Christopher Franz Kaponig, BSc

Automated Linear Performance Pricing using data analytics

Master's Thesis

to achieve the university degree of

Master of Science

Master's degree programme: Software Engineering and Management

submitted to

Graz University of Technology

Supervisor: Univ.-Prof. Dipl.-Ing. Dr.techn. Christian Ramsauer
Co-Supervisor: Dipl.-Ing. Oliver Mörth-Teo, BSc MSc

Institute for Innovation and Industrial Management

Graz, October 2021

Affidavit

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

Date

Signature

Abstract

Companies have to face an increasingly dynamic economic environment. In order to survive in competition, it is necessary to act flexibly and quickly. Due to the high degree of specialisation and the increasing outsourcing, companies have a dependency on suppliers. Therefore the amount of material costs represents a considerable part of the product costs.

Linear Performance Pricing (LPP) can be used in supplier management to overview price structures and allow more market transparency. As a result, cost reductions can be made possible with the use of LPP. The possibility of the automation of LPP is also a significant advantage. Because of this, the following thesis deals with the automation of linear performance pricing using data analytics. In the context of this work, a concept for a new process for an automated LPP process is developed and illustrated.

In the first step, the basics of data analytics and linear regression are described in detail. Then, the economical part deals with supplier management and linear performance pricing. Three different methods of LPP, as well as practical applications of LPP, are then demonstrated. Based on this, these methodologies are compared in the practical part of this thesis, and a new concept for an automated LPP process is presented. In addition, the created prototype is described.

The results of this thesis serve as a basis for the automation of an LPP process and illustrate how the possible steps for the automation of linear performance pricing can be done and in which areas further research is necessary.

Kurzfassung

Unternehmen müssen sich einem immer dynamischeren wirtschaftlichen Umfeld stellen. Um im Wettbewerb bestehen zu können, ist es notwendig, flexibel und rasch zu agieren. Aufgrund der hohen Spezialisierungsgrade und dem zunehmenden Outsourcing ist die Abhängigkeit von Lieferanten gegeben. Dabei stellt der Anteil der Materialkosten einen beträchtlichen Teil der Produktkosten dar.

Um einen Überblick über Preisstrukturen zu erhalten sowie mehr Markttransparenz zu ermöglichen, kann Linear Performance Pricing (LPP) im Lieferantenmanagement eingesetzt werden. Infolgedessen können Kostensenkungen mit Hilfe von LPP ermöglicht werden. Die Möglichkeit der Automatisierung von LPP stellt zudem einen wesentlichen Vorteil dar. Aufgrund dessen befasst sich die vorliegende Arbeit mit der Automatisierung von Linear Performance Pricing mittels Datenanalyse und präsentiert ein neues Konzept für einen automatisierten LPP Prozess.

Im ersten Schritt werden die Grundlagen der Datenanalyse sowie der linearen Regression im Detail beschrieben. Im wirtschaftlichen Abschnitt wird das Lieferantenmanagement ausgeführt sowie auf LPP eingegangen. Dabei werden drei unterschiedliche Methoden von LPP sowie praktische Anwendungen präsentiert. Auf Grundlage dessen werden im praktischen Teil dieser Arbeit diese Methodiken verglichen und ein neues Konzept für einen automatisierten LPP Prozess präsentiert. Zusätzlich wird der erstellte Prototyp erläutert.

Die Ergebnisse der vorliegenden Arbeit dienen als Grundlage für die Automatisierung eines LPP-Prozesses und veranschaulichen, wie die möglichen Schritte zur Automatisierung des Linear Performance Pricing erfolgen können und in welchen Bereichen weitere Forschungsarbeiten nötig sind.

Contents

Abstract iii

Kurzfassung iv

List of figures ix

1 Introduction 1
1.1 Motivation 1
1.2 Research methodology 3
1.3 Research question 4
1.4 Structure of this document 5

2 Data analytics 7
2.1 Definition 7
2.2 Data, variables, scales and measurements 8
2.2.1 Scales of data 8
2.2.2 Types of variables 10
2.3 General steps of data analytics 10
2.4 Feature selection 12

3 Linear regression 15
3.1 Definition 15
3.2 Steps 16
3.2.1 Step 1: Model formulation 17
3.2.2 Step 2: Estimation of the regression function 18
3.2.3 Step 3: Test regression function 21
3.2.4 Step 4: Test regression coefficients 27
3.2.5 Step 5: Test model assumptions 30
3.3 Dummy variables 38

3.4 Stepwise regression 39

4 Economical foundation 41

4.1 Supplier management 41

4.1.1 Introduction to supplier management 41

4.1.2 Supplier management process 43

4.1.3 Supplier development 47

4.2 Linear performance pricing (LPP) 50

4.2.1 Definition 50

4.2.2 LPP process according to Newman and Krehbiel 2007 51

4.2.3 LPP process according to Proch, Krampf and Schlüchtermann 2013 56

4.2.4 LPP process according to Verein Deutscher Ingenieure 2018 63

4.2.5 Practical applications of LPP 69

4.3 LPP alternatives 72

4.3.1 Non-linear Performance Pricing (NLPP) 72

4.3.2 Total cost of ownership (TCO) 73

4.3.3 Price structure analysis 73

5 Practical discussion 75

5.1 Review of manual LPP process steps 75

5.2 Levels of automation 80

5.3 Concept of a single level automated solution 81

5.3.1 Step 1: Product selection 83

5.3.2 Step 2: Data handling (gathering, cleaning & preparation, check) 84

5.3.3 Step 3: Perform pre-checks 85

5.3.4 Step 4: Calculate expected value 85

5.3.5 Step 5: Calculate market value and line 88

5.3.6 Step 6: Calculate best practice value and line 88

5.3.7 Step 7: Plan actions 89

5.3.8 Step 8: Implement actions 90

5.3.9 Step 9: Evaluate results 90

5.4 Automation concept of multiple level LPP 91

5.5 Quality criteria 93

5.6 Prototype 94
5.6.1 Functionality of the prototype 95
5.6.2 Testing of the prototype 99

6 Conclusion and outlook 100

Bibliography 104

List of Figures

1.1 Steps of research methodology 4

2.1 Example for data processing in an IoT platform 11
2.2 Big data process 11
2.3 Different methods for feature selection 12
2.4 Comparison between underfitting, good fitting, and overfitting of an model 14
2.5 Error vs complexity of a model 14

3.1 Steps of a linear regression 16
3.2 Simple scatter plot 19
3.3 Example plot of heteroscedasticity 33
3.4 Example plot of homoscedasticity 33
3.5 Autocorrelation 34
3.6 Detecting correlation using a scatter plot 37

4.1 Example representation of a market line 53
4.2 Example representation of a best practice line 54
4.3 Supplier strategy to move one product next to the best practice line 55
4.4 Proposed process steps 57
4.5 Proposed process steps 63
4.6 Example for graphical representation of different filters 68

5.1 Classification of LPP approaches 76
5.2 Steps of suggested automated single level LPP process 83
5.3 Steps of suggested automated multiple level LPP process . . . 91
5.4 Exemplary presentation of the best practice value and best practice line 97

5.5 Exemplary presentation of a correlation matrix between variables . . . 97
5.6 Exemplary presentation of a normal distribution of the residuals 98

1 Introduction

This chapter gives some introductory information concerning this thesis and starts with the motivation. Then the research methodology is presented as well as research questions, and the aims and objectives are clarified. Finally, the last part of the introduction gives an overview of the structure of this thesis.

1.1 Motivation

In the context of globalisation, digitalisation and innovation, companies have to face a more and more dynamic and rapidly changing economic environment. To stay competitive regarding costs, quality and quantity, companies have to adapt fast, stay flexible and be agile. (Helmold and Terry, 2016, p. 7) Due to differentiation of labour and specialisation, the dependence on suppliers, especially for industrial companies with a high rate of outsourced business processes or externally produced goods, is high. Therefore the part of material costs represents high costs within the product. These costs can be directly influenced by sourcing. (Arnolds et al., 2013, p. 13) Helmold and Terry (2016) state that companies must focus on their core competencies and build reliable, stable alliances with business partners and suppliers who cover the competencies and business processes that are outsourced. Obviously, the value added by lots of companies in many industries is decreasing and is usually only less than 50% (Jahns, E. Hartmann and Entchelmeier, 2007, p. 74).

According to Hendricks and Singhal (2009) a disturbance in the supply chain can reduce the stock price of a company by nearly 40% (Hendricks and Singhal, 2009, p. 51). Suppliers represent a critical resource in companies because of their input in the form of products or services. When they cannot fulfil the needs of the company, it could be beneficial to develop the capabilities of the suppliers because a change of suppliers is associated with high costs and time. Therefore the implementation of supplier development programs to improve them could be beneficial. (Modi and Mabert, 2007, p. 42)

All these aspects show that for a profitable, cost-effective product, not only product development and sales are fundamental, but also sourcing gets more crucial. Therefore sourcing strategies and supplier management obtain more and more significance. (Helmold and Terry, 2016, p. 7)

In order to gain an overview of price structures in complex material groups, a method named linear performance pricing can be applied (Verein Deutscher Ingenieure, 2018, p. 44). The goal of LPP is to compare the price and performance in order to reveal a potential for cost reduction (Gabath, 2008, p. 28). In this regard, linear performance pricing, with the focus on the price-performance ratio and the possibility of the comparability of products, is presented in the context of this master's thesis.

Due to the constantly increasing amount and availability of data, it gets more and more challenging to deal with this flood of data. Consequently, it is necessary to limit it somehow. (Bolón-Canedo, Sánchez-Maroño and Alonso-Betanzos, 2015, p. 1-2)

Currently, the buzzword data analytics is omnipresent, and many companies are collecting, storing and analysing big amounts of data. The reason for this is the potential value of data. However, only collecting and storing data does not really create a value for companies. Therefore the use of analytics becomes critical. (Watson, 2014, p. 1248) As data analytics is all about inspecting and evaluating data, the quality of the result is directly related to the quality of the collected data (Backhaus et al., 2018, p. 10).

1.2 Research methodology

This work was created with a four-step plan which is visualised in figure 1.1. The first step consists of creating the research idea and exploring the fundamentals. This includes an extensive literature investigation and finding related work. After these steps, a specific research purpose can be suggested, and research questions can be formulated. The first step is finalised by conceptional planning of the theoretical and practical parts.

The next step of the work contains the theoretical discussion of the research topic. This includes collecting the know-how about processes and understanding the methods and the domain. A more in-depth analysis of the theoretical foundations and methods are also performed.

The third step, which encloses the practical discussion, deals with the development of a software prototype. First, it is necessary to identify and evaluate the automation potential to create how LPP can be automated. After the concept phase, the planning, development and testing of the prototype can take place. The created prototype can then be used to analyse and interpret data.

The last step is all about finalising the work, which involves the evaluation and conclusion of the work. Then, finally, the master's thesis has to be written and presented.

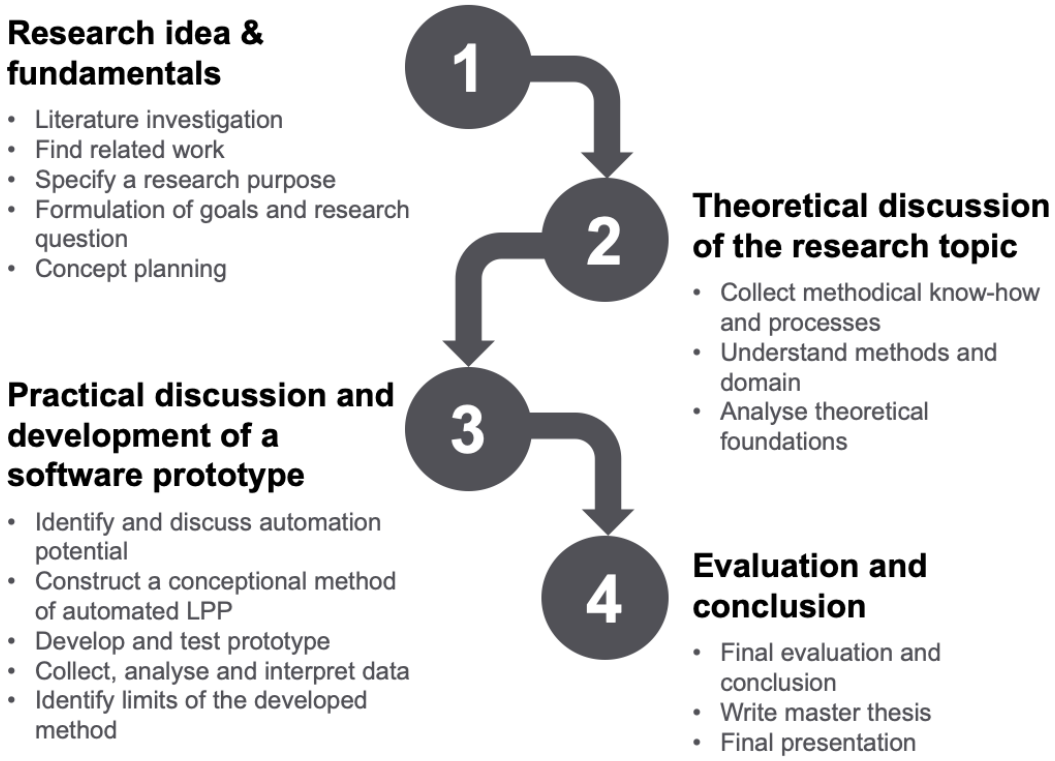

Research idea & fundamentals
• Literature investigation
• Find related work
• Specify a research purpose
• Formulation of goals and research question
• Concept planning
1
2
Theoretical discussion of the research topic
• Collect methodical know-how and processes
• Understand methods and domain
• Analyse theoretical foundations
3
Practical discussion and development of a software prototype
• Identify and discuss automation potential
• Construct a conceptional method of automated LPP
• Develop and test prototype
• Collect, analyse and interpret data
• Identify limits of the developed method
4
Evaluation and conclusion
• Final evaluation and conclusion
• Write master thesis
• Final presentation

Figure 1.1: Steps of research methodology
(own representation)

1.3 Research question

There are numerous possibilities to interpret the topic "Automated Linear Performance Pricing using data analytics". Therefore following research questions were formulated to get a more precise idea on which research topics, goals and aims the work is focused.

I. Which LPP approaches exist in the literature and do they already include steps of automation?

II. What are possible steps in order to automate linear performance pricing?

Aim and objectives

According to these research questions other questions raise up and therefore it is necessary to cover a broader expertise of the domain. Following objectives have to be fulfilled to gain a solid knowledge to answer the research questions:

- Build up domain knowledge of the economical background
- Collect methodical know-how about Linear Performance Pricing
- Identify and gather key mathematical knowledge to perform LPP
- Analyse the manual method of LPP and find out a way to automate these steps
- Observe and determine possible quality criteria for automated LPP
- Detect possible limitations and challenges of automated LPP

The title "automated Linear Performance Pricing using data analytics" contains a lot of popular buzzwords. So if separately looked at phrases in the title, like linear performance pricing and data analytics, there are numerous related works. This will extend dramatically if regression and some other mathematical foundations like correlation analysis or estimators within regression are also considered. However, it is also essential to have a look at the economic part like supplier management. This numeration shows that the thesis does not only have a specific focus on one subject. It is an interdisciplinary work that deals with a lot of different fields.

1.4 Structure of this document

This section describes the structure of the thesis and gives an overview of the chapters included. The work is divided into six main parts. After introducing the thesis in chapter 1 Introduction, the theoretical and mathematical foundations follow in chapter 2 Data analytics and 3 Linear regression. In the next chapter 4 Economical foundation supplier management and LPP are covered. Then the practical part is worked out in chapter 5 Practical discussion, and chapter 6 Conclusion and outlook presents the conclusion.

The second and third chapters consist of the theoretical and mathematical foundations of the work. Chapter 2 Data analytics deals with data analytics and first defines the term. Then it deals with variables, scales and measurements as well as general steps of data analytics. In the end, feature selection is discussed. In chapter 3 Linear regression, the fundamentals of linear regression, the steps included, dummy variables and stepwise regression are described.

The chapter 4 Economical foundation completes the theoretical part. In this chapter, all economic fundamentals are covered. On the one hand, this chapter deals with supplier management, and on the other hand, the basics of linear performance pricing are discussed. Therefore three approaches of LPP, as well as practical applications and alternatives to LPP, are presented.

Chapter five, named Practical discussion contains the practical part of the work. In the beginning, a review of the manual LPP concepts is given, and then the concept for an automated solution of LPP is presented. Furthermore quality criteria and the developed prototype are described.

The last chapter 6 Conclusion and outlook finalises the practical part and deals with final thoughts, and gives a conclusion and an outlook concerning possibilities and open points.

2 Data analytics

This chapter deals with the term data and its processing. As the topic data analytics is very extensive, this chapter will only give a rough overview of data analytics. First, a definition of data analytics is given in 2.1 Definition. Then, in 2.2 Data, variables, scales and measurements, the terms data, types of variables and scales of variables are explained. Next, in 2.3 General steps of data analytics an overview of the general process of data analytics is given. Finally, 2.4 Feature selection deals with the selection of the best features of a model.

2.1 Definition

Data analytics refers to many procedures performed on raw data to get valuable information out of it. Raw data itself does not have any value or beneficial information. The processes contain data cleansing, filtration, transformation etc. This means that through data analytics, raw data should be transformed into useful information which can be used as a basis for decision making. (Balali et al., 2020, p. 79)

There are many definitions of data quality, but in general, it can be said that data is considered high quality when it can be used for the intended operations, decision making processes or scheduled planning. The challenge is that the expectations and interpretations of data quality can vary depending on the user and across companies. (Fleckenstein and Fellows, 2018, p. 101)

2.2 Data, variables, scales and measurements

Data analytics, as the name suggests, is all about inspecting and evaluating data. So the quality of the result is directly related to the quality of the collected data. The quality of these data is depended on multiple factors like the scale of the data see section 2.2.1, how the data were collected, and the types of variables which are described in more detail in section 2.2.2. (Backhaus et al., 2018, p. 10)

2.2.1 Scales of data

Gathered data can be represented as numeric and non-numeric values. As already mentioned, the quality of data is depended on how these data were collected and how they were scaled. According to Backhaus et al. (2018), it can be distinguished between four kinds of measurements. The classification into the four scales was made because they offer a different content of information, and the application of arithmetic operations can also vary. It is possible to transform data from a larger scale to a lower scale, but not vice versa. This will be used to simplify data, but it comes with a loss of information. The four possibilities of classification are: (Backhaus et al., 2018, p. 10-12)

I. Nominal scale
II. Ordinal scale
III. Interval scale
IV. Ratio scale

I. Nominal scale

The nominal scale is the most primitive of all types, and it is just a classification of qualitative characteristics. Examples can be the gender (male - female - diverse), hair color (e.g. red - blond - brown - black) or nationalities (e.g. Austria - Germany - France - Italy). (Backhaus et al., 2018, p. 11)

At this level, no mathematical operations are possible. Just counting the number of occurrences is possible. (Backhaus et al., 2018, p. 11)

II. Ordinal scale

The next level is the ordinal scale. The values of this scale are like a ranking. Rank-based scaling can be used to see which F1 racer is faster than another F1 racer or which house is higher than another house, but it does not show how much a racer is faster or a house is higher. The distance between values in an ordinal scale can not be determined, so no arithmetic operations can be performed. (Backhaus et al., 2018, p. 11)

Just like with the nominal scale, it is possible to count the number of occurrences. In addition, it is possible to calculate the median or quantile (Backhaus et al., 2018, p. 12).

