universität wien

DIPLOMARBEIT

Lineare Regression
&
Hauptkomponentenanalyse

Verfasser:

Hofegger Manuel

Angestrebter akademischer Grad:

Magister der Naturwissenschaften

Wien, im April 2015

Studienkennzahl laut Studienblatt: **A 190 456 406**

Studienrichtung lt. Studienblatt**: Lehramt Mathematik / Geographie & Wirtschaftsk.**

Betreuer: **ao. Univ.-Prof. tit. Univ.-Prof. Dr. Hans Georg Feichtinger**

Danksagung

An dieser Stelle ist es mir ein Anliegen,
sowohl meinem Diplomarbeitsbetreuer

Herrn Prof. Dr. H. G. Feichtinger

als auch meinen Eltern und meiner Freundin Sarah

meinen Dank auszusprechen, da sie wesentlich zum
Absolvieren meines Studiums beigetragen haben.

Insbesondere möchte ich hier auch meinen Nachbar
Herrn Fritz Track erwähnen, der mir bei auftretenden
Fragen während des Studiums, immer mit gutem Rat
zur Seite stand.

Vorwort

Die unter dem Begriff *Regressionsanalyse* zusammengefassten statistischen Verfahren dienen der statistischen Analyse von Zusammenhängen zwischen zwei oder mehreren Zufallsvariablen.
Sofern eine solche qualitative Analyse den Zusammenhang zwischen zwei Zufallsvariablen behandelt, spricht man von einfacher Regressionsanalyse, handelt es sich um die qualitative Analyse eines Zusammenhangs zwischen mehr als zwei Zufallsvariablen, ist generell von mehrfacher oder multipler Regressionsanalyse die Rede.

In der vorliegenden Diplomarbeit soll im **Kapitel 1** eine Einführung in die einfache lineare Regression gegeben werden, obwohl die Einfachregression nur als Spezialfall der Mehrfachregression betrachtet werden kann. Allerdings lässt sie sich naturgemäß anschaulicher repräsentieren und eignet sich somit adäquat zur Vermittlung grundsätzlicher Überlegungen. Ausgehend von einer Stichprobenerhebung versucht die lineare Regression, die Art der Beziehung zwischen zwei Variablen festzustellen und durch eine mathematische Funktion diesen Zusammenhang zu beschreiben.
Relevanz wird dabei der Beziehung zwischen der abhängigen und der erklärenden Variable beigemessen, die von unabhängigen Parametern, den sogenannten Regressionskoeffizienten, abhängt. Als Standardinstrument für derartige Schätzungen wird die *Methode der kleinsten Quadrate* heran gezogen.
Vorab sollte man allerdings im Rahmen der Korrelationsanalyse prüfen, ob überhaupt ein linearer Zusammenhang zwischen den Variablen besteht, ehe dazu übergegangen wird, diesen zu quantifizieren und die Art des Zusammenhangs funktional zu beschreiben.

Im **Kapitel 2** wird das klassische Modell der linearen Einfachregression charakterisiert, welches ein Modell der Gesamtsituation erfordert, um ausgehend von einer Stichprobe, geeignete Schlüsse auf die Grundgesamtheit zu ermöglichen. Dazu werden notwendige Bedingungen aufgestellt, welche für die Validität des Modells erforderlich sind und es wird ausführlich auf die auftretenden Modellparameter eingegangen.
Für die Herleitung von statistischen Tests und Konfidenzintervallen wird das Modell zusätzlich durch die Normalverteilungsannahme erweitert.
Außerdem beinhaltet das Kapitel graphische Analysemethoden, die zeigen sollen, wie die Modellvoraussetzungen überprüft werden können, indem Residuen analysiert und auf Verletzungen der Normalverteilung, Linearität und Varianzhomogenität Bezug genommen wird.
Im nächsten Schritt werden zunächst die Überlegungen und Ergebnisse aus den ersten beiden Kapiteln auf die lineare Mehrfachregression erweitert bzw. notwendige Zusatzüberlegungen, bedingt durch das Vorhandensein von mehreren Variablen, angestellt.
Allerdings soll im **Kapitel 3** auch ein anderer Zugang Beachtung finden, welcher die Teilräume des R^n betrachtet, sowie die dazugehörigen orthogonalen Projektionen auf diese Teilräume. Um die einfache lineare Regression als Spezialfall der linearen

Mehrfachregression hervorzuheben und zu betonen, dass deren Anwendung problemlos auf jene der Mehrfachregression zurückgeführt werden kann, ist der Umgang mit detaillierten Beweisen in Kapitel 2 noch dürftig und wird im Kapitel 3 forciert.
Motivierend soll im Zuge dessen die Anwendung der Varianzanalyse sein, die in Form eines Vergleiches mehrerer VW - Automodelle in Hinblick auf eine quantitative Variable y durchgeführt wird.
Die Automodelle die man vergleichen möchte, können unterschiedliche Gruppen bilden (angegeben durch eine x–Variable), allerdings sind für klassische lineare Regressionen nur solche mit metrischem Skalenniveau geeignet (siehe Kapitel über Typen von Skalierungen).
Im Anschluss daran wird analog für die allgemeine Regressionsanalyse ein Maß für die Güte der Modellanpassung unter Zuhilfenahme der Quadratsummenzerlegung hergeleitet. Das Modell wird wiederum durch die Normalverteilungsannahme erweitert und ermöglicht das Herleiten von Hypothesentests und Konfidenzintervallen.

Kapitel 4 behandelt nun Abweichungen der Modellvoraussetzungen, indem die Ursachen, Gründe, bzw. eventuellen Lösungsmöglichkeiten thematisiert werden.
Im Fokus sollen hier vor allem das Problem der Kollinearität der unabhängigen Variablen bei der linearen Mehrfachregression stehen, ebenso wie mögliche Lösungen für Varianzinhomogenität.

Das **Kapitel 5** setzt sich im Wesentlichen mit einfachen und doppelten Varianzanalysen auseinander, die in allgemeiner Form auch durch Hypothesentests erfolgen können. Angestrebt wird demnach das Zerlegen einer vorliegenden Stichprobe in normalverteilte Teilstichproben, deren Mittelwerte dann miteinander verglichen werden ehe im **Kapitel 6** noch Testverteilungen und Tests für Verteilungen charakterisiert werden.

Kapitel 7 & 8 stellen primär, durch Eigenwerte/Eigenvektoren, Diagonalisierbarkeit, Orthogonalitätsprojektionen und schließlich der Singulärwertzerlegung, einen Bezug zur Linearen Algebra her und leiten schließlich zum **Kapitel 9** der Hauptkomponentenanalyse über.

Der Titel der Arbeit sagt aus, dass primär die lineare Regression diskutiert wird und somit ein linearer Zusammenhang vorliegt. Das Schlusswort soll allerdings noch einen Ausflug in nichtlineare Regressionsprobleme beinhalten. Das Ziel besteht darin, zu zeigen, dass nichtlineare Regressionsprobleme mit Hilfe der linearen Regression einfacher als auf direktem Weg lösbar sind.
Ein Anliegen dieser Arbeit ist es, die Regressionsanalyse zusätzlich zum theoretischen Hintergrund, wenn möglich mit Hilfe von Beispielen zu „untermauern“.
Die entsprechenden Datensätze für die verschiedenen Beispiele sind im Anhang angeführt, der auch noch die statistischen Verteilungen und ihre Dichtefunktionen umfasst.

Ferner wird zur statistischen Datenanalyse **SPSS - 22** und **Geogebra** verwendet.

Inhaltsverzeichnis

Kapitel 1

1. EINFACHE LINEARE REGRESSION

1.1 Einführung

Erster Schritt der deskriptiven Regressionsanalyse ist die Auswahl der interessierenden abhängigen und unabhängigen Variablen, deren Zusammenhang beschrieben werden soll. Wir gehen also von einer zweidimensionalen Stichprobe $\{(x_1.y_1),.....,(x_n,y_n)\}$ aus und nehmen die Werte der Variable X an n-Untersuchungseinheiten als fest und jene der Variable Y als zufällig an. Anders formuliert betrachten wir im folgenden X als unabhängige und Y als abhängige Variable, d.h. eine entsprechende Variation der Realisierungen von x_i wird heran gezogen, um die auftretenden unterschiedlichen y_i – Werte zu erklären. Nun wollen wir Y als Funktion von X darstellen. Im einfachsten Fall liegen alle Punkte auf einer Geraden, somit wird ein solcher Zusammenhang durch eine lineare Funktion dargestellt.

$$Y = \beta_0 + \beta_1 X \qquad (1)$$

Sofern die Datenpunkte des Stichprobenumfangs allerdings nicht genau auf einer Geraden liegen muss (1) modifiziert werden. Die Differenz zwischen dem beobachteten, exakten Y-Wert und dem Messwert der linearen Funktion $\beta_0 + \beta_1 X$ wird als ε ausgegeben. Diese Fehlervariable ε steht für eine Zufallsvariable, die eventuelle Datenfehler, Messfehler etc. umfasst. Darum kann ein plausibleres Modell durch

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad (2)$$

(1) Y: zu erklärende quantitative Größe bzw. Regressand (endogen: im Modell erklärt)
(2) X_1; X_2: erklärende Größen (Regressoren; exogen: nicht innerhalb des Modells zu erklären)
(3) ε: Fehlervariable bzw. Residuum (theoretisch unerklärter Rest)

angegeben werden, wobei β_0, $\beta_1 \in \mathbf{R}$ die Regressionskoeffizienten und ε eine Fehlervariable repräsentieren, die all jene Einflüsse auf die abhängige Variable beinhaltet, welche im Modell nicht explizit enthalten sind.
Diese Regressionsgleichung ist *linear*, weil es eine Polynomfunktion 1.ten Grades vorliegt. Zudem ist sie *einfach*, weil zur Erklärung zufälliger Schwankungen der abhängigen Variablen ein Regressor ausreicht.

Das Ziel von Regressionsgleichungen aller Art ist natürlich die zugrunde liegende Stichprobe „möglichst gut“ widerzuspiegeln. Dies erfolgt durch Anpassen einer „Ausgleichsgeraden“ an die Punktwolke der Stichprobe. Nach Augenmaß können sich verschiedene Geraden als Näherung der Punktwolke eignen, zunächst steht nicht fest, welche die Optimalste ist. Somit liegt es auf der Hand, dass eine geschätzte Bestimmung der Koeffizienten β_0 und β_1 sowie des Residuums ε unabdingbar sind.

1.2 Deskriptive lineare Regression

Im **zweiten Schritt** steht die Festlegung einer geeigneten *Funktionsform* für die Regressionsgleichung im Vordergrund, wobei die thematisierte deskriptive Regression darauf abzielt, statistische Abhängigkeiten zwischen Variablen zu beschreiben, ohne ein statistisches Modell anzuwenden.

In diesem Abschnitt lässt sich mit der Methode der kleinsten Quadrate somit schon die Standardlösungsvariante für einfache lineare Regressionen herleiten.
Betrachten wir eine zweidimensionale Stichprobe $\{(x_1.y_1),.....,(x_n,y_n)\}$, gegeben durch die Merkmale X und Y an n – Untersuchungseinheiten, so kann der Zusammenhang durch ein (x,y) – Diagramm graphisch in Form einer Punktwolke visualisiert werden.
Das dadurch entstehende Streudiagramm enthält nun alle einzelnen Punkte aus der Datenmatrix.

Beispiel 1: Die praktische Beschreibung der einfachen linearen Regression erfolgt nun durch eine im Anhang angeführte Datenmatrix, die einen Zusammenhang zwischen der Leistung in KW und dem Diesel-Kraftstoffverbrauch für VW-Standardmodelle mit Basisausstattung (aus dem Leitfaden über Kraftstoffverbrauch 2015 – Tabelle im Anhang) zeigt:

Kraftstoffverbrauch bei entsprechender Leistung in KW (siehe Tabelle 1, Anhang)

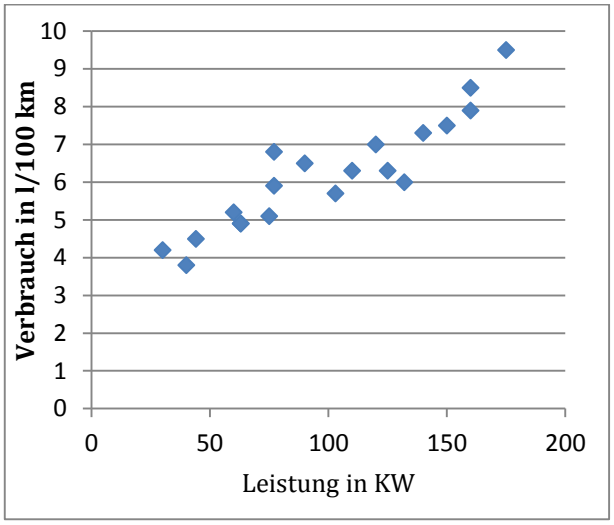

Abbildung 1: Kraftstoffverbrauch bei entsprechender Leistung in KW

In Abbildung 1 ist ersichtlich, dass die graphisch dargestellten Werte approximativ auf einer Geraden liegen und wir daher einen linearen Zusammenhang zwischen den Merkmalen X (Leistung in KW) und Y (Verbrauch in l/ 100 km) annehmen können. Gesucht ist nun jene lineare Regressionsgleichung (2), welche die vorhandene Stichprobe im Diagramm durch eine „optimale Ausgleichsgerade“ anpasst.

Definition 1.2.1: Sei e_i die Differenz zwischen dem gemessenem Wert y_i und dem durch Regressionsgerade berechneten Wert $\hat{y}_i$ (mit i = 1, ... n), dann wird

$e_i := y_i - \hat{y}_i$ als i-ter Vorhersagefehler oder i-tes Residuum definiert. Die Residuen e_i entsprechen den ε_i Fehlervariablen in (2). Die Interpretation dazu sagt aus, dass sofern sich der y_i – Wert unterhalb der „Ausgleichsgerade" befindet, das Residuum ε_i negativ und im umgekehrten Fall das Residuum positiv ist. Beim Wert 0 liegt der y_i Wert exakt auf der Geraden und somit ist auch der Fehler gleich 0.
Unser festgelegter Anspruch liegt nun darin, die Summe der Vorhersagefehler zu minimieren, indem die Summe der Residuen $\sum_{i=1}^{n} e_i$ möglichst klein wird.
Prinzipiell dagegen sprechen allerdings zwei Gründe:

- Nachdem sowohl negative als auch positive Abweichungen auftreten können, heben sich die Fehler möglicherweise auf und die dadurch festgelegte Gerade passt sich den Messwertepaaren schlecht an, obwohl die Summe der e_i annähernd oder gleich 0 ist.

- Darüber hinaus kann es passieren, dass die Eindeutigkeitseigenschaft der durch Minimierung der Summe e_i ermittelten Geraden nicht gegeben ist.

Alternativ besteht die Möglichkeit die Summe der Beträge der Residuen $\sum_{i=1}^{n} |e_i|$ zu minimieren, wogegen im Wesentlichen zwei Einwände relevant sind:

- Einerseits erweist sich die Betragsfunktion als rechentechnisch ungünstig

- Andererseits ist es sinnvoll, wie später noch veranschaulicht wird, die betragsmäßig größeren Abweichungen von der errechneten Geraden mit höherer Priorität zu behandeln und den betragsmäßig kleineren Abweichungen nicht zu viel Aufmerksamkeit zukommen zu lassen. Begründet kann dieses Vorgehen damit werden, dass betragsmäßig kleine Abweichungen des Gemessenen vom errechneten Wert oft durch zufällige Einflüsse (wie Messfehler) eintreten, betragsmäßig große Abweichungen jedoch systemischer Art sein können.

1.2.1 Die Methode der kleinsten Quadrate nach Gauß

Die gewöhnliche Methode der kleinsten Quadrate konstruiert eine Ausgleichsgerade, mit dem Fehler e_i als vertikalem Abstand des Punktes (x_i, y_i) von der Geraden, und zwar so, dass die Quadratsumme der Abweichungen aller Punkte minimal wird. (**SSE** = **S**um of **S**quares of **E**rrors). In diesem **dritten Schritt** erfolgt somit im Wesentlichen die Bestimmung der Koeffizienten der Regressionsgleichung.

Zunächst wird dazu die Bestimmung von Schätzwerten $\hat{\beta}_0$, $\hat{\beta}_1$ für die unbekannten Parameter β_0, β_1 diskutiert, bei der keine zusätzlichen Voraussetzungen über die Störgröße ε nötig sind und wir minimieren anschließend:

$$S(\beta_0, \beta_1) = \frac{1}{n}\sum_{i=1}^{n} e_i{}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

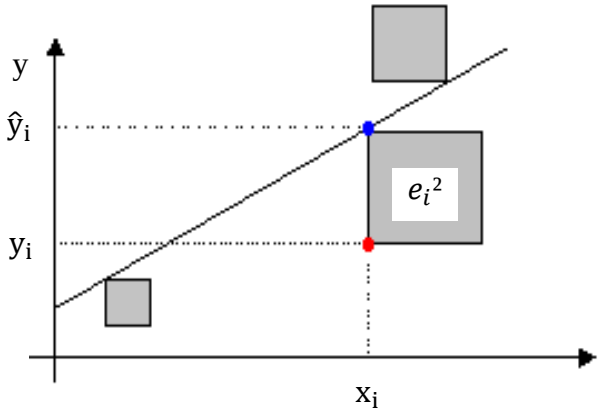

Abbildung 2: geometrische Veranschaulichung der Methode der kleinsten Quadrate

Notwendige Bedingung für die Existenz eines Minimums an einem Punkt $(\hat{\beta}_0, \hat{\beta}_1)$: Verschwinden der partiellen Ableitungen $\frac{\partial S}{\partial \beta_0}$ und $\frac{\partial S}{\partial \beta_1}$:

$$0 = \frac{\partial S}{\partial \beta_0}(\hat{\beta}_0, \hat{\beta}_1) = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \qquad (3)$$

$$0 = \frac{\partial S}{\partial \beta_1}(\hat{\beta}_0, \hat{\beta}_1) = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)\, x_i = \sum_{i=1}^{n}(y_i x_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2) \qquad (4)$$

$\Rightarrow$ **Normalgleichungen:**

$$n\bar{y} - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0 \quad \Rightarrow \quad \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \quad \text{und} \qquad (5)$$

$$\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i{}^2 = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} x_i y_i = \hat{\beta}_1 \sum_{i=1}^{n} x_i{}^2 + \hat{\beta}_0 n\bar{x} \qquad (6)$$

wobei $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ und $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ gilt.

- Durch die Überprüfung der entsprechenden Bedingungen an den zweiten partiellen Ableitungen, kann nachgewiesen werden, dass die Lösungen der partiellen Ableitungen tatsächlich an der Stelle $(\hat{\beta}_0, \hat{\beta}_1)$ ein Minimum besitzen.

- Die Normalgleichungen führen uns zu Schätzungen für die unbekannten Parameter durch Lösen des dadurch gegeben linearen Gleichungssystems in den Unbekannten β_0 und β_1.

Durch Einsetzen der umgeformten ersten Normalgleichung (5): $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
in die zweite Normalgleichung (6) erhalten wir:

$$\sum_{i=1}^{n} x_i y_i = \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 + (\bar{y} - \bar{x}\hat{\beta}_1)\, n\bar{x}$$

$$\sum_{i=1}^{n} x_i y_i = \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 + n\bar{x}\bar{y} - n\bar{x}^2\hat{\beta}_1$$

$$\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = \hat{\beta}_1 \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$

Daraus folgen die Lösungen $\hat{\beta}_0$ und $\hat{\beta}_1$ der Normalgleichungen:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (7)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \qquad (8)$$

Durch Anwendung der Definition für die empirische Varianz S_x^2 und der empirischen Kovarianz S_{xy} erhalten wir:

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = S_{xx} \text{ und } \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) = S_{xy} \quad \Longrightarrow \quad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \qquad (9)$$

Für das **Beispiel mit dem Kraftstoffverbrauch** berechnet man:

$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = 36\,838{,}2$ und $S_{xy} = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) = 1\,141{,}04$

und dadurch ist: $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{1141{,}04}{36838{,}2} = 0{,}031$ und $\hat{\beta}_0 = 6{,}19 - (0{,}031)\ 99{,}7 = 3{,}099$

Somit können wir die Gleichung der geschätzten Regressionsgeraden von y bezüglich x mit den empirischen Regressionskoeffizienten $\hat{\beta}_0$ und $\hat{\beta}_1$ festlegen:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \qquad (10)$$

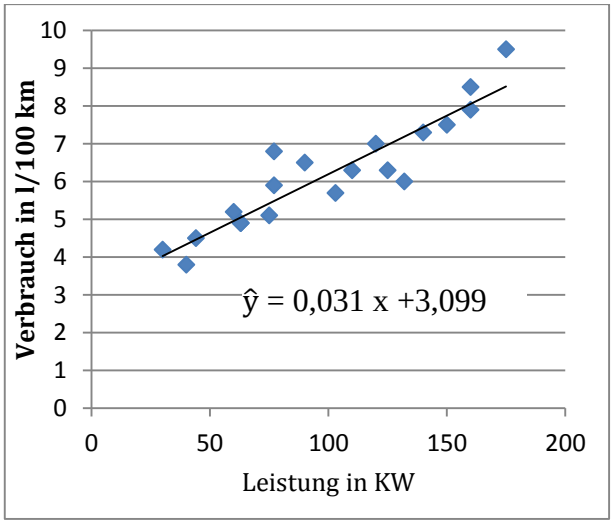

Abbildung 3: geschätzte Regressionsgerade im Streudiagramm

Bemerkungen:

- Der empirische Regressionskoeffizient $\hat{\beta}_1$ ist der Anstieg der Regressionsgeraden und $\hat{\beta}_0$ gibt den Schnittpunkt mit der y-Achse an.

