



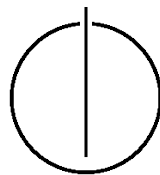
FAKULTÄT FÜR INFORMATIK

DER TECHNISCHEN UNIVERSITÄT MÜNCHEN

Abschlussarbeit in Informatik

**Effiziente statistische Methoden für  
Datenbanksysteme**

Thomas Heyenbrock







FAKULTÄT FÜR INFORMATIK

DER TECHNISCHEN UNIVERSITÄT MÜNCHEN

Abschlussarbeit in Informatik

Effiziente statistische Methoden für Datenbanksysteme

Efficient statistical methods for database systems

Autor:	Thomas Heyenbrock
Aufgabensteller:	Prof. Alfons Kemper, Ph.D.
Betreuer:	Maximilian E. Schüle, M.Sc.
Datum:	15.01.2017





Ich versichere, dass ich diese Diplomarbeit selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel verwendet habe.

München, den 27. Dezember 2017

Thomas Heyenbrock



---

## **Abstract**

An abstracts abstracts the thesis!

---



# Contents

<b>Abstract</b>	<b>vii</b>
<b>Outline of the Thesis</b>	<b>xi</b>
<b>1. Einführung und typische statistische Problemstellungen</b>	<b>1</b>
1.1. Latex Introduction . . . . .	1
<b>2. Grundlagen statistischer Methoden</b>	<b>3</b>
2.1. Lineare Regression . . . . .	4
2.1.1. Einfache lineare Regression . . . . .	5
2.1.2. Multiple lineare Regression . . . . .	5
2.2. Logistische Regression . . . . .	6
<b>3. Anwendung statistischer Methoden</b>	<b>7</b>
3.1. Latex Introduction . . . . .	7
<b>4. Statistische Methoden in Datenbanken</b>	<b>9</b>
4.1. Latex Introduction . . . . .	9
<b>5. Erweiterungspotenzial in Datenbanksystemen</b>	<b>11</b>
5.1. Latex Introduction . . . . .	11
<b>6. Fazit</b>	<b>13</b>
6.1. Latex Introduction . . . . .	13
<b>Appendix</b>	<b>17</b>
<b>A. Detailed Descriptions</b>	<b>17</b>
<b>Bibliography</b>	<b>19</b>



# Outline of the Thesis

## **Teil I: Introduction and Theory**

### CHAPTER 1: INTRODUCTION

This chapter presents an overview of the thesis and its purpose. Furthermore, it will discuss the sense of life in a very general approach.

### CHAPTER 2: THEORY

No thesis without theory.

## **Teil II: The Real Work**

### CHAPTER 3: OVERVIEW

This chapter presents the requirements for the process.



# 1. Einführung und typische statistische Problemstellungen

Here starts the thesis with an introduction. Please use nice latex and bibtex entries [1]. Do not spend time on formating your thesis, but on its content.

## 1.1. Latex Introduction

There is no need for a latex introduction since there is plenty of literature out there.



## 2. Grundlagen statistischer Methoden

Bei der Regressionsanalyse geht es im Allgemeinen darum, das Verhalten einer Größe  $Y$  in Abhängigkeit einer oder mehrerer anderer Größen  $X_1, X_2, \dots, X_n$  zu modellieren. Die Größe  $Y$  wird abhängig genannt, die Größen  $X_i$  nennt man unabhängig. Für diese Arbeit wollen wir zunächst einige Annahmen über diese voraussetzen. Diese Punkte gelten immer, falls nicht explizit etwas anderes festgelegt wird.

- Die genannten Größen sind Zufallsvariablen. Das sind Funktionen deren Werte die Ergebnisse eines Zufallsvorgangs darstellen.
- Die Zufallsvariablen sind auf der Menge  $M = \{1, \dots, m\}$  definiert und bilden in die reellen Zahlen ab:

$$Y : M \rightarrow \mathbb{R}, \quad X_1 : M \rightarrow \mathbb{R}, \quad \dots, \quad X_n : M \rightarrow \mathbb{R}$$

Das bedeutet die Zufallsvariablen sind metrisch skaliert. Die  $m$  Zahlen in der Menge  $M$  entsprechen den  $m$  Datenpunkten, die wir als Datenbasis für die Regressionsanalyse besitzen.