III. Interval level

The interval level is the first type where arithmetic operations are possible. It is divided into equal-sized segments. Therefore, the quality and quantity of information are better than from the ordinal scale because, in addition to the ranking system of the ordinal scale, the distance between data can be used as information. Examples can be the distance between two events in time or the temperatures of the morning compared to the evening. (Backhaus et al., 2018, p. 12)

Possible arithmetic operations are addition and subtraction. In addition to the possibilities of ordinal scale, the mean and variance can be calculated. (Backhaus et al., 2018, p. 12)

IV. Ratio scale

The ratio scale is the last and highest type. It contains the most information. In addition to the interval level, the ratio scale contains zero, adding another information level. Examples can be the length, weight, price, costs and so on. With variables in the ratio scale all arithmetic operations are possible. (Backhaus et al., 2018, p. 12)

2.2.2 Types of variables

Variables are numerical values that represent features of objects. These numerical values could be the results of calculations or measurements. Therefore it is necessary to classify variables into different types. Like the scale levels, the types of variables can be classified into multiple groups, which can differentiate between the content of information and how arithmetic operations can be handled. The most important classifications of variable types are metric vs. non-metric or quantitative vs. qualitative, or cardinal vs. categorial. Metric, quantitative and cardinal are used as synonyms, and non-metric, qualitative and categorial are used as well for the same kinds of variables. Nominal and ordinal scaled variables are assigned to categorical variables, whereas interval and ratio scaled variables correspond to a cardinal type of variables. Examples for categorical variables are the hair colour, nationality or gender of a person, and examples for cardinal variables are the age, height or weight of a person. (Backhaus et al., 2018, p. 11-13)

2.3 General steps of data analytics

The process of collecting and processing data is divided into five steps in the literature by Gandomi and Haider (2015) and Balali et al. (2020). Figure 2.1 shows that the first step is the creation of data followed by the collection of this data. These collected data are cleaned and transformed in order to analyse them in the next step. The result is interpreted in the last step in order to be able to make decisions. (Balali et al., 2020, p. 80)

In a slightly modified form, the process is performed as shown in figure 2.2 and is separated into two phases. The first phase of data management deals with the acquisition, cleaning, transformation and integration of data in order to be able to analyse and interpret it in the second phase. (Gandomi and Haider, 2015, p. 141)

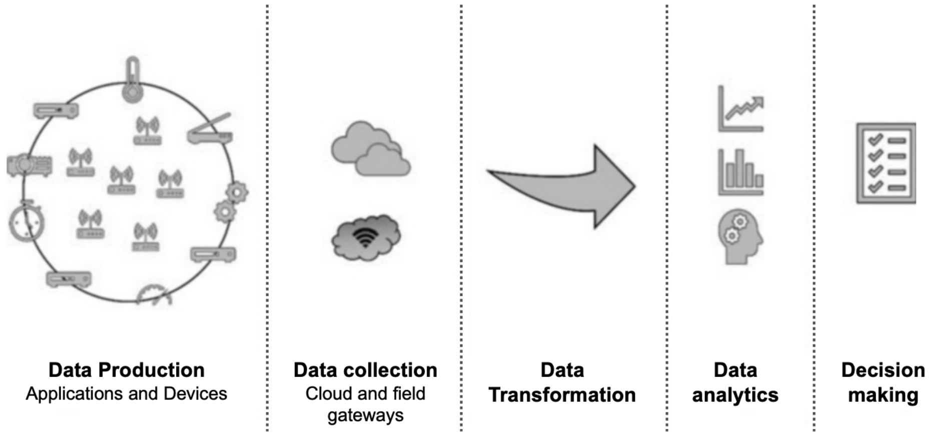


Figure 2.1: Example for data processing in an IoT platform (adapted from Balali et al., 2020, p. 80)

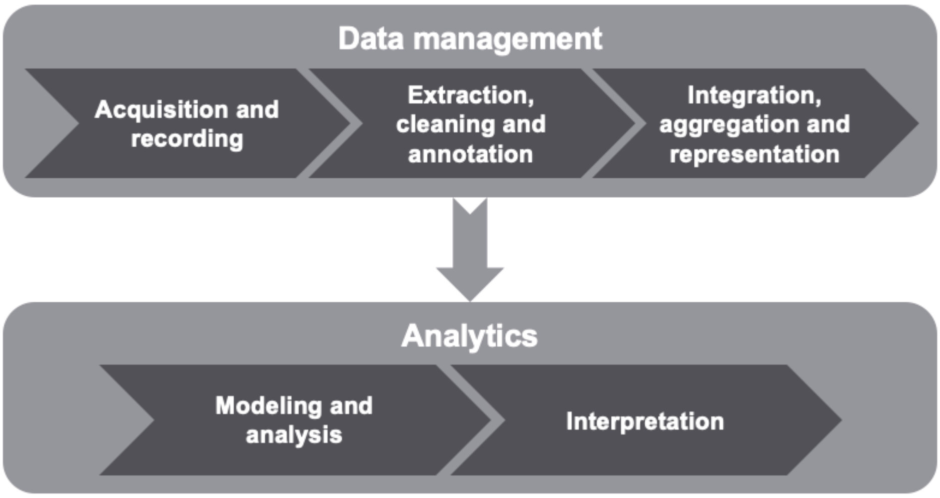
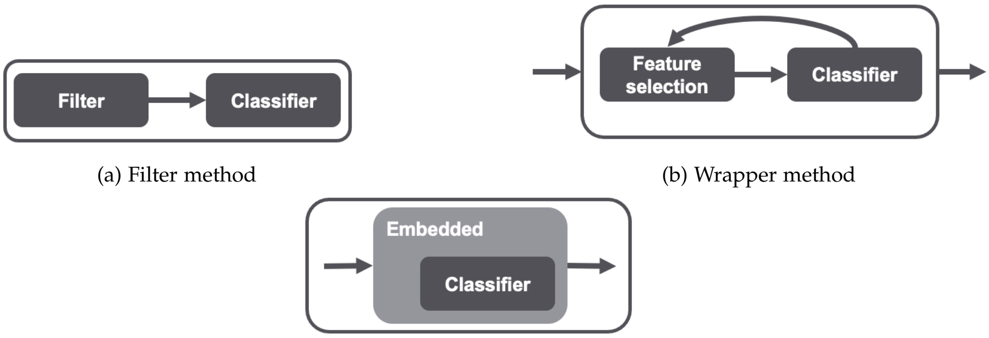

Data management
Acquisition and recording
Extraction, cleaning and annotation
Integration, aggregation and representation
Analytics
Modeling and analysis
Interpretation

Figure 2.2: Big data process by Gandomi and Haider, 2015, p. 141 (own representation)

As the amount of data is constantly increasing, it is necessary to limit it somehow. For this purpose, feature selection is needed only to include variables relevant to the problem. Feature selection is described in more detail in next chapter 2.4 Feature selection. (Bolón-Canedo, Sánchez-Maroño and Alonso-Betanzos, 2015, p. 1-2)

2.4 Feature selection

The term feature is used as a synonym for variables, and feature selection is the selection of variables to enable the best possible model. Another reason is to reduce the amount of data in order to increase performance or reduce storage requirements. Additionally, the effort to collect data is reduced. Feature selection can also help to visualise the data more efficiently and to understand the data better in a more general way. (Liebowitz, 2006, p. 2-5) It is therefore not practical to include all variables of a data set in the calculation. This leads, for example, to an unnecessary complexity of the model, reduced informative value of the model, increased effort of the calculation, or the calculation is sometimes no longer possible. For this reason, when using some algorithms, such as regression, forecasting, classification or clustering, only relevant variables should be included in the calculation. (Runkler, 2020, p. 85) There are different approaches for selecting variables or the best model, such as the filter, wrapper and embedded method (Liebowitz, 2006, p. 5).



(a) Filter method

(b) Wrapper method

(c) Embedded method

Figure 2.3: Different methods for feature selection by Bolón-Canedo, Sánchez-Maroño and Alonso-Betanzos, 2015, p. 17 (own representation)

The algorithms associated with the filter method are based only on the selection of variables from the data, are independent of the predictor and thus do not receive any feedback from it, as figure 2.3a shows. Thus this method tries to remove variables that are not useful for further processing.

Since this approach does not rely on feedback from the predictor and only selects variables once, it is typically the least computationally expensive approach. (Liebowitz, 2006, p. 89)

The wrapper method uses a learning algorithm to select the best variables. The selection of variables is made and checked using this learning algorithm. The learning algorithm calculates the quality of the selected variables without knowing the structure of the predictor. In contrast to the filter method, the wrapper method is iterative as shown in figure 2.3b and uses the output of the predictor to select variables. Since the learning algorithm is applied several times, this variant can be pretty time-consuming. (Bolón-Canedo, Sánchez-Maroño and Alonso-Betanzos, 2015, p. 25)

In contrast to the other two methods, the embedded method combines the feature selection and learning phases. This reduces the computational effort compared to the wrapper method. (Liebowitz, 2006, p. 137)

Greedy methods are the most common methods that can be linked to wrappers or embedded methods. There are different ways of implementing greedy methods, but they can generally be roughly divided into forward selection and backward elimination. With the forward selection, only the variable with the most significant influence on the result is selected at the beginning. Then, step by step, variables with the subsequent most significant influence on the result are added. (Liebowitz, 2006, p. 13) Backward elimination selects all variables first and then removes one variable with the minor influence on the result in each step (Atkinson and Riani, 2000, p. 2). Both variants are considered robust and fast (Liebowitz, 2006, p. 13). However, with both methods, it is also necessary to define termination criteria to get a suitable selection of variables. Both variants have their advantages and disadvantages. For this reason, it is also possible to combine the approaches to obtain a better result. (Verleysen, Rossi and François, 2009, p. 55-60)

After selecting the variables, it can lead to different complex models that describe the data better or worse. Figure 2.4 shows three different models that were run on the same data set. It can be seen that one of them does not

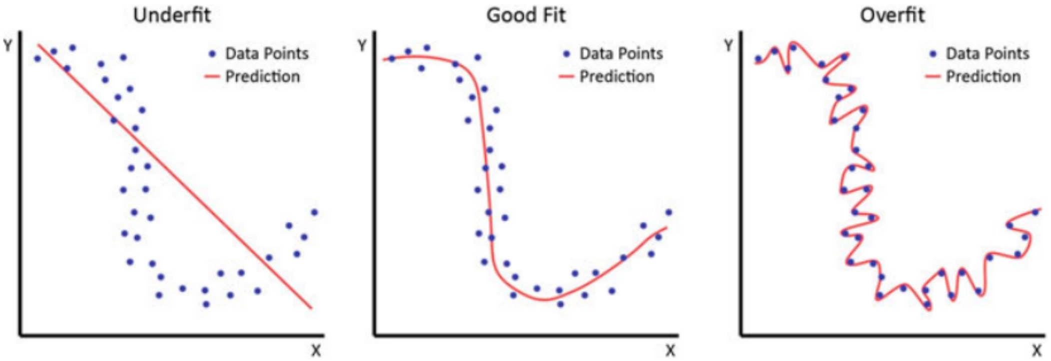


Figure 2.4: Comparison between underfitting, good fitting, and overfitting of an model (adapted from Balali et al., 2020, p. 191)

fit the data well enough. This is called underfitting. However, if the model is too complex and describes all data points because it fits the data very well, this is called overfitting. Both types are unwanted. (Balali et al., 2020, p. 120) Underfitting leads to high bias and low variance in a regression, which is described in chapter 3 Linear regression. Overfitting, on the other hand, leads to high variance but low bias. Figure 2.5 shows that it is necessary to find a balance between simple and complex models to achieve the best possible result. (Moreira, Carvalho and Horváth, 2018, p. 176-179)

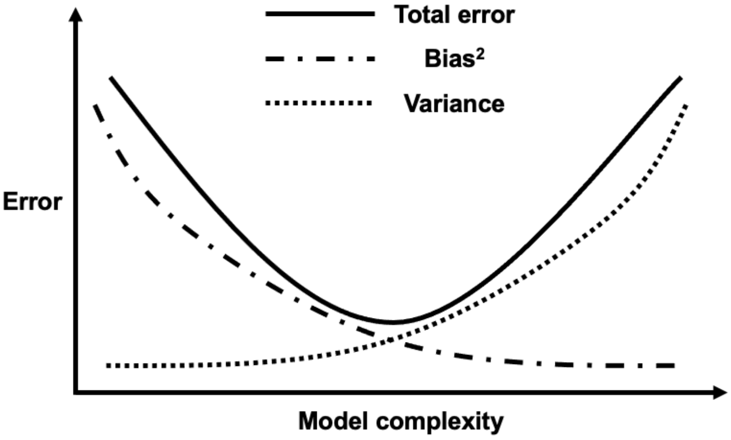


Figure 2.5: Error vs complexity of a model (adapted from Moreira, Carvalho and Horváth, 2018, p. 179)

3 Linear regression

This chapter deals with the mathematical basis of the work. In 3.1 Definition it is described what linear regression is, followed by section 3.2 Steps which describes the necessary steps to perform and test a linear regression. In section 3.3 Dummy variables the usage of categorical variables is shown and the last section 3.4 Stepwise regression deals with an extended version of linear regression.

3.1 Definition

Linear regression is an algorithm that can be used in many different ways. For example, it can describe and explain relationships of different variables or help perform predictions. According to Backhaus et al. (2018) regression and primarily linear regression is the most important and most commonly used of all multivariate statistical methods. It is possible to find and describe connections between one dependent variable and one or more independent variables with linear regression. These connections are represented in numbers, and so the connections are described relatively precisely. In other words, linear regression can describe the causal connection. It can also be used for making forecasts. (Backhaus et al., 2018, p. 16, 58)

As mentioned above, it is possible to have one or more independent variables, so there are two exceptional cases of linear regression. The simple linear regression only works with one independent variable (Backhaus et al., 2018, p. 58), and the multiple linear regression allows more than one independent variable. An elementary problem can be solved with simple linear regression, but there is more than one variable in most cases, so multiple linear regression has to be used. Multiple linear regression can also be used with just one independent variable, because it is just an extension of the simple linear regression. (Xia, 2014, p.807)

3.2 Steps

Linear regression can be divided into 5 steps as shown in figure 3.1. It starts with the formulation of a model followed by an estimation of the regression function and ends with different checks. The first check is to test the quality of the regression function. This test is used to verify how well the regression model explains the dependent variable Y. The next test, which examines the regression coefficients, determines the contribution of the variables used in the regression model for the dependent variables Y. This test shows if and how strong the variables influence the result. Finally the model assumption is tested. (Backhaus et al., 2018, p. 62-63, 74–75)

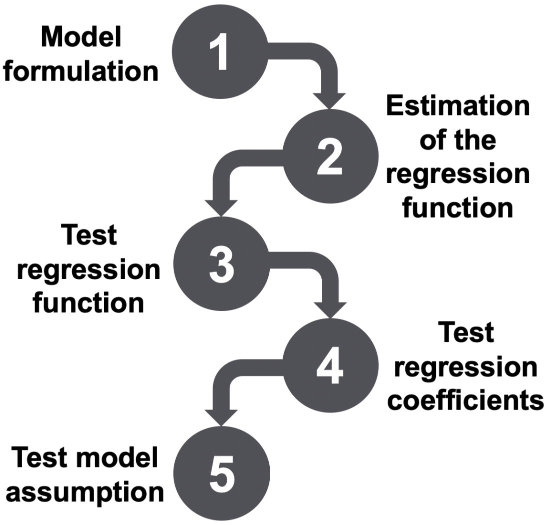


Figure 3.1: Steps of a linear regression
(own representation)

The steps shown in figure 3.1 are described in more detail in the next subchapters.

3.2.1 Step 1: Model formulation

The first step is to formulate a model. A model must contain all relevant aspects.

The model formulation starts with the creation of a mathematical model. Mathematical models are used to describe a system from reality with different mathematical concepts. This method can be used in multiple fields like natural science, medicine, economy and many more. All relevant aspects of the observed system have to be included in the mathematical model. It is essential to know that the model is becoming more complex, adding more aspects. The process is not to produce highly complex mathematical models, which describe the reality because this can lead to tricky or impractical model handling. That is the reason why it is necessary to generate a model which is as relevant and straightforward as possible for the use case. (Backhaus et al., 2018, p. 63)

There is one principle of William of Ockham called Occam's razor, also called the principle of parsimony. The principle assumes that if there are several explanations for a problem, the simplest theory is preferable. The simplicity of an explanation is achieved with as few variables as possible, few hypotheses and an understandable context. (Blumer et al., 1987, p. 377-380)

Equation (3.1) shows a model for a simple linear regression function: (Backhaus et al., 2018, p. 63)

$$\hat{Y} = b_0 + b_1 X \tag{3.1}$$

$\hat{Y}$ ... predicted dependent variable
b_0 ... intercept parameter or constant term
b_1 ... regression coefficient
X ... independent variable

Within a regression model, b_0 is treated like a constant factor which shows the value of Y if the variable X of the model is zero. This means the intersection between the regression curve and the Y-axis is marked by this constant factor. (Backhaus et al., 2018, p. 64)

b_1 is called the regression coefficient of the independent variable. It shows the influence of X on Y. In other words, it can determine how much Y is likely to change if X changes by one. (Backhaus et al., 2018, p. 64-65)

$$b_1 = \frac{\Delta \hat{Y}}{\Delta X}$$

A simple linear regression can only be used if just one variable is considered in the model. In general, most problems are more complex and need to be extended. In this case, a model for multiple linear regression is required. The equation (3.2) shows such a model. (Backhaus et al., 2018, p. 58-59)

$$\hat{Y} = b_0 + b_1 X_1 + ... + b_J X_J \tag{3.2}$$

$$\begin{aligned} \hat{Y} &\text{ ... predicted dependent variable} \\ b_0 &\text{ ... intercept parameter or constant term} \\ J &\text{ ... number of variables} \\ b_1, ..., b_J &\text{ ... regression coefficients} \\ X_1, ..., X_J &\text{ ... independent variables} \end{aligned}$$

After the model has been created, the parameters b_0, b_1, b_J have to be calculated. This process is explained in the next step. (Backhaus et al., 2018, p. 67)

3.2.2 Step 2: Estimation of the regression function

It is necessary to create a function that fits the data best, but figure 3.2 shows that it is not possible to draw a straight line through all points, and it is not enough to draw a line through the centre point to get a well-fitted regression line. The regression analysis also cannot create a straight line through all these points. The task is to search for a regression curve that fits the point distribution as well as possible. Therefore it is necessary to minimise the variances of all points as good as possible. (Backhaus et al., 2018, p. 67)

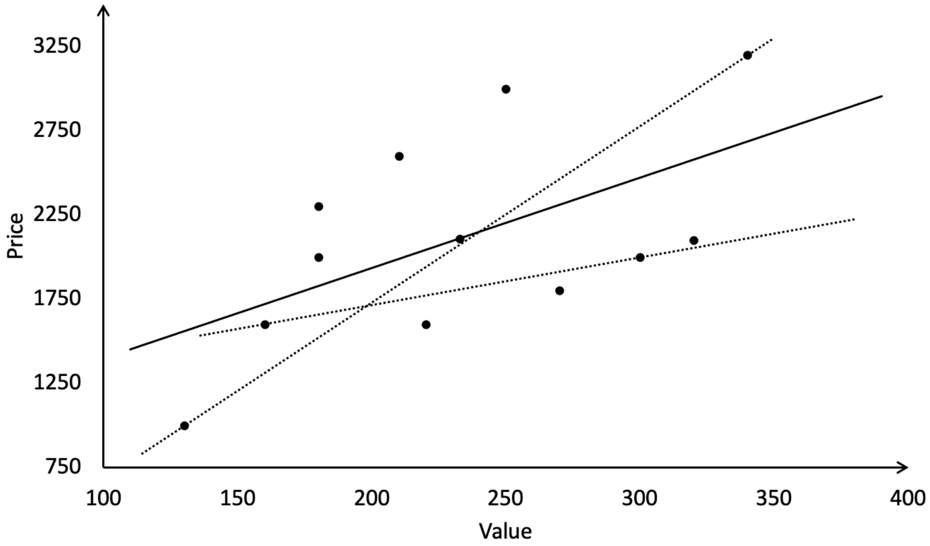


Figure 3.2: Simple scatter plot
(own representation)

The reasons why the points are not on a straight line are various. On the one hand, it is a linear regression which means that it is expected that the points are spread around the regression curve. On the other hand, not all variables that influence Y are included. Additionally, some variables are not able to be observed and included. Observation errors or measurement errors can also lead to bias or scattering of the values. (Backhaus et al., 2018, p. 67-68)

The term residuals describes the difference between the observed and the estimated value of Y. It is mathematically stated as follows: (Backhaus et al., 2018, p. 68)

$$e_k = y_k - \hat{y}_k$$

e_k ... residual of observation x_k
y_k ... real value of the dependent variable Y for x_k
$\hat{y}_k$ ... predicted value for Y for x_k
K ... number of observations
$k = 1, 2, ..., K$

In order to calculate Y mathematically, the estimated value $\hat{Y}$ must be summed with residual e. For a simple linear regression the equation looks like shown in equation (3.3) and equation (3.4) shows it for a multiple linear regression. (Backhaus et al., 2018, p. 68)

$$Y = \hat{Y} + e = b_0 + b_1 X + e \tag{3.3}$$

$$Y = \hat{Y} + e = b_0 + b_1 X_1 + ... + b_j X_j + e \tag{3.4}$$

Formula equation (3.5) illustrates the regression equation how it would look like for single observations of a simple linear regression (Backhaus et al., 2018, p. 69).

$$y_k = b_0 + b_1 x_k + e_k \tag{3.5}$$

The key to a good regression curve is to keep the residuals as small as possible. Therefore it is necessary to minimise the value of the residuals. It is now essential to find a suitable optimisation criterion so that positive and negative residuals do not eliminate each other. To achieve this, absolute values of the residuals (equation (3.6)) can be used or a squaring of the residuals (equation (3.7)) can be performed. (Backhaus et al., 2018, p. 70)

$$\sum_{k=1}^{K} |e_k| \rightarrow min! \tag{3.6}$$

$$\sum_{k=1}^{K} e_k^2 \rightarrow min! \tag{3.7}$$

The equation (3.7) shows the least square estimator (Backhaus et al., 2018, p. 70-71). It is one of the most used and important statistical methods to estimate parameters. However, it is very sensitive to the outliers, so it is important to try to detect such outliers. (Yu and Yao, 2017, p. 2) There are other numerous estimators which have various advantages and disadvantages (Backhaus et al., 2018, p. 70-71).

Identifying suitable regression parameters ($b_0, b_1, ..., b_J$) and satisfying criterion equation (3.7) requires a significant amount of effort. This can be done by using equation (3.7) and equation (3.3) or equation (3.4). (Backhaus et al., 2018, p. 72)

A further challenge is to include categorical values. This is discussed in chapter 3.3 Dummy variables described.

3.2.3 Step 3: Test regression function

Several tests have to be performed, and different quality criteria have to be evaluated to determine the quality of the regression function. The tests to verify the regression function reveal whether the regression model explains the dependent variable Y and, if so, how well the model explains Y. In order to verify the regression function, the sum of squared residuals SSR, coefficient of determination R^2, adjusted coefficient of determination R^2_{adj}, F-statistics or the standard error of the estimation can be used. (Backhaus et al., 2018, p. 74)

Step 3 / Test 1: coefficient of determination R^2

The value of R^2 indicates how much of the total variance can be explained by the variables included in the regression and range from 0 to 1. E.g. if R^2 has the value 0.7, 70% of the total variance can be explained, and 30% remains unexplained. In other words, the higher the value R^2 is, the better, and the more of the variance can be explained. (Wooldridge, 2013, p. 38)

R^2 is based on the regression variance and therefore on the sum of squared residuals SSR. In general, the smaller SSR is, the better the model fits the data, also called goodness-of-fit. However, the reliability of the SSR value is only given for exactly the specific data set since any change in the data set or concerning the variables causes that the SSR value is no longer reliable. Therefore, the SSR value is not used directly as a quality criterion but is a factor in other criteria such as the coefficient of determination. (Backhaus et al., 2018, p. 75)

In contrast to the SSR value, the R^2 value can be used to compare different datasets and partly also different models (Backhaus et al., 2018, p. 78).

The coefficient of determination can be calculated using equation (3.8) or equation (3.9) (Wooldridge, 2013, p. 37-38).

$$
\begin{aligned}
R^2 &= 1 - \frac{\text{not explained variance}}{\text{total variance}} \\
&= 1 - \frac{SSR}{SST} \\
&= 1 - \frac{\sum\limits_{k=1}^{K} e_k^2}{\sum\limits_{k=1}^{K} (y_k - \overline{y})^2}
\end{aligned}
\tag{3.8}
$$

$$
\begin{aligned}
R^2 &= \frac{\text{explained variance}}{\text{total variance}} \\
&= \frac{SSE}{SST} \\
&= \frac{\sum\limits_{k=1}^{K} (\hat{y}_k - \overline{y})^2}{\sum\limits_{k=1}^{K} (y_k - \overline{y})^2}
\end{aligned} \tag{3.9}
$$

y_k ... real value of the dependent variable Y for x_k
$\hat{y}_k$ ... predicted value for Y for x_k
$\overline{y}$ ... average value of Y
K ... number of observations
$k = 1, 2, ..., K$

An essential fact about R^2 is that it never decreases, and it usually increases when another independent variable is added to a regression (Backhaus et al., 2018, p. 78).