- Der Punkt $(\bar{x}/\bar{y})$ liegt auf der Regressionsgeraden, ersichtlich aus der ersten Normalgleichung (5):

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Wir können nun $\hat{\beta}_0$ mit dem berechneten Kleinste Quadrate Schätzer in (10) einsetzen:

$$\hat{y} = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x \quad \Longrightarrow \quad \hat{y} = \bar{y} + \hat{\beta}_1 (x - \bar{x}) \qquad (11)$$

Daraus lassen sich nun weitere Eigenschaften des Kleinsten Quadrate Schätzers ableiten, die sich unmittelbar aus den Normalgleichungen ergeben:

1. $\sum_{i=1}^{n} y_i - \hat{y}_i = \sum_{i=1}^{n} e_i = 0$ wegen (3)
2. $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i \Longrightarrow \bar{y} = \bar{\hat{y}}$
3. $\sum_{i=1}^{n} x_i e_i = 0$ wegen (4)
4. $\sum_{i=1}^{n} \hat{y}_i e_i = 0$ wegen $\sum_{i=1}^{n} \hat{y}_i e_i = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i) e_i = \hat{\beta}_0 \underbrace{\sum_{i=1}^{n} e_i}_{0} + \hat{\beta}_1 \underbrace{\sum_{i=1}^{n} x_i e_i}_{0} = 0$
5. $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i) e_i = \sum_{i=1}^{n} y_i e_i - \underbrace{\sum_{i=1}^{n} \hat{y}_i e_i}_{0} = \sum_{i=1}^{n} y_i e_i$

Beobachtete Werte, geschätzte Werte und Residuen für das Kraftstoffbeispiel

y_i	$\hat{y}_i$	e_i
4,5	4,463	0,037
4,9	5,052	-0,152
6,8	5,486	1,314
5,9	5,486	0,414
4,9	5,052	-0,152
6	7,191	-1,191
4,2	4,029	0,171
5,7	6,292	-0,592
6,5	5,889	0,611
7	6,819	0,181
7,5	7,749	-0,249
9,5	8,524	0,976
5,1	5,424	-0,324
6,3	6,509	-0,209
6,3	6,974	-0,674
5,2	4,959	0,241
3,8	4,339	-0,539
8,5	8,059	0,441
7,9	8,059	-0,159
7,3	7,439	-0,139
Σ yi = 123,8	Σ $\hat{y}_i$ = 123,8	Σ e_i = 0

Tabelle 2: beobachtete und geschätzte Werte für das Kraftstoffbeispiel

Bei der von uns gewählten Vorgangsweise wurden die Koeffizienten $\hat{\beta}_0$ und $\hat{\beta}_1$ der Regressionsgeraden durch eine zweidimensionale Messreihe (x_i,y_i) mit i = 1, ..., n bestimmt. Die x_i lagen dabei innerhalb des Intervalls $[x_{(1)}; x_{(n)}]$. Vorsicht sollte man allerdings walten lassen, sofern Werte von x_i in das Modell eingesetzt werden, die außerhalb (bzw. innerhalb) der sinnvollen Grenzen von x liegen. Sogenannte Extrapolationen (bzw. Interpolationen) sind zwar rechentechnisch einwandfrei umsetzbar, die Regressionsgerade erklärt bzw. schätzt aber nur x-Werte innerhalb des Intervalls und ist somit im Allgemeinen für entsprechende x-Werte außerhalb des Intervalls nicht zulässig. (siehe Kapitel 2.5.3)

1.3 Beurteilung der Anpassungsgüte des Modells

Als **vierten Schritt** kann man die Beurteilung der erzielten *Anpassungsgüte* & *Korrelation* auffassen, die sich durch das Bestimmtheitsmaß beurteilen lässt. Nach Berechnung der Regressionsfunktion ist es somit von Interesse, in welchem Ausmaß diese Funktion nun tatsächlich die zugrundeliegende Stichprobe widerspiegelt. Überprüft werden kann das durch Einzeichnen der Regressionsfunktion gemeinsam mit den Datenpunkten (x_i, y_i) in die Merkmalsebene. Mögliche Ursachen für Vorhersagefehler $e_i = y_i - \hat{y}_i$, i = 1,…,n können

- **a)** zufällige Abweichungen der Messwertepaare von der Regressionsgeraden und/oder
- **b)** der Mangel an Anpassung, d.h die unzulängliche Annahme eines linearen Zusammenhanges, sein.

Bei guten Anpassungen streuen die Datenpunkte in y-Richtung regellos um die Regressionsgerade, daraus folgt, dass keine systematische Tendenz der Abweichung in Abhängigkeit vom Regressor erkennbar ist. Es darf sozusagen nur Punkt a) als Verursacher der Vorhersagefehler auftreten, ist dies nicht der Fall muss ein nichtlinearer Ansatz für die Regressionsfunktion herangezogen werden.

1.3.1 Zerlegung in den von der Regressionsgerade erklärten/ unerklärten Anteil

Neben der Beurteilung der Eignung des Ansatzes lässt die in das Streudiagramm eingezeichnete Regressionsgerade auch Schlüsse über den Erklärungswert der unabhängigen Variablen für die abhängige Variable zu. Dieser ist umso größer, je geringer die Streuung der empirischen y_i - Werte um die berechneten $\hat{y}_i$ - Werte der Regressionsgeraden ist.

Jede der n Abweichungen $y_i - \bar{y}_i$ wird zerlegt in eine *unerklärte Abweichung* $y_i - \hat{y}_i$, die durch Zufallsschwankungen, den Mangel an Anpassung oder den Einfluss anderer Merkmale verursacht wird und in die durch die Regressionsgerade *erklärte Abweichung* $\hat{y}_i - \bar{y}_i$. Es ergibt sich also:

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

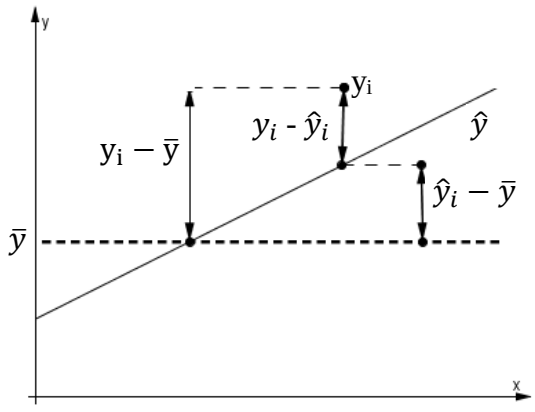

Abbildung 4: graphische Veranschaulichung der Zerlegung der Abweichung der beobachteten Werte von ihrem Mittelwert

Das Quadrat über beide Seiten und die Summation über alle n Beobachtungen, ergibt die folgende Zerlegung:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i)$$
$$= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (12)$$

$$\text{da} \sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = \sum_{i=1}^{n}\hat{y}_i\,(y_i - \hat{y}_i) - \bar{y}\sum_{i=1}^{n}(y_i - \hat{y}_i) = \underbrace{\sum_{i=1}^{n}\hat{y}_i\, e_i}_{=0 \text{ wegen (4)}} - \bar{y}\underbrace{\sum_{i=1}^{n} e_i}_{=0 \text{ wegen (2)}} = 0$$

$\sigma_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ beschreibt die gesamte Quadratsumme mit (n – 1) Freiheitsgraden, $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ den durch die Regressionsgerade erklärten Anteil, sowie $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ die unerklärte Fehlerquadratsumme mit (n - 2) - Freiheitsgraden.

$\Longrightarrow \sigma_{yy} = \sigma_R + \sigma_E$ (R...Regression, E...Residuum)

1.3.2 Bestimmtheitsmaß

Ausgehend von dieser Zerlegung, wird nun ein Maß für die Anpassungsgüte des Modells hergeleitet. Der Vergleich von $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})$ mit $\sum_{i=1}^{n}(y_i - \bar{y})$ informiert darüber, wie gut sich die geschätzte Regressionsgerade den Beobachtungspaaren im Streudiagramm anpasst, wobei die Anpassung umso besser ist, je höher der Determinationskoeffizient $\frac{S_{xy}}{S_{yy}}$ ausfällt. Das *Bestimmtheitsmaß* wird angegeben durch:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \left(= \frac{35{,}4}{41{,}7} = 0{,}85 \text{ im Beispiel}\right) \qquad (13)$$

und entspricht dem Verhältnis der erklärten Summe der Abweichungsquadrate zur gesamten Summe der Abweichungsquadrate der y-Werte von ihrem Mittelwert $\bar{y}$.

Daraus ist ersichtlich dass R^2 als Maximalwert 1 annimmt, wenn $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 0$ ist und dadurch alle Datenpunkte auf einer Geraden liegen. Umgekehrt nimmt $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2$ den Minimalwert 0 an, sofern kein linearer Zusammenhang zwischen den Merkmalen X und Y liegt.
Bei einfachen linearen Regressionen ist das Bestimmtheitsmaß das Quadrat des Korrelationskoeffizienten zwischen X und Y.

1.4 Typen von Skalierungen

Nachdem in den folgenden Kapiteln eine Reihe von Methoden der deskriptiven bzw. teilweise auch der analytischen Statistik vorkommen, sowie auch deren Anwendungsvoraussetzungen und Rahmenbedingungen beschrieben werden, ist

ausreichendes Wissen über die Art, der Skalierung und die Verteilung der Daten oder die Größe der Stichprobe erforderlich, um die dafür geeigneten statistischen Beschreibungs- und Analysemethoden wählen zu können. Statistisch betrachtet, unterscheidet man deshalb in Daten

- über die Art der Umsetzung in numerische Werte zur sinnvollen Ordnung (metrische und nichtmetrische Variable)
- über die Abstufungen ihrer Ausprägungen (stetige und diskrete Variable)

Für die lineare Regression ist zumindest Intervallskalenniveau notwendig, denn andernfalls ergibt die Datenpunktwolke keinen eindeutigen linearen Zusammenhang.

Skalierungstypen, Aussagen und Methoden

Skalenniveaus	mögliche Aussagen	mögliche Methoden	Beispiele	
Nominal (keine Ordnung der Daten möglich)	**1. Gleichheit & Ungleichheit** können festgestellt werden	(relative) Häufigkeiten, Modalwert	z.B Geschlecht, politische Orientierung, Lieblingszeitungen	NICHT METRISCH
Ordinal (größenmäßige Ordnung möglich, aber Abstände ohne Aussagekraft	1. Gleichheit & Ungleichheit **2. Rangreihung (<, >, =)**	dazu z.B kumulierte Häufigkeiten, Median	z.B Sportliche Wettkämpfe, Beliebtheitsrangliste	
Intervall (Abstände können interpretiert werden, nicht aber das Verhältnis von Größen)	1. Gleichheit & Ungleichheit 2. Rangreihung **3. Gleichheit der Unterschiede**	dazu u.a. arithmet. Mittel	z.B Temperatur, Intelligenzquotient	METRISCH
Verhältnis (die Ausprägungen haben einen absoluten Nullpunkt; das Verhältnis kann interpretiert werden)	1. Gleichheit & Ungleichheit 2. Rangreihung 3. Gleichheit der Unterschiede **4. Proportionalität** z.B $y = 2x$	dazu u.a geomet. Mittel	Alter,Preis,Größe, Inflation...	

Tabelle 3: Unterschiedliche Skalierungsformen; mögliche Aussagen und Analysemethoden

Daraus ist ersichtlich, dass grundsätzlich zwischen metrischen und nichtmetrischen Variablen unterschieden wird, wobei sich die Ausprägungen der metrischen (bzw. quantitativen) Merkmale mittels Zahlen darstellen lassen und auch Rangunterschiede und Abstände sinnvoll interpretiert werden können (z.B Körpergrößen). Als nichtmetrische Variablen werden dementsprechend alle anderen bezeichnet, deren Reihung zum Beispiel beliebig sein kann oder die sich in Form einer Rangskalierung sinnvoll reihen lassen. Ein Beispiel wäre, dass sich der Beliebteste, der Zweitbeliebteste, der Drittbeliebteste usw. sich zwar sinnvoll reihen lassen, ihre Abstände aber nicht interpretierbar sind. Demnach kann keine Aussage darüber getroffen werden, dass der Drittbeliebteste gegenüber dem Viertbeliebtesten den gleichen Abstand hat wie der Beliebteste gegenüber dem Zweitbeliebtesten. Daher sind sowohl Nominaldaten (z.B politisches Wahlergebnis) wie auch Ordinaldaten nicht metrisch.

Sogenannte *Metrische Daten* können auch wieder unterteilt werden in

- stetige oder kontinuierliche, wenn sie jeden beliebigen Wert eines bestimmten Intervalls annehmen können und

- diskrete, wenn sie nur endlich viele Werte annehmen können

1.5 Grundbegriffe der Korrelation

Bevor wir uns allerdings in das Themengebiet der linearen Regression weiter vertiefen, soll noch ein Überblick über die Annahmen erfolgen, die erfüllt sein müssen, damit die lineare Regression eine Aussagekraft hat. Prinzipiell versteht man unter einer Korrelation eine Kennzahl für den Zusammenhang zwischen Variablen. Die folgenden Zusammenhänge können bei linearer Korrelation bestehen:

- *Übereinstimmung:* je höher der Wert der Variablen A, desto höher ist oft auch der Wert der Variablen B: *positive Korrelation*
- *Gegensatz:* je höher Variable A, desto niedriger ist meist die Variable B: *negative Korrelation*
- *Unabhängigkeit:* Hohe Werte von A können relativ beliebigen Werten von B entsprechen und umgekehrt: *keine Korrelation*

Falsch wäre es zum Beispiel zu sagen, dass zwischen der Augenfarbe und der Haarfarbe eine Korrelation besteht („nominalskaliert“). Die beiden Variablen können zwar in einer Beziehung zueinander stehen, doch es handelt sich um keine quantitative Variable, weshalb diese Beziehung nicht als Korrelation bezeichnet werden kann.

Ausgehend vom Unterkapitel über die Anpassungsgüte eines Modells wird nun der Korrelationskoeffizient hergeleitet. Der Korrelationskoeffizient ist eine Zahl zwischen -1 und +1, wobei +1 eine perfekte positive lineare Beziehung angibt bzw. wenn eine Variable erhöht wird, erhöht sich die andere in perfekter Synchronisation. Ein Korrelationskoeffizient von -1 beschreibt hingegen in umgekehrter Weise eine perfekte negative lineare Beziehung. Ist der Wert der Kennzahl 0, so liegt überhaupt keine lineare Beziehung zwischen den Variablen vor. Häufig sind die Korrelationen der realen Welt nicht genau +1 , -1 oder 0 sondern liegen irgendwo dazwischen. Grundsätzlich gilt aber, je näher eine Beziehung an +1 oder -1 liegt, desto stärker ist sie. Je näher sie an 0 liegt, desto schwächer ist der Zusammenhang.

In diesem Unterkapitel liegt der Schwerpunkt unserer Betrachtungen auf der Abhängigkeit zweier Zufallsvariablen X und Y. Um ein plausibles „Abhängigkeitsmaß“ zwischen X und Y zu erhalten werden zunächst einige Begriffe definiert:

Definition 1.4.1:

a) Seien X und Y zwei Zufallsvariablen mit $E(X) = \mu_1$ und $Var(X) = \sigma_1^2$ sowie $E(Y) = \mu_2$ und $Var(Y) = \sigma_2^2$.

Falls $\sigma_{XY} = E[(X - \mu_1) (Y- \mu_2)]$ existiert, heißt
$\sigma_{XY} = Kov (X,Y)$ die Kovarianz von X und Y.

b) Falls für zwei Zufallsvariablen X und Y σ_{XY} den Wert 0 annimmt, nennt man die beiden Zufallsvariablen *unkorreliert*, gilt $\sigma_{XY} \neq 0$ bezeichnen wir die Zufallsvariablen X und Y als *korreliert*.

Satz 1.4.1: Zwei Zufallsvariable X und Y heißen stochastisch unabhängig wenn $E(X = x, Y = y) = E (X=x) \cdot E(Y=y)$. für alle möglichen Merkmalsausprägungen x und y. Unabhängige Zufallsvariable sind immer unkorreliert (**Umkehrung gilt nicht)**:

$$X,Y \text{ unabhängig } \Longrightarrow \text{ Kovarianz } (X,Y) = \text{Korrelation } (X,Y) = 0$$

Definition 1.4.2: Seien X und Y zwei Zufallsvariable mit $E(X) = \bar{x}$, $E(Y) = \bar{y}$ bzw. $Var(X) = \sigma_1^2 \neq 0$; $Var(Y) = \sigma_2^2 \neq 0$ dann ist

$$\rho_{(X,Y)} = \frac{E[(X - \bar{x})(Y - \bar{y}]}{\sigma_1 \sigma_2} = \frac{KOV(X,Y)}{\sigma_1 \sigma_2} \qquad (14)$$

der *Korrelationskoeffizient* von X und Y.

Sofern eine Stichprobe (x_1,y_1), ... (x_n,y_n) vorliegt, sind für $x = (x_1, ...,x_n)$ und $y = (y_1,...,y_n)$ die empirischen Varianzen nach (9) gegeben durch S_{xx} und S_{yy}. Die empirische Kovarianz der zweidimensionalen Stichprobe (x,y) ist S_{xy}. Daher wird der Schätzer für ρ definiert durch:

$$r_{(x,y)} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \qquad (15)$$

Kapitel 2

2. DAS LINEARE REGRESSIONSMODELL

Das Einführungskapitel sollte dazu dienen, beobachtete Daten heranzuziehen und mit Hilfe der Methode der kleinsten Quadrate einen Überblick über lineare Regressionsgleichungen zu erhalten, indem gewöhnlich eine Stichprobe an Daten statistisch bearbeitet wird. Zu den primären Aufgaben der Statistik zählen Auswertungen von Stichprobenerhebungen, um auf die Grundgesamtheit schließen zu können. Insofern findet diese Methode auch Anwendung in der Realität, weil es leichter und kostengünstiger ist, Stichprobenerhebungen von einem gewissen Umfang n durchzuführen, als die Grundgesamtheit selbst zu untersuchen.

2.1 Methodische Grundlagen

2.1.1 Die Grundannahmen des deskriptiven Modells

Nachdem es manchmal sogar schier unmöglich erscheint, die Gesamtsituation durch Beobachtung bzw. auch andere Erhebungsmethoden zu erheben, ist man an einem Modell in Form einer Regressionsanalyse interessiert, welches die Beobachtung als Stichprobe einer größeren bzw. der Gesamt- Population auffasst und die Gesamtsituation simuliert. Daraus kann beurteilt werden, ob eine „Übertragung" der Ergebnisse aus der Stichprobe auf die Gesamtsituation zulässig ist. Essentiell ist dabei, kein exaktes Abbild der Realität zu erstellen, sondern sich auf das wesentliche Mindestmaß an Grundvariablen zu beschränken, ohne das ursprüngliche Ziel, die tatsächliche Situation zu repräsentieren, aus dem Blickfeld zu verlieren.

Nun werden X, Y und ε als Zufallsvariable aufgefasst und die n X-Werte als fest vorgegebene und fehlerfrei gemessene Größen charakterisiert.
Das lineare Regressionsmodell gibt die Abhängigkeit zwischen den Variablen X und Y durch folgenden Ansatz an:

$$Y = \beta_0 + \beta_1 x + \varepsilon \qquad i = 1,\ldots,n \tag{16}$$

Hier werden die Größen β_0, β_1 nicht mehr als variierbare Koeffizienten für die beste Anpassung einer Regressionsgeraden an die Beobachtungswerte interpretiert, sondern bilden strukturelle Parameter des Modells, welche die Stärke und die Richtung des Einflusses von X auf Y ausdrücken.

Demnach werden β_0, β_1 als sogenannte *theoretische* Regressionskoeffizienten bezeichnet und die Zufallsvariable ε heißt Störkomponente.

Bei n Beobachtungen der Variablen X und Y, sollen die n Werte x_i der unabhängigen Variablen X feste Größen sein, d.h. die x_i sollen nicht durch zufällige Messfehler beeinflusst werden und exakt bleiben.

Durch fortlaufende Wiederholung der Messung an der Stelle x_i können die Werte $e_{i1},\ldots,e_{ij}$ beobachtet werden, die man als Realisationen der Zufallsvariablen ε_i ansieht und als Störvariablen bezeichnet. Dementsprechend setzt sich jede Messung an der Stelle x_i aus dem beobachteten Wert x_i und dem aus der Gleichung erhaltenen Wert y_i zusammen.
Folglich unterscheidet man auch in eine systematische Komponente ($\beta_0 + \beta_1 x_i$) und eine stochastische Komponente (ε_i).
Nachdem lineare Regressionsfunktionen sich auf die notwendigsten Variablen x_i beschränken um y_i zu beschreiben, erfassen die ε_i die Wirkung aller übrigen Variablen, welche die abhängige Variable beeinflussen, aber nicht explizit in die Regressionsfunktion aufgenommen wurden.

Folgende Annahmen werden nun für die Modellvoraussetzungen getroffen:

a. $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, \dots, n$ (Linearität)

b. Bei der nicht beobachtbaren Fehlervariablen ε wird davon ausgegangen, dass sie den Erwartungswert 0 und die Varianz σ^2 hat. (Homoskedastizität)

c. Die ε_i alle die selbe Varianz σ^2 haben (Varianzhomogenität der Fehler)

d. Die Störvariablen unkorreliert sind und somit gilt:

$$E(\varepsilon_i \cdot \varepsilon_j) = Kov(\varepsilon_i \cdot \varepsilon_j) = 0 \quad \text{für } i \neq j \qquad i, j = 1, \dots, n$$

e. ε_i ist für alle i normalverteilt

f. die x_i sind linear unabhängig

Um die Funktionstüchtigkeit eines regressionsanalytischen Algorithmus zu gewährleisten wird zusätzlich noch vorausgesetzt:

- dass die n Werte x_i nicht alle paarweise gleich sind
- und n größer als zwei ist.

Daraus ergibt sich im Weiteren der Erwartungswert, die Varianz und die Kovarianz von Y_i im Punkt x_i:

$$E(Y_i) = E(\beta_0 + \beta_1 x_i + \varepsilon_i) \rightarrow E(Y_i) = \beta_0 + \beta_1 \cdot E(x_i) + E(\varepsilon_i)$$

$$Var(Y_i) = Var(\beta_0 + \beta_1 x_i + \varepsilon_i) = Var(\varepsilon_i) = E(\varepsilon_i^2) - E(\varepsilon_i)^2 = \sigma^2 \quad \text{für alle } i = 1, \dots, n$$

$$Kov(Y_i) = E[(Y_i - \beta_0 - \beta_1 x_i)(Y_j - \beta_0 - \beta_1 x_j)] = E(\varepsilon_i \varepsilon_j) = \sigma_{Yi} = \begin{cases} 0 \text{ für } i \neq j \\ \sigma^2 \text{ für } i = j \end{cases}$$

Annahme e. fordert die Normalverteilung der Störvariablen ε_i mit Erwartungswert 0 und Varianz σ^2 als Voraussetzung für die später behandelten statistischen Verfahren. Zudem lassen sich Messfehlerverteilungen häufig durch Normalverteilungen approximieren und somit folgt aus der Gleichung $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, dass auch die Y_i normalverteilt mit Erwartungswert $\mu_i = \beta_0 + \beta_1 x_i$ und Varianz σ^2 sein müssen.

2.1.2 Erweiterungen für das stochastische Modell

Durch das Setzen bestimmter Annahmen gelangt man vom Grundmodell der Regressionsanalyse auf das stochastische Regressionsmodell. Das Erklärungsziel verändert sich dadurch nicht, jedoch lassen sich einige zusätzliche Ergebnisse ableiten. So geht in etwa das lineare stochastische Regressionsmodell von der Annahme der Existenz eines „datengenerierenden Prozesses“ aus, der sich aus einer „deterministischen“ linearen Beziehung zwischen der abhängigen Variable Y und der unabhängigen Variable x_i zusammen setzt, die durch eine stochastische Komponente ε

überlagert wird. Sofern angenommen wird, dass die Werte der unabhängigen Variablen gegeben (also keine Zufallsvariablen) sind und man die Scheinvariable X_0 berücksichtigt, so ergibt sich Y_i als Linearkombination der X_i, ergänzt um den stochastischen Term ε_i.

Y_i selbst ist somit eine Zufallsvariable. Da nur bestimmte Realisationen des postulierten Prozesses beobachtbar sind, ist es das Ziel der stochastischen Regressionsanalyse, Schätzwerte $\hat{\beta}_0$, $\hat{\beta}_1$ für die Koeffizienten β_0, β_1 und die Störvariable ε zu ermitteln. Die Schätzer sind ebenfalls Zufallsvariable.
Die Schätzungen machen Annahmen bezüglich der **stochastischen Eigenschaften der Störvariablen** erforderlich. Es wird also wie schon erwähnt, angenommen, dass sich die stochastischen Störeinflüsse im Mittel ausgleichen, dass der Erwartungswert von ε bei gegebenem x_i, also Null und ε_i damit, hinsichtlich seines Erwartungswertes, auch unabhängig von den x_i ist.