- Wir verwenden die folgenden Abkürzungen für die Werte der Zufallsvariablen:

$$\begin{aligned} y_i &:= Y(i) \quad \text{für alle } i \in M, \\ x_{i,j} &:= X_j(i) \quad \text{für alle } i \in M \text{ und } 1 \leq j \leq n \end{aligned}$$

- Einen Datenpunkt aus unserer Datenbasis fassen wir als Vektor der Länge  $(n + 1)$  auf. Damit lässt sich die Datenbasis schreiben als:

$$(y_1, x_{1,1}, \dots, x_{1,n}), \dots, (y_m, x_{m,1}, \dots, x_{m,n})$$

Das Modell definieren wir anhand einer Funktion  $f$ , welche für Werte der unabhängigen Variablen einen geschätzten Wert für die abhängige Variable liefert. Idealerweise existiert eine Funktion, die zum Einen eine einfache Darstellung (z.B. durch eine arithmetische Formel) besitzt und zum Anderen alle unabhängigen Werte der Datenmenge exakt prognostiziert. Das bedeutet:

$$y_i = f(x_{i,1}, \dots, x_{i,n}) \quad \text{für alle } 1 \leq i \leq m$$

Falls eine Formel wie hier für alle Datenpunkte gelten soll, verwenden wir als Abkürzung auch die Zufallsvariablen selbst, also:

$$Y = f(X_1, \dots, X_N)$$

Im Allgemeinen ist es nicht möglich eine Funktion  $f$  zu finden, die beide Eigenschaften erfüllt. Man versucht also eine Funktion mit einer möglichst einfachen Form zu finden, die die Datenmenge möglichst gut approximiert. Wir definieren für jeden Datenpunkt den Fehler  $e_i$ , der sich durch die nicht exakte Modellfunktion  $f$  ergibt:

$$e_i = y_i - f(x_{i,1}, \dots, x_{i,n})$$

Ziel der Regressionsanalyse ist es nun eine Funktion  $f$  zu finden, die diese Fehlerterme minimiert. Diese Optimierung geschieht global, also für die gesamte Datenmenge und nicht nur für einzelne Datenpunkte.

### 2.1. Lineare Regression

Bei der linearen Regression geht man von einem linearen Zusammenhang zwischen der abhängigen und den unabhängigen Variablen aus. Die Funktion  $f$  ist also von folgender Form:

$$f(x_1, \dots, x_n) = \alpha + \sum_{i=1}^n \beta_i \cdot x_i \quad \text{mit } \beta_i \in \mathbb{R}$$

Das Maß für die Qualität einer Funktion  $f$  definiert durch die Parameter  $\alpha, \beta_1, \dots, \beta_n$  ist die Summe der quadrierten Fehlerterme:

$$E(\alpha, \beta_1, \dots, \beta_n) = \sum_{j=1}^m e_j^2 = \sum_{j=1}^m (y_j - f(x_{j,1}, \dots, x_{j,n}))^2 = \sum_{j=1}^m \left( y_j - \alpha - \sum_{i=1}^n \beta_i \cdot x_{i,j} \right)^2$$

Wir suchen also die Parameter  $\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_n$  für die gilt:

$$E(\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_n) = \min \{ E(\alpha, \beta_1, \dots, \beta_n) \mid \alpha \in \mathbb{R}, \beta_1 \in \mathbb{R}, \dots, \beta_n \in \mathbb{R} \}$$

Um dieses Minimierungsproblem zu lösen berechnen wir die partiellen Ableitungen von  $E$ .

$$\begin{aligned} \frac{\partial E}{\partial \alpha} &= -2 \cdot \sum_{j=1}^m (y_j - f(x_{j,1}, \dots, x_{j,n})) = -2 \cdot \sum_{j=1}^m \left( y_j - \alpha - \sum_{i=1}^n \beta_i \cdot x_{i,j} \right) \\ \frac{\partial E}{\partial \beta_k} &= -2 \cdot \sum_{j=1}^m x_{k,j} \cdot (y_j - f(x_{j,1}, \dots, x_{j,n})) \\ &= -2 \cdot \sum_{j=1}^m x_{k,j} \cdot \left( y_j - \alpha - \sum_{i=1}^n \beta_i \cdot x_{i,j} \right) \quad \text{für } 1 \leq k \leq n \end{aligned}$$

Durch Nullsetzen der partiellen Ableitungen erhält man ein lineares Gleichungssystem mit  $(n+1)$  Gleichungen und ebensovielen Unbekannten.

$$\frac{\partial E}{\partial \alpha} = 0, \quad \frac{\partial E}{\partial \beta_1} = 0, \quad \dots, \quad \frac{\partial E}{\partial \beta_n} = 0$$

Die Lösung dieses Gleichungssystems (falls eine existent) ist das gesuchte Minimum.