The coefficient of determination can be used as one of many quality criteria for linear regression, but there is a drawback. The complexity of a model is not taken into account in R^2. It can result in a better fit (higher R^2) for a complex model with multiple variables, even if the estimated values do not improve significantly. (Wooldridge, 2013, p. 80) The reason for this is that each parameter included in the regression consumes one degree of freedom. Thus, the number of degrees of freedom $df = K - J - 1$ for the estimation is reduced as soon as the number J of independent variables increases. (Backhaus et al., 2018, p. 78) Simplified SSR never decreases. It can only increase if another variable is added to the model. (Wooldridge, 2013, p. 80) Therefore it is necessary to find a way for a correction or adjustment of the coefficient of determination. This method is called adjusted R^2 or R^2_{adj} and is shown in equation (3.10). (Backhaus et al., 2018, p. 78-79)

$$R_{adj}^2 = R^2 - \frac{J \cdot (1 - R^2)}{K - J - 1} \tag{3.10}$$

R_{adj}^2 ... adjusted coefficient of determination
K ... number of observations
J ... number of explanatory variables / independent variables
$K - J - 1$ ... degree of freedom

The equation shows that when the number of independent variables increases, R_{adj}^2 decreases. It can thus be said that a high model complexity is penalised by this adjustment. Through this correction, the adjusted coefficient of determination may be only equal to or smaller than the simple coefficient of determination. However, R_{adj}^2 can have a negative value due to this correction. As with the simple coefficient of determination, a maximum value of 1 can be achieved, and the higher the value, the better. (Backhaus et al., 2018, p. 78-79)

Step 3 / Test 2: F-statistics

The quality measure of the coefficient of determination of regression only indicates the quality of the regression function compared to the observed data. Therefore, it is only a measure of quality for a specific sample. In addition, the quality of the regression function beyond this sample should also be checked. In order to verify the validity of the population, on the one hand, the representativeness of the sample (R^2) and, on the other hand, the significance of the estimated model has to be evaluated. A test of significance can be performed using F-statistics. (Backhaus et al., 2018, p. 79)

Linear regression is based on the assumption that a causal relationship exists between the dependent variable Y and the independent variables X_j. So the valid values of the regression coefficients β_j have to be non-zero. Therefore it is necessary to refuse the null hypothesis H_0 - see equation (3.11). The test of the null hypothesis can be done by using an overall F-test. (Backhaus et al., 2018, p. 80)

$$H_0 : \beta_1 = \beta_2 = ... = \beta_J = 0 \tag{3.11}$$

The F-statistic includes different values such as various variances, the sample size K and the number of independent variables J. The implementation of the F-test is divided into four steps, which are explained in more detail below. (Backhaus et al., 2018, p. 81)

In the beginning, the empirical F-value F_{emp} has to be calculated by applying the formula equation (3.12) or equation (3.13) (Backhaus et al., 2018, p. 81).

$$\begin{aligned} F_{emp} &= \frac{\text{explained variance}/J}{\text{not explained variance}/(K-J-1)} \\ &= \frac{\sum\limits_{k=1}^{K} (\hat{y}_k - \overline{y})^2/J}{\sum\limits_{k=1}^{K} (y_k - \hat{y}_k)^2/(K-J-1)} \end{aligned} \tag{3.12}$$

$$F_{emp} = \frac{R^2/J}{(1-R^2)/(K-J-1)} \tag{3.13}$$

The next step is to specify a significance level α in the range between zero and one and indicates how likely the null hypothesis is to be rejected, even if it would be correct. E.g. with a value of 0.05, H_0 is rejected 5% of the time even although it should not be rejected. The probability of confidence $1 - \alpha$ indicates the likelihood of the null hypothesis being correctly rejected. The null hypothesis is correctly rejected in 95% of all cases in this example. α can be seen as an error tolerance. (Backhaus et al., 2018, p. 81)

In the third step, the theoretical F-value F_{tab} is taken from an F-table using the values of J and $(K - J - 1)$. Each confidence probability has its F-table. This extracted value is needed in the next step to compare with the calculated value F_{emp}. (Backhaus et al., 2018, p. 81)

In the fourth and last step, the found theoretical F-value (F_{tab}) is compared with the calculated empirical F-value (F_{emp}). If $F_{emp} > F_{tab}$ holds, the null hypothesis must be rejected (equation (3.14)). In this case, it can be assumed that a relationship between the dependent and independent variables is statistically significant. However, if $F_{emp} \leq F_{tab}$ holds, then the null hypothesis cannot be rejected (equation (3.15)). Therefore it cannot be confirmed that there is a statistically significant relationship between the dependent and independent variables. This does not mean there is no correlation. According to the result, the lack of correlation can occur if the sample size is too small and other random influences hide the influence of the variables. Another possibility is that those relevant variables are not added to the model, and consequently, the explained variance is too low. (Backhaus et al., 2018, p. 81-82)

$$F_{emp} > F_{tab} \text{ ... reject } H_0 \rightarrow \text{correlation is significant} \tag{3.14}$$

$$F_{emp} \leq F_{tab} \text{ ... don't reject } H_0 \tag{3.15}$$

p-Value of the F-test

The classical F-test is a bit unhandy with the use of such a table, and only a yes or no decision can be made. Therefore the F-test can be performed using a so-called p-value (in Latin "probabilitas"). If the p-value of the F-statistic is used, it is not necessary to work with such a table. As with the classical F-test, a significance level α also needs to be selected. The significance level is then compared with the p-value, and if $p < \alpha$ applies, then the null hypothesis can be rejected. (Backhaus et al., 2018, p. 83)

The p-value is always in the range between 0 and 1. The p-value is an easier way to interpret the F-statistics and has higher information content.

Using the p-value makes it even possible to make a statement without first selecting an α. In addition, the classical F-test only states whether H_0 is to be rejected or not. It is possible to determine how close this value is to α by using the p-value. In contrast to the classical F-test, it is possible to make not only a yes or no decision. (Backhaus et al., 2018, p. 83-84)

Step 3 / Test 3: Standard error of the estimation

The last quality benchmark to check the regression function is the standard error of the estimation. The standard error describes the average error when estimating the dependent variable Y. It is calculated by the square root of SSR divided by the number of degrees of freedom. This is calculated from the square root of SSR divided by the number of degrees of freedom, see equation (3.16). (Backhaus et al., 2018, p. 84)

$$s = \sqrt{\frac{\sum e_k^2}{K - J - 1}} \tag{3.16}$$

The standard error can be compared with the mean $\overline{y}$ and from this a percentage estimate of the standard error is obtained. This makes it possible to decide whether the value is good (low % value) or bad (high % value) - see equation (3.17). (Backhaus et al., 2018, p. 84)

$$\frac{100}{\overline{y} \cdot s} \tag{3.17}$$

3.2.4 Step 4: Test regression coefficients

After the global check of the regression has been performed, the individual coefficients can be checked. This can be done by checking the t-statistics and calculating the confidence intervals of the regression coefficients. (Backhaus et al., 2018, p. 84, 88)

Step 4 / Test 1: t-statistics

The test using t-statistics checks whether the null hypothesis can be rejected or not and is similar to the F-test. Compared to the F-test, which tests several variables, the t-test only includes one variable, and therefore the null hypothesis is $H_0 : \beta_j = 0$. A suitable test criterion to perform the t-statistic is equation (3.18). (Wooldridge, 2013, p. 121-122)

$$t_{emp} = \frac{b_j}{s_{bj}} \tag{3.18}$$

t_{emp} ... empirical t-value of the independent variable
b_j ... calculated / estimated regression coefficient
s_{bj} ... standard error of the regression coefficient b_j

The t-test is based on the same four steps as the F-test, which has previously been described. First, the empirical t-value is calculated with the equation (3.18). After that, a significance level α is set, then the theoretical t-value t_{tab} is searched in a table using the number of degrees of freedom $K - J - 1$ and $1 - \alpha$. In the last step, the theoretical t-value is compared with the empirical t-value. For example, the following applies when comparing the two values: (Backhaus et al., 2018, p. 86)

$|t_{emp}| > t_{tab}$ ... reject $H_0 \rightarrow$ correlation is significant
$|t_{emp}| \leq t_{tab}$ ... don't reject H_0

Like the F-test, the t-test can also be performed using a p-value. This has the advantage that the value p can be interpreted immediately and does not have to be compared with a value from a table. With the t-test using a p-value, the assumption that the null hypothesis can be rejected at $p < \alpha$ can also be applied. (Backhaus et al., 2018, p. 87)

Step 4 / Test 2: Confidence interval of the regression coefficient

The real value of the regression coefficients β_j is not given and cannot exactly be calculated or estimated. Nevertheless, it is possible to form a confidence interval to estimate a possible range for β_j. The basis for this estimate is the already calculated regression coefficient b_j. A range is now formed around b_j in which the real regression coefficient β_j is estimated. This area is formed by using the standard error of the regression coefficient and the previously determined t-value. In equation (3.19) it is described how the range around the real regression coefficient β_j is formed. (Backhaus et al., 2018, p. 88)

$$b_j - t \cdot s_{b_j} \leq \beta_j \leq b_j + t \cdot s_{b_j} \tag{3.19}$$

β_j ... real regression coefficient which is unknown"

b_j ... calculated / estimated regression coefficient

t_{tab} ... t-value of the t-statistic from the t-table

s_{b_j} ... standard error of the regression coefficient b_j

This equation can be used to determine the range of the true regression coefficient β_j with the used confidence probability $1 - \alpha$, which is used for calculating the t-statistic. The size of this range is also of significant interest in terms of the quality of regression because a smaller range expresses a better estimation of the calculated regression coefficient b_j. However, when the range is huge, it contains much uncertainty, and if there is also a change in sign, b_j should be checked very closely as this indicates poor quality. For example if the range of the real regression coefficient looks like $-1 \leq \beta_j \leq 2$. (Backhaus et al., 2018, p. 89)

3.2.5 Step 5: Test model assumptions

In the previously described steps 1 to 4, the assumption has been made that all prerequisites for the regression calculation and the tests executed are fulfilled. This has only been assumed and has to be proven in this step. Since a linear regression has been carried out in Step 2: Estimation of the regression function using the linear squared estimator, only the requirements for the linear squared estimator and not the requirements for any other possible estimators are discussed in this section. In order to complete the verification, it is essential to ensure that six conditions are fulfilled in order to achieve the best possible and optimal result. (Backhaus et al., 2018, p. 90) If these conditions are not given, unforeseen problems and biases can occur (Stoetzer, 2017, p. 133-203). It is also beneficial if another seventh assumption is fulfilled because this is advantageous and relevant when performing the significance test: (Backhaus et al., 2018, p. 91)

1. correct specified model
2. expected value of the error term not equal to zero
3. homoscedasticity
4. no autocorrelation of error terms u_k
5. no correlation between the explaining variables and the error term
6. no multicollinearity
7. normal distribution of the error terms u_k

The assumptions are separated into different categories. A general check of the model hypothesis is performed by using assumption 1. Additionally, assumptions 2, 3, 4, 5 and 7 are used to check the error terms and residuals. Finally, the explanatory variables are checked by performing assumptions 5 and 6. (Backhaus et al., 2018, p. 90)

Step 5 / Assumption 1: Correct specified model

The first assumption is universal. The basic idea is that a model is correctly specified. A model is correctly specified if: (Stoetzer, 2017, p. 182)

- it is linear in the parameters β_0 to β_j
- there is no interdependence between the variables x_j used to estimate the variable Y
- all relevant variables have been included in the model
- the number of parameters to be estimated is smaller than the number of data sets $J + 1 < K$
- irrelevant variables are not included if possible
- errors in the data are avoided as best as possible

The consequences of the case that these points are not fulfilled are numerous and can lead to biased or inconsistent estimation, excessive or unpredictable variance or loss of efficiency (Stoetzer, 2017, p. 183-203).

As these points are very wide-ranged and an explanation, analysis and description of the consequences and possibilities for detection would go beyond the scope, the reader is referred to other documents such Wooldridge (2013), Stoetzer (2017) and Backhaus et al. (2018).

Step 5 / Assumption 2: Expected value of the error term not equal to zero

Suppose the model is correctly specified as assumed under assumption 1. In that case, only random factors are included in the error term u, which have a positive and negative influence on the estimated values $\hat{Y}$. Assumption 2 assumes that these influences neutralise each other and result in a total of zero. (Verein Deutscher Ingenieure, 2018, p. 24) The reason for such positive and negative variances can be measurement errors. (Backhaus et al., 2018, p. 93)

The consequence of not satisfying this assumption can be that the constant regression coefficient b_0 is biased. However, a bias of b_0 does not cause a problem in every situation because the other regression coefficients b_j are not biased, and the weighting of these is still the same. The only difference is that the measurement error shifts the entire regression function. However, suppose the regression does not contain a constant term. In that case, this measurement error is included in the other regression coefficients b_j and can lead to a bias of these coefficients. So the angle of the regression line also changes. (Backhaus et al., 2018, p. 93)

Step 5 / Assumption 3: Homoscedasticity

Homoscedasticity means that the residuals are randomly distributed and do not follow a certain pattern or structure. When the variance of the residuals depends on a variable, it is called heteroscedasticity. (Stoetzer, 2017, p. 135)

In the case of homoscedasticity, the least square estimator is still unbiased and consistent, but the occurrence of heteroscedasticity means that the estimation is no longer efficient. In other words, the t-values are unreliable because the efficiency of estimating the standard errors is limited. This also reduces the reliability of estimating the significance of a coefficient estimation. The reason for this problem is that the residuals e are squared, and if the deviations are getting bigger, the estimation becomes inaccurate. (Stoetzer, 2017, p. 135-136)

Various tests can be performed to check for homoscedasticity or heteroscedasticity. It is possible to test visually by comparing the residuals and the dependent variables Y_k. Figure 3.3 shows a possible representation of heteroscedasticity. Figure 3.4 shows a constant dispersion of the residuals, and this means there is a homoscedasticity. (Backhaus et al., 2018, p. 95) Mathematically several tests can be used to identify homoscedasticity or heteroscedasticity. Possible tests for this are the Goldfeld-Quandt, Glesjer (Backhaus et al., 2018, p. 95-96), Breusch-Pagan or White test (Wooldridge, 2013, p. 296).

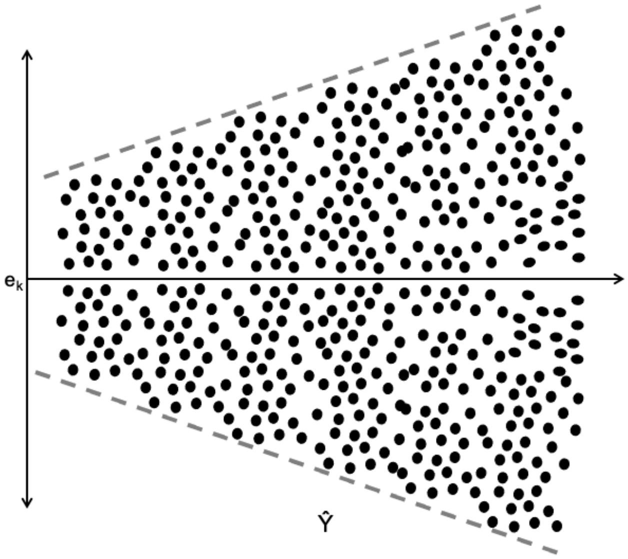


Figure 3.3: Example plot of heteroscedasticity by Backhaus et al., 2018, p. 95 (own representation)

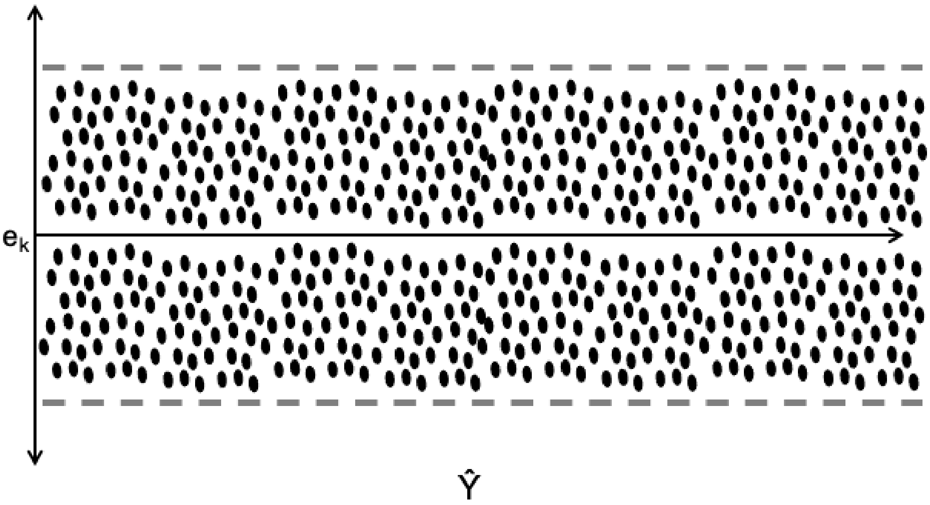


Figure 3.4: Example plot of homoscedasticity by Stoetzer, 2017, p. 137-138 (own representation)

Step 5 / Assumption 4: Autocorrelation

The assumption of autocorrelation implies that the error terms u_k are uncorrelated, see equation (3.20). If this is not the case, the observations are not in a random order. Figure 3.5 shows a case of autocorrelation. (Backhaus et al., 2018, p. 96)

$$Cov(u_k, x_{k+r}) = 0 \text{ where } r \neq 0 \tag{3.20}$$

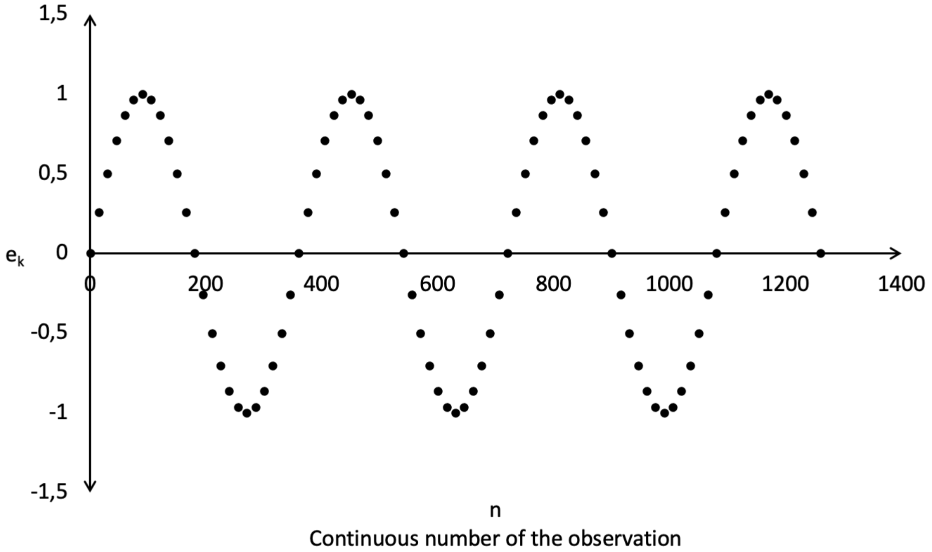


Figure 3.5: Autocorrelation by Stoetzer, 2017, p. 148 (own representation)

The effect of autocorrelation is, as with heteroscedasticity, once again, that the estimation of the standard deviation is too large or small. The values of the coefficients are not biased in this case and remain consistent. Hence the t-values are again unreliable. (Stoetzer, 2017, p. 149)

To find autocorrelation, in a visual inspection, the residuals of the different independent variables can be plotted against the sequence number k of the data as shown in figure 3.5 (Stoetzer, 2017, p. 149). Furthermore, autocorrelation can be statistically examined using the Breusch-Godfrey test, the Durbin-Watson test or with an extended version the Durbin-h test (Wooldridge, 2013, p. 422).

Step 5 / Assumption 5: No correlation between the explaining variables and the error term

The assumption $Cov(u_k, x_{jk}) = 0$ is partially covered by assumption 1 - selection of the correct variables. But it is not always possible to include all relevant variables in the model. Examples for missing variables can be too much effort to collect data or problems while collecting the desired variables. As a result, one part of assumption one is violated, which can lead to biased estimations. But by assuming that there is no correlation between the variables x_{jk} and the error term u_k, the partial violation of assumption one becomes less relevant under this special condition. The reason for this is that the missing variables only lead to a partial bias of b_0. This corresponds to a constant measurement error and therefore does not disturb the regression coefficients b_j. The detection of such a problem is analogous to assumption 2, except that now a correlation between the error terms and the variables is checked. (Stoetzer, 2017, p. 189-191)

Step 5 / Assumption 6: No multicollinearity

Multicollinearity in a linear regression analysis means that the independent variables are linearly dependent on each other. There are two cases of collinearity. The first type is that one variable is dependent on another variable, such as that the interest rate on a home loan is dependent on the Euribor interest rate. In the second case, one variable may depend on several other variables. An example can be if the variables length, width, height and the volume of an object are included in the regression, where volume is dependent on the other three variables. (Shikhman and Müller, 2021, p. 116-117)

There is also a distinction between weak, strong, extreme and perfect multicollinearity. These various characteristics can lead to different interpretations of the results of linear regression. (Stoetzer, 2017, p. 161-162)

Weak multicollinearity does not lead to a problem since a low dependence of the variables is normal. However, from strong multicollinearity onwards, the significance of the regression coefficients b_j and various checks performed on the regression, such as the F test, are limited. This increases the standard error of the coefficients and, therefore, to losing or reducing the significance of the hypothesis tests and confidence intervals for the independent variable. The term extreme multicollinearity is used when a calculation is only possible with significant effort and when small changes in the data or variables have a significant influence on the estimated regression coefficients. Extreme multicollinearity can lead to a result that is difficult or even impossible to interpret. Perfect multicollinearity has the consequence that the calculation of coefficients of the independent variables is impossible. (Stoetzer, 2017, p. 161-162)

In the case of multicollinearity, it can also happen that a significant value for the coefficient of determination R^2 is found, although none of the regression coefficients indicates significance. Other problems can be that important independent variables are insignificant because the standard errors of these variables are high, or another problem could be that a small change in the data can lead to a significant change of the regression coefficients. (Pochiraju and Kollipara, 2019, p. 217-218)

In order to detect collinearity, a correlation matrix can be created (Stoetzer, 2017, p. 162) or two variables can be compared in pairs in a graph. An example is shown in figure 3.6. By using these methods, it is possible to detect only collinearities between two variables. (Verein Deutscher Ingenieure, 2018, p. 17) In order to identify multicollinearity, a regression can be performed for each independent variable $y_j = x_j$ using the other independent variables without x_j. Moreover, it is essential to check for a high R_j^2, which corresponds to a high level of collinearity. (Wooldridge, 2013, p. 95) Furthermore, a variance inflation factor (VIF) or the procedure of the condition number can also be applied for detecting potential multicollinearity (Stoetzer, 2017, p. 162-163), but it is difficult to define an absolute value where multicollinearity becomes a problem (Wooldridge, 2013, p. 95).