Verteilung der Epsilons bei linearer Einfachregression

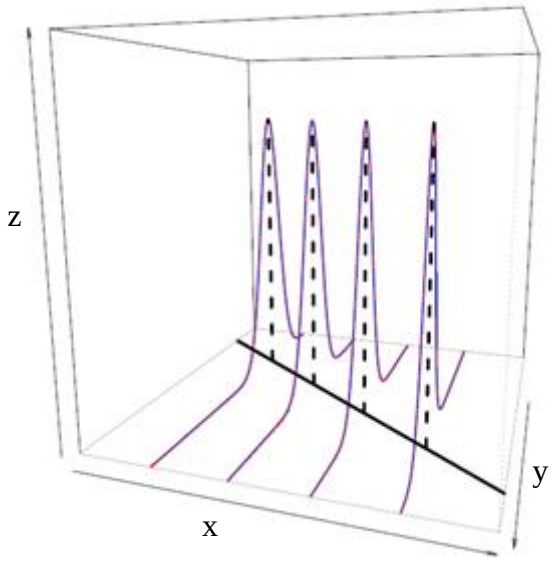

Abbildung 5: Verteilung der Epsilons bei linearer Einfachregression

Einige der Annahmen lassen sich, falls nur eine unabhängige Variable vorliegt, der Abbildung entnehmen. Auf der Geraden der XY –Ebene liegen die Erwartungswerte $E(Y_i|X_i) = \hat{Y}_i = \beta_0 + \beta_1 x_i$. Auf diesen Erwartungswerten sind die bedingten Wahrscheinlichkeitsdichten der Störgrößen ε_i errichtet.

Die Ermittlung der Koeffizienten des stochastischen Regressionsmodells kann in gleicher Weise erfolgen wie bei der deskriptiven Regressionsrechnung, also zum Beispiel mit Hilfe der Methode der kleinsten Quadrate. Auch das Bestimmtheitsmaß kann analog berechnet werden. Wenn die getroffenen Annahmen gelten, so können auch weitere Maßzahlen ermittelt werden, mit denen die Güte des Regressionsmodells beurteilt werden kann.

Bezogen auf die stochastischen Maßzahlen sind zunächst die *Standardabweichungen* der errechneten Regressionskoeffizienten $\hat{\beta}_0$, $\hat{\beta}_1$ erwähnenswert, denn diese drücken die zufallsbedingte Streuung der $\hat{\beta}_j$ um ihre Erwartungswerte β_j aus.
Liegt nun Homoskedastizität vor, so lassen sich die Standardabweichungen der $\hat{\beta}_j$ schätzen als:

$$s_{\hat{\beta}_0} = s_\varepsilon \cdot \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{n\sigma_x^2}\right)} \quad \text{und} \quad s_{\hat{\beta}_1} = \frac{s_\varepsilon}{\sqrt{n} \cdot \sigma_x}, \tag{17}$$

Dabei ist n die Anzahl der Beobachtungen, $\bar{x}$ der Mittelwert der unabhängigen Variablen dieser n Beobachtungen, s_ε die geschätzte Standardabweichung um die Regressionslinie (Schätzer für σ) und σ_x die Standardabweichung der unabhängigen Variablen in den n Beobachtungen. Als Schätzwert für die Standardabweichung der Störgröße σ kann zwar

$$S_Y = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} \varepsilon_i^2} \tag{18}$$

verwendet werden, allerdings ist dieser nicht erwartungstreu. Die nachstehende Maßzahl, die jene durch die Koeffizientenschätzung verlorengegangene Zahl an Freiheitsgraden v berücksichtigt, ist hingegen erwartungstreu:

$$s_\varepsilon = \sqrt{\frac{1}{n - v - 1} \cdot \sum_{i=1}^{n} \varepsilon_i^2}. \tag{19}$$

Die so erhaltenen Standardabweichungen können darin Verwendung finden, die errechneten Koeffizientenwerte einem Signifikanztest auf den Wert Null zu unterziehen. Außerdem wird der so „beobachtete“ Wert des t-Tests bei der Ergebnispräsentation häufig zusammen mit den Koeffizientenwerten angegeben. Für die Variable x_i berechnet er sich als

$$t_i = \frac{\beta_i - 0}{s_{\beta_i}} = \frac{\beta_i}{s_{\beta_i}}$$

Sofern sich nun die Hypothese eines wahren Koeffizientenwertes von Null zu einem gegebenen Signifikanzniveau ablehnen lässt, wird dies als Indiz dafür gesehen, dass die dem entsprechenden Koeffizienten zugeordnete Variable einen signifikanten Einfluss auf die abhängige Variable ausübt. Als problematisch gilt allerdings, dass die zugrundeliegende Hypothese eine Punkthypothese darstellt und somit bei genügend großen Fallzahlen immer verwerfbar ist.

2.1.3 Durbin-Watson-Test

Die Validität der Ergebnisse eines linearen Regressionsmodells ist primär von der Einhaltung der Modellvoraussetzungen abhängig. Nachfolgend werden Einblicke in verschiedene Tests gegeben.
Der Durbin-Watson-Test ist ein Test auf Autokorrelationsfreiheit der Störvariablen, welche die Korrelation einer Funktion mit sich selbst zu einem früheren Zeitpunkt

beschreibt und Aussagen darüber ermöglicht, ob die benachbarten Ausprägungen der Störvariablen einer linearen autoregressiven Beziehung der folgenden Art unterliegen:

$$e_t = \rho \cdot e_{t-1} + \varepsilon_t \qquad (20)$$

mit: $|\rho| < 1$

e Fehler im Modell zur Zeitperiode t

ε_t unabhängige N(0,σ_a)-verteilte Zufallsvariable

$|\rho| < 1$ wird als Autokorrelationskoeffizient bezeichnet und gibt die Korrelation benachbarter Werte der Störvariablen an. ε_t ist wiederum die stochastische Störkomponente, die normalverteilt mit Erwartungswert 0 und fester Varianz ist. Anschließend tritt die Frage auf, welche Werte die Gültigkeit besitzen sie als benachbart zu betrachten. Diess ist allein bei Zeitreihendaten bzw. bei aufeinanderfolgenden Periodenwerten naheliegend. Querschnittsdaten zum Beispiel erfordern zunächst die Bestimmung eines adäquaten Ordnungskriteriums. Wenn den Berechnungen Querschnittsdaten zugrunde liegen, welche nicht nach einem geeigneten Kriterium sortiert sind, so ist der hier beschriebene Test sinnlos.

Daraus folgen einige interessante Eigenschaften der Fehler ε_t:

$e_t = \sum_{i=1}^{\infty} \rho^i \, a_{t-i}$ $\quad$ $\mathrm{Cov}(e_t, e_{t+i}) = \rho^{|i|}\sigma_\varepsilon^2(\frac{1}{1-\rho^2})$

$\mathrm{E}(e_t) = 0$ und $\quad$ $\mathrm{Var}(e_t) = \sigma_a^2(\frac{1}{1-\rho^2})$

D.h die Fehler haben Erwartungswert 0 und konstante Varianz, sind aber autokorreliert, außer für $\rho = 0$.

Es wird somit $H_0: \rho = 0$ gegen $H_1(a): \rho \neq 0$ bzw. $H_1(b): \rho > 0$ *bzw.* $H_1(c): \rho < 0$ getestet. Als Testgröße („Durbin-Watson-Statistik“) wird der folgende Ausdruck heran gezogen:

$$\mathrm{DW} = \frac{\sum_{i=2}^{n}(e_t - e_{t-1})^2}{\sum_{i=1}^{n} e_t^2}. \qquad (21)$$

Die Variable e_t charakterisiert den mit der ermittelten Regressionsgleichung errechneten Wert der Störvariable für die Beobachtung t und n ist die Gesamtzahl der Beobachtungen.

Zwischen der Testgröße DW und dem Autokorrelationskoeffizienten gilt näherungsweise die Beziehung:

$$\mathrm{DW} = 2 \cdot (1 - \rho)$$

Beweis: $\mathbf{DW} = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} = \frac{\sum_{t=2}^{n} e_t^2 - 2\sum_{t=2}^{n} e_t e_{t-1} + \sum_{t=2}^{n} e_{t-1}^2}{\sum_{t=1}^{n} e_t^2} =$

$$= \frac{2\sum_{t=2}^{n-1} e_t^2 + e_1^2 + e_n^2}{\sum_{t=1}^{n} e_t^2} - 2\frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2} \approx 2 - 2\hat{\rho} = 2\,(1 - \hat{\rho})$$

Der Wert der Testgröße informiert nun über das Ausmaß der Autokorrelation erster Ordnung. Ist ρ=0 (bei kompletter Abwesenheit der Autokorrelation) so ist DW =2, der Wert ρ= +1 wird hingegen bei vollständig positiver Autokorrelation angenommen, daraus ergibt sich DW = 0 und vollständig negative Autokorrelation ρ = -1 führt zu DW = 4.
In Abhängigkeit von der vorzugebenden Irrtumswahrscheinlichkeit, der Anzahl der Beobachtungen und der Anzahl der unabhängigen Variablen treten für die Testgröße DW jedoch zwei Unbestimmtheitsbereiche auf. Nimmt das berechnete DW einen Wert in diesen Intervallen an, so kann die Hypothese abwesender Autokorrelation weder bestätigt noch verworfen werden. Die Unbestimmtheitsbereiche ergeben sich über die nachstehenden tabellarischen Werte d_u und d_o. Die folgenden Fälle lassen sich nun unterscheiden:

Fall	DW liegt im Intervall	Aussage (zu gegebener Irrtumswahrscheinlichkeit
1	$[0,d_u[$	Positive Autokorrelation
2	$[d_u,d_o]$	Keine Aussage möglich
3	$]d_o,4-d_o[$	Keine Autokorrelation
4	$[4-d_o,4-d_u]$	Keine Aussage möglich
5	$]4-d_u,4]$	Negative Autokorrelation

Tabelle 4: Durbin-Watson-Test - Interpretationshilfe

Die oberen und unteren kritischen Werte $\mathbf{d_u}$ und $\mathbf{d_o}$ liegen in tabellierter Form für verschiedene k Werte (Zahl der erklärenden Variablen) und n vor.
Sofern der Durbin-Watson-Test auf Autokorrelation der Störvariablen hindeutet, muss eine Fehlspezifikation des Regressionsmodells in Betracht gezogen werden, die auf Nichtberücksichtigung wichtiger unabhängiger Variablen oder eine falsche Funktionsform zurückzuführen ist.

Beispiel:
Ein Getränkeabfüllunternehmen möchte die jährlichen regionalen Einkünfte für ein bestimmtes Produkt durch eine Funktion der jährlichen regionalen Werbeausgaben für dieses Produkt voraussagen. Dazu werden die Verkaufsdaten der letzten 20 Jahre (siehe Tabelle) herangezogen und ausgehend von der Annahme einer geeigneten linearen Beziehung, werden die gewöhnlichen Kleinsten-Quadrate verwendet.
Nachdem die Regressorvariable t eine Zeitspanne ist, nimmt man an, dass Autokorrelation vorliegt, die bei näherer Betrachtung der gegebenen Daten tatsächlich bewiesen wird, sofern man in einem Koordinatensystem die Residuen gegen die Zeit aufträgt. Augenscheinlich ist dieser *Plot* nicht linear, sondern weißt zuerst einen Aufwärtstrend und anschließenden Abwärtstrend der Residuen auf. Für so ein Muster kann Autokorrelation verantwortlich sein
Wir werden nun auch den Durbin – Watson Test verwenden für:

$$H_0: \rho = 0$$
$$H_1: \rho > 0$$

$$d = \frac{\sum_{t=2}^{20}(e_t - e_{t-1})^2}{\sum_{t=1}^{20} e_t^2} = \frac{8195{,}21}{7587{,}92} = 1{,}08$$

Wenn nun eine Irrtumswahrscheinlichkeit α = 0,05 vorgegeben wird, so kann man aus der Tabelle für den DW-Test die kritischen Werte ablesen, die mit n = 20 und einem Regressor korrespondieren und diese sind d_u =1,2 und d_o = 1,41. Nachdem allerdings der beobachtete Wert von d = 1,08 kleiner als d_u ist, verwerfen wir H_0 und folgern, dass die Fehler positiv autokorreliert sind.

	t	jährl. regionale Einkünfte y_t	jährl. Werbeausgaben x_t (in € mal 1000)	kl. Quadrate Residuen e_t	e_t^2	$(e_t - e_{t-1})^2$	jährl.regionale Bevölkerung z_t
1990	1	3083	75	-32,330	1045,229		825 000
1991	2	3149	78	-26,603	707,720	32,799	830 445
1992	3	3218	80	2,215	4,906	830,477	838 750
1993	4	3239	82	-16,967	287,879	367,949	842 940
1994	5	3295	84	-1,148	1,318	250,241	846 315
1995	6	3374	88	-2,512	6,310	1,860	852 240
1996	7	3475	93	-1,967	3,869	0,297	860 760
1997	8	3569	97	11,669	136,166	185,940	865925
1998	9	3597	99	-0,513	0,263	148,401	871 640
1999	10	3725	104	27,032	730,729	758,727	877 745
2000	11	3794	109	-4,422	19,554	989,354	886 520
2001	12	3959	115	40,032	1602,561	1976,158	894 500
2002	13	4043	120	23,577	555,875	270,767	900 400
2003	14	4194	127	33,940	1151,924	107,392	904 005
2004	15	4318	135	-2,787	7,767	1348,873	908 525
2005	16	4493	144	-8,606	74,063	33,861	912 160
2006	17	4683	153	0,575	0,331	84,291	917 630
2007	18	4850	161	6,848	46,895	39,351	922 220
2008	19	5005	170	-18,971	359,899	666,621	925 910
2009	20	5236	182	-29,063	844,658	101,848	929 610
			$\sum_{t=1}^{20} e_t^2 = 7587{,}915$		$\sum_{t=2}^{20}(e_t - e_{t-1})^2 = 8195{,}21$		

Tabelle 5: Daten der Getränkeumsätze einer Region

Parameter	Schätzung	Standardfehler	t-Statistik
β_0	1608,508	17,022	94,49
β_1	0,143	0,143	140,71
n = 20	R^2= 0,991	$\hat{\sigma}$ = 421,549	

Tabelle 6: Statistik für das kl. Quadrate Modell des Beispiels

2.1.4 Test auf Homoskedastizität

Homoskedastizität bedeutet, dass die Varianz der Residuen und dadurch die Varianz der erklärten Variablen selbst, für alle Ausprägungen der anderen Prädiktorvariablen nicht signifikant unterschiedlich ist. Heteroskedastizität liegt in der Statistik dagegen bei unterschiedlicher Streuung innerhalb einer Datenmessung vor. Bei diesem Test auf

Homoskedastizität wird zuerst so vorgegangen, dass man das Datenmaterial in zwei sachgerechte Teilbereiche A und B aufteilt. Bei Einfachregressionen kann als Aufteilungskriterium die Höhe des Wertes der exogenen Variable herangezogen werden, wobei dann im Teilbereich A die Beobachtungen mit den niedrigeren Werten, im Teilbereich B jene mit den höheren Werten der exogenen Variable lägen. Bei Zeitreihenanalysen ist die Zeit das Zerlegungskriterium, welches eine einfache Durchführung des Tests zulässt, nachdem die Daten bereits sortiert nach dem Kriterium Zeit vorliegen. Bei multivariaten Querschnittsanalysen muss hingegen zuerst ein adäquates Kriterium festgelegt werden und die Möglichkeit bestehen, die Beobachtungen nach der Höhe der Kriteriumsvariable zu ordnen.

Im Anschluss an die Berechnung des eigentlichen Modells sind bei der Vermutung auf Heteroskedastizität (z.B nach Ansicht der Grafiken der berechneten Residuen), weitere Berechnungen durchzuführen. Aus den n_A Daten des Datenbereichs A wird eine Regressionsfunktion berechnet und die geschätzten Störgrößen e_{i_A} lassen die Ermittlung folgender Größe zu:

$$e_A = \sum_{i_A=1}^{n_A} e_{i_A}^2$$

In einer zweiten Regression berücksichtigt man alle n_B Beobachtungen des Teilbereichs B und ermittelt den Wert

$$e_B = \sum_{i_B=1}^{n_B} e_B^2$$

Wenn nun die Anzahl der exogenen Variablen mit v bezeichnet wird, folgt daraus die Prüfgröße für den F-Test aus dem Vergleich der beiden geschätzten Varianzen der Störgrößen als

$$F = \frac{s_{e,B}^2}{s_{e,A}^2} = \frac{\frac{e_B}{n_B - v - 1}}{\frac{e_A}{n_A - v - 1}} \qquad (22)$$

Aus einer Tabelle der F-Verteilung (vergleiche Anhang) ist für eine gegebene Irrtumswahrscheinlichkeit und die Freiheitsgrade $v_1 = n_B\text{-}v\text{-}1$ und $v_2 = n_A\text{-}v\text{-}1$ der kritische F-Wert zu ermitteln. Wenn nun

$$F < F_{\alpha,v_1,v_2},$$

so kann bei der gewählten Irrtumswahrscheinlichkeit α die Nullhypothese gleicher Varianzen nicht verworfen werden und es darf von Homoskedastizität ausgegangen werden.

2.1.5 Test auf Strukturkonstanz

Strukturkonstanz ist dann gegeben, sobald die unterstellte Regressionsbeziehung für alle Beobachtungen gleichermaßen zutreffend ist. Beim sogenannten Strukturbruchtest wird das Beobachtungsmaterial wiederum in zwei Teile zerlegt, wobei Homoskedastizität vorausgesetzt wird. Insofern gilt die Empfehlung, zuerst den entsprechenden Test

durchzuführen nachdem beim Strukturkonstanztest auch die gleichen Ordnungskriterien wie beim Test auf Homoskedastizität gelten. Die Nullhypothese des Strukturbruchtests behauptet, dass die Regressionskoeffizienten, die aus den beiden Teilen des Beobachtungsmaterials gewonnen werden, gleich sind. Der Test kann auch nur auf einige interessierende Regressionskoeffizienten beschränkt werden.
Im Rahmen von zwei Regressionsrechnungen sind die Werte ε_A, ε_b, n_A, n_B analog zum Vorabschnitt zu bestimmen, dabei werden ε und n der eigentlichen Regressionsrechnung entnommen, die beide Teilbereiche berücksichtigt und v ist die Zahl der exogenen Variablen. Als Prüfgröße für den F-Test folgt dann:

$$F = \frac{\frac{e - e_A - e_B}{v + 1}}{\frac{e_A - e_B}{n - 2v - 2}} \tag{23}$$

Stellt man dieser Größe wiederum den aus der F-Tabelle gewonnenen kritischen F-Wert gegenüber so kann bei vorgegebener Irrtumswahrscheinlichkeit die Nullhypothese gleicher Koeffizienten nicht abgelehnt werden und man darf von Gleichheit der Koeffizienten in beiden Beobachtungsgruppen ausgehen. Wenn die Nullhypothese verworfen wird, so unterscheidet sich mindestens ein Koeffizient beider Regressionen in signifikantem Ausmaß.

2.2 Eigenschaften der kleinste Quadrate Schätzer

Nachdem von einer theoretisch linearen Regression ausgegangen wird und die deskriptive Regression des voran gehenden Kapitels sich durch einen linearen Ansatz an die empirischen Datenpunkte anpasst, besteht die Möglichkeit, die Parameter β_i durch empirische Regressionskoeffizienten zu schätzen, welche die Lösungen der Normalgleichungen bilden.

Wie bereits gezeigt, sind $\hat{\beta}_0$ und $\hat{\beta}_1$ Linearkombinationen der Beobachtungen y_i, somit gilt: $\hat{\beta}_1 = \frac{S_{xy}}{S_x^2} = \sum_{i=1}^{n} c_i(y_i - \bar{y})$, mit $c_i = \frac{x_i - \bar{x}}{S_x^2}$ für i = 1, … , n und

$$\hat{\beta}_0 = \hat{y} - \hat{\beta}_1 \bar{x}.$$

Y wird als Zufallsvariable aufgefasst und $\hat{\beta}_0$ und $\hat{\beta}_1$als Zufallsvariable bzw. Schätzer für β_0 und β_1.

