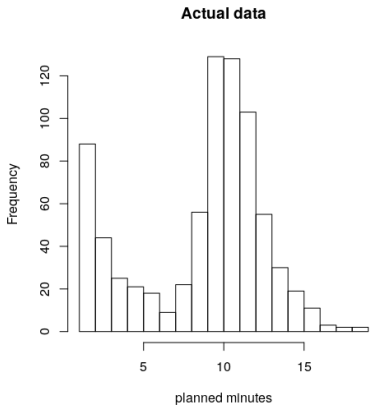

Figure 4.55: Histogram planned minutes

minutes of the orders in the reference data respectively the actual data. We can see that the distributions are relatively similar but in the reference data there are more orders where the planned time is between 8 and 10 minutes whereas in the actual data there are more orders where the planned time is between 10 and 13 minutes. This means that in the actual data there are more orders where the planned time is longer and which therefore seem to be more complex. In the next two figures, Figure 4.56 and Figure 4.57, histograms for the number of planned units per order in the two different datasets are shown. The two distributions seem to be very similar. As the histograms for the reference data and the actual data are relatively similar for the planned minutes as well as for the planned units, its seems to be appropriate to use the data of the reference day to estimate the throughput for the "new" day.
As already mentioned above, 16975 units have to be done by the internal operators on the 5^{th} of March, the day for which we make the prediction. We use the actual working times of the operators and the units actually processed by them to calculate the actual throughputs as we did in the evaluation of the short-term model (see Chapter 4.4.2.3). If we multiply the actual throughputs by the working times of the operators and sum up the units, we get the units that can be done within the shift. For the originally planned working time we get a number of 14104 units that can be done, for the increased working time we get a number of 17929

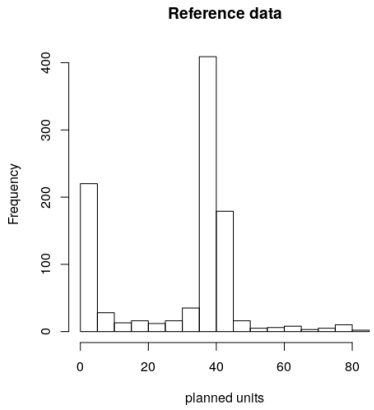

Figure 4.56: Histogram planned units reference data

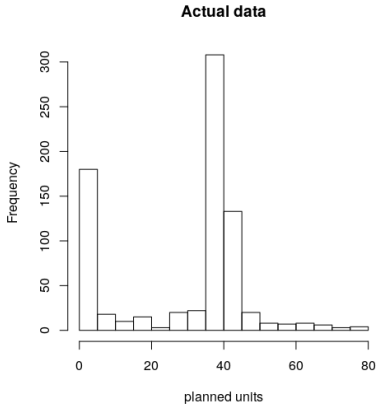

Figure 4.57: Histogram planned units

units. Thus, it was correct to increase the working time but we increased it too much because our reference dataset contained more orders than the actual data.
In this example we did not multiply the estimated possible units by any factor as it was proposed in the previous chapter. If we use a factor $f = 0.83$ which seemed to be a reasonable value in the short-term prediction, we increase the working times of the operators by 32 hours in total and can receive a number of 20655 units that can be done according to the actual throughputs. Thus, we add even more unnecessary working time.

4.4.3.2 Conclusion long-term prediction

If we make a prediction one or several weeks in advance when the real orders are not known yet, we of course have to cope with some uncertainty. We can replace the real orders by reference data but as we have seen in our example, the reference data does not always perfectly represent the real data. E.g., in the example the reference data contained a lot more orders than the real data and we therefore added too many hours of working time which causes unnecessary personnel costs. But our basic prediction, i.e. that more working time than scheduled is needed, was correct and by using this prediction we can at least avoid having to look for operators that are willing to stay longer on a short-term basis.
Of course, one example is not enough to make a valid statement about the use of the approach described above to make a prediction during the planning phase. But our example indicated that it can bring some benefits but it is of course not as reliable as a prediction when the real orders are already available. It is thus advisable to use the short-term prediction at the beginning of the shift to check if the allocation made based on the long-term prediction is appropriate for the actual orders and to make adjustments at the beginning of the shift if the planned working times seem to be too high or too low.