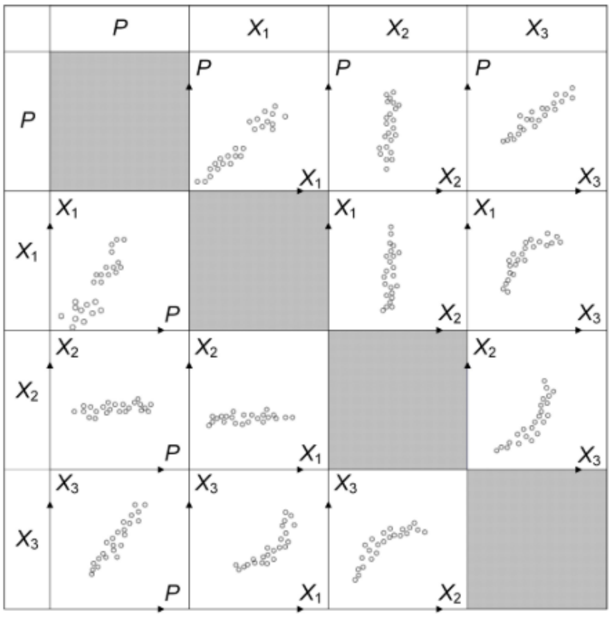


Figure 3.6: Detecting correlation using a scatter plot (adapted from Verein Deutscher Ingenieure, 2018, p. 19)

Step 5 / Assumption 7: Normal distribution of error terms u_k

The assumption of a normal distribution of the error terms is not a prerequisite for a linear regression using the least square estimator. The normal distribution is of interest concerning the hypothesis tests (F-test, t-test). In these tests, it is assumed that the regression coefficients b_0 to b_j have a normal distribution, and therefore also the error terms are normally distributed. If this is not the case, the hypothesis test has no significance, so the test is invalid. The calculation of the regression coefficients is consistent and unbiased even without fulfilling the assumption of a normal distribution. (Stoetzer, 2017, p. 153-154)

The check for the presence of a normal distribution can be done using a quantile plot or a graphical check of the residuals or variables via a histogram. A visual check can only be determined whether there is a substantial deviation from the normal distribution. Similar to the other assumptions, a better determination of a violation of the normal distribution can be established by applying mathematical tests. According to Stoetzer (2017) the Shaprio-Wilk and Kolmogorov-Smirnov tests are widely used for verification. (Stoetzer, 2017, p. 154)

3.3 Dummy variables

As already described in 2.2 Data, variables, scales and measurements variables and data can be classified in different ways. The simplest classification is between categorical and non-categorical. Non-categorical variables can be used directly in regression, but it is not possible to use categorical variables directly in a regression. To use categorical variables in a regression, they must be converted into so-called dummy variables. (Stoetzer, 2020, p. 21-22, 30) A dummy variable can be either 0 or 1. The conversion of a categorical variable into dummy variables is done by creating a separate dummy variable for each category, which is set to the value 1. However, it is possible to save one dummy variable because this can be represented by all dummy variables set to 0. This means that $n-1$ dummy variables exist for n categories. (Berger, Maurer and Celli, 2018, p. 518)

In order to understand this better, an example is given. A category variable of a product could be the material. Possible materials of the product are copper, zinc and iron. Therefore, two dummy variables X_1 and X_2 are created. If $X_1 = 1$ and $X_2 = 0$ then it is iron, if $X_1 = 0$ and $X_2 = 1$ it is zinc and copper is represented by $X_1 = X_2 = 0$. A regression equation for this example is shown in equation (3.21) and a tabular represenation is shown in table 3.1. (Berger, Maurer and Celli, 2018, p. 518)

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 \tag{3.21}$$

Categorical representation	Cardinal representation	
Product material	X_1	X_2
Copper	0	0
Zinc	0	1
Iron	1	0

Table 3.1: Example of an dummy variable - product material

3.4 Stepwise regression

Stepwise regression is an extended form of linear regression. This form of regression is a simple method that is broadly accepted by analysts. (Hwang and Hu, 2015, p. 1794) In a conventional linear regression, it is assumed that the variable selection has already been performed in advance and that the regression is only performed once. If this is the case, the approach can be called a filter method, which is described in 2.4 Feature selection. Because this does not always reflect the natural world and a selection of the variables should be performed by an algorithm, a so-called stepwise regression is used. (Jank, 2011, p. 107-111) Stepwise regression is separated into the selection of variables, the execution of the regression and the evaluation of the result (Heiberger and Holland, 2015, p. 297-298). Based on this procedure, it is evident that this is an iterative algorithm with a learning phase. Therefore, this can be assigned to the wrapper or embedded method of feature selection, which are described in 2.4 Feature selection. (Liebowitz, 2006, p. 6)

There are different ways to perform stepwise regression, but all of them are based on forward selection, backward elimination or a hybrid form of feature selection. As already mentioned in 2.4 Feature selection, the different variants add or remove variables step by step until a stopping condition is given. There are different approaches in the literature to select which variables are added or removed or what the stopping condition looks like. This is further described by Żogała-Siudem and Jaroszewicz (2020), Hwang and Hu (2015), Campobasso and Fanizzi (2012) and many others. (Heiberger and Holland, 2015, p. 297) A simple way to perform a stepwise regression using backward elimination is by evaluating the p-values of the t-statistics. For this purpose, a regression with all variables is performed at the beginning. Then the variable with the least significant p-value is removed, and a new regression is performed. This is done until no p-value of a variable exceeds a previously specified threshold. (Jank, 2011, p. 108)

4 Economical foundation

This chapter covers, on the one hand, supplier management in section 4.1 Supplier management and gives a deeper insight into the process of supplier management with a focus on supplier development. On the other hand, this chapter focuses on linear performance pricing, which is described in section 4.2 Linear performance pricing (LPP). After providing a definition and keywords of LPP, three LPP approaches are discussed. Then practical applications of LPP are given. Finally, to complete this chapter, alternatives to LPP are presented in section 4.3 LPP alternatives.

4.1 Supplier management

To understand the meaning of the term supplier management, first, a definition and goals of supplier management are presented in section 4.1.1 Introduction to supplier management. After the introduction, the process of supplier management and the steps are discussed in section 4.1.2 Supplier management process. The subsequent section 4.1.3 Supplier development deals with supplier development in more detail.

4.1.1 Introduction to supplier management

Due to outsourcing, increasing use of system suppliers and international supplier networks, suppliers are increasingly turning from pure suppliers into strategic business partners (Lorenzen and Krokowski, 2018, p. 93). To gain a competitive advantage and differentiate from competitors, the know-how of suppliers gets more and more crucial (Hofbauer, Mashhour and Fischer, 2012, p. 23). The performance of a supplier directly influences the company's success. It is essential to acquire new suppliers and to develop existing suppliers in a targeted manner. Particularly strategically relevant suppliers in the international environment require better supplier management characterised by regular care and promotion and a supplier relationship based on partnership. Cooperation with suppliers goes far beyond the unilateral demand for price reductions and is aimed at mutual cost management with a mutual win-win partnership. (Lorenzen and Krokowski, 2018, p. 93-94)

The management of suppliers builds the core area of sourcing and is crucial for successful sourcing. Supplier management refers to the interface between buyer and supplier and in this regard refers to the arrangement of the relationship between these two stakeholders. (Irlinger, 2012, p. 22-23) Janker (2008) defines supplier management as a process which starts with supplier identification followed by supplier limitation, -analysis, -rating and supplier selection or -controlling and leads to the supplier relationship management (Janker, 2008, p. 33).

The goal of supplier management is to enable the analysis of existing and potential suppliers to make strategic decisions based on the analysis results. On a strategic level, this means optimising the suppliers in a medium- and long-term period to increase the quality of supply from vendors and promote relationships with essential and difficult-to-replace suppliers through cooperation. In addition, the aim is to minimise supply risks and dependencies of single suppliers. (Helmold and Terry, 2016, p. 31)

Operational supplier management aims to increase supplier performance and reduce procurement costs. The focus should always be on the best suppliers in order to stop cooperation with non-competitive suppliers. A comparability of the suppliers is therefore necessary. Targeted information on supplier performance for negotiations, supplier evaluation and the associated uncovering of optimisation potential, as well as supplier development measures, enable continuous improvement of supplier performance and quality. (Helmold and Terry, 2016, p. 53)

Instruments of the operational supplier management are, for example, supplier selection, supplier rating, supplier development, etcetera (Helmold and Terry, 2016, p. 57-93). These instruments are explained in more detail in the next section 4.1.2 Supplier management process.

4.1.2 Supplier management process

Supplier management represents the process starting with supplier identification and is followed by supplier limitation. These initial steps of the process are also called supplier pre-qualification or pre-selection. (Lasch and Janker, 2005, p. 410) Based on these steps, supplier analysis, -rating, and supplier selection or -controlling can follow. Finally, the process ends with supplier relationship management which includes, for example, supplier care or supplier development. (Arnold et al., 2008, p. 1003-1004)

Supplier identification

The initial step of the supplier management process is the identification of suppliers for a specific product. In this step, a large number of potential suppliers is considered. It is crucial to identify those suppliers who already offer or can produce the required product. It can be helpful for buyers to look through their existing suppliers if there is already a supplier who can fulfil the requirements for a particular product. (Arnold et al., 2008, p. 1004) If this is not the case, it is necessary to search for new or potential suppliers who produce the desired product or who can do so (Koppelmann, 2000, p. 239).

Supplier limitation

After supplier identification, it is necessary to reduce many potential suppliers for a specific product in the step of supplier limitation. The goal is to get a shortened number of suppliers which fulfil the requirements. This is achieved by a quick and rough pre-check of all suppliers. A possible way to reduce the number of suppliers step-by-step is to make a selection process based on further information about the suppliers. A self-information of a supplier in the form of a questionnaire, possible supplier certificates, or specific knockout criteria help limit the supplier base successively. Afterwards, just a few suppliers should remain for a more detailed analysis and rating. (Arnold et al., 2008, p. 1004-1005)

Supplier analysis

The next step is supplier analysis, where the gathered information concerning the economic, ecological and technical capacity of potential suppliers is analysed (H. Hartmann, Orths and Kössel, 2017, p. 16). Finally, if the buyer still needs further information concerning the supplier, there is the possibility of audits executed by the buyer. Through these audits, the buyer wants to examine the potential supplier in detail and intends to find out the strengths and weaknesses of the supplier. (Arnold et al., 2008, p. 1005-1006)

Supplier rating

Based on the results of the supplier analysis, the supplier rating takes place (Koppelmann, 2000, p. 233). In this step, a systematic evaluation of the efficiency of the supplier is made. Therefore it is necessary to define specific criteria for evaluation. The outcome of the supplier rating builds the basis for the supplier selection or supplier controlling and for the supplier relationship management. (Arnold et al., 2008, p. 1005-1006)

Supplier selection

Once suppliers are identified, narrowed down, analysed, and the rating is performed, the supplier selection is the final point of decision making (Arnold et al., 2008, p. 1007). The selection of a supplier is an essential point in the supplier management process because it is all about choosing the ideal or best supplier (Koppelmann, 2000, p. 234).

Supplier controlling

This step refers to the standard control and monitoring of the supplier's performance and is thus closely linked to the maintenance of the supplier relationship management. Depending on the supplier, the extent and the effort of controlling varies. A classification of the suppliers can help to determine the intensity of control. Problem suppliers are increasingly monitored, and best practice suppliers serve as a benchmark for the development and promotion of new suppliers. (Arnold et al., 2008, p. 1007)

Supplier relationship management

Suppliers play an important role in companies - comparable to the role of employees or customers. Especially for companies with a high rate of outsourcing or low-value creation, a good relationship with suppliers is of significance and essential for a company's success. The maintaining of existing, successful supplier relationships is very likely to be more cost-effective than dealing with supply problems or building new supplier relationships. (Koppelmann, 2000, p. 256) Therefore supplier relationship management builds the basis for a good partnership (Arnold et al., 2008, p. 1008). The following activities can be used to ensure a stable and cooperative buyer-supplier relationship: (Arnolds et al., 2013, p. 229)

- Supplier care
- Supplier integration
- Supplier development
- Supplier promotion
- Supplier education

The buyer's reputation with its suppliers is primarily shaped by how the sourcing department treats its suppliers. In supplier care, the sourcing department wants to enable a trusting relationship with its suppliers and convince the supplier that the buyer is a fair and reasonable business partner who keeps agreements. Consequently, supplier care helps build a cooperative partnership between buyer and supplier and aims to maintain the supplier's performance. Companies that maintain good relations with their suppliers can rely on them to ensure delivery even under challenging situations, in contrast to companies where this is not the case. (Arnolds et al., 2013, p. 229-230)

Supplier integration refers to the cooperation with suppliers and the inclusion of the supplier in the company of the buyer (Arnold et al., 2008, p. 1008).

Supplier development allows the development of the capabilities of the supplier through the buyer in order to be able to fulfil the needs of the buyer (Modi and Mabert, 2007, p. 42). In this thesis, the subject of supplier development is needed, and besides, it is discussed in some facets in the chapter of LPP. Therefore the supplier development is described in detail in the following section 4.1.3 Supplier development.

In the context of supplier promotion, the buyer supports the supplier if difficulties occur at the supplier's side and if he cannot manage the problems by himself. Due to supplier promotion, an improvement of the supply performance can be achieved. Supplier promotion is mainly applied to small and medium-sized companies. Support is offered in various areas, but it is principally used in the area of production. Examples for support can be proposals for rationalisation, technical support for the supplier in production, training or development of employees from the supplier. Therefore, supplier promotion, on the one hand, brings improvements for the supplier itself. However, on the other hand, it helps the buyer to enable the supplier to fulfil his specific requirements. (Arnolds et al., 2013, p. 239-241)

Through supplier education, the buyer can motivate the supplier to achieve extraordinary performance. Therefore the buyer has a range of incentives that may help to influence the supplier's performance. Such incentives can include rewards, giving a bonus or ordering a higher amount of products. But if the supplier does not perform as desired, sanctions such as more control through the buyer or reducing the number of orders by the buyer can follow. (Arnold et al., 2008, p. 1008)

4.1.3 Supplier development

Products and processes get more complex because the outsourcing of parts of the supply chain significantly increases. Thus, the dependence of the buyer on suppliers rises. The supplier has a central role in the buyer's company due to the input through products or services. Hence, it is crucial to ensure a good performance of the supplier in order to fulfil the needs of the buyer. (Hofbauer, Mashhour and Fischer, 2012, p. 85)

However, if the supplier's performance is not as expected by the buyer, there are the following options for the buyer: (Krause, Scannell and Calantone, 2000, p. 34)

- change of supplier
- own production of the previously purchased product
- supplier development, such as investing in the supplier in order to improve the performance of the supplier for the company's benefit

As a supplier change is related to high costs and high investment in time and the production of a previously purchased product is not that easy to manage, supplier development becomes more critical. Moreover, especially in times of crisis, activities in supplier development show that the buyer can rely on a stable partnership with suppliers in contrast to companies that do not invest in supplier development. (Durst, 2011, p. 2)

Definition

Hofbauer, Mashhour and Fischer (2012) state that supplier development is used to optimise or improve suppliers in terms of price, quality, technology and time. Supplier development refers to the improvement of potential suppliers as well as present suppliers in order to intensify the existing buyer-supplier relationship. (Hofbauer, Mashhour and Fischer, 2012, p. 84-85)

Proch, Worthmann and Schlüchtermann (2017) specify supplier development as a range of activities the buyer can take for new, potential suppliers as well as for the improvement of existing suppliers to intensify the long-term cooperative alliance. The measures should ensure the performance of the supplier concerning quality, costs and time to fulfil the short-term as well as the long-term supply needs of a company to stay competitive. (Proch, Worthmann and Schlüchtermann, 2017, p. 17-18)

Benefits of supplier development

In the context of supplier development, competitive and cost advantages can be achieved along the supply chain (Rüdrich, Meier and Kalbfuß, 2016, p. 72). If supplier development succeeds, the buyer has a good selection of top suppliers for cooperation. The benefits of buyers if they develop suppliers' capabilities can be the following: (Hofbauer, Mashhour and Fischer, 2012, p. 86-87)

- ensuring supply even in challenging times
- maximising the supplier performance
- having a pre-selected number of suppliers, which fulfils the requirements of the buyer
- cooperating with top suppliers
- having a sustainable supplier-network
- minimising risk regarding costs, quality and time aspects

According to Watts and Hahn (1993) supplier development programs aim to improve the quality of the material of the supplier and enhance the technical capability of the supplier among the advantages mentioned above (Watts and Hahn, 1993, p. 14-15).

Activities of supplier development

Supplier development is particularly relevant for buyers when the following points are present in the company: (Rüdrich, Meier and Kalbfuß, 2016, p. 72)

- value chain that is highly linked
- products that are particularly innovative
- goods that are technologically sophisticated
- products that have high-quality requirements

When some of the points mentioned above are present, supplier development can be applied in different forms. The basis for the definition of activities in supplier development, however, is formed by the results of the supplier evaluation. (Rüdrich, Meier and Kalbfuß, 2016, p. 72) Building on this, potentials for improvement can be identified, and the goals for supplier development can be formulated. Finally, the activities for supplier development can be specified. (Hofbauer, Mashhour and Fischer, 2012, p. 87)

Büsch (2011) presents the following options for actions that buyers can implement to develop the supplier: (Büsch, 2011, p. 243)

- process-oriented, operational consulting for the supplier
- know-how transfer to the supplier
- consulting of the supplier on strategic issues
- support for market entry of the supplier
- transfer of human resources to the supplier
- financial support for the supplier

Rüdrich, Meier and Kalbfuß (2016) demonstrate various activities of supplier development that depend on different requirements. The requirements represent the intensity of the buyer-supplier relationship and the complexity of a product. Depending on these two factors, the effort in supplier development varies. In general, it can be said that the more complex a product and the higher the intensity of the buyer-supplier relationship, the more effort must be invested in the supplier development. Possible activities in supplier development dependent on these two requirements represent, for example, supplier discussions, supplier workshops or intensive collaborations with suppliers through partnerships. (Rüdrich, Meier and Kalbfuß, 2016, p. 74)

Supplier discussions (e.g. annual supplier talk) are applied for suppliers who have hardly any relationship with the buyer and for those suppliers who deliver standard products like raw materials or goods that are easy to substitute. As there is hardly any relationship and low dependencies on the supplier, there is little effort for the buyer in supplier development. (Rüdrich, Meier and Kalbfuß, 2016, p. 72-73)

Intensive collaboration (e.g. joint ventures) with the supplier takes place in a partnership-based buyer-supplier relationship. Such intensive partnerships occur with suppliers who provide particularly innovative products or goods that are highly important for the buyer. (Rüdrich, Meier and Kalbfuß, 2016, p. 74)

4.2 Linear performance pricing (LPP)

This chapter deals with the topic of Linear Performance Pricing. The beginning of this chapter provides a definition of LPP in 4.2.1 Definition. In the following three approaches are presented in 4.2.2 LPP process according to Newman and Krehbiel 2007, 4.2.3 LPP process according to Proch, Krampf and Schlüchtermann 2013 and 4.2.4 LPP process according to Verein Deutscher Ingenieure 2018. Finally, 4.2.5 Practical applications of LPP describes some possible applications.

4.2.1 Definition

There are two theories in the literature about the origin of linear performance pricing. According to Verein Deutscher Ingenieure (2018) performance pricing has its origin in the unit-kilogram price method, which is very common in industrial practice. The unit-kilogram price method is a one-dimensional performance pricing, in which the price of a product is evaluated according to only one criterion, which is the weight of the product. (Verein Deutscher Ingenieure, 2018, p. 6)

Other sources in literature state that linear performance pricing was first used in the 1990s by the consulting company McKinsey & Company[1] as a tool for short-term cost reductions within sourcing, which was first mentioned in a journal of McKinsey & Company in 1997 (Chapman et al., 1997; Proch, Krampf and Schlüchtermann, 2013, p. 517). Consequently, it was further developed by Newman and Krehbiel (2007) and Proch, Krampf and Schlüchtermann (2013) into a tool of supplier management for cost optimisation.

Linear Performance Pricing is considered a particular form of pricing structure analysis and is used in sourcing. LPP is based on the regression analysis, which is described in chapter 3 Linear regression. The goal is to compare the price and performance in order to reveal a potential for cost reduction. (Gabath, 2008, p. 28)

4.2.2 LPP process according to Newman and Krehbiel 2007

Newman and Krehbiel (2007) describe a way to include not only a product from a tier one supplier but also intermediate products from a tier two supplier in the supply chain. The LPP process is described in concrete terms below. (Newman and Krehbiel, 2007, p. 152)

The strategy of Newman and Krehbiel (2007) assumes multiple steps, which includes several linear regressions and some managerial implications (Newman and Krehbiel, 2007, p. 155-164).

Before the process can be performed, the products must be divided into product groups. In the multi-stage process of Newman and Krehbiel (2007),

[1] https://www.mckinsey.com/

a specific product group is considered as the main component. This main component is composed of several other parts or products, which are called subcomponents. Therefore, it is crucial to find all the necessary data about the main components and their subcomponents. Once the identification and classification of this data have been carried out, the process can begin. However, the identification and classification of these main components and subcomponents are not described and is therefore not listed as a separate process step. (Newman and Krehbiel, 2007, p. 152-154)

Calculate the expected price of all subcomponents

At the beginning of the process, the value added by the tier two supplier is examined. Therefore suitable performance parameters have to be found for all subcomponents. However, the performance parameters can be selected separately for each subcomponent and do not have to be the same. These performance parameters are developed in collaboration with tier 1 and tier 2 suppliers. Based on the previously collected data and the determination of the performance parameters, multiple linear regression can now be performed per subcomponent. The defined performance parameters are considered as independent variables, and the subcomponent's actual price is considered the dependent variable. The result of this regression is the expected price for each subcomponent. This value is intended to represent the actual value of the product and is therefore also called desired or technical value. (Newman and Krehbiel, 2007, p. 156-159, 162)

Calculate the expected value of the main component

Next, the expected value of the main component has to be determined. This also includes the value added by the tier one supplier. Newman and Krehbiel (2007) assume that the performance parameter of the main component is the value of all subcomponents plus the value added by the tier 1 suppliers. Therefore, the expected value of the main component is calculated as the sum of the expected values of the subcomponent plus the value added of the tier one supplier. (Newman and Krehbiel, 2007, p. 156-157)

The calculation of the added value of the tier one supplier is a bit complex. An essential factor is to include all costs, which are in addition to the acquisition of the subcomponents. This includes, e.g. overheads and a reasonable profit. The value added can be calculated in cooperation with the suppliers to get a reliable value, or it can also be a commonly used market or industry value. (Newman and Krehbiel, 2007, p. 157)

Determine the market price and market line

The next part focuses on calculating the market price, which is calculated from the expected price and the actual price of the main component. For this purpose, another linear regression is performed with the actual price as the dependent variable and the expected price of the main component as the independent variable. Finally, by plotting a so-called market line as shown in figure 4.1, the spread of the products around the market price becomes visible. (Newman and Krehbiel, 2007, p. 156-57) The market price can be described as the expected purchase price in contrast to its functionality (Proch, 2017, p. 92).