1. **Erwartungswert**

   $E(\hat{\beta}_0) = \beta_0$, $E(\hat{\beta}_1) = \beta_1$

   d.h $\hat{\beta}_0$ und $\hat{\beta}_1$ sind erwartungstreue Schätzer von β_0 und β_1. (24)

   Beweis: $E(\hat{\beta}_1) = E(\sum_{i=1}^{n} c_i Y_i) = \sum_{i=1}^{n} c_i E(Y_i) = \sum_{i=1}^{n} c_i(\beta_0 + \beta_1 x_i) =$
   $= \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i = \beta_1$ wegen: $\sum_{i=1}^{n} c_i = 0$ und $\sum_{i=1}^{n} c_i x_i = 1$

Außerdem gilt nach (1.7)für Y als Zufallsvariable: $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\bar{x}$. Daraus folgt:

$$E(\hat{\beta}_0) = E(\overline{Y} - \hat{\beta}_1\bar{x}) = E(\overline{Y}) - \underbrace{\bar{x}E(\hat{\beta}_1)}_{\beta_1} = \frac{1}{n}\sum_{i=1}^{n} E(Y_i - \bar{x}\beta_1) =$$

$$= \frac{1}{n}[\sum_{i=1}^{n}\beta_0 + \beta_1 x_i] - \beta_1\bar{x} = \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} = \beta_0$$

2. Varianz

$$\text{Var}(\hat{\beta}_0) = E(\hat{\beta}_0 - \hat{\beta}_0)^2) = \frac{\sigma^2}{n}\cdot\frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i - \bar{x})^2}$$

$$\text{Var}(\hat{\beta}_1) = E(\hat{\beta}_1 - \hat{\beta}_1)^2) = \frac{\sigma^2}{n}\cdot\frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (25)$$

Beweis: $\text{Var}(\hat{\beta}_1) = \text{Var}(\sum_{i=1}^{n} c_i Y_i) = \sum_{i=1}^{n} c_i^2\ \text{Var}(Y_i) = \sigma^2\sum_{i=1}^{n} c_i^2 =$

$$= \frac{\sigma^2\sum_{i=1}^{n}(x_i - \bar{x})^2}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}}$$

$$\text{Var}(\hat{\beta}_0) = \text{Var}(\overline{Y} - \beta_1\bar{x}) = \text{Var}(\overline{Y}) + \bar{x}^2\text{Var}(\hat{\beta}_1) - 2\bar{x}\,\text{Cov}(\overline{Y}, \hat{\beta}_1)$$

$$= \text{Var}(\overline{Y}) + \bar{x}^2\text{Var}(\hat{\beta}_1) = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})$$

Weil:

$$\text{Cov}(\overline{Y}, \hat{\beta}_1) = E[(\overline{Y} - E(\overline{Y}))\,(\hat{\beta}_1 - E(\hat{\beta}_1))] = E(\bar{\varepsilon}(\hat{\beta}_1 - \beta_1)) = E(\bar{\varepsilon}\hat{\beta}_1) =$$

$$\sum_{i=1}^{n} c_i E(\bar{\varepsilon}Y_i) = \sum_{i=1}^{n} c_i \frac{1}{n}\sum_{j=1}^{n} E(\varepsilon_j\varepsilon_i) = \frac{\sigma^2}{n}\sum_{i=1}^{n} c_i = 0$$

$$\text{Var}(\bar{Y}) = \text{Var}\left(\frac{1}{n}(Y_1 + \ldots + Y_n)\right) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

3. Varianz für (a) die Vorhersagevariable $\hat{Y}$ und (b) die Varianz der Residuen

a) $\text{Var}(\hat{Y}_i) = \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_i) = \text{Var}\left(\overline{Y} + \hat{\beta}_1(x_i - \bar{x})\right) =$

$$= \text{Var}(\overline{Y}) + (x_i - \bar{x})^2\,\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{n} + (x_i - \bar{x})^2\frac{\sigma^2}{S_{xx}} = \sigma^2\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right)$$

b) $\text{Var}(E_i) = \text{Var}(Y_i - \hat{Y}_i) = \text{Var}(Y_i) + \text{Var}(\hat{Y}_i) - 2\text{Cov}(Y_i, \hat{Y}_i)$

$$= \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right] - 2\text{Cov}(Y_i, \hat{Y}_i). \qquad (26)$$

Weil nach (11) gilt:

$$Cov(Y_i, \hat{Y}_i) = Cov\left(Y_i, \bar{Y} + \hat{\beta}_1(x_i - \bar{x})\right) = Cov(Y_i, \bar{Y}) + Cov\left(Y_i, \hat{\beta}_1(x_i - \bar{x})\right).$$

4. Konsistenz

Falls für $n \longrightarrow \infty$ der Ausdruck $\sum_{i=1}^{n}(x_i - \bar{x})^2 \rightarrow \infty$ strebt, dann gilt

$$\hat{\beta}_0 \underset{p}{\rightarrow} \beta_0 \text{ und } \hat{\beta}_1 \underset{p}{\rightarrow} \beta_1 \tag{27}$$

5. Verteilung

Falls $\varepsilon_i \sim N(0, \sigma^2)$, so erhält man:

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2}{n} \cdot \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i - \bar{x})^2}\right)$$
$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n} x_i - \bar{x})^2}\right) \tag{28}$$

Anmerkung:

Für großes n bleiben die angegebenen Verteilungen auch dann im Allgemeinen *approximativ* gültig, wenn die ε_i *nicht normalverteilt* sind (zentraler Grenzwertsatz)

- i. Allg. wichtigster Parameter: β_1 – Steigung der Geraden
  $$\hat{\beta}_1 \sim N(\beta_1, Var(\hat{\beta}_1))$$
- Die Varianz von $\hat{\beta}_1$ ist umso kleiner je
  - kleiner σ^2, die Varianz des Fehlerterms
  - größer n, die Anzahl der Beobachtung
  - größer S_X die Streuung der $x_1 \ldots x_n$

6. Kovarianz von $(Y_i, \hat{Y}_i)$

$$\text{Cov}(Y_i, \hat{Y}_i) = \text{Var}(\hat{Y}_i) = \frac{\sigma^2}{n} + \frac{(x_i - \bar{x}^2)}{S_{xx}}\sigma^2 = \sigma^2\left[\frac{1}{n} + \frac{(x_i - \bar{x}^2)}{S_{xx}}\right]. \tag{29}$$

weil gilt:

$$\text{Cov}(Y_i, \bar{Y}) = \text{Cov}(Y_i, \frac{1}{n}(Y_1 + \cdots + Y_n)) = \frac{1}{n}\sum_{j=1}^{n}\text{Cov}(Y_i, Y_j)$$
$$= \frac{1}{n}\sum_{j=1}^{n}\text{Cov}(\varepsilon_i, \varepsilon_j) = \frac{\sigma^2}{n}$$

$$\text{Cov}\left(Y_i, \hat{\beta}_1(x_i - \bar{x})\right) = \frac{x_i - \bar{x}}{S_{xx}}\text{Cov}(Y_i, S_{xY}) = \frac{x_i - \bar{x}}{S_{xx}}\text{Cov}(Y_i, \sum_j (x_j - \bar{x})Y_j) =$$
$$= \frac{x_i - \bar{x}}{S_x^2}$$

Daraus kann nun die Var(E_i) gefolgert werden:

$$\text{Var}(E_i) = \text{Var}(Y_i - \hat{Y}_i) = \sigma^2 \left[1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right)\right]$$

Nach dem „*Satz von Gauss-Markov*“ sind $\hat{\beta}_0$und $\hat{\beta}_1$ sogar die wirksamsten Schätzer von β_0 und β_1 in der Klasse aller linearen und erwartungstreuen Schätzer.
Sei β_1^* also ein linearer, erwartungstreuer Schätzer von β_1 so gilt:

$$\text{Var}(\hat{\beta}_1) \leq \text{Var}(\beta_1^*)$$

2.3 Das klassische normalverteilte Modell der linearen Einfachregression

Die Wahrscheinlichkeitsverteilung der Fehlervariablen ε waren bis jetzt auf Annahmen über den Erwartungswert und die Varianz-Kovarianzmatrix beschränkt. Dieses Unterkapitel setzt nun zusätzlich zu den bisherigen Modellannahmen die Normalverteilung der Zufallsvariablen ε voraus, wodurch verschiedene Tests und Konfidenzintervalle hergeleitet werden können.

2.3.1 Erwartungstreue Schätzer der theoretischen Regressionskoeffizienten

Es wird vorausgesetzt, dass das lineare Regressionsmodell wie bisher beschrieben in den Variablen x und y vorliegt und eine Stichprobe die Wertepaare $\big((x_1, y_1), \dots, (x_n, y_n)\big)$ liefert. So dann kann die empirische Regressionsgleichung mit normalverteilten Fehlern ermittelt werden:

Lineares Regressionsmodell mit normalverteilten Fehlern

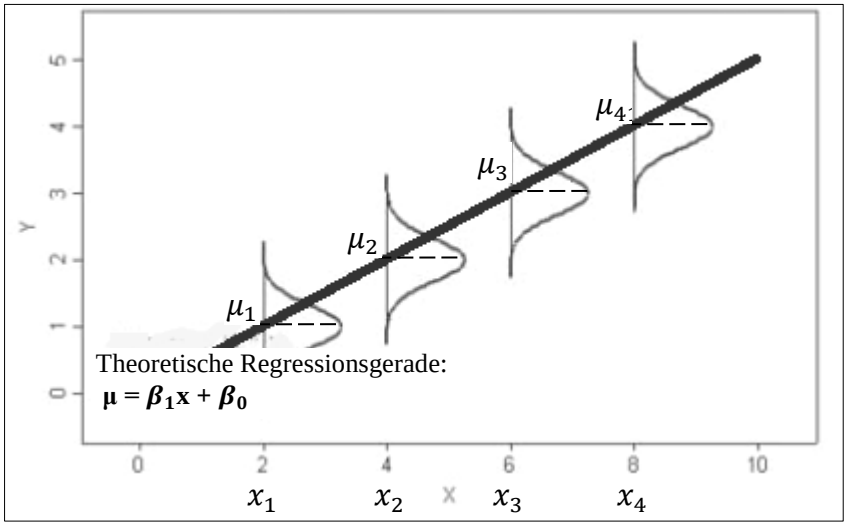

Abbildung 6: Lineares Modell der Einfachen Regression. Bedingte Verteilung der abhängigen Variable Y. Die Dichte von Y bei gegebenen $\boldsymbol{x_1}$ ist die Dichte der N ($\boldsymbol{\beta_1 x + \beta_0, \sigma^2}$) - Verteilung

Der Anstieg β_1 und der Achsenabschnitt β_0 konnte mit Hilfe der Methode der kleinsten Quadrate berechnet werden:

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \qquad \text{bzw.} \quad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

Somit können β_0 und β_1 als Realisation der beiden Zufallsvariablen angesehen werden:

$$B_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \qquad \text{und} \quad B_0 = \bar{Y} - B\bar{X}$$

Nachdem die Werte x_i fix sind, werden nur die y_i als Zufallsvariablen angenommen. B_0 und B_1 sind dadurch als Funktionen der n Zufallsvariablen y_i darstellbar und auch wieder Zufallsvariablen. Außerdem sind B_0 und B_1 Linearkombinationen der Zufallsvariablen Y_i wodurch B_0 und B_1 zu linearen Schätzfunktionen für β_0 und $\beta_1$werden.

Definition: Die Schätzfunktion $S_n = s_n (Y_1, \ldots , Y_n)$ für den Parameter υ heißt erwartungstreu, wenn sie den folgenden Erwartungswert annimmt:

$$E(S_n) = E\big(s_n(Y_1, \ldots, Y_n)\big) = \upsilon \tag{30}$$

Satz: Die Zufallsvariable B_1 ist eine linear erwartungstreue Schätzfunktion für den im klassischen Modell der linearen Einfachregression auftretenden Parameter β_0. (31)

Beweis
Sofern dem linearen Regressionsmodell die Variablen x und y zugrunde liegen, gilt die theoretische Regressionsgleichung:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

überdies gilt:

$$\bar{Y} = \beta_0 + \beta_1 \bar{x} + \bar{\varepsilon}$$

mit $\bar{Y} = \beta_0 + \beta_1 \bar{x} + \bar{\varepsilon}$, $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ und $\bar{\varepsilon} = \frac{\sum_{i=1}^{n} \varepsilon_i}{n}$

Daraus lässt sich B_1 nun folgendermaßen bestimmen:

$$B_1 = \frac{\sum_{i=1}^{n} x_i(\beta_0 + \beta_1 x_i + \varepsilon_i) - n\bar{x}(\beta_0 + \beta_1 \bar{x} + \bar{\varepsilon})}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} =$$

$$= \frac{\sum_{i=1}^{n} x_i \beta_0 + \sum_{i=1}^{n} \beta_1 x_i^2 + \sum_{i=1}^{n} x_i \varepsilon_i - n\bar{x}\beta_0 - n\beta_1 \bar{x}^2 - n\bar{x}\bar{\varepsilon}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} =$$

$$= \beta_1 + \frac{\sum_{I=1}^{n} x_i(\beta_0 + \varepsilon_i) - n\bar{x}(\beta_0 + \bar{\varepsilon})}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} =$$

$$= \beta_1 + \frac{\sum_{i=1}^{n} x_i(\beta_0 + \varepsilon_i - \beta_0 - \bar{\varepsilon})}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

und somit folgt:

$$E(B_1) = E\left(\beta_0 + \frac{\sum_{i=1}^{n} x_i(\varepsilon_i - \bar{\varepsilon})}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\right) = \beta_0 \qquad \text{weil } E(\varepsilon_i - \bar{\varepsilon})=0 \text{ ist}$$

Satz: Die Zufallsvariable B_0 ist eine linear erwartungstreue Schätzfunktion für den Parameter β_0, der im klassischen Modell der linearen Einfachregression auftritt.

Beweis: $$B_0 = \bar{y} - B_1\bar{x} \tag{33}$$

Es wird für $\bar{y}$ eingesetzt:

$$\begin{aligned} B_0 &= \beta_0 + \beta_1\bar{x} + \bar{\varepsilon} - B_1\bar{x} \\ &= \beta_0 + \bar{\varepsilon} + \bar{x}\,(\beta_1 - B_1) \end{aligned}$$

Nachdem $E(\bar{\varepsilon}) = 0$ und $E(B_1) = \beta_1$ ist, kann gefolgert werden: $E(A) = \beta_0$

Jetzt können die vorhergesagten $\hat{y}_i$ der empirischen Regressionsgleichung als Realisierung der Zufallsvariablen $\hat{Y}_i$ betrachtet werden und dadurch gilt für den Erwartungswert $\hat{y}_i$:

$$E(\hat{y}_i) = E(\beta_0 + \beta_1 x_i) = \beta_0 + \beta_1 x_i$$
$$\Longrightarrow \qquad E(\hat{y}_i) = E(y_i)$$

2.3.2 Schätzung von σ²

In diesem Unterkapitel ist es das Ziel auch für σ^2 einen Schätzwert zu finden, um den im vorhergehenden Kapitel erhaltenen Schätzer wirklich anwenden zu können. Aus den Residuen bzw.der Fehlerquadratsumme erhält man einen erwartungstreuen Schätzer σ^2:

$$s_e^2 = \sum_{i=1}^{n} e_i^2 (y_i - \hat{y}_i) = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 \tag{34}$$

e_i, y_i und $\hat{y}_i$ werden wieder als Realisationen der Zufallsvariablen E, Y und $\hat{Y}$ gedeutet und nachdem $E(s_e^2) = E(\sum_{i=1}^{n} E_i^2) = (n-2)\sigma^2$ ist, kann ein unverzerrter Schätzer für σ^2 angegeben werden durch:

$$\hat{\sigma}^2 = \frac{s_e^2}{n-2}. \tag{35}$$

Beweis*:*

Es gilt $$Var(\varepsilon_i) = (1 - v_i)\,\sigma^2 \qquad \text{mit } v_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Anschließende Summation der v_i über alle n führt zu:

$$\sum_{i=1}^{n} v_i = \sum_{i=1}^{n} \frac{1}{n} + \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Aus dem ersten Summanden ergibt sich $\sum_{i=1}^{n} \frac{1}{n} = 1,$

ebenso für den zweiten $\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 1$

Daraus kann gefolgert werden, dass:

$$\sum_{i=1}^{n} v_i = v_1 + \ldots + v_n = 2$$

Summation der Varianzen Var (ε_i) über alle n ergibt:

$$\sum_{i=1}^{n} \mathrm{Var}(\varepsilon_i) = \sum_{i=1}^{n} (1 - v_i)\sigma^2 =$$
$$= \sum_{i=1}^{n} \sigma^2 - \sigma^2 \sum_{i=1}^{n} v_i =$$
$$= n\sigma^2 - 2\sigma^2 =$$
$$= (n - 2)\,\sigma^2$$

Aus $E(\varepsilon_i) = 0$ folgt $\mathrm{Var}(\varepsilon_i) = E(\varepsilon_i^2)$ und somit ist:

$$\sum_{i=1}^{n} E(\varepsilon_i^2) = (n-2)\sigma^2$$
$$E\left(\sum_{i=1}^{n} \varepsilon_i^2\right) = (n-2)\sigma^2$$

Beispiel: Um σ² für die Daten aus dem Kraftstoffverbrauchsbeispiel zu schätzen, bestimmen wir zuerst:

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = \sum_{i=1}^{n} y_i^2 - \frac{\sum_{i=1}^{n} (y_i)^2}{n}$$
$$= 808{,}02 - \frac{(123{,}8)^2}{20} = 41{,}7$$

Die Fehlersumme der Quadrate ist:

$$s_e^2 = S_{yy} - \hat{\beta}_1 S_{xy}$$
$$= S_{yy} - \hat{\beta}_1 S_{xy}$$
$$= 41{,}7 - (0{,}031)(1141{,}04)$$
$$= 6{,}327$$

Deshalb ergibt das geschätzte σ^2:

$$\hat{\sigma}^2 = \frac{s_e^2}{n-2} = \frac{6{,}327}{18} = 0{,}352$$

2.3.3 Eine alternative Form des Modells

Es existiert eine alternative Form des einfachen linearen Regressionsmodells welches sich gelegentlich als nützlich erweist. Angenommen man definiert die Regressor-Variable x_i als die Abweichung von ihrem eigenen Durchschnitt folgendermaßen: $x_i - \bar{x}$. Das Regressionsmodell wird dann zu:

$$y_i = \beta_0 + \beta_1(x_i - \bar{x}) + \beta_1\bar{x} + \varepsilon_i$$
$$= (\beta_0 + \beta_1\bar{x}) + \beta_1(x_i - \bar{x}) + \varepsilon_i$$
$$= \beta_0' + \beta_1(x_i - \bar{x}) + \varepsilon_i \qquad (36)$$

Zu beachten ist, dass die Regressor-Variable den Ursprung der x - Werte von Null zu $\bar{x}$ verschoben hat. Um die geschätzten Werte gleich zu halten im originalen wie im transformierten Modell, ist es notwendig, den originalen Abschnitt zu modifizieren. Die Beziehung zwischen dem originalen und dem transformierten Abschnitt kann wie folgt angegeben werden:

$$\beta_0' = \beta_0 + \beta_1 \bar{x}$$

Die kleinsten Quadrate Normalgleichungen für diese Form des Modells sind:

$$n\hat{\beta}_0' = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (y_i - \bar{y})\,(x_i - \bar{x})$$

und die resultierenden kleinste - Quadrate - Schätzer sind:

$$\hat{\beta}_0' = \bar{y}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i (x_i - \bar{x})}{(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

Demnach wird bei dieser Form des Modells der Abschnitt geschätzt durch $\bar{y}$ und die Neigung bleibt unbeeinflusst durch die Transformation.
Vorteile die sich nun durch dieses alternative Modell der linearen Regression ergeben sind:

a) Die Normalgleichungen sind leichter zu lösen, weil die Kreuzproduktterme verschwinden.
b) Die kleinste Quadrate Schätzer $\beta_0' = \bar{y}$ und $\hat{\beta}_1 = \frac{S_{xy}}{S_x^2}$ sind unkorreliert, sodass $\mathrm{Cov}(\hat{\beta}_0', \hat{\beta}_1) = 0$. Dadurch werden einige Anwendungen des Modells einfacher, wie z.B das Festlegen von Konfidenzintervallen um y.

Schlussendlich ist das geschätzte Modell: $\hat{y} = \bar{y} + \hat{\beta}_1 (x - \bar{x})$

Obwohl $\hat{y}$ äquivalent ist zu (10), erinnert dieses $\hat{y}$ den Analysten direkt daran, dass das Regressionsmodell nur gültig ist über den Bereich der x-Werte, wobei dieses Gebiet zentriert ist um $\bar{x}$.

2.4 Hypothesentest für die Steigung β_1 und Verschiebung auf der y-Achse β_0

Es besteht oft Interesse an Hypothesentests und Konfidenzintervallen bei den Modell-parametern. Diese Verfahren erfordern die zusätzliche Annahme, dass die Modellfehler ε_i normalverteilt sind. Daher setzt man normalverteilte, unabhängige Fehler $N(0, \sigma^2)$ voraus. Derartige Tests zur statistischen Überprüfung von Hypothesen sind sogenannte Signifikanztests. Sie gehen von dem Problem aus, dass:

- der Forscher/ die Forscherin eine Hypothese über einen Zusammenhang zwischen zwei Merkmalen (alternativ auch über Differenzen zwischen Gruppen hinsichtlich eines Merkmals) erstellt.

- Stichproben-Daten erhoben werden, in denen sich der vermutete Zusammenhang zeigt (das muss nicht unbedingt eintreten – wenn jedoch der Zusammenhang in der Stichprobe nicht vorhanden ist, erübrigt sich der Signifikanztest mehr oder weniger, ABER: Irgendein – wenn auch nur ein schwacher- Zusammenhang existiert meistens in den Daten)

- die Frage, ob die Annahme, dass der Zusammenhang auch in der Grundgesamtheit besteht, gültig ist.

Der „*Signifikanztest"* ermittelt die Wahrscheinlichkeit, mit der das gefundene empirische Ergebnis sowie noch extremere Ergebnisse auftreten können, wenn die Populationsverhältnisse der Nullhypothese entsprechen. Sofern diese Wahrscheinlichkeit < α % ist, bezeichnet man das Stichprobenergebnis als statistisch signifikant. Dabei werden für α per Konvention die Werte 5 % bzw. 1% festgelegt. So sind zum Beispiel Stichprobenergebnisse, deren bedingte Wahrscheinlichkeit bei Gültigkeit der H_0 kleiner als 5% ist, auf dem 5% (Signifikanz-)Niveau „signifikant".

Ein (sehr) signifikantes Ergebnis ist also ein Ergebnis, das sich mit der Nullhypothese praktisch nicht vereinbaren lässt, weshalb die Nullhypothese praktisch verworfen und die Alternativhypothese im Gegenzug akzeptiert wird. Andernfalls, also bei nicht signifikanten Ergebnis, wird die Nullhypothese beibehalten und die Alternativhypothese verworfen.

Angenommen man möchte jetzt die Hypothese testen, dass die Steigung gleich einer Konstanten, z.B c ist. Eine geeignete Hypothese wäre dann

$$H_0: \beta_1 = c$$
$$H_1: \beta_1 \neq c$$

wo eine zweiseitige Alternative angeführt wird. Da die Fehler $N(0,\sigma^2)$ verteilt sind, sind die Beobachtungen y_i $N(\beta_0 + \beta_1 x_i, \sigma^2)$ verteilt. Somit ist $\hat{\beta}_1$ eine Linearkombination der Beobachtungen, mit Erwartungswert β_1 und Varianz $\frac{\sigma^2}{S_x^2}$.
Somit ist die Teststatistik:

$$P_0 = \frac{\hat{\beta}_1 - c}{\sqrt{\frac{\sigma^2}{S_x^2}}} \qquad (37)$$

nach (2.3) normalverteilt mit N(0,1), sofern die Nullhypothese H_0: $\beta_0 = c$ zutrifft.
Wenn wir σ^2 kennen, können wir P_0 verwenden um die Hypothese zu testen. Andernfalls ist das mittlere Residuum zum Quadrat ($\hat{\sigma}^2$) ein erwartungstreuer Schätzer

von σ^2 und die Verteilung von $\frac{(n-2)\hat{\sigma}^2}{\sigma^2}$ ist χ^2_{n-2} – verteilt. Zudem sind $\hat{\sigma}^2$ und $\hat{\beta}_1$ unabhängige Zufallsvariable, was impliziert, dass sobald σ^2 in P_0 durch σ^2 ersetzt wird, die Statistik:

$$t_0 = \frac{\hat{\beta}_1 - c}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \tag{38}$$

t-verteilt ist, mit n-2 Freiheitsgraden, sofern die Nullhypothese $H_0: \beta_1 = c$ erfüllt wird. Die Freiheitsgrade von t_0 sind die Anzahl der Freiheitsgrade die mit $\hat{\sigma}^2$ verbunden werden. Der statistische t_0- Wert wird verwendet, um $H_0: \beta_1 = c$ zu testen und zwar durch einen Vergleich der beobachteten Werte von t_0 mit dem oberen $\frac{\alpha}{2}$ – Prozentpunkt der t_{n-2} Verteilung ($t_{\alpha/2, n-2}$). Verworfen wird die Nullhypothese, falls

$$|t_0| > t_{\alpha/2, n-2}$$

Um die Hypothese des y – Achsenabschnitts zu testen, kann genauso vorgegangen werden:

$$H_0: \beta_0 = d$$
$$H_1: \beta_0 \neq d$$

Es wird folgende Statistik verwendet:

$$t_0 = \frac{\hat{\beta}_0 - d}{\sqrt{(\hat{\sigma}^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}} \tag{39}$$

und die Nullhypothese wird verworfen, wenn $|t_0| > t_{\frac{\alpha}{2}, n-2.}$

Ein wichtiger Spezialfall von $H_0: \beta_1 = c, \quad H_1: \beta_1 \neq c$ ist:

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

Diese Hypothese bezieht sich auf die *Signifikanz der Regression*, wenn es verabsäumt wird H_0: β_1=0 zu verwerfen, wird impliziert, dass kein linearer Zusammenhang zwischen x und y besteht. Diese Situation wird in Abbildung 9 gezeigt, wobei hervorgehoben werden soll, dass das entweder impliziert, dass x kleine Werte annimmt um die Variation in y zu erklären und der beste Schätzer von y für irgendein x ist $\hat{y} = \bar{y}$ (Abbildung 7a) oder dass die richtige Beziehung zwischen x und y nicht linear ist (Abbildung 7b).