### 2.1.1. Einfache lineare Regression

Man spricht von einfacher linearer Regression, wenn man mit nur eine unabhängige Variable arbeitet. Anschaulich möchte man hier die bestmögliche Schätzgerade durch eine gegebene Punktwolke legen.

Wir nennen die unabhängige Variable in diesem Kapitel statt  $X_1$  einfach nur  $X$ . Ebenso schreiben wir  $\beta_1 = \beta$  und  $x_{1,j} = x_j$ . Dann können wir das lineare Gleichungssystem zum Auffinden des Minimums explizit aufschreiben:

$$\begin{aligned} 0 &= -2 \cdot \sum_{j=1}^m (y_j - \alpha - \beta \cdot x_j) \\ 0 &= -2 \cdot \sum_{j=1}^m x_j \cdot (y_j - \alpha - \beta \cdot x_j) \end{aligned}$$

Für dieses Gleichungssystem kann die Lösung explizit angegeben werden, wobei wir hier nicht näher auf die Herleitung dieses Ergebnisses eingehen wollen:

$$\begin{aligned} \hat{\beta} &= \frac{\sum_{j=1}^m (x_j - \bar{x})(y_j - \bar{y})}{\sum_{j=1}^m (x_j - \bar{x})^2} \\ \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} \end{aligned}$$

Dabei bezeichnen  $\bar{x}$  und  $\bar{y}$  die Mittelwerte von  $X$  respektive  $Y$ .

### 2.1.2. Multiple lineare Regression

Bei multibler linearer Regression existieren mindestens zwei unabhängige Variablen. Hier ist es nicht mehr zweckmäßig eine explizite Lösung anzugeben. Hier sind alternative Methoden zur Berechnung der Parameter nötig.

Neben einer Vielzahl von Algorithmen, die ein Optimierungsproblem iterativ lösen, gibt es auch die Möglichkeit die Parameter durch Matrizenmultiplikation zu berechnen. Definieren wir dazu die folgenden Matrizen und Vektoren:

$$\begin{aligned} X &= \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m,1} & \dots & x_{m,n} \end{pmatrix} \in \mathbb{R}^{m \times (n+1)} \\ y &= \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} \in \mathbb{R}^{m \times 1}, \quad b = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_n \end{pmatrix} \in \mathbb{R}^{(n+1) \times 1} \end{aligned}$$

Dabei ist  $b$  der Vektor mit gesuchten Parametern für die Minimierung der kleinsten Quadrate. Falls die Matrix  $X^T X$  invertierbar ist, gilt die folgende Formel für die Berechnung der gesuchten Parameter:

$$b = (X^T X)^{-1} X^T y$$

## 2.2. Logistische Regression

Die logistische Regression findet Anwendung im Falle, dass die abhängige Variable eine binäre Variable ist, also eine Variable, die nur zwei Werte annehmen kann. Oft handelt es sich um eine Eigenschaft, die ein bestimmter Datensatz besitzt oder nicht, wie zum Beispiel ein Premium-Abonnement für eine Web-Service oder der Besitz eines Auto. Auch das Geschlecht einer Person ist ein Beispiel für eine binäre Variable. Wir bezeichnen die beiden möglichen Werte einer solchen Variablen hier immer mit 0 und 1. Die Zuordnung vom Merkmal zur Zahl ist frei wählbar.