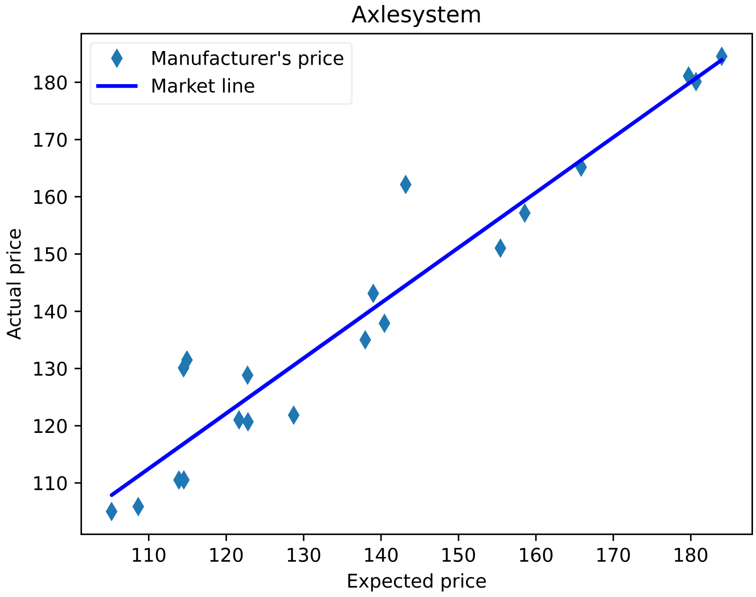


Figure 4.1: Example representation of a market line (own representation)

Determine the best practice price and best practice line

Then the best practice price is calculated. For this calculation, another linear regression is performed. For this calculation, only the 20%-30% of products with the most significant negative difference between expected price and market price are used. The actual price of such best-performing products is used as the dependent variable and the market price as the independent variable in linear regression. Using the obtained equation of the regression, the best practice prices of all products can now be determined. This can also be displayed visually as shown in figure 4.2 to get a better overview. In this graphic, the best practice line is also visible, which expresses the potential cost savings. (Newman and Krehbiel, 2007, p. 156)

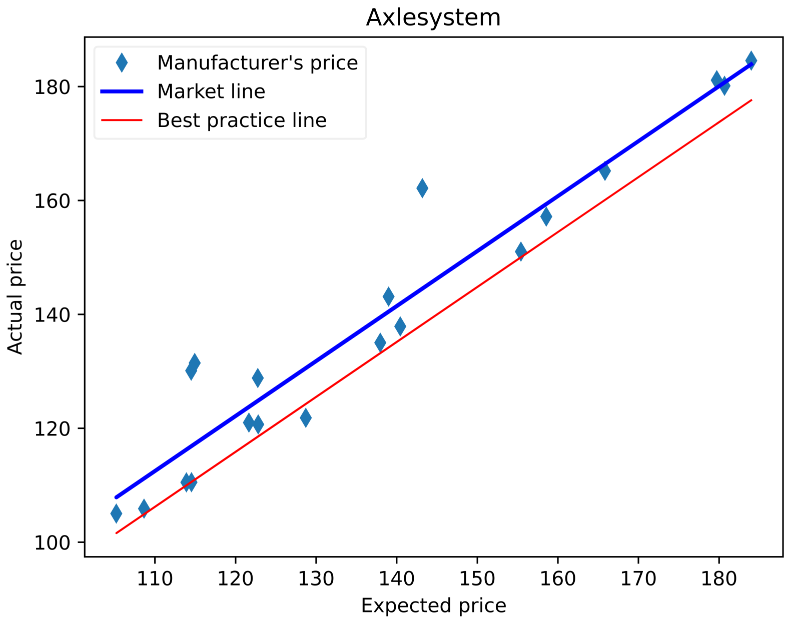


Figure 4.2: Example representation of a best practice line by Newman and Krehbiel, 2007, p. 161 (own representation)

LPP analysis and supplier strategy development

After the expected price, the market line and the best price line of the main component have been calculated, these values can be analysed and evaluated in detail. Based on this evaluation, different strategies can be considered. In general, Newman and Krehbiel (2007) speak of three possible strategies. Since the approach presupposes cooperation with the suppliers, these three strategies are designed for behaviour that enables longer-term cooperation and does not only aim for quick wins. Newman and Krehbiel (2007) do not give these strategies exact names. They are numbered from one to three. Figure 4.3 illustrates these three strategies. (Newman and Krehbiel, 2007, p. 160-164)

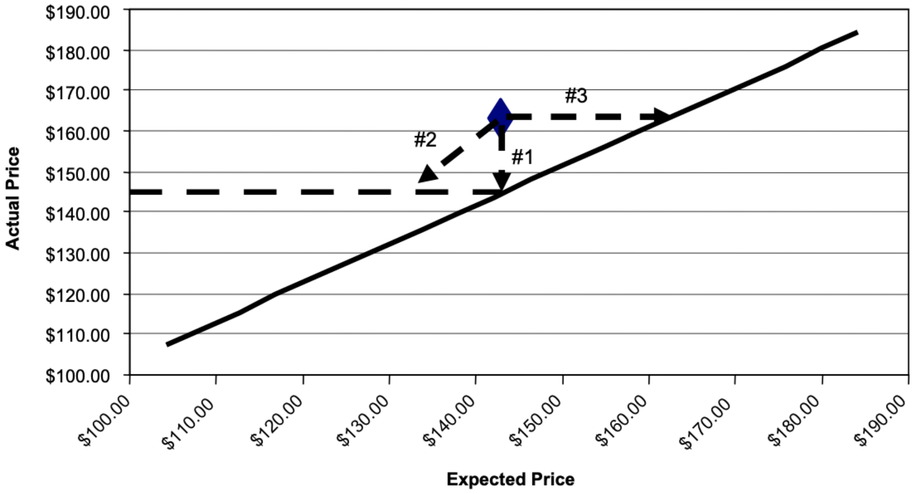


Figure 4.3: Supplier strategy to move one product next to the best practice line (adapted from Newman and Krehbiel, 2007, p. 163)

The first strategy is the simplest and is based on that the supplier reduces the price of the product in order to realise a price close to the best price line. However, this approach is not a typical supplier development as it involves reducing the price while maintaining the same level of performance. Such an approach should be well considered and is not possible for every supplier, as it is often impossible to reduce costs and deliver the same quality and performance. (Newman and Krehbiel, 2007, p. 160)

The second strategy also requires a reduction of the price by the supplier. However, in contrast to the first strategy, the quality or performance which is delivered is also reduced. This reduction also affects the expected value of the product, and therefore this value should be lower. With this strategy, the change in quality or performance must significantly reduce the supplier's actual price. With this method, it is, therefore, possible to achieve a better price-performance ratio and ensure that the product is on the best price line. This strategy is based on a situation in which both the buyer and the supplier can profit. There are many ways to implement this strategy. One possibility can be the adaptation of the product so that, for example, cheaper subcomponents are selected, or the product design is changed. Therefore this strategy is a win-win situation. (Newman and Krehbiel, 2007, p. 162)

The third strategy can be seen in contrast to the second strategy. The strategy assumes that the supplier's actual price is constant or only increases slightly, but the performance and, therefore, the expected value increases significantly. In figure 4.3 it can be seen that this strategy also has the goal of approaching the best price line. In order to achieve this, it is necessary to intensify the cooperation with the suppliers to implement the strategy to the benefit of both parties. Like the second strategy, this is a win-win situation in which both buyer and supplier benefit. (Newman and Krehbiel, 2007, p. 162)

4.2.3 LPP process according to Proch, Krampf and Schlüchtermann 2013

Proch, Krampf and Schlüchtermann (2013) build on Newman and Krehbiel (2007) LPP approach, but especially highlight supplier management. The newly developed approach is based similarly to other approaches on cooperative behaviour and is designed to lead to a win-win situation. In order to enable a smooth process, the process is divided into seven steps. Figure 4.4 shows that the process begins with the selection or definition of a suitable product group and is finalised with a cooperative definition of actions with the supplier. (Proch, Krampf and Schlüchtermann, 2013, p. 518-519)

The process in figure 4.4 is additionally classified by the authors into two domains. The first domain covers the first four steps and can be performed within the buyer's company if the necessary data is available. This can be used for price negotiations with potential suppliers, and therefore it is a potential for cost reduction in sourcing. In the second domain, which includes the last three steps of the process, the supplier should be involved in the process to achieve a satisfying optimisation in the supply chain. (Proch, Krampf and Schlüchtermann, 2013, p. 518-519)

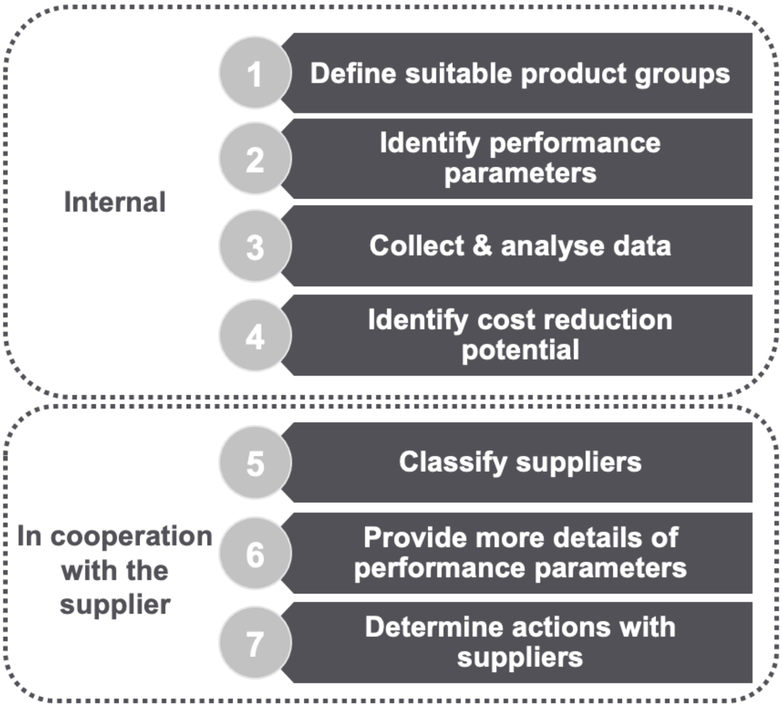

Internal
1 Define suitable product groups
2 Identify performance parameters
3 Collect & analyse data
4 Identify cost reduction potential
In cooperation with the supplier
5 Classify suppliers
6 Provide more details of performance parameters
7 Determine actions with suppliers

Figure 4.4: Proposed process steps by Proch, Krampf and Schlüchtermann, 2013, p. 519 (own representation)

Step 1: Define suitable product group

LPP is a good tool for sourcing, but it causes some effort and costs. For this reason, it is essential to make a preselection, as the application of LPP should lead to a cost-saving and not to an increase in costs. The first preselection is ensured by defining a suitable product group. For this purpose, the products to be sourced must be limited according to various criteria, such as the purchasing volume or expected sales in the future. Now the products must be grouped together to obtain a product group. The grouping of products can be done, for example, by their functionality or production technology. (Proch, Krampf and Schlüchtermann, 2013, p. 519)

Step 2: Identify performance parameters

This step is a crucial aspect of the process. In order to establish a price-performance ratio, a definition of price and performance is crucial. The definition of the price of a product is easy because it is the purchase price. However, the definition of the performance of a product is not always easy to answer unless it is a raw material, and even then, there are several performance parameters such as quality, quantity or purity. Therefore, it is necessary to identify the functionality of the product group and also to be able to evaluate it. Proch, Krampf and Schlüchtermann (2013) describe that it is not always possible to have access to a variety of potential performance criteria. Therefore, in the beginning, an attempt should be made to find a performance parameter by analysing the primary function of a product group. If the price difference within a product group can be explained only by this parameter, it can be defined as a performance parameter. If, however, the price difference cannot be explained, other properties or functionalities must be converted into performance parameters. Then, the parameter with the strongest correlation with price must be selected. The reason for performing the LPP analysis with a single performance parameter, in the beginning, is that it is not always possible to collect all performance parameters of a product, or it is connected with considerable effort and costs. (Proch, Krampf and Schlüchtermann, 2013, p. 519-520)

Step 3: Collect and analyse data

The third step of the process is divided into two parts. First, data from all suppliers must be collected. This includes the purchase price as well as the selected performance parameter. After the data has been collected, the analysis can begin. The analysis is performed in the same way as in steps Calculate the expected price of all subcomponents, Determine the market price and market line and Determine the best practice price and best practice line in the approach of Newman and Krehbiel shown in chapter 4.2.2 LPP process according to Newman and Krehbiel 2007. With the difference that the data of the main component are used for the analysis. The product's purchase price is defined as the dependent variable and the performance parameter as the independent variable. The LPP process is used to calculate an expected, market and best practice price for each product. (Proch, Krampf and Schlüchtermann, 2013, p. 521-522)

Step 4: Identify cost reduction potentials

In this step, first, the statistical potential to reduce costs is identified, and then actions to implement these cost reduction potentials are developed (Proch, Krampf and Schlüchtermann, 2013, p. 522). The causes for price differences between the actual price and the expected price or best price are various. Examples can be high material costs, a too high profit margin, poor product design or an over-specification of the product. (Wildemann, 2008, p. 5)

The statistical potential is calculated as the difference between the actual price and the calculated best price of a product. After the statistical potential has been determined, the causes must be analysed. One possible cause is if the product has a special functionality that is not reflected in the selected performance parameter. This additional functionality should now be analysed, and if necessary, the statistical potential should be adjusted by the value of this additional functionality. This adjusted value represents the real cost reduction potential. (Proch, Krampf and Schlüchtermann, 2013, p. 523)

Step 5: Classify suppliers

In this step possible patterns of a supplier can be recognised. Initially, the suppliers should be classified in order to adjust the procedure with the suppliers. The classification is done per suppliers across products. A distinction is made between challenger, low performer and outlier. (Proch, Krampf and Schlüchtermann, 2013, p. 524)

The challengers are the best performers, and their products are, below or very close to the best practice line. The recommendation of such suppliers is easy because the cooperation with those partners should be intensified, or they should be won as new suppliers. However, it is important to pay attention to various factors such as switching and ramp-up costs or possible degressive effects from the involved suppliers. If a complete shift to challenger suppliers is not possible, the positive factors should be transferred to other suppliers in the context of supplier development. The use of LPP is a win-win situation for the buyer and supplier, as more transparency is created and open communication promotes a sustainable and cooperative partnership. (Proch, Krampf and Schlüchtermann, 2013, p. 524-525)

A low-performer supplier can be seen in exact contrast to a challenger. The prices of this supplier are significantly above the best-price price and therefore provide a bad price-performance ratio. In this case, a distinction must be made between willingness and no willingness to work in partnerships. If the supplier shows no willingness, the previously performed LPP analysis can only be used to reduce the price, and it should be considered to reallocate to other suppliers. If the supplier shows willingness, the exact reasons for the price discrepancy must be analysed. As it is not a question of individual products but rather a large number of products, a general pattern can usually be assumed. Possible examples of such a pattern could be deficits in the manufacturing process, an excessively high profit margin or overpriced raw materials of the supplier. Once the causes have been identified, further steps can be considered with the supplier. The success factors of challenger suppliers can be applied to these suppliers to achieve improvement. (Proch, Krampf and Schlüchtermann, 2013, p. 525-526)

A supplier can be classified as an outlier if only a few products significantly differ between the actual and best prices. However, the causes, in this case, are very diverse and must be analysed per product. Therefore, the functions of the product's main components should be analysed and evaluated in detail to find possible causes. Such causes can range from sub-optimal resources or inefficient resource allocation to a lack of know-how of the supplier for specific processes. (Proch, Krampf and Schlüchtermann, 2013, p. 526)

The developed actions should be critically reviewed and questioned in any case, as they only relate to a previously identified performance parameter (Proch, Krampf and Schlüchtermann, 2013, p. 524).

Step 6: Provide more details of performance parameters

In this step, the focus is no longer on the functionality or performance parameter of the main component but instead on the functionality of the subcomponents. Therefore, a more detailed examination and analysis of the product is possible. To achieve this, the main component must first be divided into subcomponents to identify the functionalities subsequently. The functionalities are then converted into quantifiable performance parameters. However, each component can now have multiple performance parameters, as the collaboration of the supplier simplifies this. Once the necessary performance parameters have been collected, price information of the subcomponents needs to be collected in order to calculate the market price using multiple linear regression for each subcomponent. The performance parameters are used as independent variables and the price information as a dependent variable in the multiple linear regression. In the next step, a benchmark value or best practice price can be determined. This can be done by another linear regression or by using a parallel shift of the regression line in the direction of the data set with the best price-performance ratio. (Proch, Krampf and Schlüchtermann, 2013, p. 526-527)

Step 7: Determine actions with suppliers

The potential savings identified in step 6, which are based on the price difference of the subcomponents between the actual price and the best practice price, can now be used to analyse the causes per product and develop measures together with the supplier. For the analysis, certain aspects should be considered. These include whether the functionality, quality or scope of the product is required by the buyer, whether there is some over-engineering or whether cheaper components can replace components of the product. Such over-specification is of little importance to the buyer and can lead to unnecessary costs. It can also be determined whether the average variance is above or below in comparison to other suppliers. This can be taken into account when considering the actions to be taken. However, it is essential to discuss each action individually with the supplier for each problem and product to achieve a win-win situation. (Proch, Krampf and Schlüchtermann, 2013, p. 527-529)

4.2.4 LPP process according to Verein Deutscher Ingenieure 2018

This approach proposes a 7-step model and also includes possible loops as shown in figure 4.5. It is not only designed for products but also for services. (Verein Deutscher Ingenieure, 2018, p. 7-8)

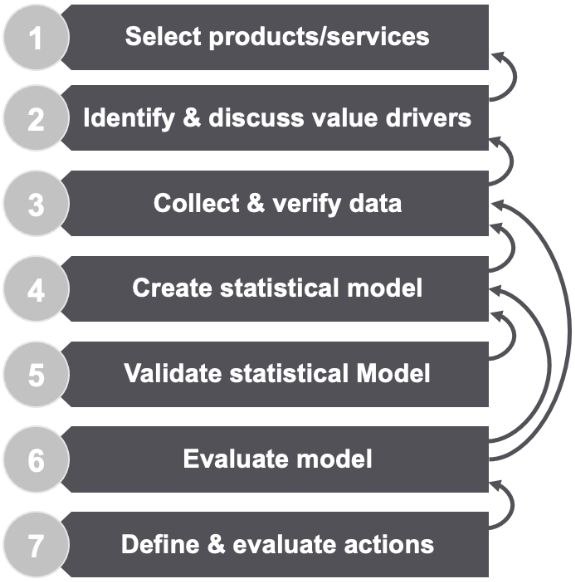


Figure 4.5: Proposed process steps by Verein Deutscher Ingenieure, 2018, p. 7 (own representation)

Step 1: Select products/services

In the first step of this process, a selection of products or services is performed. However, various aspects have to be taken into account in this selection, including the importance and scope of the products. In addition, the value added of the suppliers, the raw material, and currency fluctuations must be considered. Another critical factor is that there are enough products to compare. Otherwise, the validity of the analysis procedures is not guaranteed. When selecting products, it is also essential to ensure that they have objectively comprehensible criteria for evaluation. It is also an advantage to consider products with a specific purchasing volume, as the method can take a few person-days to complete even if the data is good. (Verein Deutscher Ingenieure, 2018, p. 8-9)

Step 2: Identify and discuss value drivers

The second step is to identify and discuss variables or so-called value drivers of the products. This is best found and discussed by interdisciplinary teams. Methods to find value drivers are brainstorming, 635 brainwriting, or using an Ishikawa diagram. (Verein Deutscher Ingenieure, 2018, p. 9)

The objective in this step is not to find as many value drivers as possible but to find a reasonable level since the effort required to collect data increases with the number of value drivers. Therefore, when selecting the value drivers, the relevance concerning the product, potential dependencies between the value drivers, the effort and the possibility to collect the data of the value driver should also be taken into account. (Verein Deutscher Ingenieure, 2018, p. 10)

Step 3: Collect and verify data

This step focuses on the collection of data. Furthermore, the quality of the data must be ensured. Otherwise, the validity of the statistical model is reduced or not given. (Verein Deutscher Ingenieure, 2018, p. 12) However, before data can be collected, the quality of the value drivers must be ensured. In order to provide good data quality, it is essential to guarantee error-free, accuracy, complete and objective data. (Verein Deutscher Ingenieure, 2018, p. 14-15)

In addition to the data of value drivers of a product, other supplementary information should also be collected. Such supplementary information can provide a better picture of the products or be seen as a value driver in certain situations. Examples of additional information are the quantity, various filters (country of manufacture, country of production) or the product status. The product status can be pre-series, series, post-series or a special price status. It should be noted that the price of a product is also strongly influenced by the current product status, and if this influence is not taken into account, the statistical evaluation may be less reliable. (Verein Deutscher Ingenieure, 2018, p. 15-16)

The data can also include historical data or rapidly changing data, such as fluctuating raw material prices. Therefore, price indexation should also be applied in this step. Standardisation of prices is necessary to make prices comparable. (Verein Deutscher Ingenieure, 2018, p. 15)

When all data have been collected, their quality must be checked and verified. For example, it can be checked for a variance around the mean value to examine outliers more closely. In addition, the correct notation of categorical values and the use of the exact dimensions within a parameter should be examined. (Verein Deutscher Ingenieure, 2018, p. 16)

Step 4: Create statistical model

This step deals with the creation of the model and includes steps one and two of linear regression, which has already been explained in the previous chapters 3.2.1 Step 1: Model formulation and 3.2.2 Step 2: Estimation of the regression function, so they are not discussed in detail in this step. In addition, VDI emphasises that regression is an iterative model and also refers explicitly to the possibility of a forward selection, backward elimination or stepwise procedure for selecting the variables of the model. (Verein Deutscher Ingenieure, 2018, p. 16-21)

Step 5: Validate statistical model

In this step, the statistical reliability of the model is verified. This includes steps three, four and five of a linear regression. This has already been described in chapters 3.2.3 Step 3: Test regression function, 3.2.4 Step 4: Test regression coefficients and 3.2.5 Step 5: Test model assumptions. (Verein Deutscher Ingenieure, 2018, p. 21-28)

Verein Deutscher Ingenieure (2018) also describes the possibility of non-linear behaviour and presents three approaches. The first variant is to enable linear behaviour by transforming the value drivers. However, it can also be attempted to divide the model into several models to achieve linearity in these models. The last method is to use a non-linear regression method, but it should be noted that the goal of LPP is not to produce a highly complex equation that fits all points. Instead, the goal is to obtain an approximation of the price equation that is suitable for practical application. (Verein Deutscher Ingenieure, 2018, p. 21-28)

Step 6: Evaluate model

After the model has been identified as statistically valid and reliable, the next step is to evaluate the model in an application context. For this purpose, different checks can be performed, which are described below. (Verein Deutscher Ingenieure, 2018, p. 28-36)

Check extreme values: This checks whether the products with the lowest and highest expected price reflect the calculated market value of the product (Verein Deutscher Ingenieure, 2018, p. 29).