Alternativ, wenn H_0: β_1=0 verworfen wird, kann man implizieren, dass x von Wert ist um die Variabilität in y zu erklären, was in Abbildung 8 gezeigt wird. Allerdings kann das bedeuten, wenn H_0: β_0=0 verworfen wird, dass das geradlinige Modell passend ist (Abbildung 8a) oder dass, obwohl eine lineare Wirkung von x vorliegt, bessere Resultate erreicht werden können, wenn Polynomfunktionen höheren Grades zur

Näherung verwendet werden (Abbildung 8b).

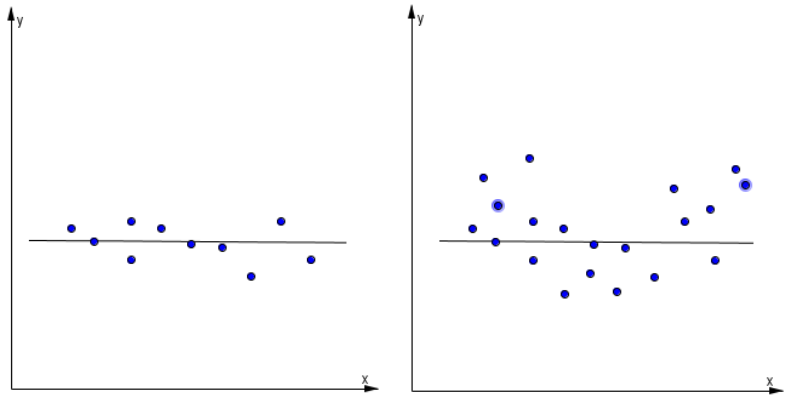

Abbildung 7: Situationen wo die Hypothese H0: $\boldsymbol{\beta_1}$= 0 nicht verworfen wird.

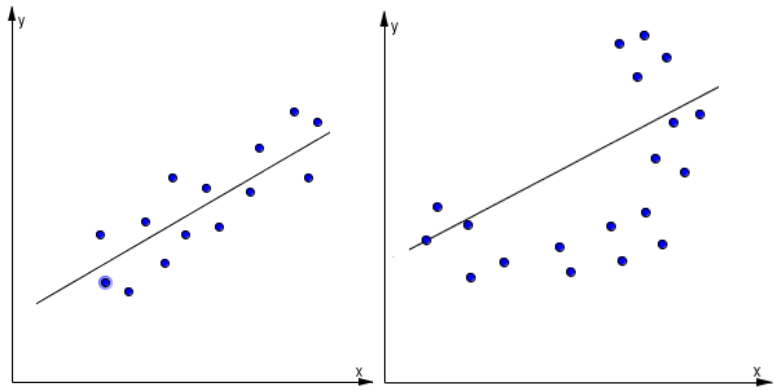

Abbildung 8: Situationen wo die Hypothese H_0: β_1 = 0 verworfen wird.

Um die Hypothese H_0: β_1= 0 zu testen, wird eine *„ analysis of variance“* – Methode verwendet. Die Teststatistik ist

$$F = \frac{\sum_{i=1}^{n} Y_i - \overline{Y}}{\frac{1}{(n-2)}\sum_{i=1}^{n}(Y_i - \widehat{Y}_1)^2} \qquad (40)$$

F – verteilt, und spiegelt das Verhältnis des Anteils, der durch die Regressionsfunktion erklärt wird und dem unerklärten Anteil wieder. Nachdem dieser Wert für die Gültigkeit von H_0 möglichst groß sein sollte, sprechen kleine Werte gegen H_0. Deshalb wird bei einem Test der Hypothese H_0: β_1= 0 die Teststatistik F berechnet und H_0 verworfen wenn F > $F_{\alpha,1,n-2}$.

2.5 Intervallschätzung bei einfachen linearen Regressionen

Dieser Abschnitt befasst sich mit Überlegungen zu Konfidenzintervallschätzungen von den Regressionsmodellparametern und dem Erwartungswert E(y) für gegebene x-Werte, wobei wiederum die Normalverteilungsannahmen der letzten Kapitel vorausgesetzt werden.

2.5.1 Konfidenzintervalle von β_0, β_1 und σ^2

Zusätzlich zu den Punktschätzungen von β_0,β_1 und σ^2 werden nun auch die beobachteten geschätzten Konfidenzintervalle dieser Parameter charakterisiert, weil die *Breite* dieser Intervalle eine bedeutende Maßzahl für die Qualität der Regressionslinie ist. Wenn die Fehler normalverteilt und unabhängig sind, sind beide Statistiken

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} \quad \text{und} \quad \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \tag{41}$$

t - verteilt mit n-2 Freiheitsgraden. Somit sind die *(1 - α)* - *Konfidenzintervalle* von β_1 und β_0 (und damit die Wahrscheinlichkeit, dass β_1 und β_0 im mit Wahrscheinlichkeit 1-α im Intervall liegt) gegeben durch:

$$\left[\hat{\beta}_1 - t_{\frac{\alpha}{2},n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \leq \quad \boldsymbol{\beta_1} \quad \leq \hat{\beta}_1 + t_{\frac{\alpha}{2},n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}\right]$$
$$\left[\hat{\beta}_0 - t_{\frac{\alpha}{2},n-2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)} \leq \quad \boldsymbol{\beta_0} \quad \leq \hat{\beta}_0 + t_{\frac{\alpha}{2},n-2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}\right] \tag{42}$$

Wenn man die Auswahl einer Stichprobe mit demselben Umfang bzw. einem ähnlichen Format der x-Werte, wiederholt, dann würden immerhin 95% dieser Intervalle den wahren Wert von β_1 beinhalten. Die Quantität

$$s_e(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

des Konfidenzintervalls vorhin wird als Standardfehler der Steigung $\hat{\beta}_1$ bezeichnet. Dies ist eine Maßzahl dafür, wie präzise der Anstieg der Regressionsgerade geschätzt wurde. Fast ident kann auch der Standardfehler $s_e(\hat{\beta}_0)$ des oberhalb beschriebenen Konfidenzintervalls bestimmt werden:

$$s_e(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}$$

Insofern ist die ausgewählte Verteilung von $\frac{(n-2)\,\hat{\sigma}^2}{\sigma^2}$ die Chi-Quadrat Verteilung mit n - 2 Freiheitsgraden (Anm.: bei normalverteilten und unabhängigen Variablen)

Deshalb ist:

$$P\left(\chi^2_{1-\frac{\alpha}{2},\ n-2} \leq \frac{(n-2)\hat{\sigma}^2}{\sigma^2} \leq \chi^2_{\frac{\alpha}{2},\ n-2}\right) = \quad 1-\alpha \qquad (43)$$

woraus weiterführend durch umformen das Konfidenzintervall für σ^2 ableitbar ist.

Beispiel: Hier sollen die 95 % Konfidenzintervalle für β_1 und σ² aus den Kraftstoffverbrauchdaten (siehe Tabelle im Anhang) bestimmt werden. Der Standardfehler von $\hat{\beta}_1$ ist

$$s_e(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = \sqrt{\frac{0{,}352}{36838{,}2}} = 0{,}0031$$

und die T-Tabelle liefert für $t_{0,25,18}$ den Wert 2,101. Als Konfidenzintervall ergibt sich also hier:

$$0{,}031 - (2{,}101)\cdot(0{,}0031) \ \leq \ \beta_1 \ \leq \ 0{,}031 + (2{,}101)\cdot(0{,}0031)$$
$$0{,}025 \leq \ \boldsymbol{\beta_1} \ \leq \ 0{,}0038$$

Statistik bei einer Stichprobe

	H	Mittelwert	Standardabweichung	Standardfehler Mittelwert
Leistung	20	**99,70**	44,032	9,846
Verbrauch	20	**6,1900**	1,48143	,33126

Test bei einer Stichprobe

	Testwert = 0					
					95% Konfidenzintervall der Differenz	
	t	df	Sig. (2-seitig)	Mittelwertdifferenz	Unterer	Oberer
Leistung	10,126	19	,000	99,700	79,09	120,31
Verbrauch	18,686	19	,000	6,19000	5,9067	6,4733

Tabelle 7: Konfidenzintervalle für das Kraftstoffverbrauchbeispiel

2.5.2 Intervallschätzung des Erwartungswertes

Eine der Hauptanwendungen des Regressionsmodells ist jene der Schätzung des Erwartungswerts E(y) für einen speziellen Wert der unabhängigen Variable x. Es wird zum Beispiel x_0 als jene unabhängige Variable x gewählt, für welche der Erwartungswert geschätzt werden soll. Zudem soll ein x_0 gewählt werden, das innerhalb des Datenbereiches der Originaldaten von x liegt. Ziel ist es nun einen erwartungstreuen Schätzer von Y im Punkt x_0 ($E(y|x_0)$) zu bestimmen, der durch das folgende Modell: $\widehat{E(Y|x_0)} = \hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ beschrieben werden kann.

Um ein (1-α)-Konfidenzinterfall von $E(y|x_0)$ zu erhalten, muss beachtet werden, dass $\hat{y}_0$

eine normalverteilte Zufallsvariable ist, weil eine Linearkombination der Beobachtungen y_i vorliegt. Die Varianz von $\hat{y}_0$ ist nun:

$$\begin{aligned}\mathrm{Var}(\hat{y}_0) &= \mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0)\\ &= \mathrm{Var}\left[\bar{y} + \hat{\beta}_1(x_0 - \bar{x})\right]\\ &= \frac{\sigma^2}{n} + \frac{\sigma^2(x_0-\bar{x})^2}{S_{xx}}\\ &= \sigma^2\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]\end{aligned}$$

Folglich ist die Beispiel-Verteilung von:

$$\frac{\hat{y}_0 - E(y|x_0)}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}}$$

t-verteilt mit n-2 Freiheitsgraden und daher kann ein (1 - α)-Konfidenzintervall des Erwartungswerts beim Punkt $x = x_0$ angegeben werden durch:

$$\left[\hat{y}_0 - t_{\frac{\alpha}{2},n-2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)} \le E(y|x_0) \le \hat{y}_0 + t_{\frac{\alpha}{2},n-2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}\right] \quad (44)$$

Die „Intervallbreite“ ist somit minimal für $x_0 = \bar{x}$ und wird größer wenn sich $|\, x_0 - \bar{x}\,|$ erhöht.

Beispiel: Jetzt wollen wir ein 95 % - Konfidenzintervall von $E(y,x_0)$ für die Kraftstoffverbrauchdaten bestimmen. Dazu wird in die gerade aufgestellte Formel dementsprechend eingesetzt

$$\hat{y}_0 - 2{,}101\sqrt{0{,}352\left(\frac{1}{20} + \frac{(x_0-99{,}7)^2}{36838{,}2}\right)} \le \mathbf{E(y|x_0)} \le \hat{y}_0 + 2{,}101\sqrt{0{,}352\left(\frac{1}{20} + \frac{(x_0-99{,}7)^2}{36838{,}2}\right)}$$

Ist zum Beispiel $x_0 = \bar{x} = 99{,}7$, so ist $\hat{y}_0 = 6{,}19$ und wir erhalten folgendes 95% Intervall

$$5{,}91 \le \mathbf{E(y|99,7)} \le 6{,}47$$

2.5.3 Interpolation und Extrapolation neuer Beobachtungen

Eine Extrapolation entspricht der Schätzung von Datenpunkten auf der Regressionsgeraden über den gesicherten Bereich der vorgegebenen x-Werte hinaus wohingegen die Interpolation jene Herangehensweise beschreibt, bei welcher innerhalb des Bereichs gesicherter Werte, auch jene Funktionswerte von x durch die Gerade geschätzt werden, die gar nicht untersucht wurden.

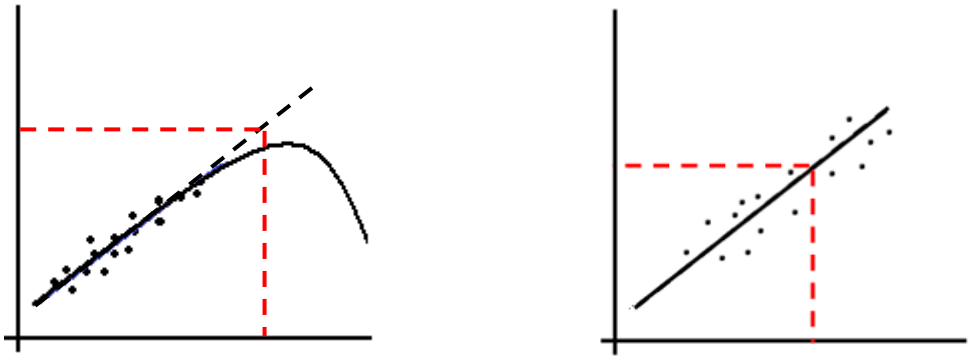

Abbildung 9: Beispiel für Extrapolation / Beispiel für Interpolation

Obwohl diese Verfahren fehlerbehaftet sind, dienen derartige Anwendungen des Modells zur Vorhersage neuer Beobachtungen. Allerdings ist das Konfidenzintervall des Erwartungswertes bei $x = x_0$ ungeeignet, um die zukünftige Beobachtung y_0 zu schätzen, weil dies eine Intervallschätzung des Erwartungswertes von y, also einem Parameter und nicht von einer Wahrscheinlichkeitsaussage über Zukunftsbeobachtungen dieser Verteilung ist. Ein Vorhersageintervall für die Zukunftsbeobachtungen kann allerdings dennoch durch:

$$\hat{y}_0 - t_{\frac{\alpha}{2},n-2}\sqrt{\hat{\sigma}^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_x^2}\right)} \leq y_0 \leq \hat{y}_0 + t_{\frac{\alpha}{2},n-2}\sqrt{\hat{\sigma}^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_x^2}\right)}$$

angegeben werden. Dieses Vorhersageintervall ist wiederum minimal bei $x_0 = \bar{x}$ und wird größer wenn sich $|x_0 - \bar{x}|$ erhöht. Ein Vergleich mit dem Konfidenzintervall darüber deutet darauf hin, dass das Vorhersageintervall bei x_0 immer größer ist als das Konfidenzintervall bei x_0, weil das Vorhersageintervall zum einen vom Fehler des beobachteten Modells und zum anderen von jenem Fehler, der in Zusammenhang mit den Zukunftsbeobachtungen steht, abhängt.

2.5.4 Maximum-Likelihood Schätzung

Die Methode der kleinsten Quadrate kann angewendet werden, um die Parameter eines linearen Regressionsmodells zu schätzen und zwar ohne Rücksicht auf die Art der Verteilung der Fehler ε. Andere statistische Verfahren wie Hypothesentests und die Konstruktion von Konfidenzintervallen setzen hingegen sehr wohl die Normalverteilung der Fehler voraus. In komplexeren Fällen, wenn zum Beispiel der zu schätzende Zusammenhang nicht linear ist, kann jedoch in etwa mit der Maximum-Likelihood-Methode auf ein alternatives Verfahren zur Schätzung unbekannter Parameter zurückgegriffen werden.

Im Wesentlichen geht es bei einer Maximum-Likelihood Methode darum, eine konkrete Stichprobe vorliegen zu haben und die Frage zu beantworten, für welche Parameterwerte (z.B. Mittelwert und/oder Varianz) das Zustandekommen dieser konkreten Stichprobe „am wahrscheinlichsten“ ist. Dazu muss allerdings a-priori

bekannt sein, aus welcher Verteilung diese Stichprobe gezogen wurde. In dieser Hinsicht ist somit die kleinste-Quadrate-Methode weniger restriktiv.

Wir betrachten die Methode zuerst für den einfachsten Fall, nämlich für die Verteilung einer Zufallsvariablen X mit der Wahrscheinlichkeitsfunktion bzw. Dichte f(x), die von einem einzelnen Parameter u abhängt. Das betreffende Experiment werde dann n-mal ausgeführt und die so erhaltene Stichprobe von n voneinander unabhängigen Werten sei $x_1, x_2, ..., x_n$.

Im Fall einer diskreten Variablen ist dann die Wahrscheinlichkeit, eine Stichprobe zu erhalten, die gerade aus den obigen Werten besteht, durch das folgende Produkt gegeben

$$\mathbf{L} = f(x_1)f(x_2) \dots f(x_n) \tag{44}$$

denn $f(x_1)$ ist die Wahrscheinlichkeit, mit der X den Wert x_1 annimmt, usw. Ist X stetig verteilt, so ist dementsprechend die Wahrscheinlichkeit, eine Stichprobe zu erhalten, die sich gerade aus n Werten zusammen setzt, die in den kleinen Intervallen

$$x_1 \le x \le x_1 + \Delta x, \dots, x_n \le x \le x_n + \Delta x$$

liegen, gegeben durch $f(x_1)\Delta x \cdot f(x_2)\Delta x \dots \cdot f(x_n)\Delta x = l(\Delta x)^n$

Nachdem die Werte $f(x_1)$,..., $f(x_n)$ von u abhängen und L demnach von x_1, ..., x_n und auch von u abhängt, so ist $L = L(x_1, ..., x_n, u)$ und wird als **Likelihood-Funktion** bezeichnet.

Für die Maximum-Likelihood-Methode gilt nun, als Näherung für den unbekannten Parameter *u* einen Wert zu nehmen, für den (die Wahrscheinlichkeit) L möglichst maximal wird.

Dazu bildet man die partielle Ableitung von L nach u

$$\frac{\partial l}{\partial u} = 0$$

und zwar deshalb die partielle Ableitung, weil l auch von den Größen $x_1,\dots,x_n$ abhängt.

Da f(x) nicht negativ ist, so ist L an der Stelle eines Maximums i. A. positiv. Der natürliche Logarithmus ln L ist genauso eine monoton wachsende Funktion von L, die dort ein Maximum hat, wo L ein Maximum hat, dadurch verwenden wir:

$$\frac{\partial \ln L}{\partial u} = 0 \tag{45}$$

Dementsprechend erhält man bei einer Verteilung mit mehreren, z. B z-Parameter $u_1,\dots u_z$, die z-Gleichungen

$$\frac{\partial L}{\partial u_1} = 0, \dots, \frac{\partial L}{\partial u_z} = 0 \quad \rightarrow \quad \frac{\partial \ln L}{\partial u_1} = 0, \dots, \frac{\partial \ln L}{\partial u_z} = 0$$

Somit hat man anstatt lästiger Differentiation von Produkten nur Summen zu differenzieren.

Wird nun eine Stichprobe aus einer gegebenen Verteilung gezogen, so gibt die Wahrscheinlichkeitsfunktion (charakterisiert durch einen unbekannten Parameter τ die Wahrscheinlichkeit an, mit der die Realisationen gezogen werden und hängt natürlich von den Parametern der Grundgesamtheit ab, z.B. dem Mittelwert.

$$f(y_1, \dots, y_n|\tau) = f(Y = y_1|\tau) \cdot \dots \cdot f(Y = y_n|\tau) = \prod_{i=1}^{n} f(Y = y_i|\tau)$$

gibt die Wahrscheinlichkeit der Realisation dieser Stichprobe für gegebene Parameter τ an und die Likelihoodfunktion L interpretiert nun diese gemeinsame Wahrscheinlichkeitsfunktion als Funktion unbekannter Parameter τ für gegebene Beobachtungen.

$$L(\tau|Y) = l(\tau|Y = y_1) \cdot \dots \cdot l(\tau|Y = y_n) = \prod_{i=1}^{n} l(\tau|y_i)$$

Aus der Dichtefunktion von Y:

$$f(y_i, \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(y_i-\mu)}{2\sigma^2}}$$

folgt die Likelihoodfunktion einer Stichprobe vom Umfang n:

$$\Rightarrow \qquad L(\mu, \hat{\sigma}^2|y) = \prod_{i=1}^{n} \frac{1}{2\pi^2\sigma^2} e^{-\frac{-(y_i-m)^2}{2\sigma^2}}$$

Aufgrund der Monotonieeigenschaften des Logarithmus bildet man nun die Log-Likelihood Funktion:

$$\ln L = -n \cdot ln\sigma - n \cdot ln\sqrt{2\pi} - \frac{1}{2\sigma^2} - \frac{1}{2\sigma^2}\sum (y_i - m)^2$$

Auch wenn sich die Form dieser Log-Likelihood Funktion bei wiederholten Ziehungen von Stichprobe zu Stichprobe unterscheidet, kann für eine gegebene Stichprobe ein Schätzwert für den Parameterwert μ berechnet werden, der eben diese konkrete Stichprobe „am wahrscheinlichsten macht“.