5 Conclusion

The aim of this Master's Thesis was to find factors that are influencing the throughput, i.e. the number of units processed during a certain time, of manual order picking and to create a model to predict the throughput for this process step for a certain time given a certain configuration of input parameters.
In the first chapters of this thesis a general introduction to mathematical modelling has been given and regression analysis has been presented as a method to analyse the relationship between an output variable and one or several input parameters. Different types of regression as well as methods to analyse the results and compare different models have been presented. In the fourth chapter these methods have been used to analyse the relationship between the throughput and different input parameters. At first, the initial situation and the available data have been described. Afterwards the data from the different sources has been analysed. The analysis of the data from the redPILOT database has lead to the conclusion that the operators have a significant influence on the throughput but that this dataset does not contain enough information to make a reliable prediction of the throughput. Therefore, data from a different source, namely the customers warehouse management system (WMS), has been used. In this dataset the planned time for an order and the articles contained in that order are given which gives information about the complexity of an order. We have analysed different models to predict the actual time of an order based on the planned time and other input parameters. Finally we have decided that a linear model for each operator using the planned time and the number of distinct articles as predictors to estimate the actual time is the most adequate. As our cost function is not exactly the same as the cost function that is minimised for the estimation of the coefficients of the linear model, we have varied those coefficients a little bit to see if this leads to an improvement of the prediction.
The coefficients we received in this process have then been used to predict the units that can be done during the working time of specific operators given a list of orders containing the planned time, the number of planned units and the number of distinct articles for an order. Measures that are taken based on the result of this estimation have been defined and the quality and the benefits of the prediction have been evaluated. We have seen that the use of the procedure can lead to a small reduction of operator costs and can help to improve the satisfaction of the employees by informing them earlier about changes in their working time. But the procedure can only be applied if the real orders are already available which is why it can only be used on a very short-term basis. Therefore, additionally an approach to estimate

the throughput longer in advance has been presented.
In the evaluation process we have made a lot of assumptions which are not valid in real working conditions. E.g., the shift manager could take much more different decisions than those two we used in our evaluation. Furthermore, already now the shift managers make some decisions to send people home earlier or to let them work longer than planned. These decisions are made based on the order list and based on their personal experience. One of the benefits of the procedure described above could be that it provides a standardised procedure to estimate how much time is needed to fulfil the orders that have to be done and to give a guideline for the decisions that are made based on this estimation. Currently the decisions made depend a lot on the experience and the abilities of the shift manager. A standardised procedure can help to keep a constant quality of decisions and therefore reduce instabilities and irregularities. But, as already mentioned above, it is necessary to evaluate the procedure in practice to see if it really brings some benefits. It has to be used by shift managers in their everyday work to see if it can help them to make appropriate decisions.
To make a prediction more time in advance, which can lead to much higher cost savings than a prediction on a very short-term basis, more reference data has to be collected. We need at least the data for one full year to have reference data for all weeks. Using this data we can apply the procedure described in Chapter 4.4.3 and evaluate the quality and the benefits of this prediction.

Glossary

$(\tilde{x}_1, \tilde{y}_1), ..., (\tilde{x}_n, \tilde{y}_n)$ test data with only one input parameter.

$(x_1, y_1; x_2, y_2; ...; x_n, y_n)$ training data with only one input parameter.

$\boldsymbol{\beta}$ vector of coefficients of a linear model.

$\boldsymbol{\theta}$ vector of input parameters.

$\boldsymbol{\varepsilon}$ vector of residuals.

$\hat{f}(\mathbf{x}_i)$ predicted value for the ith observation.

$\mathbf{X}$ design matrix of a linear model.

$\mathbf{x}_i$ vector containing the values of all the input parameters for observation i.

$\mathbf{y}$ vector of all response values.

n number of observations (number of entries in a dataset used for creating a model; in application: number of orders).

p number of input parameters of a model.

y_i response value for the ith observation.

actual time Time (minutes) that has actually been needed by the operator to fulfil an order.

planned time Time (minutes) to fulfil an order that is estimated by the WMS.

revised planned time Value of planned time that has been updated after short term changes in the order.

throughput number of units processed during a certain timespan (usually one hour).

Bibliography

[1] Shervin Asadzadeh and Abdollah Aghaie. "Cause-selecting control charts based on Huber's M-estimator". In: *International Journal of Advanced Manufactoring Technology* 45 (3-4 2009), pp. 341–351

[2] Vivek S. Borkar. "There's No Such Thing as a Free Lunch. The Bias-Variance Dilemma". In: *Resonance* 3 (6 1998), pp. 40–51

[3] Ludwig Fahrmeir, Thomas Kneib, and Stefan Lang. *Regression. Modelle, Anwendungen und Methoden*. Springer-Verlag Berlin Heidelberg, 2007

[4] Frank Haußer and Yury Luchko. *Mathematische Modellierung mit MATLAB. Eine praxisorientierte Einführung*. Spektrum Akademischer Verlag Heidelberg, 2011

[5] Michael Ten Hompel and Thorsten Schmidt. *Warehouse Management. Automatisierung und Organisation von Lager- und Kommissioniersystemen*. 2nd ed. Springer-Verlag Berlin Heidelberg, 2005

[6] Gareth James et al. *An Introduction to Statistical Learning. with Applications in R*. Springer New York Heidelberg Dordrecht London, 2013

[7] Ron Kohavi. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection". In: *International Joint Conference on Artificial Intelligence (IJCAI)*. Vol. 2. 1995, pp. 1137–1143

[8] David J. Lilja. *Linear Regression Using R: An Introduction to Data Modeling*. University of Minnesota Libraries Publishing, 2016

[9] Marco Riani et al. "Monitoring Robust Regression". In: *Electronic Journal of Statistics* 8 (2014), pp. 646–677

[10] G. K. Robinson. "That BLUP Is a Good Thing: The Estimation of Random Effects". In: *Statistical Science* 6.1 (1991), pp. 15–51

[11] Vincenzo Verardi and Christophe Croux. "Robust Regression in Stata". In: *The Stata Journal* 9.3 (2009), pp. 439–453