Check leverage points: In this case, it is checked whether possible leverage points lead to substantial variances in the model. If this is not the case, these points should be left in the model, as they increase the quality, otherwise, they should not be included in the calculation. (Verein Deutscher Ingenieure, 2018, p. 29)

Check cluster: If the data form clusters, these clusters should be separated and split into separate models (Verein Deutscher Ingenieure, 2018, p. 29).

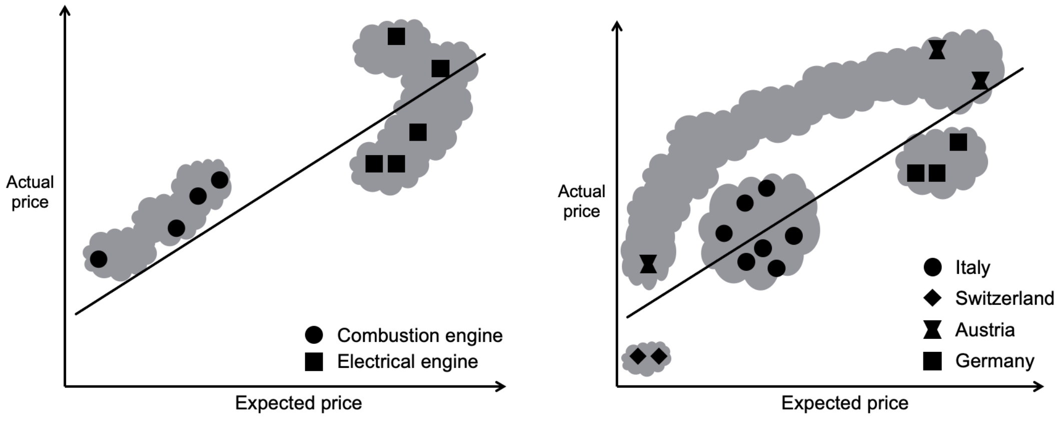
Check regression equation qualitatively: In order to perform a qualitative test, the signs, coefficients and their influence on the price and price function should be evaluated and verified (Verein Deutscher Ingenieure, 2018, p. 29-30).

Check regression equation quantitatively: If a cost structure analysis has been performed, the coefficients should reflect this in order to perform a quantitative check (Verein Deutscher Ingenieure, 2018, p. 30).

Determine plausibility based on cost structure analysis: In this case, the values of the LPP are validated by using cost structure analysis. For this purpose, the target price is calculated for three products that best match the market line and cover the entire range of values. The calculated value of the cost structure analysis is compared with the actual price. Different cases can now occur. If both values are identical, this indicates a reliable model. If the comparison values are systematically lower or higher, a recalibration of the market line should be carried out by a parallel shift based on the difference of both values. In this case, the model is also reliable. However, if both values are different and no systematic variance can be found, the model needs to be fundamentally and critically reviewed. (Verein Deutscher Ingenieure, 2018, p. 31-32)

Check supplier's price level: In this case, it is checked whether experience shows that cheaper or more expensive suppliers are reflected in the model. This is another sign of a reliable model. (Verein Deutscher Ingenieure, 2018, p. 32-34)

Graphically check value drivers and filters: In this step, all value drivers and filters are checked graphically. For this purpose, a graph (actual price vs expected price) is created, in which the data points are marked differently, see figure 4.6. Now it is possible to see whether patterns or structures are formed. If a significant correlation is found, the value driver should remain in the model. If a filter correlates with the price, it should be included if it improves the model. (Verein Deutscher Ingenieure, 2018, p. 34)



(a) Engine type filter

(b) Manufactor country filter

Figure 4.6: Example for graphical representation of different filters by Verein Deutscher Ingenieure, 2018, p. 33 (own representation)

Check price ratios in relation to manufacturing quantities: Analogous to the previous step, the production quantity is checked here graphically. It should be observable that lower quantities lead to higher costs and vice versa. (Verein Deutscher Ingenieure, 2018, p. 35)

Check result regarding sourcing rules of the company: In this case, it should be checked whether all sourcing rules of the company were applied to all products in the model, otherwise this can lead to a variance in the model. Therefore, the price of the product should be adjusted for such indirect costs. (Verein Deutscher Ingenieure, 2018, p. 36)

Step 7: Define and evaluate actions

Once a model has been developed, tested and identified as reliable, concrete actions can be developed. Three steps can be used to identify and define potential actions. (Verein Deutscher Ingenieure, 2018, p. 36)

The first step is to calculate the statistical potential. The potential of a product is the price difference between the actual price and the established target price. The target price can be determined by various methods, including the market line, best practice line, best in class line, mixed potential determination, product grouping or clustering. (Verein Deutscher Ingenieure, 2018, p. 36-40)

After the potential has been determined, Pareto analysis is used to classify the products. The focus for determining actions should now be on those products with the most significant statistical potential. These are A-potential products and represent 80% of the total potential. (Verein Deutscher Ingenieure, 2018, p. 40)

In the third and last step, actions for the selected products are determined by using a graphical analysis. A distinction can be made between price reduction, performance improvement, simplification and strategic development of suppliers. (Verein Deutscher Ingenieure, 2018, p. 41-43)

In addition, the developed model can also be used for newly offered products that are not included in the regression or newly developed products. Therefore, the offered price of new products or new suppliers can be compared to the target price. It should be noted that the target price is valid due to the methodology, but the validity must be checked, as this product is not included in the regression model. (Verein Deutscher Ingenieure, 2018, p. 42-43)

4.2.5 Practical applications of LPP

Linear Performance Pricing is mainly used in sourcing (Schmidt, 2019, p. 69). The method has been widely used in the automotive industry in order to achieve cost reductions in the context of supplier management (Newman and Krehbiel, 2007; Proch, Krampf and Schlüchtermann, 2013). Rüdrich, Meier and Kalbfuß (2016) also emphasises the focus of the use of LPP, especially in the area of sourcing, and mentions that management consultancies often use this method in the context of consulting projects in this field (Rüdrich, Meier and Kalbfuß, 2016, p. 56).

Performance pricing makes it possible to quickly gain an overview of price structures in complex material groups (Verein Deutscher Ingenieure, 2018, p. 44). Thus, LPP provides the area of sourcing with significantly greater market transparency. The results of LPP can be used mainly as a tool for short-term price reductions as well as for building up good buyer-supplier relationships. (Rüdrich, Meier and Kalbfuß, 2016, p. 55-56)

Concerning the application of LPP connected with the achievement of price reductions, the following can be summarised. LPP creates transparency, and outliers can be quickly identified. Hence, the results are a good basis for discussions or negotiations with suppliers. These results can be used in the course of price negotiations with the supplier for price reductions as well as for joint discussions with the supplier to find the cause of deviations and to eliminate them if necessary. (Rüdrich, Meier and Kalbfuß, 2016, p. 64-67) Therefore, it is possible to put suppliers under price pressure with benchmark values when negotiating with them (Proch, Krampf and Schlüchtermann, 2013, p. 517).

Particularly in the corporate areas such as sourcing, marketing and product development, LPP can be used for price reduction. For products with the same technical value but with increased price, a possible portfolio adjustment can take place. (Verein Deutscher Ingenieure, 2018, p. 41)

In the short term, LPP can be used to implement price reductions and savings successfully, but an excellent long-term partnership with suppliers should not be disregarded. LPP can also help the sourcing department to build good, long-term competitive advantages. (Rüdrich, Meier and Kalbfuß, 2016, p. 67)

Especially Proch, Worthmann and Schlüchtermann (2017) sees the application of LPP in supplier management with a particular focus on supplier development to identify development candidates as well as to derive actions for development (Proch, Worthmann and Schlüchtermann, 2017, p. 91). Especially with suppliers whose prices are close to the best practice line, an intensive, long-term partnership with benefits for both sides should be established. These positive effects of such best performers, named as challenger suppliers, should be transferred to other suppliers in the context of supplier development. (Proch, Krampf and Schlüchtermann, 2013, p. 524-525) This cooperative behaviour in the buyer-supplier relationship leads to a win-win situation for both buyer and supplier (Proch, Krampf and Schlüchtermann, 2013, p. 518-519).

To sum up, it can be said that the benefits of LPP include transparency of cost drivers, internal and external resource optimisation, better communication between tier suppliers and more focused negotiations with suppliers. LPP as well supports the buyer in the negotiation concerning the initial cost for components of new products or currently sourced components. (Newman and Krehbiel, 2007, p. 164)

Moreover LPP is collaborative concerning developing the supply chain and not only focuses the area where improvement is needed but also uses the network (like for example benchmarks, the tier two cost understanding etc.) in order to support suppliers to adjust their prices, processes etc. to better fit the market (Newman and Krehbiel, 2007, p. 162).

Schuh et al. (2012) describes in comparison to other authors like Newman and Krehbiel (2007), Proch, Krampf and Schlüchtermann (2013) and Verein Deutscher Ingenieure (2018) that LPP should only be applied to simple products. It is assumed that LPP is only practicable with only one cost driver (Schuh et al., 2012, p. 186-187) and according to Heß (2008) LPP should only be used if the price depends on only few cost drivers (Heß, 2008, p. 223).

Münch (2018) describes LPP as universally applicable, which offers a wide range of possible applications in addition to the classic application in sourcing. LPP can be used in areas where saving potentials can be achieved by determining target prices and benchmarks. Despite sourcing, LPP can also support product development and sales. Product development can be used to optimise the cost impact of components or create price forecasts in early development phases. In sales, it can be used in the same way as in sourcing to check price consistency, making it possible to evaluate each offer better. In addition, it can be used to accelerate the preparation of offers or to use customer-specific value-based pricing. (Münch, 2018, p. 36-38)

4.3 LPP alternatives

This chapter shows some possible alternatives to LPP in the area of cost optimisation. These include, non-linear performance pricing (NLPP), total cost of ownership (TCO) or price structure analysis (Heß and Laschinger, 2019, p. 94).

4.3.1 Non-linear Performance Pricing (NLPP)

Many relationships, in reality, are not linear and therefore, models that assume linearity, such as LPP, cannot adequately represent reality (Rüdrich, Meier and Kalbfuß, 2016, p. 66). Non-linear performance pricing therefore forms the further development of LPP. This method takes several cost drivers for which there is no linear relationship between performance parameters and price into account. (Locker and Grosse-Ruyken, 2019, p. 119)

NLPP assumes that prices do not always develop linearly. This is due to the correlation of product properties to each other. (Godek, 2021)

NLPP can be used like LPP. It compares the price-performance of any number of products. It is analysed how much a product may cost for given product characteristics (e.g. length, diameter, delivery time) in a market comparison. In this way, savings potentials can be uncovered in the area of sourcing. (Online-Magazin für Procurement, Beschaffung, Supply-Chain-Management (SCM) & Digitalisierung, 2018)

NLPP enables reliable performance and produces stable and robust results even when data are few and not perfect (Saphirion AG, 2021, p. 15).

4.3.2 Total cost of ownership (TCO)

The concept of TCO was first developed in the 1980s by the consulting company Gartner Group[2] for the evaluation of alternative investment projects (Stollenwerk, 2016, p. 154). Ellram (1995, p. 4) defines TCO as ″a purchasing tool and philosophy which is aimed at understanding the true cost of buying a particular good or service from a particular supplier″. This means that not only the price of a product offered by a supplier is important, but also the additional costs which arise during the whole life cycle of a product (Bremen, 2010, p. 25). TCO is therefore intended to represent all costs associated with sourcing, ownership, use, maintenance or repair of a product (Ellram, 1995, p. 4). The price of a product may be low, but the total cost over the life cycle may be high. A popular example of this is an office printer, where the price is low. However, the toners required are expensive. The consumption of toners and their price influence the costs which arise for the printer over its entire life cycle. Therefore, when making a sourcing decision, not only the price, but also the additional costs of the product should be considered. (Stollenwerk, 2016, p. 153)

TCO can provide a basis for decision-making concerning supplier selection, as the use of it makes it possible to list the total costs per unit of a product that arise along the product life cycle (Bremen, 2010, p. 33). There is no universal calculation method for TCO because many different models that differ in terms of the cost categories exist (e.g. sourcing costs, quality costs, logistic costs or maintenance costs) (Stollenwerk, 2016, p. 154-155).

4.3.3 Price structure analysis

The primary purpose of the price structure analysis is to check the appropriateness of the supplier's price to find out whether the supplier's profit is justified. This can be helpful not only in the sourcing of new products but also in existing contracts with suppliers, for example, if they want to increase prices. (Arnolds et al., 2013, p. 100-101)

[2] https://www.gartner.com/

Using the price structure analysis, it is possible to find out the supplier's profit share by calculating the unit costs. For this purpose, all relevant cost elements of a product have to be determined and evaluated. In the calculation, costs concerning material, production, development, administration and distribution have to be considered. Usually, the unit costs of the product are then calculated based on full cost accounting by adding up the relevant direct and overhead costs of these cost elements. The difference between the selling price and the unit cost thus represents the supplier's profit. (Stollenwerk, 2016, p. 146)

However, the execution of the price structure analysis is not quite simple and usually associated with difficulties. One reason for this is that the information concerning the cost elements required to calculate the unit costs is often unavailable. Further reasons are that it is not possible to carry out the analysis because the manufacturing process is too complex or the employees do not have the appropriate qualifications. (Arnolds et al., 2013, p. 103-104)

5 Practical discussion

In the previous chapters, the theoretical foundations are explained more in depth. These theoretical fundamentals are used in this chapter to develop an approach for the automation of LPP.

This chapter is divided into six sections. In the beginning, a review of manual LPP process steps is given to analyse and compare the presented LPP methods. Then the levels of automation are explained. This is necessary because, in the following chapter, a concept for an automated single level LPP is illustrated and presented. In addition, an automation concept of multiple level LPP is demonstrated. Finally, quality criteria and the prototype are described in more detail.

5.1 Review of manual LPP process steps

In the chapter 4.2 Linear performance pricing (LPP) three methods are described which show how LPP can be performed. These three methods differ in some aspects, but they also have similarities. This subchapter deals with an in depth review and comparison of the presented methods.

In terms of time, the approach of Newman and Krehbiel (2007) is the first and Proch, Krampf and Schlüchtermann (2013) partly builds on the previous method and ideas. Several elements of both methods can be found in the method of Verein Deutscher Ingenieure (2018), which does not explicitly refer to both sources.

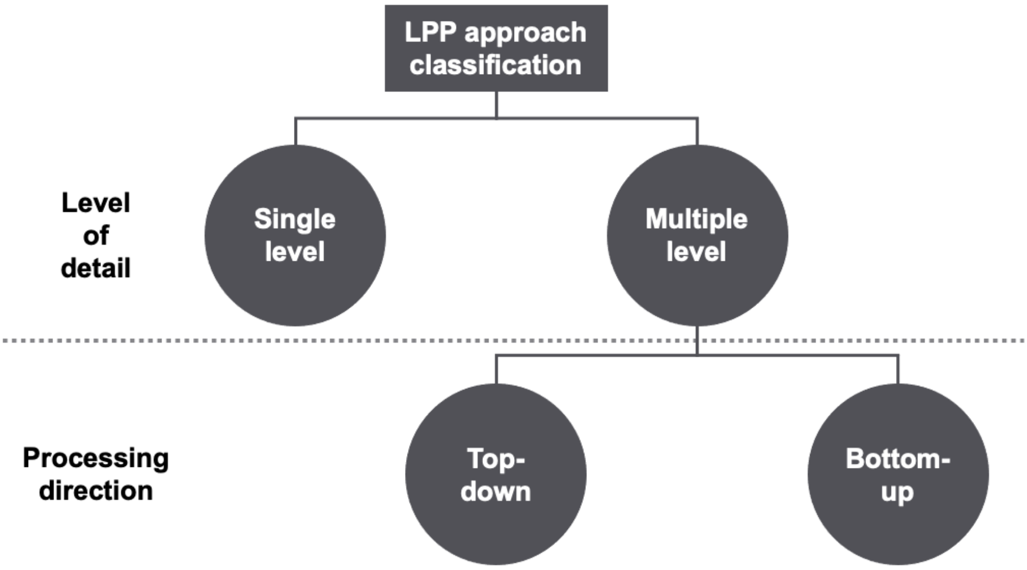


Figure 5.1: Classification of LPP approaches
(own representation)

As illustrated in figure 5.1, the presented methods can be classified differently. One possibility is to classify the procedures according to their level of detail. The classification according to the level of detail can be divided into single level and multiple level. An approach such as Verein Deutscher Ingenieure (2018), which only considers one level, only includes the main component in the calculation. However, if a method, such as Newman and Krehbiel (2007) or Proch, Krampf and Schlüchtermann (2013), takes subcomponents into account as it can be called multiple level. In addition, procedures that include more than one level can be classified into top-down and bottom-up approaches. This classification refers to the processing direction. The approach of Proch, Krampf and Schlüchtermann (2013) follows the top-down approach and Newman and Krehbiel (2007) processes the components using the bottom-up principle. The bottom-up approach of Newman and Krehbiel (2007) assumes that linear regressions of the subcomponents, i.e. the lowest level, is performed first, and the main component consists of the sum of all subcomponents plus some value added by the supplier. In order to allow further analysis, two more regressions are performed on the main component to get the final result. However, the top-down approach of Proch, Krampf and Schlüchtermann (2013) first runs two regressions on the main component to determine potential supplier strategies and afterwards, a detailed analysis are performed using the results of further regressions on the subcomponents. With this detailed analysis, specific actions for the individual products are determined.

In the approaches of Newman and Krehbiel (2007) and Proch, Krampf and Schlüchtermann (2013) the main component and also its subcomponents are considered. However, the approach of Verein Deutscher Ingenieure (2018) is only based on one level and therefore only considers the main component. The consideration of only one level results in a lower level of detail in the analysis. This elimination of multiple levels leads to less complexity and effort of data acquisition and data processing. This simplification can also improve the quality of the data. The quality of the data is a very crucial criterion in an LPP because every analysis procedure depends on the number of data sets and their quality.

Now that a general classification, description and first comparison of the methods has been made, table 5.1 compares the methods of Newman and Krehbiel (2007), Proch, Krampf and Schlüchtermann (2013) and Verein Deutscher Ingenieure (2018) presented in the theoretical part. The different approaches have to be divided into more general steps before a comparison is possible. Therefore, the whole process is divided into six areas. This grouping is shown in the column process areas in table 5.1 and includes select products, select performance parameters, gather data, perform regression, develop strategy and define & evaluate actions. The other three columns contain the numbers of the respective steps of the methods which can be assigned to the areas. For example, in Newman's approach, a regression is performed in steps one, three and four, while in Proch's approach, it is only performed in two steps, four and six.

It is also important to mention that the process of Newman and Krehbiel (2007) is not directly separated into steps, but in order to ensure the possibility of a comparison, the process of Newman and Krehbiel (2007) is separated into five steps. For this purpose, Calculate the expected price of all subcomponents is the first step, Calculate the expected value of the main component the second step, Determine the market price and market line is the third step, Determine the best practice price and best practice line is the fourth step and LPP analysis and supplier strategy development is the fifth step. The collection of data and the selection of performance characteristics are not described in detail in the approach of Newman and Krehbiel (2007) and therefore this is referred in the following as step zero, because this is already done before the process.

Process	**Steps in approach of**		
area	**Newman 2007**	**Proch 2013**	**VDI 2018**
Select products	0	1, 6	1
Select performance parameters	0	2, 6	2
Gather data	0	3, 6	3
Perform regression	1, 3, 4	4, 6	4, 5, 6
Develop strategy	5	5, 7	7
Define & evaluate actions	5	7	7

Table 5.1: Comparison of steps of the different approaches

In terms of the general process steps, the table 5.1 shows that there is a similarity between the process flows of Newman and Krehbiel (2007) and Verein Deutscher Ingenieure (2018). However, Verein Deutscher Ingenieure (2018) describes in addition to Newman and Krehbiel (2007) the selection of products up to the collection of the necessary data. Furthermore, it can be seen that Proch, Krampf and Schlüchtermann (2013) approach is a more general approach with an iteration loop that is focused more on the whole process and not on performing a regression.

Table 5.1 illustrates that Newman and Krehbiel (2007) and Verein Deutscher Ingenieure (2018) both have a continuous process, whereas the approach of Proch, Krampf and Schlüchtermann (2013) has an iteration loop built in. Although Proch, Krampf and Schlüchtermann (2013) and Newman and Krehbiel (2007) are based on a two-stage approach with main components and subcomponents, it is obvious that there is a difference between the two approaches. In Newman and Krehbiel (2007), the subcomponents are calculated first, then the main components are calculated, and finally, the further actions are defined. In the multi-stage approach of Proch, Krampf and Schlüchtermann (2013), which as mentioned above is iterative, the focus of the first iteration is entirely on the main component. A detailed analysis using the data of subcomponents is performed in the second iteration. This has several practical reasons. The separated method of Proch, Krampf and Schlüchtermann (2013) makes it easier and also quicker to define supplier strategies in an early stage of the whole process. The advantage is that it is not necessary to analyse the products and especially their subcomponents in detail and collecting the whole data of all products and subcomponents early with this technique. This reduces the effort of data collection at the beginning significantly. Another reason is that the approach of Newman and Krehbiel (2007) assumes that the buyer has already all data of the products and their subcomponents, but in practice, this is very unlikely without prior involvement of the suppliers. For this reason, the detailed analysis of the products and subcomponents in the approach of Proch, Krampf and Schlüchtermann (2013) is performed in a later step with the cooperation of the suppliers.

If just the effort of regressions is now compared, the approach of Newman and Krehbiel (2007) is to be preferred, since, with one main component and n subcomponents, only $n + 3$ regressions have to be performed. In contrast $n * 3 + 3$ regressions have to be performed in the approach of Proch, Krampf and Schlüchtermann (2013). This means that with an increasing number of subcomponents, several regressions, their prerequisite tests, and the selection of the model must be performed in the approach of Proch, Krampf and Schlüchtermann (2013). On the other hand, the approach of Newman and Krehbiel (2007) requires that all data (main components and inclusive subcomponents) are already available in good quality at the beginning of the process, even without possible supplier cooperation. This is not necessary using the approach of Proch, Krampf and Schlüchtermann (2013), as only the data of the main component need to be available at the beginning, and the data of the main component should be available to the buyer in almost full range. Only after a supplier strategy has been defined and cooperation with suppliers is desired, a detailed analysis is performed in cooperation with the supplier in the approach of Proch, Krampf and Schlüchtermann (2013). Through the cooperation, the buyer has better access to the data of the subcomponent, and it is easier and of better quality to obtain the data for the detailed analysis.