Das Maximum dieser Funktion erhält man, durch Null-setzen der ersten (partiellen) Ableitungen der Log-L.-Funktion:

$$\frac{\partial \ln L}{\partial \mu} = \frac{2}{2\sigma^2}\sum_{i=1}^{n}(y_i - m) = 0$$

$$\sum_{i=1}^{n} y_i = n \cdot m$$

$$m = \frac{\sum y_i}{n} = \bar{y}$$

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{n}{s} + \frac{1}{s^3}\sum_{i=1}^{n}(y_i - m)^2 = 0$$

$$\sum_{i=1}^{n}(y_i - m)^2 = ns^2$$

$$s^2 = \frac{\sum(y_i - \bar{\mu})^2}{n}$$

Beispiel (Poisson-Verteilung):
Unter Verwendung einer Stichprobe $x_1,...,x_n$ gewinnt man eine Maximum-L.-Schätzfunktion für den Parameter μ der Poissonverteilung $f(x) = \frac{\mu^x}{x!} e^{-\mu}$. Für L ergibt sich durch (44) folgendes Produkt:

$$L = \frac{\mu^{x_1}}{x_1!} e^{-\mu} \cdot \frac{\mu^{x_2}}{x_2!} e^{-\mu} \dots \frac{\mu^{x_n}}{x_n!} e^{-\mu}$$

Durch zusammenfassen der Exponentialfaktoren und auch der Potenzen folgt

$$L = \frac{1}{x_1! \dots x_n!} \mu^{x_1+\cdots+x_n} e^{-n\mu} = \frac{1}{x_1! \dots x_n!} \mu^{n\bar{x}} e^{-n\mu}$$

→ durch logarithmieren: $\ln L = -\ln(x_1! \dots x_n!) + n\bar{x} \ln \mu - n\mu$

Also hat hier (45) die Form: $\frac{\partial \ln L}{\partial \mu} = \frac{n\bar{x}}{\mu} - n = 0$ und die Schätzfunktion $\tilde{\mu}$:

$$\tilde{\mu} = \bar{x} = \frac{1}{n}(x_1 + \cdots x_n)$$

2.5.5 Simultane Rückschlüsse auf die Modellparameter

In den vorangegangenen Unterkapiteln wurden einige Typen von Konfidenz- und Vorhersageintervallen definiert und es zeigte sich, dass einige Probleme entstehen, wenn derartige Intervalle für ein und dieselbe Stichprobe bestimmt werden. In diesem Fall ist der Analyst für gewöhnlich an einem speziellen Faktor bzw. Koeffizient interessiert, der simultan auf eine Auswahl von Intervallschätzungen zutrifft. Diese Auswahl von Intervallen, die gleichzeitig mit Wahrscheinlichkeit 1- α stimmen, heißen simultane Konfidenz-/Vorhersageintervalle.
Betrachtet man nun die Schätzung für β_0 und β_1 mit einem ausgewählten Konfidenzbereich, so dass mit einer Überzeugung von 100(1- α) % beide Schätzungen korrekt sind, so ist das Modell gegeben durch:

$$y = \beta_0 + \beta_1 x + \varepsilon = \beta_0' + \beta_1(x - \bar{x}) + \varepsilon$$

Die kleinste-Quadrate Schätzer von β_0 und β_1 sind $\hat{\beta}_0' = \bar{y}$ und $\hat{\beta}_1' = \frac{S_{xy}}{S_{xx}}$ mit:

$\mathrm{Var}(\hat{\beta}_0') = \frac{\sigma^2}{n}$ bzw. $\mathrm{Var}(\hat{\beta}_1') = \frac{\sigma^2}{S_{xx}}$ und Standardnormalverteilungen zum Quadrat:

$$\left[\frac{\hat{\beta}_0' - \beta_0'}{\sqrt{\frac{\sigma^2}{n}}}\right]^2 = \frac{n(\hat{\beta}_0' - \beta_0')^2}{\sigma^2} \sim \chi_1^2 \quad \text{und} \quad \left[\frac{\hat{\beta}_1 - \beta_1'}{\sqrt{\frac{\sigma^2}{n}}}\right]^2 = \frac{S_{xx}(\hat{\beta}_0' - \beta_0')}{\sigma^2} \sim \chi_1^2 \qquad (46)$$

Die Additivitätseigenschaft von Chi-Quadrat und die Unabhängigkeit der beiden eben betrachteten Chi-Quadrat-verteilten Zufallsvariablen $\hat{\beta}_0'$ und $\hat{\beta}'_1$ lässt auf folgendes schließen:

$$\frac{n(\hat{\beta}_0' - \beta_0')^2}{\sigma^2} + \frac{S_{xx}(\hat{\beta}_0' - \beta_0')}{\sigma^2} \sim \chi_2^2$$

Jetzt ist die Verteilung von $\frac{(n-2)\hat{\sigma}^2}{\sigma^2}$, χ^2_{n-2}-verteilt und $\hat{\sigma}^2$ unabhängig von $\hat{\beta}_0^{'}$ und $\hat{\beta}_1$

$$\Rightarrow \qquad \frac{\frac{1}{2}\left[\frac{n\left(\hat{\beta}_0^{'}-\beta_0^{'}\right)^2+S_{xx}\left(\hat{\beta}_1-\beta_1\right)^2}{\sigma^2}\right]}{\left[\frac{(n-2)\hat{\sigma}^2}{\sigma^2}\right]}=\frac{n\left(\hat{\beta}_0^{'}-\beta_0^{'}\right)^2+S_{xx}(\hat{\beta}_1-\beta_1)^2}{2\hat{\sigma}^2}$$

Substituieren $\hat{\beta}_0^{'}=\hat{\beta}_0+\hat{\beta}_1\bar{x}$ und $\beta_0^{'}=\beta_0+\beta_1\bar{x}$:

$$P\left(\frac{n\left(\hat{\beta}_0-\beta_0\right)^2+2\sum x_i\left(\hat{\beta}_0-\beta_0\right)\left(\hat{\beta}_1-\beta_1\right)+\sum x_i^2(\hat{\beta}_1-\beta_1)^2}{2\hat{\sigma}^2}\leq F_{\alpha.2.n-2}\right)=1-\alpha$$

Diese Gleichung definiert dann sogar eine Ellipse, welche bei wiederholtem Ziehen einer Stichprobe aus der Grundgesamtheit, β_0 und β_1 in 100(1-α)% der Fälle gleichzeitig enthält.

Beispiel: Für eine Konstruktion dieser Konfidenzregion werden wiederum die Kraftstoffverbrauchdaten herangezogen. Um eine 95 % Konfidenzregion für β_0 und β_1 bestimmen zu können, setzen wir $\hat{\beta}_0=3{,}099, \hat{\beta}_1=0{,}031, \sum_{i=1}^{20} x_i^2=235\,640$, $\hat{\sigma}^2=0{,}352$ und $F_{0{,}05,\ 2,\ 18}=\mathbf{3,55}$ in die Ungleichung oben ein und erhalten folgendes

$$[\,20\,(3{,}099-\beta_0)^2+2\,(1994)(3{,}099-\beta_0)\,(0{,}031-\beta 1)+ \\ +(235640)(0{,}031-\beta 1)^2]\,/\,[2(0{,}352)]=\mathbf{3,55}$$

als Außengrenze der Ellipse.

Anzumerken ist, dass die Ellipse nicht parallel zur β_1 – Achse und die Schiefe der Ellipse eine Funktion der Kovarianz zwischen $\hat{\beta}_0$ und $\hat{\beta}_1$ ist, welche durch $-\bar{x}\sigma^2/S_{xx}$ beschrieben wird. Liegt eine positive Kovarianz vor, so wird angenommen, dass die Fehler in den Punktschätzungen von β_0 und β_1 voraussichtlich in derselben Richtung liegen während eine negative Kovarianz indiziert, dass diese Fehler wahrscheinlich in entgegengesetzte Richtungen liegen. In unserem Beispiel ist $\bar{x}$ positiv, weshalb die Kovarianz Cov($\hat{\beta}_0, \hat{\beta}_1$) negativ ist. Die Ausdehnung der Region hängt von den relativen Größen der Varianzen von β_0 und β_1 ab.

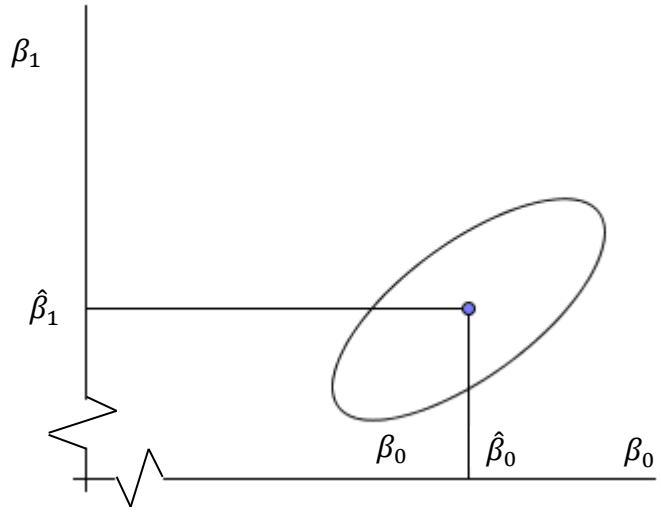

Abbildung 10: 95 % - Konfidenzregion für $\boldsymbol{\beta_0}$ und $\boldsymbol{\beta_1}$ für die Kraftstoffverbrauchsdaten

Zusätzlich gibt es eine andere allgemeine Annäherung, um ähnliche Intervallschätzungen der Parameter in einem einfachen linearen Regressionsmodell zu erhalten. Diese Konfidenzintervalle können nun so konstruiert werden, indem man $\boldsymbol{\beta_j \mp \Delta s_e(\beta_j)}$ **mit j = 0,1**verwendet, wo das konstante Δ derart gewählt wird, dass beide Intervalle korrekt sind. Nun können einige Methoden gewählt werden um Δ zu bestimmen:

Die Bonferroni Methode

Die Bonferroni-Konfidenzintervalle sind irgendwie gewöhnliche Konfidenzintervalle basierend auf der t- Verteilung, außer dass jedes einzelne Konfidenzintervall für β_0 den Konfidenzkoeffizienten 1 – α/2 anstelle von 1- α hat. Bei dieser Approximation setzen wir $\Delta = t_{\alpha/4,\ n-p}$ so dass dies eingesetzt $\boldsymbol{\beta_j \mp t_{\frac{\alpha}{4},n-p}\, s_e(\beta_j)}$ **mit j = 0,1** ergibt. Um zu verifizieren, dass diese Approximation zu korrekten Aussagen führt, wird angenommen, dass E_0 das Ereignis für ein falsches Konfidenzintervall für β_0 ist und E_1 jenes Ereignis, dass das Konfidenzintervall für β_1 inkorrekt ist, so dass $P(E_0) = P(E_1) = \alpha/2$
Die Wahrscheinlichkeit dass entweder eines oder beide Ereignisse inkorrekt sind ist:

$$P(E_0 \cup E_1) = P(E_0) + P(E_1) - P(E_0 \cap E_1) \qquad (47)\ (I)$$

und

$$1 - P(E_0 \cup E_1) = 1 - P(E_0) - P(E_1) + P(E_0 \cap E_1) \qquad (II)$$

Nachdem $1 - P(E_0 \cup E_1) = P(\overline{E_0 \cup E_1}) = P(\bar{E}_0 \cap \bar{E}_1)$, ist die linke Seite von (II) die Wahrscheinlichkeit, dass beide Konfidenzintervalle korrekt sind. Nachdem außerdem $P(E_0 \cap E_1) \geq 0$ ist, können wir (II) folgendermaßen schreiben:

$P(\bar{E}_0 \cap \bar{E}_1) = P$ (beide Intervalle sind korrekt)
$\geq 1 - P(E_0) - P(E_1)$
$\geq 1 - \alpha/2 - \alpha/2 \geq 1 - \alpha$ (Dieser Ausdruck nennt sich Bonferroni Ungleichheit)

Es muss β_0 und β_1 mit Konfidenzintervallen geschätzt werden, so dass der gwählte Koeffizient zumindest 1 – α ist und dann werden 100(1 – α/2) % Konfidenzintervalle gebildet, sowohl für β_0 als auch β_1.

Beispiel für die Kraftstoffverbrauchdaten (siehe Tabelle im Anhang)
Bilden eines 90 % Konfidenzintervalls für β_0 und β_1, indem ein 95 % Intervall für jeden Parameter aufgestellt wird wird.

$$\hat{\beta}_0 = 3{,}099\,, \qquad s_e(\hat{\beta}_0) = 0{,}335$$
$$\hat{\beta}_1 = 0{,}031\,, \qquad s_e(\hat{\beta}_1) = 0{,}352$$

und $t_{0,05/2,\ 18} = 2{,}101$, die Konfidenzintervalle sind allgemein

$$\hat{\beta}_0 - t_{0,025,18}\, s_e(\hat{\beta}_0) \leq \beta_0 \leq \hat{\beta}_0 + t_{0,025,18}\, s_e(\hat{\beta}_0)$$

$$\hat{\beta}_1 - t_{0,025,18}\, s_e(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{0,025,18}\, s_e(\hat{\beta}_1)$$

Allerdings ist die Bonferroni Methode nicht die einzige Approximation um Δ passend zu wählen, sondern andere Methoden wie die Scheffe S- Methode mit $\Delta = (2F_{\alpha,2,n-2})^{1/2}$ oder das Maximum-Modul t - Verfahren mit $\Delta = u_{\alpha,\,2,\,n-2}$, wo $u_{\alpha,\,2,\,n-2}$ der obere Ausläufer der Verteilung vom absoluten Maximalwert zweier unabhängig verteilter student-t Zufallsvariablen gewählt wird, sind ebenso geeignet.

Kapitel 3

3. MASSZAHLEN FÜR DIE MODELLADÄQUANZ

Die wesentlichen Annahmen die bislang behandelt wurden, sind die folgenden:

- linearer Zusammenhang zwischen x und y oder zumindest eine näherungsweise Beziehung durch eine Gerade
- der Fehler ε hat Erwartungswert 0
- der Fehler ε hat eine konstante Varianz σ^2
- die Fehler sind unkorreliert
- die Fehler sind normalverteilt

Nun sollen einige Typen von Modellunzulänglichkeiten diskutiert werden, die potentiell ernstzunehmende Folgen haben und dazu führen können, dass verschiedene Stichproben ein komplett anderes Modell mit gegenteiligen Schlussfolgerungen ergeben. Für gewöhnlich können Abweichungen von den zugrundeliegenden Annahmen nicht durch Überprüfung der Standardstatistiken (wie T-Statistik, F-Statisitk oder R^2) geschützt werden, weil diese „globale“ Modelleigenschaften sind und als solche die Angemessenheit des Modells nicht garantieren. Deshalb ist es das ausgewiesene Ziel, hier einige nützliche Methoden für die Diagnose und den Umgang mit Verletzungen der einfachen Regressionsannahmen vorzustellen.

3.1 Residualanalyse

Die Residualanalyse ist im Rahmen der Regressionsmodelle ziemlich bedeutend, darum werden zuerst kurz die zentralen Ziele und Anwendungsgebiete der Untersuchung von Residuen vorgestellt werden, um dann die Residuen im Klassischen Linearen Modell zu definieren und anzuwenden. Das darauffolgende Kapitel bezieht sich dann auf den allgemeinen Gebrauch der geschätzten Störterme im Kontext der Generalisierten Linearen Modelle (GLM).

Die Vorteile der unterschiedlichen Formulierungen der Residualanalyse sind vor allem im Hinblick auf die Modelldiagnostik zu untersuchen. Zudem wird die Residualanalyse in der multiplen linearen Regression zur Annahmenprüfung (z.B von Varianzhomogenität oder Unkorreliertheit der Fehlerterme) verwendet, um die Gestalt des linearen Prädiktors zu diskutieren, der die systematische Komponente charakterisiert. Zudem kann auch das Untersuchen von Ausreißern ein Anwendungsgebiet für die Residualanalyse sein.

3.1.1 Definition der Residuen

Die Residuen wurden definiert durch:

$$e_i = y_i - \hat{y}_i \qquad \text{i} = 1, \ldots, \text{n} \tag{48}$$

wobei y_i eine exakte Beobachtung und $\hat{y}_i$ der entsprechende geschätzte Wert ist. Das Residuum kann somit betrachtet werden als die Differenz zwischen exaktem und geschätztem Wert und stellt eine Maßzahl für die Schwankungen dar, die nicht im Modell erklärt werden. Darum sollen einige Abweichungen der angenommenen Annahmen der Fehler in den Residuen aufgezeigt werden, wie zum Beispiel ein Erwartungswert gleich Null oder eine geschätzte durchschnittliche Varianz von $\hat{\sigma}^2$:

$$\frac{\sum_{i=1}^{n}(e_i-\bar{e})^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \hat{\sigma}^2 \tag{49}$$

Nachdem die Residuen nicht unabhängig sind, führt das zu Auswirkungen auf die Modelladäquanz, solange *n* nicht zu klein ist. Darum ist es vorteilhaft, manchmal „*standardisierte Residuen*“ anzuwenden:

$$d_i = \frac{e_i}{\sqrt{\hat{\sigma}^2}}, \quad \text{i= 1, ... ,n} \tag{50}$$

Die *standardisierten Residuen* haben den Erwartungswert Null und näherungsweise einheitliche Varianz. Zudem unterteilt diese Gleichung die Residuen in Gruppen mit einheitlicher mittlerer Standardabweichung, weil in einigen (einfachen) linearen Regressionsdatensätzen Residuen auftreten können, deren Standardabweichungen sich markant unterscheiden.

$$\begin{aligned} \text{Var}\,(\text{e}_\text{i}) &= \text{Var}\,(\text{y}_\text{i} - \hat{\text{y}}_\text{i}) \\ &= \text{Var}(\text{y}_\text{i}) + \text{Var}(\hat{\text{y}}_\text{i}) - 2\text{Cov}(\text{y}_\text{i}, \hat{\text{y}}_\text{i}) \\ &= \sigma^2 + \sigma^2\left[\frac{1}{\text{n}} + \frac{(\text{x}_\text{i}-\bar{\text{x}})^2}{\text{S}_{\text{xx}}}\right] - 2\text{Cov}(\text{y}_\text{i}, \hat{\text{y}}_\text{i}) \end{aligned}$$

$$\Longrightarrow \qquad \begin{aligned} \text{Cov}(\text{y}_\text{i}, \hat{\text{y}}_\text{i}) &= \text{Cov}[\text{y}_\text{i}, \bar{\text{y}} + \frac{\text{S}_{\text{xy}}}{\text{S}_\text{x}^2}(\text{x}_\text{i} - \bar{\text{x}})] \\ &= \sigma^2[\frac{1}{\text{n}} + \frac{(\text{x}_\text{i}-\bar{\text{x}})^2}{\text{S}_{\text{xx}}}] \end{aligned}$$

Nachdem eine Hauptaufgabe im Linearen Modell darin besteht, die Modellannahmen zu prüfen und insbesondere die Residuen zu betrachten, gibt es dafür verschiedene

Definitionen der zu untersuchenden Residuen, wobei die intuitivste Form der Unterschied zwischen dem durch die Regression vorhergesagten Wert und dem beobachteten Wert ist. Außerdem sollten im Idealfall die Residualplots keine (bzw. nur geringe) Varianzhomogenität aufweisen bzw. keine Autokorrelationsstruktur haben.
Die Varianzhomogenität kann somit nicht durch die normalen Residuen graphisch diskutiert werden, weil diese Residuen Varianzheterogenität aufweisen, auch wenn die Annahmen der Regression erfüllt sind. Aus diesem Grund wird eine mögliche *Standardisierung* eingeführt. Daraus wiederum folgt die Varianz des i-ten Residuums:

$$\mathrm{Var}(e_i) = \sigma^2\left[1-\left(\frac{1}{n}+\frac{(x_i-\bar{x})^2}{S_{xx}}\right)\right]$$

Die *„studentisierten Residuen"* können dann definiert werden durch:

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2\left[1-\underbrace{\left(\frac{1}{n}+\frac{(x_i-\bar{x})^2}{S_{xx}}\right)}_{p_{ii}}\right]}}, \quad i=1,\dots,n \qquad (51)$$

Im Nenner der Formel der standardisierten Residuen existiert die geschätzte Standardabweichung der Residuen, welche wiederum von den e_i abhängt. Aus diesem Grund lässt sich bei (50) keine Verteilung der standardisierten Residuen angeben, mit den studentisierten Residuen gelingt dies jedoch. In (51) wird das gewöhnliche kleinste-Quadrate Residuum e_i geteilt durch den exakten Standardfehler. Die Anwendung des studentisierten Residuums bei Regressions- diagnosen ist vor allem bei kleinen Datensätzen ziemlich nützlich, weil dadurch oft eine geeignetere Gruppierung der Varianzen gegeben ist, im Gegensatz zu den Standardresiduen, und die Differenzen bei Residuenvarianzen deutlicher sind. Bei großem *n* tritt nur ein kleiner Unterschied zwischen den zwei Methoden der kleinste-Quadrate Residuen auf. Im nächsten Abschnitt werden nun einige Residuenplots vorgestellt, die nützlich sind, um Unangemessenheiten des Modells aufzudecken.

Standardisierte und studentisierte Residuen der Kraftstoffdaten:

e_i	d_i	r_i
0,037	0,062	0,064
-0,152	-0,256	-0,263
1,314	2,215	2,272
0,414	0,698	0,716
-0,152	-0,256	-0,263
-1,191	-2,007	-2,061
0,171	0,288	0,295
-0,592	-0,998	-1,024
0,611	1,030	1,056
0,181	0,305	0,313
-0,249	-0,420	-0,431
0,976	1,645	1,690
-0,324	-0,546	-0,560
-0,209	-0,352	-0,361
-0,674	-1,136	-1,166
0,241	0,406	0,417
-0,539	-0,908	-0,931
0,441	0,743	0,763
-0,159	-0,268	-0,275
-0,139	-0,234	-0,241

Tabelle 8: Standardisierte und studentisierte Residuen der Kraftstoffdaten

3.1.2 Formen der Residualanalyse im linearen Modell

Die Residualanalyse ermöglicht eine Untersuchung der Modellannahmen bezogen auf die Störgröße durch graphische Methoden. Besondere Beachtung wird dabei dem Normalverteilungs-Plot, dem Plot von Residuen gegen $\hat{y}_i$ und dem Plot von Residuen gegen x_i, geschenkt.

Normalverteilungs-Plot

Obwohl kleine Abweichungen von der Normalverteilung das Modell nicht so stark beeinflussen, sind Abweichungen der Normalverteilung wesentlich ernstzunehmender als die T- oder F-Statistiken, denn Konfidenz- und Vorhersageintervalle hängen von der Normalverteilungsannahme ab. Außerdem können die kleinste-Quadrate-Schätzer auf eine kleine Teilmenge der Daten empfindlich reagieren, wenn die Fehler von einer Verteilung mit dickerem / größerem Rest als bei der Normalverteilung herrühren.
Eine einfache Methode um die Normalverteilungsannahme zu überprüfen besteht darin, die Residuen auf Normalwahrscheinlichkeitspapier zu plotten. Dieses Papier ist so formatiert, dass die kumulative Normalverteilung als Gerade geplottet wird.

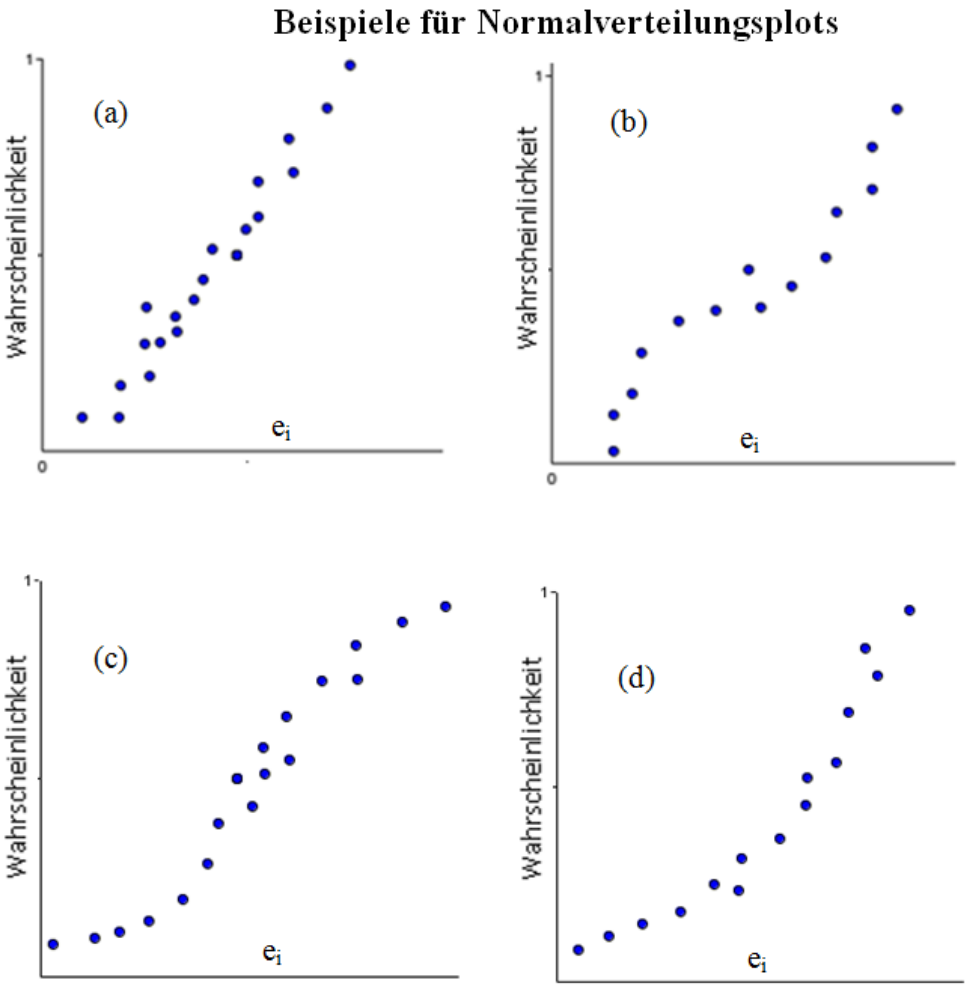

Abbildung 11: Beispiele für Normalverteilungsplots: (a) ideal; (b) "heavy tailed" Verteilung; (c) „light-tailed" Verteilung; (d) positive Schiefe

Angenommen $e_1 < e_2 < ... < e_n$ seien die Residuen, die in aufsteigender Form geordnet sind. Wenn nun e_i gegen die kumulative Wahrscheinlichkeit (bzw. erwarteter Normalverteilungswert) $P_i = (i - 1/2) / n$ auf Normalverteilungspapier *geplotet* wird, so sollten die Punkte näherungsweise auf einer Geraden liegen. Das folgt aus der Tatsache, dass $E(e_i) = \Phi^{-1}[(i - 1/2) / n]$ angenommen wird. Beträchtliche Abweichungen von einer Gerade indizieren, dass die Verteilung nicht *normal* ist.