Lineare Regression eignet sich oft nicht zur Modellierung einer binären Variablen, da eine lineare Funktion in der Regel unbeschränkt ist, also insbesondere Werte größer als 1 und kleiner als 0 annimmt. Um diesem Problem abzuweichen wählen wir eine Funktion, die beliebige Zahlen auf das Intervall  $[0, 1]$  abbildet. Im Falle der logistischen Regression verwendet man die gleichnamige logistische Funktion:

$$l : \mathbb{R} \rightarrow (0, 1), \quad x \mapsto \frac{1}{1 + e^{-x}}$$

Diese Funktion wendet man nun auf die Linearkombination aller unabhängigen Variablen mit Parametern  $\beta_1, \dots, \beta_n$  und konstantem Term  $\alpha$  an. Das Ergebnis für den  $i$ -ten Datensatz bezeichnen wir mit  $\pi_i$

$$\pi_i = \pi_i(\alpha, \beta_1, \dots, \beta_n) := l\left(\alpha + \sum_{j=1}^n \beta_j \cdot x_{i,j}\right) = \left(1 + \exp\left(-\alpha - \sum_{j=1}^n \beta_j \cdot x_{i,j}\right)\right)^{-1}$$

Anschaulich repräsentiert  $\pi_i$  die Wahrscheinlichkeit dafür, dass die abhängige Variable eines Datensatzes mit unabhängigen Variablen  $x_{i,1}, \dots, x_{i,n}$  gleich 1 ist, also:

$$\pi_i = P(Y_i = 1 | X_1 = x_{i,1}, \dots, X_n = x_{i,n})$$

Man möchte die Parameter  $\alpha, \beta_1, \dots, \beta_n$  nun so schätzen, dass die Wahrscheinlichkeit für das Auftreten der vorhandenen Datenbasis maximiert wird. Diese Wahrscheinlichkeit ist gegeben durch:

$$L(\alpha, \beta_1, \dots, \beta_n) = \prod_{i=1}^m P(Y_i = y_i | X_1 = x_{i,1}, \dots, X_n = x_{i,n}) = \prod_{i=1}^m y_i \cdot \pi_i(\alpha, \beta_1, \dots, \beta_n)$$

Dieses Verfahren bezeichnet man auch als Maximum-Likelihood-Methode. Die Funktion  $L$  nennt man dementsprechend auch Likelihoodfunktion. Oft maximiert man nicht  $L$  direkt, sondern eher  $\ln(L)$ . Der Sinn ist, dass man das Produkt damit in eine Summe einzelner Logarithmen umwandeln kann, welche wiederum einfacher abzuleiten ist. Das darf man machen, da der Logarithmus eine stetig wachsende Funktion ist und die Werte von  $L$  stets zwischen 0 und 1 liegen.

Wir suchen also die Parameter  $\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_n$  mit:

$$L(\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_n) = \max \{L(\alpha, \beta_1, \dots, \beta_n) \mid \alpha \in \mathbb{R}, \beta_1 \in \mathbb{R}, \dots, \beta_n \in \mathbb{R}\}$$

In diesem Fall kommt man leider nicht mehr an einer iterativen Lösung vorbei, da die partiellen Ableitungen und das entstehende lineare Gleichungssystem nicht mehr exakt lösbar sind. Eine der einfachsten Methoden zur Lösung von Optimierungsproblemen ist das Gradientenverfahren, welches im kommenden Teilkapitel kurz eingeführt wird.

## **3. Anwendung statistischer Methoden**

Here starts the thesis with an introduction. Please use nice latex and bibtex entries [1]. Do not spend time on formating your thesis, but on its content.

### **3.1. Latex Introduction**

There is no need for a latex introduction since there is plenty of literature out there.



## **4. Statistische Methoden in Datenbanken**

Here starts the thesis with an introduction. Please use nice latex and bibtex entries [1]. Do not spend time on formating your thesis, but on its content.

### **4.1. Latex Introduction**

There is no need for a latex introduction since there is plenty of literature out there.



## **5. Erweiterungspotenzial in Datenbanksystemen**

Here starts the thesis with an introduction. Please use nice latex and bibtex entries [1]. Do not spend time on formating your thesis, but on its content.

### **5.1. Latex Introduction**

There is no need for a latex introduction since there is plenty of literature out there.





## 6. Fazit

Here starts the thesis with an introduction. Please use nice latex and bibtex entries [1]. Do not spend time on formating your thesis, but on its content.

### 6.1. Latex Introduction

There is no need for a latex introduction since there is plenty of literature out there.



# Appendix



## **A. Detailed Descriptions**

Here come the details that are not supposed to be in the regular text.



# Bibliography

- [1] Leslie Lamport. *LaTeX : A Documentation Preparation System User's Guide and Reference Manual*. Addison-Wesley Professional, 1994.