In addition, there is another challenge in the approach of Newman and Krehbiel (2007). The added value by the supplier has to be calculated, but this is very difficult to do it correctly and also difficult to implement without the cooperation of the suppliers. In this case, some approaches can be used, like using an average value of the industry. Without the supplier's cooperation, this is just an approximate value that just reflects the whole industry but not a single supplier.

In contrast to the other two methods, the approach of Verein Deutscher Ingenieure (2018) explicitly mentions that the use of the LPP process can also be applied to services and not primarily to manufactured products.

5.2 Levels of automation

Linear performance pricing and its mathematical foundation linear regression are good ways to compare products from different manufacturers. The mathematical method of linear regression is already widely used in science. However, the practicable use of linear performance pricing has its limits, as the calculation is often only applied to a few essential products. It is necessary to enable better automation of this method in the future. Therefore, it is essential, as in other areas, to move from manual processing through semi-automation to full automation.

In this work, the degrees of automation are divided into four levels:

- Level 1 - Manual execution
  Complete manual execution, calculation and evaluation by the user.

- Level 2 - Supported manual execution
  Calculation and providing the final result and various criteria using a programme. Interpretation and selection of variables by the user.

- Level 3 - Semi-automated execution
  Calculation and providing the final result and all criteria using a programme. The programme already selects suitable models and makes the results available to the user in order to perform the final evaluation.

- Level 4 - Fully automated execution
  Calculation and preparation of the final result and all criteria using a programme, with additional interpretation, variable selection and evaluation.

5.3 Concept of a single level automated solution

As shown in the previous chapter 5.1 Review of manual LPP process steps, there are various similarities between the presented methodologies, and all of the approaches have their advantages and disadvantages. Therefore, a concept for an automated single level LPP is developed within this thesis that combines different methods to enable a higher degree of automation of the process steps of LPP.

The newly developed approach focuses more on the less considered point of the prerequisite tests of a linear regression compared to the other presented approaches. These approaches only consider the precondition check in broad terms and state it as a given. This check is also a significant point for automation because, in an automated approach, a large number of regressions and thus also a large number of prerequisite checks must be executed.

Figure 5.2 shows the newly developed approach for better automation, which is divided into four phases and a total of nine steps. The first phase, called preparation, includes three steps covering data acquisition, cleansing and verification. In the second phase, the technical implementation of LPP is done. This includes the mathematical part, which is implemented in steps four to six. The next and third phase deals with the economic aspects in steps seven and eight, which includes implementing the results found in the previous phase. The process is finalised with the evaluation phase, in which an evaluation of the process and, if necessary, a plan to adapt the process for the next run takes place. These four phases do not represent a closed process, they are recursive, and if they are well automated, they can be performed continuously.

An automation of the process can be achieved in different steps. In the following subchapters, the phases and steps are explained in detail. In addition, the automation potential of the respective step is also discussed.

During the development of this process, it was ensured that the various process steps were as flexible as possible. This enables using a different mathematical algorithm while retaining the developed phase-based process flow and solely adapting individual process steps.

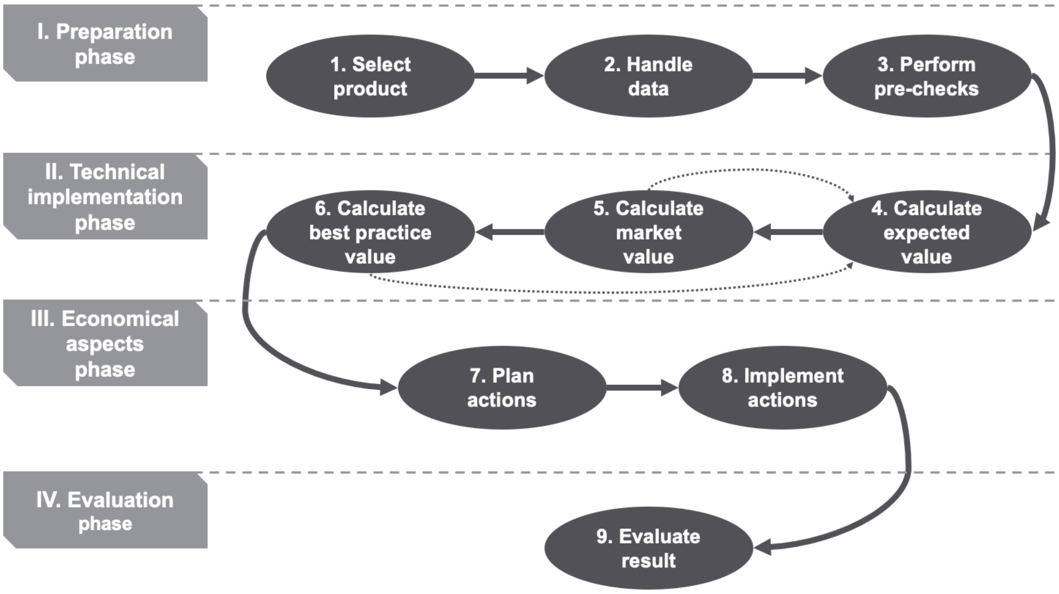

I. Preparation phase
II. Technical implementation phase
III. Economical aspects phase
IV. Evaluation phase
1. Select product
2. Handle data
3. Perform pre-checks
4. Calculate expected value
5. Calculate market value
6. Calculate best practice value
7. Plan actions
8. Implement actions
9. Evaluate result

Figure 5.2: Steps of suggested automated single level LPP process (own representation)

5.3.1 Step 1: Product selection

The first step, the selection of products or product groups, is analogous to the processes presented by Proch, Krampf and Schlüchtermann (2013) and Verein Deutscher Ingenieure (2018).

In the beginning, a team of domain experts divides products into certain product groups. For this purpose, the products and their functionality are compared, and an attempt is made to identify and group similar products. However, the scope of the product group should not be too broad. Otherwise, there will be an insufficient amount of common variables. It must be ensured that there are many more data sets per product group than variables included in the regression.

In addition to the pure manual selection, an automated pre-selection of products can support the domain experts. For this purpose, an existing data set of products is analysed and used to find products with similar variables to group them. This serves as an assistance for the team of experts to identify possible similarities more easily and quickly.

5.3.2 Step 2: Data handling (gathering, cleaning & preparation, check)

This step focuses on the processing of the data. First, all data from the products to be verified are collected. Next, a cleaning of the data takes place. This is done by reviewing the individual data and then deleting records with missing data. Another possible method is to fill in missing data by using the mean value. Non-stochastics usually use such an approach to fill in the missing data to generate multiple data sets. However, this is not recommended as the regression can be biased by adding such data. Therefore, the best way is not to include data sets with lacking data in the calculation of a regression.

In most cases, data is obtained from different sources, which can lead to problems during further processing and must therefore be converted into a standardised format. Therefore, data preparation and data adaptation is necessary.

Finally, categorical variables or the so-called dummy variables must be converted. The reason for doing the conversion is because dummy variables cannot be used directly as textual values in the regression. Instead, they can only be used as numerical values. In chapter 3.3 Dummy variables the procedure dealing with dummy variables is described in more detail.

The collection and processing of data can be very time-consuming, but in terms of the field of application, this step can be automated for the most part.

5.3.3 Step 3: Perform pre-checks

In order to complete the preparation phase, pre-checks are necessary. The prerequisite checks of a regression are done after a regression. So other checks have to be performed in this step.

A review of the collected data must be done. Various techniques can be used to identify problems. It is important to pay attention to the correct notation of variables with a textual value, such as dummy variables, or to detect outliers in the data and check whether they have arisen from measurement errors, for example. Checking for outliers before a regression can also reveal problems with the units of measurement and thus poor data preparation and data fitting.

A check for a linear correlation between the actual value and the variables can also be conducted. However, this check only indicates that there may be problems with the data or selected variables. The reason is that only pairwise correlations are found between the actual value and the variables and not between the actual value and variable combinations. This means that variable combinations can still have a decisive linear influence on the actual value.

5.3.4 Step 4: Calculate expected value

After the preparatory phase, the second phase of the technical implementation can be initiated. Since the step of calculating the expected value is complex, it can be divided into three parts:

- calculation of the expected value
- checking the requirements of a linear regression
- selection of the best model for the expected value

The reason why these different tasks are combined in one step is that they form a unit and are performed together in order to determine the best model.

In the beginning, a linear regression is performed to *"calculation of the expected value"*. The expected value or price is also called the technical value or target price of a product. This value indicates how much the specific product should cost according to the regression or, in other words, what the value of the product is. However, it is not possible to speak of an exact value because, as with every mathematical method, a certain degree of variance is always present. This variance is expressed by the standard error. To calculate the expected value, the actual value is used as the dependent variable. As shown in the first part of this step, the selected variables of a product are used as independent variables in a linear regression. The calculation of a linear regression is described in detail in chapter 3.2.1 Step 1: Model formulation and 3.2.2 Step 2: Estimation of the regression function. The selection of variables is an essential part and is described in more detail in the third part *"selection of the best model for the expected value"* of this step. To ensure a reliable selection of the best model, the prerequisite test must be completed positively - see second part *"checking the requirements of a linear regression"* of this step.

The second part of this step, after the regression has been done, is the *checking the requirements of a linear regression*. This is necessary because otherwise, the regression and its quality criteria are not meaningful. The checking of the criteria is described in detail in chapter 3.2.3 Step 3: Test regression function, 3.2.4 Step 4: Test regression coefficients and 3.2.5 Step 5: Test model assumptions.

The third part of this step is to identify the best model. It is, therefore, necessary to create and compare a series of models. The independent variables determine a model, and it can be said that it is necessary to make an optimal selection of independent variables. Depending on the number of variables, this can be very complex and time-consuming. To ensure an optimal combination of variables, there are different possibilities, which are described in the chapter 2.4 Feature selection.

The most obvious method for selecting the best model is to compare all the different combinations and select the best model. However, this approach is not really practical due to the increasing number of variables, being highly time-consuming and increasing complexity. With a selection of just five variables, there are already 31 different possibilities, and with ten variables, this number rises to 1,023 different possibilities. Using the sum of the binomial coefficient see equation equation (5.1) it is possible to calculate the exact number of possible variations. Ten variables of a product seem a lot at first, but when a dummy variable is converted, the number of variables increases. If it is assumed that a dummy variable can take on four values, three variables are generated from such a dummy variable. Thus, it can be said that $n - 1$ variables are generated from a dummy variable that has n different values.

$$\sum_{k=1}^{n} \binom{n}{k} = \sum_{k=1}^{n} \frac{n!}{(n-k)!\,k!} \tag{5.1}$$

$$n = \text{Number of independent variables}$$

The part of selecting the best model is essential in order to automate LPP. However, to do this, quality metrics and boundaries need to be defined in advance. As science has already shown, there are no universal quality criteria. Specific quality criteria are set too high for certain use cases and too low for others. Furthermore, the result is influenced by the method used to select the best model. This is why the variable selection part should not be fully automatic without further research. To avoid problematic variable combinations, the selection of variables should be verified by an expert person. Semi-autonomy is recommended to find several optimal models by the algorithm, which a qualified person then evaluates. Semi-autonomy has a considerable advantage over the manual method, as finding and discussing potential variables can be very demanding and therefore time-consuming. Additionally, with manual processing, possible superior variable combinations can be overlooked.

The different quality criteria indicate the quality of the regression performed. More details on quality criteria are explained in the chapter 5.5 Quality criteria.

5.3.5 Step 5: Calculate market value and line

The fifth step deals with the calculation of the market value and the market line. The market value reflects the usual market value of the product. In other words, it can be said to be an average value or price for the offered performance. For this purpose, the previously determined expected value is used as the independent variable and the actual value of the product as the dependent variable. Using the regression equation, a market line can also be created to represent the model graphically. These values and graphs show the deviation of the actual value from the average value on the market. From this, initial conclusions can already be drawn, such as whether the value of a product or a supplier's product range is above or below the average market value.

A check of the prerequisites and quality criteria must also be performed for this regression. If this check of the performed regression of the market value fails, the next better model of the step 5.3.4 Step 4: Calculate expected value must be selected, and the calculation of the market value must be repeated.

5.3.6 Step 6: Calculate best practice value and line

The fifth step deals with the calculation of the market value and the market line. The market value reflects the usual market value of the product. In other words, it can be said to be an average value or price for the offered performance. For this purpose, the previously determined expected value is used as the independent variable and the actual value of the product as the dependent variable. Using the regression equation, The sixth step of the process is also the last step of the technical implementation phase. The last regression is performed in this step, and it deals with the calculation of the best practice value for each product. For this purpose, similar to calculating the market value, a regression is done using the expected value as the independent variable and the actual value as the dependent variable. The difference in the calculation of the market value is that only the top performer products are included in the regression calculation. A product is called a top performer if it is among the $20 - 30\%$ of products with the smallest difference between actual and market values. The resulting regression equation can then be applied to all products in the dataset to calculate the best practice value.

As in the previous step, it is mandatory that checks of the assumptions and quality criterion of the regression performed to determine the best practice value are successful. If this is not the case, the next better model must be selected of the step 5.3.4 Step 4: Calculate expected value and jump back to the step 5.3.5 Step 5: Calculate market value and line.

This step concludes the mathematical phase, and in the next step 5.3.7 Step 7: Plan actions the analysis and implementation can be started using the calculated market value and best practice value.

5.3.7 Step 7: Plan actions

The seventh step of the model is the start of the phase in which the economic aspects are examined in more detail. The suppliers or individual products can be classified using the previously calculated values of market value, best practice value, and their difference to the actual value. Based on the willingness to cooperate and the identified problems of the suppliers and products, a plan for improvements is developed in this step. The concept of challenger, outlier and low performer suppliers can be used here, among others. This concept is already a well-described method, which is explained in more detail in chapter 4.2.3 Step 5: Classify suppliers. Based on this information, actions are planned together with the potential supplier.

During the planning of the actions, it is crucial to define some key performance indicators to evaluate the success after implementing these actions. Possible key performance indicators are highly diverse and depend on the planned measurements. In the simplest case, this can be a price reduction achieved through an optimised or simplified production process.

5.3.8 Step 8: Implement actions

Finally, the plans developed in chapter 5.3.7 Step 7: Plan actions are evaluated in order to be implemented in cooperation with the suppliers. After the implementation is completed, the economic aspect phase ends.

The implementation of the actions depends strongly on the field of application, and therefore it is not discussed further in this chapter.

5.3.9 Step 9: Evaluate results

The last phase, consisting of only one step, is the evaluation phase. This step is essential for the further development and improvement of the performed process and the achieved result. This is why this step is all about evaluating the steps previously taken, the measures derived, and the goals achieved. Furthermore, it is also of interest to track possible problems, errors and best practices in order to be able to apply them again when the process is repeated or to apply them to other suppliers.

In order to be able to perform the evaluation step, data from the previous steps must be available. In addition, it is advantageous if specific key performance indicators and their targets are already defined during the planning of the actions.

Possible ways for evaluation are:

- Measurement of supply shortages
- Measuring of better JIT delivery
- Price reduction measurement

In addition to the evaluation of the KPI's, the selection of variables in step 5.3.4 Step 4: Calculate expected value should be reviewed and if it is necessary to select different variables in a new run.

5.4 Automation concept of multiple level LPP

As already described, it is possible to perform the LPP process in a single level or multiple level mode. In the previous chapter 5.3 Concept of a single level automated solution a concept for the automation of a single level LPP process was discussed. This chapter deals with the automation of a multiple level LPP process. The multi-level approach includes the supplier's subcontractor. In order to implement the more complex multiple level process, the discussed method in chapter 5.3 Concept of a single level automated solution has to be adapted.

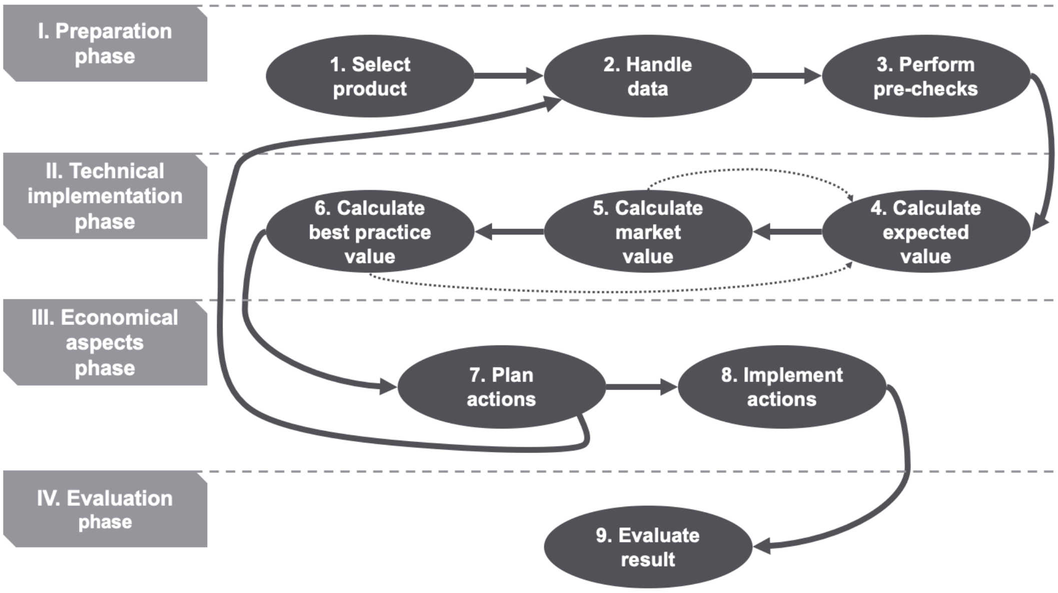

I. Preparation phase
II. Technical implementation phase
III. Economical aspects phase
IV. Evaluation phase
1. Select product
2. Handle data
3. Perform pre-checks
4. Calculate expected value
5. Calculate market value
6. Calculate best practice value
7. Plan actions
8. Implement actions
9. Evaluate result

Figure 5.3: Steps of suggested automated multiple level LPP process (own representation)

The multi-level LPP approach of Proch, Krampf and Schlüchtermann (2013) is considered to be good and practicable since it is possible first to do a general review of the suppliers (1st level) and with the cooperation of willing suppliers to do a detailed analysis (2nd level) using the top-down approach. Therefore, the presented automation process in chapter 5.3 Concept of a single level automated solution was extended. The changes of the process are visualised in figure 5.3. To enable the possibility of two-stage processing, an iteration is introduced in the process, which involves the process steps from data handling to calculate market value. The first iteration focuses on the main product. The second iteration deals with the subcomponents and is done together with the supplier. The changes are explained in detail below.

First iteration - Main component

In the first iteration, only the process step plan actions changes. This is where based on the analysed data, a pattern among the products is tried to be identified. As Newman, Proch, and VDI have already described, suppliers can be divided into three groups - Challenger, Low Performer and Outlier. It is easiest and best to work with challenger suppliers because they already have an outstanding price-performance ratio. However, this case is not always possible, and therefore outlier and low performer suppliers should be preferred. However, outlier suppliers are preferable to low performers because it is more easily possible to turn these suppliers into challenger suppliers with an effective and targeted supplier development. However, this does not exclude that supplier development of a low performer can and should be made. Under certain circumstances, it may make sense to develop this type of supplier as well. If the selected suppliers are cooperative, the second iteration can be initiated.

Second iteration - Subcomponent

This iteration deals with an in-depth analysis, considering the subcomponents. The process steps from data handling to calculate best practice value are repeated the same way as in the simple automated process with the exception that each process step is executed per subcomponent individually. The second execution of the process step plan actions is changed to focus on the individual subcomponents and provide a detailed analysis of the price differences between the actual value and the best practice value. This is done in close cooperation with the supplier, and therefore it can be even more precise than the first iteration.

5.5 Quality criteria

The quality criteria are an essential aspect of a linear regression. They describe the quality of the calculated model and therefore also serve as comparison criteria for an automatically performed linear regression. The quality of the data goes hand in hand with the quality criteria of a linear regression. Incorrect or corrupted data can lead to variances in the calculation of a linear regression. Therefore, it is important to ensure that data are collected and checked carefully. Furthermore, the amount of data also plays a key role. If the data are appropriate, the accuracy of an LPP increases with the number of data. There are two different ways to increase the amount of data. On the one hand, it is possible to include more products, i.e. data sets, and on the other hand, it is possible to include more variables or value drivers. The more data sets the model has, the better and more generally valid the model is. The more potential value drivers are included, the more possibilities exist to perform a linear regression and the higher the effort. When selecting the value drivers, care should be taken to provide an appropriate number, so the principle of KISS - keep it simple and stupid - should be applied in this case.

The quality criteria are described in the chapters 3.2.3 Step 3: Test regression function, 3.2.4 Step 4: Test regression coefficients and 3.2.5 Step 5: Test model assumptions. While chapter 3.2.5 Step 5: Test model assumptions does not directly describe quality criteria, but it contains assumptions that have to be fulfilled in order to obtain a meaningful and correct linear regression.

For the quality criteria such as R^2 or R^2_{adj} there are different limit values. It depends on the specific application. Some areas consider a value of $> 50\%$ to be acceptable and others consider a value of at least 70% to be an acceptable limit. In general, it can be said that the value of R^2 and adj. R^2 is better the closer it gets to 100% or 1.0. For the p-value of the F-statistics, this value must be smaller than α in order to be fulfilled. In most cases, α is chosen as 5% or 0.05. For the quality criteria, a distinction can be made between criteria that affect the entire regression - 3.2.4 Step 4: Test regression coefficients - and individual coefficients - 3.2.5 Step 5: Test model assumptions.

5.6 Prototype

A prototype was developed in Python to test certain aspects of the automation potential and to see how this can be implemented in a simple prototype. It uses the library openpyxl[1], as it offers a good range of functions regarding the processing and reading of files such as Excel files. Matplotlib[2] is used for the graphical output. Pandas[3] is a library that is generally used for data analysis and as a manipulation tool, in the prototype the library is mainly used because of its data structure DataFrame. Numpy[4] is additionally used for the preparation for graphical output. The library called statsmodels[5] is used because it offers a very good and broad functionality in terms of regression and various statistical calculations.

The prototype is divided into several files and eight classes to make customisation and extension as easy as possible. The classes include a logger, which is responsible for the log output, an exception class called MyCustomError and a class called DataImporter, which has the task of reading in the data and creating the necessary data structures. The total control of the process, from reading in the data to initiating the calculations and finally creating the output, is performed within the Prototype class. The read data is stored using the class Component, and this class reflects one component. All necessary data, such as the dependent variables and calculation results, are stored in this class and used for further processing. Furthermore, this class also represents the graphical and textual output functionality. After the data is read into the program, a regression is performed. For this purpose, a linear regression or a stepwise regression is executed using the class RegressionManager, which serves as the central instance for selecting the different types of regressions. These two types of regression are mapped in the classes LinearRegression and StepwiseRegression.