Abbildung 11 a) zeigt einen „idealisierten" Normalverteilungsplot bei dem die Punkte annähernd um eine Gerade streuen. Die Darstellungen b) – d) zeigen andere typische Probleme. So sieht man in Beispiel b) eine stark ansteigende Kurve, die sich dann abflacht, ehe sie wieder stärker steigt, was indiziert, dass die Enden dieser Verteilung zu heftig abweichen, um als Normalverteilung klassifiziert zu werden. Umgekehrt zeigt c) eine Abflachung an den Enden, eine typische Sorge bei Beispielen mit Verteilungen, die dünnere Enden haben, als die normale. Das Studieren derartiger *Plots*, trägt insgesamt dazu bei, ein Gefühl dafür zu bekommen, wie viel Abweichung von der Geraden akzeptabel ist. Außerdem kann angemerkt werden, dass Normalverteilungsplots oft gar kein ungewöhnliches Verhalten zeigen, sogar dann, wenn die Fehler ε_i nicht normalverteilt sind. Dieses Problem entsteht, weil die Residuen keine einfache

Zufallsstichprobe sind; vielmehr sind sie ein Überbleibsel eines Parameter Schätzprozesses, die sich als Linearkombinationen der Modellfehler ε_i darstellen lassen. Deshalb führt ein Anpassen der Parameter zur Vernichtung der Beweislage für *Nichtnormalität* in den Residuen und folglich können wir uns nicht immer auf Normalverteilungsplots stützen, um Abweichungen von der Normalverteilung aufzudecken. Ein gängiger Defekt, welcher in Normalverteilungsplots aufgezeigt wird, ist das Auftreten von ein bis zwei großen Residuen, die manchmal ein Indiz dafür sind, dass die korrespondierenden Beobachtungen Ausreißer sind. (siehe Abschnitt 3.2).

3.1.3 Plot von Residuen gegen $\hat{y}_i$

Ein Plot der Residuen e_i (oder der skalierten Residuen d_i oder r_i) versus der korrespondierenden beobachteten Werte $\hat{y}_i$, ist dazu nützlich, um einige gängige Typen von Modellunangemessenheiten aufzudecken. Wenn ein Plot der unten dargestellten Abbildung a) ähnelt, was indiziert, dass die Residuen um ein horizontales Band streuen, dann liegen keine offensichtlichen Modelldefekte vor. Plots von e_i versus $\hat{y}_i$ die einer der Musterdarstellungen b) - d) ähneln, sind symptomatisch für Modelldefizite.

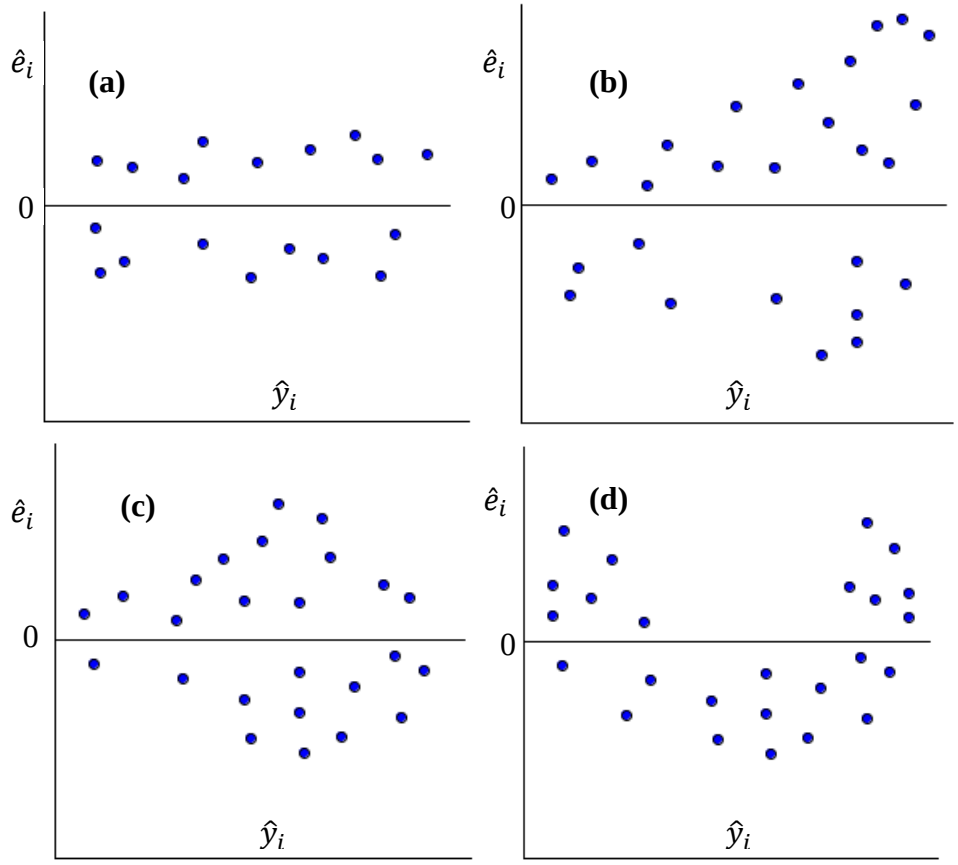

Abbildung 12: Muster für Residuenplots

Die Muster in den Darstellungen b) und c) indizieren, dass die Varianz der Fehler nicht konstant ist. Das nach außen offene Trichtermuster in b) sagt aus, dass die Varianz eine ansteigende Funktion von y ist. Die Darstellung c) tritt oft auf, wenn y ein Maß zwischen null und eins ist. Die varianz der Binomialverteilung nahe 0,5 ist größer als eine nahe 0 oder 1. Die gewöhnliche Annäherung an derartige Varianz Ungleichmäßigkeiten erfolgt durch die Wahl einer geeigneten Transformation für die

abhängige oder unabhängige Variable bzw. durch die Methode der Gewichtung der kleinsten Quadrate. Ein Kurvenplot wie in d) ist ein Indiz für Nichtlinearität, was bedeuten kann, dass andere Regressorvariablen im Modell zusätzlich gebraucht werden. Ein Plot von Residuen gegen $\hat{y}_i$ kann auch das eine oder andere ungewöhnlich große Residuum enthüllen. Diese Punkte sind natürlich potentielle Ausreißer. Große Residuen die als extreme $\hat{y}_i$ Werte auftreten könnten auch bedeuten, dass entweder die varianz nicht konstant ist oder die wahre Beziehung zwischen y und x nicht linear ist. Diese Möglichkeiten sollten untersucht werden, bevor man Punkte als Ausreißer betrachtet.

3.1.4 Plot von Residuen gegen x_i

Die Residuen gegen die korrespondierenden Werte der Regressorvariable zu *ploten* ist auch hilfreich, weil diese Plots oft Muster wie jene in der Abbildung oben aufweisen, mit dem Unterschied, dass die horizontale Skalierung nicht $\hat{y}_i$ ist, sondern x_i. Wiederum ist der Anblick eines horizontalen Bandes, um das die Residuen streuen wünschenswert.

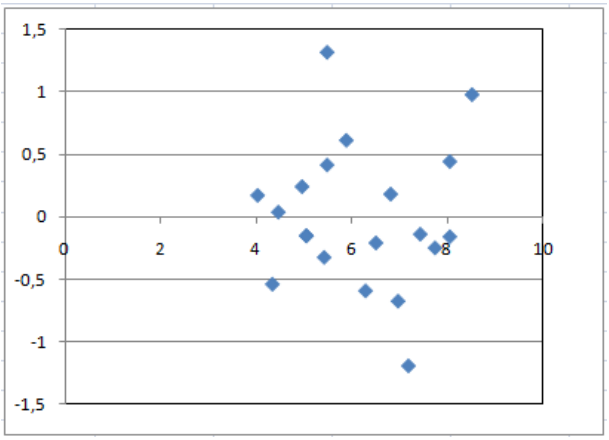

Abbildung 13: Plot der Residuen e_i gegen die geschätzten $\hat{y}_i$

3.1.5 Andere Residuenplots

Zusätzlich zu diesen standardmäßigen Residuenplots gibt es einige andere die gelegentlich sinnvoll sein können. Wenn die Zeitspanne zum Beispiel bekannt ist, in der die Daten gesammelt wurden, kann es sinnvoll sein, die Residuen und verschiedenen Zeitpunkte in einem Koordinatensystem zusammen zu *ploten*. Sofern die entstehende Punktwolke ähnlich zu einem Muster oben ist, ist das ein Indiz dafür, dass sich die Varianz mit der Zeit ändert oder dass mit der Zeit lineare oder quadratische Terme mit der Zeit zum Modell hinzugefügt werden. Dieser Zeitsequenz-Plot der Residuen kann indizieren, dass die Fehler einer Zeitperiode mit Fehlern anderer Zeitperioden korrelieren. Die Korrelation zwischen Modellfehlern zu unterschiedlichen Zeitperioden nennt man Autokorrelation.

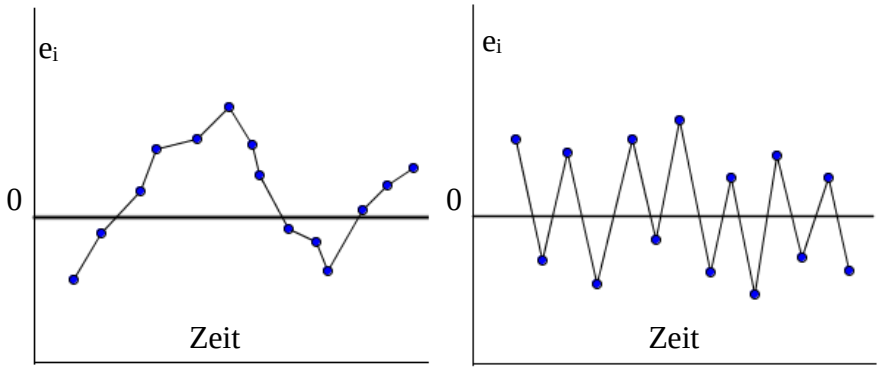

Abbildung 14: ein Prototyp von Residuenplots gegen die Zeit zeigt Autokorrelation in den Fehlern a) positive Autokorrelation; b) negative Autokorrelation

Die Präsenz von Autokorrelation in den Fehlern ist ein ernstzunehmendes Vergehen gegen die Basisregressionsannahmen.
Außerdem können manchmal Modellunangemessenheiten aufgedeckt werden, indem man Residuen gegen irgendwelche weggelassenen Regressoren *plotet*. Natürlich ist ein derartiger Plot nur möglich, wenn die Ebenen der weggelassenen Regressoren bekannt sind. Irgendein systematisches Muster, dass sich dadurch ergibt, indiziert, dass das Modell durch Hinzufügen des neuen Regressors verbessert werden kann.

3.2 Erkennung bzw. Umgang mit Ausreißern

Datenanalysen sollten nach Möglichkeit, neben der Überprüfung der Modellannahmen, die Erkennung sowie den Umgang mit extremen/ weit entlegenen Punkten, sogenannten Ausreißern sowie die Suche nach deren Ursachen umschließen. Residuen die größere absolute Werte als die anderen haben, sagen wir drei oder vier Standardabweichungen vom Mittelwert, sind potentielle Ausreißer. Abhängig vom x-Wert, können Ausreißer moderate bis sehr ernstzunehmende Effekte auf das Regressionsmodell haben. Residuenplots gegen $\hat{y}_i$ und der Normalverteilungsplot sind hilfreich zum identifizieren von Ausreißern. Sie sollten sorgfältig untersucht werden, um einen eventuellen Grund für ihr ungewöhnliches Verhalten zu finden. Manchmal sind Ausreißer „schlechte“ Werte, die als Resultat ungewöhnlicher aber erklärbarer Ereignisse auftreten. Beispiele können mangelhafte Messungen oder Analysis, eine inkorrekte Datenerhebung und Fehler des Messinstrumentes sein. Wenn das der Fall sein sollte, dann ist es angebracht, den Ausreißer (wenn möglich) zu korrigieren oder aus dem Datensatz zu löschen. Klarerweise ist es wünschenswert *schlechte* Werte sofort zu verwerfen, weil die Kleinsten Quadrate die angepasste Gleichung verfälschen können, so wie wenn sie die Quadratsumme der Residuen minimiert. Bei der einfachen linearen Regression kann man diese Punkte durch betrachten des Streudiagramms der Wertepaare (x_i, y_i) aufdecken. Allerdings nehmen wir an, dass ein strenger nichtstatistischer Beweis vorliegen sollte, dass der Ausreißer ein schlechter Wert ist, bevor man ihn degradiert.
In den nachfolgenden Abbildungen sieht man, dass x-Werte die abseits der anderen x-Werte liegen, relativ starken Einfluss auf das Regressionsmodell ausüben. In der Darstellung wurde die Regressionsgerade mit („strichlierte Linie“) und ohne die extremen Punkte („durchgezogene Linie“) eingezeichnet.

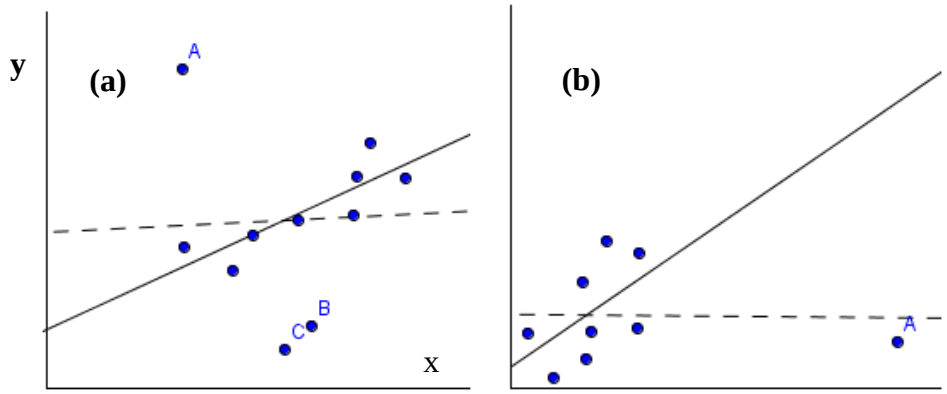

Abbildung 15: a) β_1 hängt stark von einem oder beiden Punkten A,B und C ab und die übrigen Datenpunkte würden eine andere Schätzung ergeben, wenn diese Punkte entfernt würden. b) β_1 wird zum Großteil vom extremen Punkt A bestimmt; durch Weglassen dieses Punktes würde β_1 womöglich null sein.

Derartige Situationen treten in der Praxis häufig auf und aus den Abbildungen ist zu erkennen, dass wir im Wesentlichen zwei Arten (siehe Kapitel 3.1) von Ausreißérn unterscheiden:

- Ausreißer in y-Richtung (Abbildung a)
- Ausreißer in x-Richtung

Natürlich kann ein Punkt auch beides erfüllen, allerdings trifft diese Unterteilung der Ausreißer in x- bzw. y-Richtung nur für die einfache lineare Regression zu.
Für die lineare Mehrfachregression ist es hingegen nicht mehr so einfach, Ausreißer durch die graphische Veranschaulichung der Datenpunkte der abhängigen/unabhängigen Variablen zu erkennen, weshalb die Residuen sowie die Projektionsmatrix zur Analyse herangezogen werden. (siehe multiple Regression)

3.3 Test für den Mangel an Anpassung

Hier soll ein formaler statistischer Test für mangelnde Anpassung eines Regressionsmodells vorgestellt werden. Dieses Verfahren geht davon aus, dass die Normalitäts-. Unabhängigkeits- und konstanten Varianzannahmen erfüllt sind und nur der *„first order“* bzw. der geradlinige Charakter der Beziehung angezweifelt werden. Betrachten wir zum Beispiel die Datenpunkte der Abbildung unten, so sprechen einige Indizien dafür, dass diese geradlinige Regressionsgerade nicht zufriedenstellend ist und es hilfreich sein könnte, ein Testverfahren anzuwenden, welches auf systematische Anpassungsfehler der linearen Regression aufmerksam macht.
Der *Test auf Anpassungsmangel* erfordert, dass man die Anpassung von y für eine Auswahl von x wiederholen sollte, wobei zu betonen ist, dass diese wiederholten Anpassungen tatsächliche Abgleichungen der Messungen sein sollten und nicht nur Duplikate der Ergebnisse. Angenommen in einem Beispiel sei y die Viskosität und x die Temperatur, so besteht ein korrekter Abgleich im Durchlaufen von n_i separaten Experimenten bei $x = x_i$ und angepasster Viskosität und nicht darin, indem man ein einziges Experiment bei x_i durchlaufen lässt und die Viskosität n_i mal misst. Denn die beobachteten Angaben dieses Verfahrens liefern nur Information für die Veränderlichkeit der Methode beim Messen der Viskosität. Die Fehlervarianz σ^2

inkludiert diesen Messfehler und die Variabilität die verbunden wird mit dem Erreichen bzw. Beibehalten des gleichen Temperaturlevels in unterschiedlichen Experimenten.

Diese aufgezählten Punkte werden nun verwendet, um eine modellunabhängige Schätzung von σ^2 zu erhalten. Angenommen wir haben n_i Beobachtungen als Reaktion der i-ten Stichprobe x_i mit i =1,..,m. Wir bezeichnen nun mit y_{ij} die j-te Beobachtung als Reaktion auf x_i mit j = 1,... n_i. Somit gibt es n =$\sum_{i=1}^{n} n_i$ Beobachtungen insgesamt. Das Testverfahren involviert wieder eine Untergliederung der Quadratumme von Residuen in zwei Komponenten

$$\sum_{i=1}^{n} e_i^2 = \mathbf{SS_{PE}} + \mathbf{SS_{LOF}}$$

wobei SS_{PE} die Quadratsumme des reinen Fehlers („pure error“) und SS_{LOF} die Quadratsumme des Anpassungsmangels („lack of fit“) *(siehe auch Kapitel 1*).

Zur Entwicklung dieser Partitionierung von $\sum e_i^2$ wird angemerkt, dass das ij – te Residuum jenes ist:

$$y_{ij} - \hat{y}_i = \left(y_{ij} - \bar{y}_i\right) + (\bar{y}_i - \hat{y}_i)$$

wo $\bar{y}_i$ der Durchschnitt der n_i Beobachtungen bei x_i ist. Das Quadrieren beider Seiten dieser Gleichung und die Summation über i und j führt zu

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(y_{ij} - \hat{y}_i\right)^2 = \sum_{\mathbf{i=1}}^{\mathbf{m}}\sum_{\mathbf{j=1}}^{\mathbf{n_i}}\left(\mathbf{y_{ij}} - \bar{\mathbf{y}}_\mathbf{i}\right)^\mathbf{2} + \sum_{\mathbf{i=1}}^{\mathbf{m}}\mathbf{n_i}(\bar{\mathbf{y}}_\mathbf{i} - \hat{\mathbf{y}}_\mathbf{i})^2$$

nachdem der Kreuzprodukt-Term gleich null ist.

Die linke Seite der Gleichung misst wiederum die gewöhnliche Quadratsumme der Residuen und die zwei Komponenten der rechten Seite messen den „pure error“ und den „lack of fit“. Man sieht, dass die reine Fehlerquadratsumme beobachtet werden kann, indem man die korrigierte Quadratsumme der wiederholten Beobachtungen auf jedem Level von x berechnet und dann über die m - Levels von x zusammenfasst.

Die Quadratsumme für den Anpassungsmangel ist dann eine gewichtete Quadratsumme von Abweichungen zwischen dem mittleren beobachteten Wert $\bar{y}_i$ bei jedem x Level und dem korrespondierenden, beobachteten Wert. Wenn die angepassten $\hat{y}_i$ Werte nahe den korrespondierenden durchschnittlichen $\bar{y}_i$ Werten liegen, so ist das ein signifikantes Indiz dafür, dass die Regressionsfunktion linear ist. Folgende Teststatistik lässt sich dadurch bilden

$$F_0 = \frac{SS_{LOF}/(m-2)}{SS_{PE}(n-m)} = \frac{MS_{LOF}}{MS_{PE}} \qquad (52)$$

Beispiel

x	1,0	2,0	3,3	3,3	4,0	4,0	4,0	4,7	5,0
y	10,84	16,35	22,88	24,35	24,56	25,46	29,16	24,59	22,25
x	5,6	5,6	5,6	6,0	6,0	6,5	6,9	1,0	
y	25,9	27,2	25,61	25,45	26,56	21,03	21,46	9,30	

Die angepasste Gerade ist $\hat{y} = 13{,}301 + 2{,}108\,x$ mit S_{yy} = 487,613, SS_R=234,71 und $\sum e_i^2 = 252{,}90$. Zusätzlich kann angemerkt werden, dass 10 individuelle Levels von x

vorkommen, mit Wiederholungspunkten bei x = 1,0; x = 3,3; x= 4,0; x = 5,6 und x = 6,0. Die reine Fehlerquadratsumme wird berechnet, indem die wiederholten Punkte wie folgt verwendet werden

Level von x	$\sum_j (y_{ij} - \bar{y}_i)^2$	Freiheits-grade
1,0	1,186	1
3,3	1,08	1
4,0	11,247	2
5,6	1,434	2
6,0	0,616	1
Total	15,563	7

Varianzanalyse (ANOVA) für dieses Beispiel

	Quadrat summe	Freiheits-grade	Mittlere Quad. Abweichung	F_0
Regression	1,186	1	234,789	
Residuum	1,08	1	16,860	
„lack of fit“	11,247	2	29,668	13,34
„pure error“	1,434	2	2,223	
Total	0,616	1		

Tabelle 9 a,b und c: Varianzanalyse

→ $SS_{LOF} = \sum_{i=1}^{n} e_i^2 - SS_{PE} = 252{,}9 - 15{,}56 = 237{,}34$ mit 10 – 2 = 8 Freiheitsgraden.

Dieser Test für den Mangel an Anpassung hat eine F - Statistik von 13,34 und nachdem $F_{0,25,8,7} = 1{,}7$ ist, verwerfen wir die Hypothese, dass das Modell die Daten adäquat beschreibt.