[1] https://openpyxl.readthedocs.io/

[2] https://matplotlib.org/

[3] https://pandas.pydata.org/

[4] https://numpy.org/

[5] https://www.statsmodels.org/

Besides the classes listed above, a file named const.py with different parameters is also needed for the configuration and contains all possible settings for the prototype. By changing the parameters, it is possible, among other things, to change the type of regression. The choice is between linear regression and a simplified stepwise regression model. The α and p-limits for the stepwise regression can also be set in this file.

5.6.1 Functionality of the prototype

As this is a first prototype, it does not provide the full range of features to perform an automated linear regression at the moment. However, the first steps towards automation have already been taken. The range of functions includes:

- linear multiple regression
- stepwise regression
- Calculation of various quality criteria
- Performing various graphical and textual tests
- Performing a multiple level regression using a bottom-up approach
- textual as well as graphical output

The prototype offers the possibility to perform a linear regression as well as a stepwise regression. The difference between the two variants is the selection of variables. The linear regression is currently not automated, and the user makes the selection of variables. In the implementation of Stepwise Regression, the selection of variables is made by a simple algorithm. This performs several linear regressions and evaluates the p-value of the coefficients, and determines whether the variables remain in the regression in the next step or not.

With regard to chapter 5.2 Levels of automation, the implemented prototype can be considered to be a level 2 of automation. It already calculates some quality criteria and provides the means for the prerequisite check. These are then interpreted by the user and used for variable selection.

The implementation of a multiple level LPP can be done using a top-down or bottom-up approach. The prototype offers the possibility to run both approaches. The top-down approach is divided into two iterations and so it is necessary to use two separate files as input files in the prototype. The first file contains all data of the main component and the second file contains all data of the subcomponents. In both calculations, the expected, market and best practice values and various criteria and tests are calculated and displayed for each component. In the bottom-up approach, only one input file is needed, which contains all the data of the main component and its subcomponents. The prototype then automatically calculates the expected values of the subcomponent and the main component. Using these values, the prototype can further calculate the market value and best practice value of the main component. Also for this approach, various criteria and tests are calculated and output.

Figure 5.4, figure 5.5 and figure 5.6 show sample outputs of the prototype. The first figure figure 5.4 presents the result of the regression visually. It shows the expected values as a rhombus symbol and illustrates the market and best practice line. This allows the user to check the result quickly and easily.

A check of the correlation between the variables can be done using figure 5.5. It shows a comparison of all the variables. The correlation can be determined visually with the help of the colours or in detail with the help of the calculated correlation values shown in the graph.

Figure 5.6 can be used to visually check the normal distribution of the residuals. The bars show the amount of different residuals and the orange line indicates an auxiliary line which shows the target value of the normal distribution.

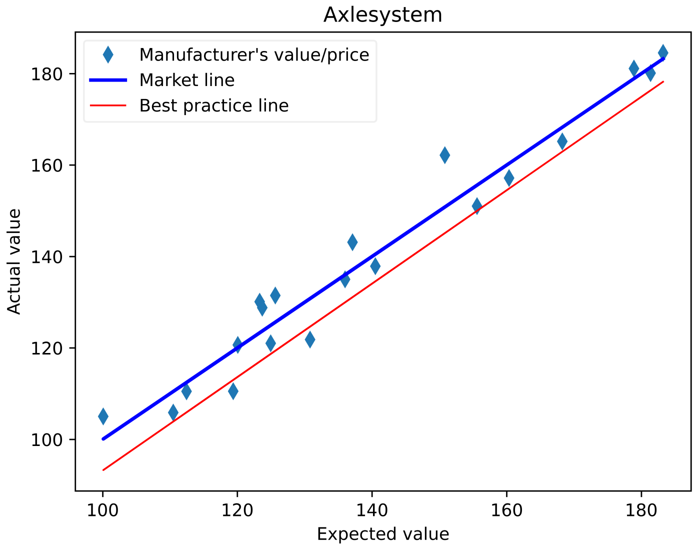


Figure 5.4: Exemplary presentation of the best practice value and best practice line (own representation)

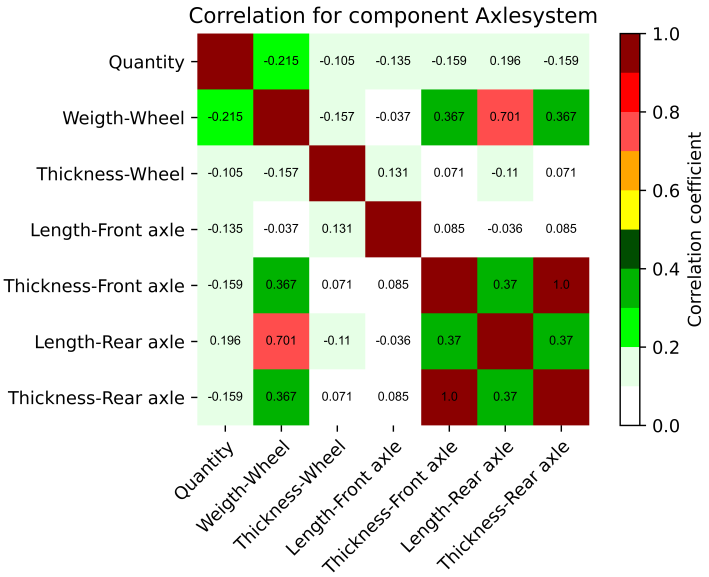


Figure 5.5: Exemplary presentation of a correlation matrix between variables (own representation)

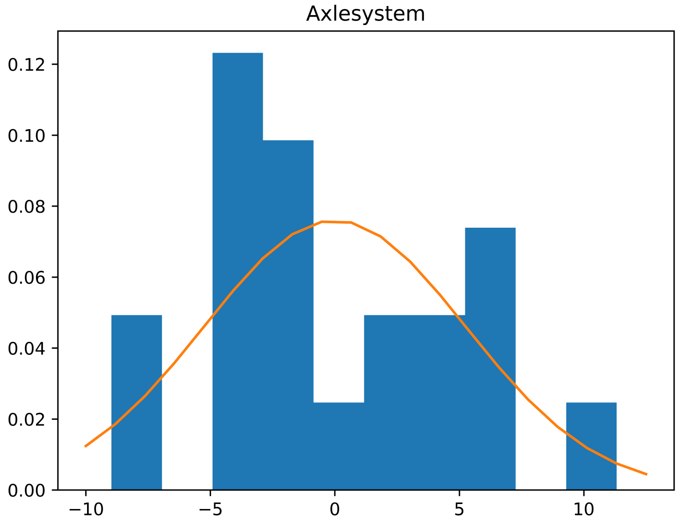


Figure 5.6: Exemplary presentation of a normal distribution of the residuals (own representation)

5.6.2 Testing of the prototype

The scope of functions and the goal of the prototype were not fully defined at the beginning. This was constantly adapted and expanded during the working process. As a result, the focus of the testing was not on the code, but on its functionality. For this reason, the method black box testing was chosen. An advantage of this testing method is that the extension of test cases is also possible by non-developers and that not every single step in the programme has to be tested separately. A disadvantage of this method is that in the case of a negative test case, it is not immediately obvious in which part of the code the error occurs. If the functionality of the prototype is extended, this must also be reflected in the test cases.

In order to be able to perform black box testing, the final result of various test data is specified previously. The prototype is tested by running the regression again using this test data and comparing the result with a predefined final result. If the result deviates from this value, there is an error in the code.

Specifically, the final results for three data sets were calculated for linear regression and for stepwise regression. Six test scenarios were then created from this, which can be checked if the code is changed.

6 Conclusion and outlook

This thesis aimed to identify possible steps in order to automate linear performance pricing. First, research questions were defined, and a methodology to meet the objectives of this work was elaborated.

Data analytics and linear regression were explained to provide a theoretical basis and mathematical foundation of the work. First, in the context of data analytics, a definition of data and a rough overview of the topic were given, and the steps of feature selection were described. Then linear regression was discussed in detail in order to show the necessary steps of linear regression. Afterwards, it was presented how a variable transformation of the dummy variables can be done and how the extension of linear regression to a stepwise regression can be made possible. Based on the literature reviewed and a closer look in the context of this work, it can be said that, that the calculations of a linear regression in the first step are relatively simple. However, in this assumption, the check of the quality criteria and the prerequisites of a regression are mostly disregarded. If these checks are included and done correctly, linear regression becomes more complex than initially assumed.

The economical foundation covered supplier management and linear performance pricing. As Linear Performance Pricing is mainly used in supplier management to achieve cost reductions, supplier management was discussed in detail. Therefore, a definition of supplier management was given, and its process was described with a focus on supplier development. As a change of suppliers is related to high costs and high investment in time, supplier development gets more and more critical. In the next step, LPP was discussed. Therefore a definition of LPP was given, and then three LPP processes were presented. Moreover, practical applications of LPP were discussed. LPP was first mentioned in the journal of McKinsey & Company in 1997. Further approaches by Newman and Krehbiel (2007), Proch, Krampf and Schlüchtermann (2013) and the standardisation of Verein Deutscher Ingenieure (2018) were taken up much later. In addition Verein Deutscher Ingenieure (2021) recently presents a paper about grouping of possible product groups to simplify the manual process. As shown, there is already literature with regard to LPP, whereas the aspect of automation has not yet found its way into the literature. Consequently, the focus of this work was on LPP automation, and possible approaches of automation were illustrated. This also made the difficulties and challenges of automation of LPP obvious.

In the practical part of this work, a review of the manual LPP process steps was given, and then a concept of an automated solution of single and multiple level LPP was presented. First, the review showed that the three presented LPP methods differ in some aspects but also have similarities. Then a concept of a new process for automated LPP was developed and presented. This new approach of automated LPP consists of three stages and a 9-step process and can be seen as a combination of the presented approaches in this work, with the additional focus on a higher degree of automation. In the course of this work, it can be determined that the essence of the automated LPP process is a better and more precise structure into the phases of preparation, technical implementation, economical aspects and evaluation. Furthermore, this structure makes it possible to record individual steps in more detail and thereby automate them.

Moreover, quality criteria were shown, and more details about the prototype were given. Some points of the presented automated process have already been implemented in the prototype developed in the course of this work. If implemented correctly, it is a very extensive program with much statistical functionality. Therefore the first steps of linear regression (single level and multiple level) were implemented in the prototype. Some checks like the Durbin-Watson test and the possibility of a stepwise regression have already been implemented. Since it is not yet a complete collection of tests, there is still no automatic selection of different models. The prototype can be successively expanded with further checks in the following steps to ensure greater automation in the future.

In general, it can be said that the thesis provides the basis for the automation of the LPP process and shows how a concept for the automation of the LPP process can look like. Since automation is very much based on statistical analyses and quality criteria, further research is necessary for this field. Through additional research or the targeted use of artificial intelligence, quality criteria must be found to use it for particular business areas. Besides, diverse statistical methods considered for the calculation in the LPP process have to be compared. It has to be defined which quality criteria have to be selected based on the different dependencies or conditions of the calculation.

This document is set in Palatino, compiled with pdfLaTeX2e and `Biber`.

The LaTeX template from Karl Voit is based on KOMA script and can be found online: https://github.com/novoid/LaTeX-KOMA-template

Bibliography

‘Koordination und Organisation der logistischen Leistungserstellung’ (2008). In: *Handbuch Logistik*. Ed. by Dieter Arnold, Heinz Isermann, Axel Kuhn, Horst Tempelmeier and Kai Furmans. VDI-Buch. Springer Berlin Heidelberg, pp. 971–1015.

Arnolds, Hans, Franz Heege, Carsten Röh and Werner Tussing (2013). *Materialwirtschaft und Einkauf*. Springer Fachmedien Wiesbaden.

Atkinson, Anthony and Marco Riani (2000). *Robust Diagnostic Regression Analysis*. Vol. 97. Springer Series in Statistics 457. Springer New York, pp. 365–366.

Backhaus, Klaus, Bernd Erichson, Wulff Plinke and Rolf Weiber (2018). *Multivariate Analysemethoden*. Springer Berlin Heidelberg.

Balali, Farhad, Jessie Nouri, Adel Nasiri and Tian Zhao (2020). *Data Intensive Industrial Asset Management*. Springer International Publishing.

Berger, Paul D., Robert E. Maurer and Giovana B. Celli (2018). *Experimental Design*. Springer International Publishing.

Blumer, Anselm, Andrzej Ehrenfeucht, David Haussler and Manfred K. Warmuth (Apr. 1987). ‘Occam’s Razor’. In: *Information Processing Letters* 24.6, pp. 377–380.

Bolón-Canedo, Verónica, Noelia Sánchez-Maroño and Amparo Alonso-Betanzos (2015). *Feature Selection for High-Dimensional Data*. Artificial Intelligence: Foundations, Theory, and Algorithms. Springer International Publishing.

Bremen, Philipp Maximilian (2010). 'Total Cost of Ownership'. PhD thesis. ETH Zürich.

Büsch, Mario (2011). *Praxishandbuch Strategischer Einkauf*. Gabler.

Campobasso, Francesco and Annarita Fanizzi (2012). 'A Proposal for a Stepwise Fuzzy Regression: An Application to the Italian University System'. In: *Lecture Notes in Computer Science*. Vol. 3590, pp. 71–87.

Chapman, Timothy, Jack Dempsey, Glenn Ramsdell and Michael Reopel (1997). 'Purchasing: No Time for Lone Rangers'. In: *The McKinsey Quarterly* 2, p. 31.

Durst, Sebastian M. (2011). *Strategische Lieferantenentwicklung*. Vol. 53. Gabler, pp. 1689–1699.

Ellram, Lisa M. (Oct. 1995). 'Total cost of ownership'. In: *International Journal of Physical Distribution & Logistics Management* 25.8, pp. 4–23.

Fleckenstein, Mike and Lorraine Fellows (2018). *Modern Data Strategy*. Springer International Publishing, pp. 1–263.

Gabath, Christoph Walter (2008). *Gewinngarant Einkauf*. Gabler.

Gandomi, Amir and Murtaza Haider (Apr. 2015). 'Beyond the hype: Big data concepts, methods, and analytics'. In: *International Journal of Information Management* 35.2, pp. 137–144.

Godek, Manfred (2021). *LPP und NLPP: Mit Performance Pricing besser verhandeln*. URL: `https://www.technik-einkauf.de/einkauf/strategien/lpp-und-nlpp-mit-performance-pricing-besser-verhandeln-244.html` (visited on 01/10/2021).

Hartmann, Horst, Heinrich Orths and Nina Kössel (2017). *Lieferantenbewertung – aber wie?* 6. Duncker & Humblot.

Heiberger, Richard M. and Burt Holland (2015). *Statistical Analysis and Data Display*. Springer Texts in Statistics. Springer New York.

Helmold, Marc and Brian Terry (2016). *Lieferantenmanagement 2030*. Springer Fachmedien Wiesbaden.

Hendricks, Kevin B. and Vinod R. Singhal (Jan. 2009). 'An Empirical Analysis of the Effect of Supply Chain Disruptions on Long-Run Stock Price Performance and Equity Risk of the Firm'. In: *Production and Operations Management* 14.1, pp. 35–52.

Heß, Gerhard (2008). *Supply-Strategien in Einkauf und Beschaffung*. Gabler.

Heß, Gerhard and Manfred Laschinger (2019). *Strategische Transformation im Einkauf*. Springer Fachmedien Wiesbaden.

Hofbauer, Günter, Tarek Mashhour and Michael Fischer (2012). *Lieferantenmanagement*. Oldenbourg Wissenschaftsverlag.

Hwang, Jing Shiang and Tsuey Hwa Hu (June 2015). 'A stepwise regression algorithm for high-dimensional variable selection'. In: *Journal of Statistical Computation and Simulation* 85.9, pp. 1793–1806.

Irlinger, Wolfgang (2012). *Kausalmodelle zur Lieferantenbewertung*. Gabler.

Jahns, Christopher, Evi Hartmann and Aiko Entchelmeier (July 2007). 'Strategisches Supply Performance Measurement'. In: *Controlling & Management* 51.S2, pp. 74–82.

Jank, Wolfgang (2011). *Business Analytics for Managers*. Springer New York.

Janker, Christian G. (2008). *Multivariate Lieferantenbewertung*. Gabler.

Koppelmann, Udo (2000). *Beschaffungsmarketing*. Springer-Lehrbuch. Springer Berlin Heidelberg.

Krause, Daniel R., Thomas V. Scannell and Roger J. Calantone (Mar. 2000). 'A Structural Analysis of the Effectiveness of Buying Firms' Strategies to Improve Supplier Performance'. In: *Decision Sciences* 31.1, pp. 33–55.

Lasch, Rainer and Christian G. Janker (July 2005). 'Supplier selection and controlling using multivariate analysis'. In: *International Journal of Phys-*

ical Distribution & Logistics Management 35.6. Ed. by Herbert Kopfer, pp. 409–425.

Liebowitz, Jay (2006). *Feature Extraction*. Ed. by Isabelle Guyon, Masoud Nikravesh, Steve Gunn and Lotfi A. Zadeh. Vol. 207. Studies in Fuzziness and Soft Computing. Springer Berlin Heidelberg.

Locker, Alwin and Pan Theo Grosse-Ruyken (2019). *Chefsache Finanzen in Einkauf und Supply Chain*. Springer Fachmedien Wiesbaden, pp. 1–35.

Lorenzen, Klaus Dieter and Wilfried Krokowski (2018). *Einkauf*. Studienwissen kompakt. Springer Fachmedien Wiesbaden.

Modi, Sachin B. and Vincent A. Mabert (Jan. 2007). 'Supplier development: Improving supplier performance through knowledge transfer'. In: *Journal of Operations Management* 25.1, pp. 42–64.

Moreira, João Mendes, André C. P. L. F. de Carvalho and Tomáš Horváth (June 2018). *A General Introduction to Data Analytics*. John Wiley & Sons, Inc.

Münch, Robert M. (Jan. 2018). 'Präzise Preisprognosen'. In: *Bi-spektrum*, pp. 36–39.

Newman, W. Rocky and Timothy C. Krehbiel (Mar. 2007). 'Linear performance pricing: A collaborative tool for focused supply cost reduction'. In: *Journal of Purchasing and Supply Management* 13.2, pp. 152–165.

Online-Magazin für Procurement, Beschaffung, Supply-Chain-Management (SCM) & Digitalisierung (2018). *Mit Non-Linear Performance Pricing zum besten Preis*. URL: `https://einkauf-und-management.at/preisfindung-im-einkauf-mit-non-linear-performance-pricing/` (visited on 01/10/2021).

Pochiraju, Bhimasankaram and Hema Sri Sai Kollipara (2019). 'Statistical Methods: Regression Analysis'. In: pp. 179–245.

Proch, Michael (2017). *Optimale Steuerung der Lieferantenentwicklung*. Springer Fachmedien Wiesbaden.

Proch, Michael, Peter Krampf and Jörg Schlüchtermann (2013). 'Linear Performance Pricing als Instrument zur Kostenoptimierung in der Supply Chain'. In.

Proch, Michael, Karl Worthmann and Jörg Schlüchtermann (Jan. 2017). 'A negotiation-based algorithm to coordinate supplier development in decentralized supply chains'. In: *European Journal of Operational Research* 256.2, pp. 412–429.

Rüdrich, Gerold, Alexander E. Meier and Werner Kalbfuß (2016). *Materialgruppenmanagement: Strategisch einkaufen*. Springer, p. 208.

Runkler, Thomas A. (2020). *Data Analytics*. Vol. 42. 6. Springer Fachmedien Wiesbaden, pp. 385–386.

Saphirion AG (2021). *Non-Linear Performance Pricing (NLPP) - Die smarte Preisanalyse-Lösung*. URL: `https://www.saphirion.com/` (visited on 01/10/2021).

Schmidt, Fabian (2019). 'Application of cost management methodologies for sustainable cost reductions of purchased parts: Comparison and evaluation of different methods in the industrial application'. In: *Journal of Applied Leadership and Management* 7, pp. 66–80.

Schuh, Christian, Joseph L. Raudabaugh, Robert Kromoser, Michael F. Strohmer and Alenka Triplat (2012). *The Purchasing Chessboard*. Springer New York.

Shikhman, Vladimir and David Müller (2021). *Mathematical Foundations of Big Data Analytics*. Springer Berlin Heidelberg.

Stoetzer, Matthias-W. (2017). *Regressionsanalyse in der empirischen Wirtschafts- und Sozialforschung Band 1*.

Stoetzer, Matthias-W. (2020). *Regressionsanalyse in der empirischen Wirtschafts- und Sozialforschung Band 2*.

Stollenwerk, Andreas (2016). *Wertschöpfungsmanagement im Einkauf*. Springer Fachmedien Wiesbaden.

Verein Deutscher Ingenieure (2018). *VDI 2817 Blatt 1 - Performance Pricing (PP) - Grundlagen und Anwendung.*

Verein Deutscher Ingenieure (2021). *VDI 2817 Blatt 2 - Performance Pricing (PP) - Materialgruppenbibliothek.*

Verleysen, Michel, Fabrice Rossi and Damien François (2009). 'Advances in Feature Selection with Mutual Information'. In: ed. by Michael Biehl, Barbara Hammer, Michel Verleysen and Thomas Villmann. Vol. 5400. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 52–69.

Watson, Hugh J. (2014). 'Tutorial: Big Data Analytics: Concepts, Technologies, and Applications'. In: *Communications of the Association for Information Systems* 34.

Watts, Charles A. and Chan K. Hahn (1993). 'Supplier Development Programs: An Empirical Analysis'. In: *International Journal of Purchasing and Materials Management* 29.1, pp. 10–17.

Wildemann, Horst (2008). *Einkaufspotentialanalyse: Programme zur partnerschaftlichen Erschliessung von Rationalisierungspotentialen.* 2. Auflage. TCW, Transfer-Centrum.

Wooldridge, J.M. (2013). *Introductory Econometrics: A Modern Approach.* Cengage Learning.

Xia, Chun Yan (2014). 'Construction and Application of Multivariate Linear Regression Model on Road Cost'. In: *Applied Mechanics and Materials* 556-562, pp. 807–811.

Yu, Chun and Weixin Yao (2017). 'Robust linear regression: A review and comparison'. In: *Communications in Statistics - Simulation and Computation* 46.8, pp. 6261–6282.

Żogała-Siudem, Barbara and Szymon Jaroszewicz (2020). 'Fast Stepwise Regression Based on Multidimensional Indexes'. In: *Information Sciences* 549, pp. 288–309.