Kapitel 4

4. MULTIPLE LINEARE REGRESSION

Bislang wurde immer die lineare Abhängigkeit zweier Variablen behandelt, doch viele praktische Anwendungen erfordern die simultane Berücksichtigung von mehr als nur einer unabhängigen Variablen. Soll nun also der Erwartungswert einer Zielgröße Y als lineare Funktion mehrerer Einflussgrößen x_1, x_2 … x_k beschrieben werden, so kommt die multiple bzw. mehrfache lineare Regression zur Anwendung, die eine Verallgemeinerung der einfachen linearen Regression darstellt.

Sind nun $x_1, \ldots, x_k$ mit $k \geq 2$ die Regressoren bzw. Einflussgrößen und Y die Zielgröße, so vermutet man einen linearen Zusammenhang zwischen den Regressoren (X_i) und dem Regressand (Y) und legt folgendes Modell zugrunde:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

mit für gewöhnlich unbekannten Parametern $\beta_0, \beta_1, \ldots, \beta_k$ und dem Einfluss des Fehlerterms ε_i, wobei für ε_i Erwartungswert 0 und Kovarianzmatrix $\sigma^2 I$ vorausgesetzt wird. Außerdem müssen die vorliegenden Gleichungen in den β_j linear sein. Dank der Matrixschreibweise kann das multiple Regressionsmodell sehr kompakt präsentiert werden, indem man eine Stichprobenerhebung vom Umfang n, mit den Werten der unabhängigen Merkmale X und des abhängigen Merkmals Y heran zieht:

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon} \qquad (\mathbf{54})$$

Dabei enthält der n-Vektor y die Beobachtungen des abhängigen Merkmals, während die [n x (k + 1)] – Matrix bzw. auch *Designmatrix*, die Beobachtungen der k unabhängigen Merkmale charakterisieren und als erste Spalte einen Vektor mit lauter Einsen als Multiplikatoren für das Interzept beinhaltet. Der (k+1) -Vektor β enthält die Regressionskoeffizienten und der n -Vektor ε die Störgrößen der Beobachtungen:

$$X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Voraussetzung für die Designmatrix ist, die Beobachtungen an den Punkten ($x_{i1}, \ldots x_{ik}$) zu betrachten, um zu garantieren, dass die Designmatrix vollen Rang hat. Denn andernfalls kann es passieren, dass die Schätzer der Regressionskoeffizienten nicht eindeutig sind, weil die x_{ij} nicht zufällige Größen sein müssen. Unter Heranziehung des Vektors $x_i = (1, x_{i1}, \ldots x_{ik})'$ kann die Modellgleichung der i-ten Beobachtung auch wie folgt angeschrieben werden:

$$Y_i = x_i'\beta + \varepsilon_i.$$

Beispiel:

Es besteht die Vermutung, dass der Umsatz der Filialen einer Supermarktkette neben der Verkaufsfläche auch vom durchschnittlichen Einkommen der Haushalte im Einzugsbereich der Filiale bestimmt wird. Die Tabelle zeigt den entsprechenden Datensatz, der bearbeitet werden soll:

Aus dieser Tabelle wurde zum einen der Korrelationskoeffizient zwischen Einkommen und Umsatz berechnet, der mit 0,222 bedeutend kleiner ist, als jener zwischen Umsatz und Verkaufsfläche (0,969). Dadurch gibt es kein markantes Indiz auf einen möglichen Erklärungsbeitrag des Einkommens für den Umsatz. Trotzdem weist das Streudiagramm der Residuen gegen die Variable Einkommen einen Korrelationskoeffizient von 0,99 auf, weshalb es sich anbietet, das Modell für den „Umsatz“ um die

Filiale	Umsatz	Fläche	Einkommen
1	7,48	157	169,9
2	2,19	109	153,6
3	13,6	279	156,5
4	3,25	120	141,1
5	6,7	172	144,4
6	8,87	196	139,4
7	4,51	120	155,3
8	11,04	247	153,5
9	8,81	231	130,7
10	4,24	128	154,6
11	12,25	259	155,2
12	4,92	116	162
13	6,87	189	149,7
14	11,44	242	144,8

Tabelle 10: Vergleich von Korrelationskoeffizienten

4.1 Geometrischer Zugang zur multiplen linearen Regression

Ausgehend von $y = X\beta + \varepsilon$, soll der Fehlerterm ε und somit $|| y - X\beta||^2$ minimiert werden um dementsprechend $y \approx X\beta$ zu erhalten. Dazu sei V ein linearer Unterraum von R^n der Dimension $d < n$ und die orthogonale Projektion eines Punktes $x \in R^n$ auf V eine lineare Abbildung $x \rightarrow Px$ (P ist eine n x n Matrix).

Satz: Die Vektoren $v_1, v_2, \ldots, v_d$ bilden eine Basis von V und X ist jene (n x d)-Matrix mit den Spalten $v_1, \ldots, v_d$. Sei P eine (n x n)-Projektionsmatrix definiert durch: $P = X (X' X)^{-1} X'$, dann gilt:

(a) $Px \in V \qquad \forall\, x \in R^n$ (55)

(b) $x - Px \perp V \qquad \forall\, x \in R^n$ (d.h P ist die orthogonale Projektion auf V)

Beweis: zuerst wird die Invertierbarkeit der (d x d)-Matrix $X^T X$ gezeigt:

$$Xy = \sum_{j=1}^{d} y_j v_j \neq 0 \text{ mit } y \in R^d \backslash \{0\}, \text{weil die Spalten } v_j \text{ von X linear unabhängig sind}$$

$$\Rightarrow \qquad \langle y, X^T X y \rangle = \langle Xy, Xy \rangle = ||Xy||^2 > 0$$

Ist die Invertierbarkeit für X´X nicht gegeben, so würde ein Vektor $y \in R^d \backslash \{0\}$ existieren mit $X^T X y = 0$ und somit führt $\langle y, X^T X y \rangle = 0$ zu einem Widerspruch.

Ist nun $x \in R^n$ und $y = (X'X)^{-1} X^T \in R^d$, so ist

$$Px = Xy = \sum_{j=1}^{d} y_j v_j \in V. \qquad (\mathbf{a})$$

Ist $x \in R^n$ und $w \in V$, so ist $y_1, \ldots, y_d$ mit $w = \sum_{j=1}^{d} y_j v_j = Xy$.

$$\Rightarrow \qquad \langle x - Px, w \rangle = \langle x - Px, Xy \rangle = \langle X^T x - X^T Px, y \rangle$$

$$= \langle x - X^T X (X^T X)^{-1} X^T x,\ y\rangle = \langle X^T x - X^T x, y \rangle = 0 \qquad \textbf{(b)}$$

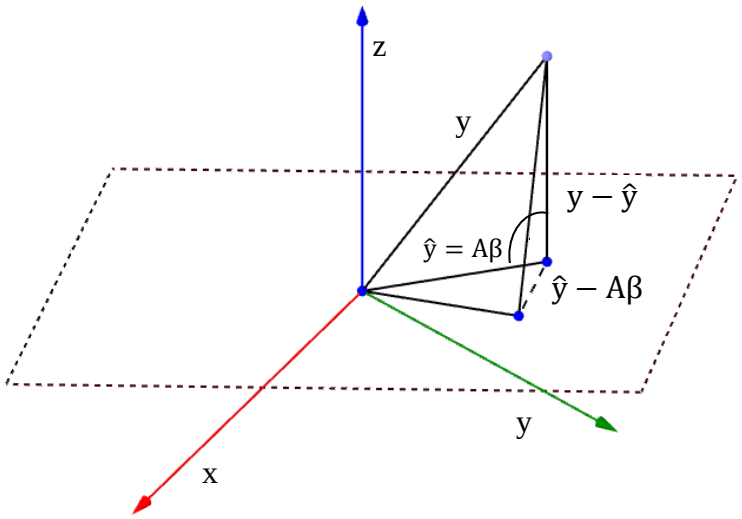

Abbildung 16: Kleinste Quadrate Schätzung durch Orthogonalprojektion

Nachdem $y = X\beta + \varepsilon$ zutrifft und ε der Fehlerterm ist, soll möglichst $y \approx X\beta$ erreicht werden, indem man den Fehler bzw. $||y - X\beta||^2$ minimiert.

Die Ebene V_2 der Skizze sei der von den Spalten von X aufgespannte Unterraum vom R^n und P_2 die Projektion auf den Vektorraum V_2. Dadurch gilt, dass $P_2 y - AX \in V_2$, weil sowohl $X\beta \in V_2$ als auch $P_2 y \in V_2$. Aus Satz 4.1 kann gefolgert werden, dass $\langle y - P_2 y, P_2 y - X\beta\rangle = 0$ zutrifft und somit ist:

$$||y - X\beta||^2 = ||y - P_2 y + P_2 y - X\beta||^2 = ||y - P_2 y||^2 + ||P_2 y - X\beta||^2$$

$$\Rightarrow \quad ||y - X\beta||^2 \text{ minimal} \quad \Leftrightarrow \quad X\beta = P_2 y$$

$$\Leftrightarrow \quad X\beta = X(X^T X)^{-1} X^T y \quad | \cdot (X^T X)^{-1} X^T \text{ von links}$$

$$\Leftrightarrow \quad \beta = (X^T X)^{-1} X^T y \quad (= \text{Schätzer für } \beta = \hat{\beta})$$

Denn für die Schätzung der Regressionskoeffizienten β zieht man im Prinzip wieder die Kleinste-Quadrate Methode heran und durch Ableiten der Summe der quadrierten Abweichungen, $S(\beta) = \varepsilon'\varepsilon = (y - X\beta)(y - X\beta)$ nach β ergeben sich die Normalgleichungen $(\mathbf{X'X})\,\beta = \mathbf{X'}\,y$.

Die dabei auftretende symmetrische Matrix X´X ist:

$$X^TX = \begin{pmatrix} n & \sum_i x_{i1} & \sum_i x_{i2} & \cdots & \sum_i x_{ik} \\ \sum_i x_{i1} & \sum_i x_{i1}^2 & \sum_i x_{i1}x_{i2} & \cdots & \sum_i x_{i1}x_{ik} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_i x_{ik} & \sum_i x_{ik}x_{i1} & \sum_i x_{ik}x_{i2} & \cdots & \sum_i x_{ik}^2 \end{pmatrix}$$

und man nennt sie die Matrix der Summe und Kreuzprodukte. Daraus können nun die Kleinste-Quadrate Schätzer für β als Lösung der Normalgleichungen gefolgert werden:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X^TX})^{-1}\,\mathbf{X^T}y. \tag{56}$$

Somit lautet die empirische Regressionsfunktion

$$\hat{\mathbf{y}} = X\hat{\beta} = X\,(X^TX)^{-1}\,X^Ty \qquad \text{mit} \quad \hat{y} = \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{pmatrix}.$$

Beispiel: Für die lineare Einfachregression mit k = 1 und $X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$ ergibt sich für

$X^TX\hat{\beta} = X^Ty$ das Gleichungssystem für die Normalgleichungen der Einfachregression:

$$\begin{pmatrix} \sum_{i=1}^n x_i^2 & n\bar{x} \\ n\bar{x} & n \end{pmatrix}\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n x_i y_i \\ n\bar{y} \end{pmatrix}$$

4.2 Modellspezifikation und Eigenschaften

Es wird eine Stichprobe vom Umfang n vorausgesetzt, wobei als i-te Beobachtung i=1,…,n neben dem Wert y_i der abhängigen Variablen die erklärenden Variablen $x_{i1},\ldots x_{ik}$ beobachtet wurden. Außerdem nimmt man an, dass die x_{ij} keine zufälligen Größen sind und die Punkte $(x_{i1}, \ldots, x_{ik})$, i = 1, … , n der jeweiligen y_i geeignet angeordnet sind, um das Schätzproblem zu lösen. Für die Störgrößen ε_i werden dieselben Bedingungen wie beim einfachen, linearen Regressionsmodell vorausgesetzt:

- $E(\varepsilon_i) = 0$.
- $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2 & i = j \\ 0 & i \neq j \end{cases}$
- $\varepsilon_i \sim N(0, \sigma^2)$
- Unkorreliertheit

Zudem kann man auch die stochastischen Eigenschaften der ε_i in Matrixschreibweise angeben: $\varepsilon \sim N(0, \sigma^2 I)$ mit:

$$\boldsymbol{\sigma^2}\ \mathrm{I} = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix}$$

die als Varianzen der ε_i in der Hauptdiagonale σ^2 und wegen der Unkorreliertheit der Störgrößen als Nichthauptdiagonalelemente Nullen hat. Die Matrixschreibweise (mit k=1) unterscheidet sich also in keiner Weise vom einfachen linearen Modell.

Für die Existenz der Eindeutigkeit der Lösung, muss die Voraussetzung r(X) = k + 1 erfüllt sein, womit die (k+1)x(k+1) – Matrix (X´X) dann den Rang k + 1 besitzt und invertierbar ist.

Die $\hat{\beta}_i$ sind nicht unkorreliert, weil X´X und somit $(X´X)^{-1}$ keine Diagonalmatrix ist. (Ausnahme: Spalten von X sind orthogonale Vektoren)

Ein erwartungstreuer Schätzer für die Varianz der Störgrößen ist gegeben durch:

$$\hat{\sigma}^2 = \frac{\hat{\varepsilon}´\hat{\varepsilon}}{n-(k+1)}$$

und der n-te Vektor $\hat{\varepsilon}$ enthält die Residuen $\hat{\varepsilon} = y - \overbrace{X\hat{\beta}}^{\hat{y}}$ und damit die Differenzen zwischen beobachteten Werten der y_i und den prognostizierten Werten $x_i´\hat{\beta}$, die aus dem geschätzten Modell folgen mit:

$$\frac{\hat{\varepsilon}´\hat{\varepsilon}}{\sigma^2} \sim \chi^2(n-k-1)$$

Im nächsten Schritt werden nun die Eigenschaften der Kleinste-Quadrate Schätzer $\widehat{\beta}$ in Matrixform untersucht:

Definition : Die Kovarianzmatrix eines beliebigen Zufallsvektors Z ist definiert durch (57)

$$\sum_Z (\mathrm{Cov}(Z_i, Z_j))_{1 \le i,j \le n} \text{ mit:}$$

$$\mathrm{Cov}(Z_i,Z_i) = \mathrm{E}((Z_i - \mathrm{E}(Z_i))^2) = \mathrm{Var}(X_i)$$

$$\mathrm{Cov}(Z_i, Z_j) = \mathrm{E}[(Z_i - \mathrm{E}(Z_i))(Z_j - \mathrm{E}(Z_j))]$$

Satz: Ist die Zufallsvariable $\hat{\beta} = (X^T X)^{-1} X^T Y$ ist eine linear erwartungstreue Schätzfunktion für den Spaltenvektor β dann gilt $\mathrm{E}(\hat{\beta}) = \beta$. (58)

Beweis:

$$\mathrm{E}(\hat{\beta}) = \mathrm{E}[(X^T X)^{-1} X^T y] = \mathrm{E}[(X^T X)^{-1} X^T (X\beta + \varepsilon)]$$

$$= E[\beta + (X^T X)^{-1} X^T \varepsilon] = \beta \qquad \text{wobei } E(\varepsilon) = 0$$

Satz : Die Kovarianzmatrix der erwartungstreuen Schätzfunktion für $\hat{\beta}$ ist

$\text{Cov}(\hat{\beta}) = \sum_{\hat{\beta}} = \left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T\right]$.

Beweis:

$$\sum_{\hat{\beta}} = E\{[(X^T X)^{-1} X^T \varepsilon]\,[(X' X)^{-1} X\,\varepsilon]\}$$

$$= E\,[\underbrace{(X^T X)^{-1} X^T}_{I} X \varepsilon\varepsilon' (X^T X)^{-1}] = \sigma^2 (X^T X)^{-1}$$

Seien die Varianzen der geschätzten Zufallsvariablen $\hat{\beta}_j$ per Konvention die Elemente der Hauptdiagonale von $\sum_{\hat{\beta}}$, wobei c_{jj} das j-te Diagonalelement von $(X'X)^{-1}$ ist, so gilt $\text{Var}(\hat{\beta}_j) = c_{jj}\sigma^2$.

Die Herleitung der Varianzen der geschätzten Regressionskoeffizienten für die lineare Einfachregression lässt sich nun wie folgt zeigen (vgl mit Kapitel 2):

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad \Rightarrow \quad X'X = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum x_i^2 \end{pmatrix} \quad \Rightarrow \quad \det(X^T X) = n \sum_{i=1}^{n} x_i^2 - n^2 \bar{x}^2$$

$$\Rightarrow \quad (X'X)^{-1} = \frac{1}{n \sum_{i=1}^{n} x_i^2 - n^2 \bar{x}^2} \begin{pmatrix} \sum x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix} = \text{Inv}\,(X'X)$$

$$\Rightarrow \qquad \text{Var}(\hat{\beta}_0) = \sigma^2 c_{00} = \frac{\sigma^2 \sum x_i^2}{n \sum x_i^2 - n^2 \bar{x}^2} = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum x_i^2 - n\bar{x}^2}\right)$$

$$\text{Var}(\hat{\beta}_1) = \hat{\sigma}^2 c_{11} = \frac{\sigma^2}{\sum x_i^2 - n\bar{x}^2}$$

Eigenschaften von P und Q = I − P

Nun werden einige Eigenschaften der (n x n)-Projektionsmatrix $P = X(X' X)^{-1} X'$ angewendet (mit λ_i gleich i-ter Eigenwert): (59)

Symmetrie von P: $P^T = [X(X' X)^{-1} X']' = X(X' X)^{-1} X' = P$

Idempotenz von P: $P^2 = X(X' X)^{-1}(X' X)(X' X)^{-1} X' = X(X' X)^{-1} X' = P$

Spur von P: $sp(P) = sp(X(X' X)^{-1} X') = sp(X'X(X' X)^{-1})$

$= sp(I_{k+1}) = \sum \lambda_i = k + 1$ (Summe d. Eigenwerte)

Dabei treten die Eigenwerte 1 (k+1 -fach) und 0 (n-k-1 -fach) auf, weil der Eigenraum zum Eigenwert 1 der von den Spalten von A aufgespannte Raum ist, und der Eigenraum von 0 dazu orthogonal liegt.

Für Q = I – P gilt, dass aus $\hat{Y} = PY$ dann $Y - \hat{Y} = (I - P)Y = QY$ folgt, mit den Eigenschaften a) Q´ = Q bzw. b) $Q^2 = Q$

Wiederum gilt, dass 0 ein (k+1) - facher Eigenwert von Q und 1 ein (n-k-1)- facher Eigenwert ist, weil aus Px=x folgt, so dass Q x = 0 und P x = 0 ist mit Q x = x.

4.3 Hypothesentests bei multipler linearer Regression

Bei multiplen Regressionsproblemen sind gewisse Tests von Hypothesen über die Modellparameter sinnvoll, um die Eignung des Modells zu messen. In diesem Abschnitt werden nun einige wichtige Hypothesentest – Verfahren beschrieben. Wiederum wird hier die Normalverteilungsannahme der Fehler vorausgesetzt.

4.3.1 Test auf Signifikanz der Regression

Der Test auf Signifikanz der Regression ist ein Test um festzustellen, ob eine lineare Beziehung zwischen der abhängigen Variable y und irgendeiner der Regressorvariablen $x_1, x_2, \ldots, x_k$ vorliegt. Dafür geeignete Hypothesen sind:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

$$H_1 : \beta_j \neq 0 \quad \text{für mindestens ein j}$$

Die Ablehnung von H_0: $\beta_j = 0$ impliziert, dass zumindest einer der Regressoren $x_1, x_2, \ldots, x_k$ einen signifikanten Beitrag zum Modell leistet. Dieses Test-Verfahren ist eine Verallgemeinerung jener Methode, die bei der einfachen linearen Regression verwendet wurde. Die Gesamtsumme der Quadrate S_{yy} ($bzw. \sum_{i=1}^{n}(y_i - \bar{y})^2$) wird aufgeteilt in eine Summe von Quadraten die durch Regression erklärt wird: SS_R (bzw. $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$) und eine Rest-/Störgröße von Quadratsummen SS_E (bzw. $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$) zum Beispiel:

$$\Longrightarrow \quad \sigma_{yy} = \sigma_R + \sigma_E$$

Wenn nun H_0: $\beta_j = 0$ wahr ist, so ist $\sigma_R / \sigma^2 \sim \chi_k^2$ und die dabei auftretende Anzahl der Freiheitsgrade ist äquivalent zur Anzahl der Regressorvariablen im Modell.
Zudem kann gezeigt werden, dass $SS_E/\sigma^2 \sim \chi_{n-k-1}^2$ und dass SS_E und SS_R unabhängig sind.

Beim Testverfahren für H_0: $\beta_j = 0$ berechnet man

$$F_0 = \frac{\sigma_R/k}{\sigma_E/(n-k-1)} = \frac{\bar{\sigma}_R}{\bar{\sigma}_E} = \frac{\text{mittleres Quadrat von } \sigma_R}{\text{mittleres Quadrat von } \sigma_E}$$

und verwirft H_0 wenn $F_0 > F_{\alpha,\, k,\, n-k-1}$ ist. Für gewöhnlich wird dieser Prozess in einer Varianzanalysetabelle zusammengefasst.

Eine Formel für σ_R erhält man, indem man ausgeht von

$$\sigma_E = \mathbf{y^T y} - \hat{\boldsymbol{\beta}}^{\mathbf{T}}\mathbf{X^T y} \tag{60}$$

und anschließend die bereits bekannte Darstellung für S_{yy} verwendet:

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n} = \mathbf{y^T y} - \frac{(\sum_{i=1}^{n} y_i)^2}{n}$$

und dadurch erhält man für die Gleichung σ_E oben:

$$\sigma_E = \mathbf{y^T y} - \frac{(\sum_{i=1}^{n} y_i)^2}{n} - \left[\hat{\boldsymbol{\beta}}^{\mathbf{T}}\mathbf{X^T y} - \frac{(\sum_{i=1}^{n} y_i)^2}{n}\right] \quad \text{bzw.} \quad \sigma_E = \sigma_{yy} - \sigma_R$$

Deshalb ist die Quadratsumme der Regression:

$$\sigma_R = \hat{\beta}^T X^T y - \frac{(\sum_{i=1}^{n} y_i)^2}{n} \tag{61}$$

die Quadratsumme der Residuen:

$$\sigma_E = y^T y - \hat{\beta}^T X^T y \tag{62}$$

und die gesamte Quadratsumme:

$$\sigma_{yy} = y^T y - \frac{(\sum_{i=1}^{n} y_i)^2}{n} \tag{63}$$

Beispiel:
Ein Limonadenabfüller möchte die Servicerouten für die Getränkeautomaten in seinem Vertriebssystem analysieren. Darum ist er daran interessiert, die benötigte Zeit für eine Tagesroute (auffüllen, warten etc. der Automaten) vorherzusagen. Der für die Zeitkalkulierung verantwortliche Mitarbeiter nimmt an, dass die zwei wichtigsten Parameter für die Auslieferungszeit die Anzahl der aufzufüllenden Produkte und die zurückgelegte Distanz sind. Aus 25 gesammelten Stichproben bzw. realen Beobachtungen soll nun eine Schätzung der Modellparameter und anschließende Durchführung eines Signifikanztests erfolgen (siehe Tabelle 8: eines Getränkelieferanten; im Anhang)

Die $X^T X$ Matrix ist

$$\mathbf{X^T X} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 7 & 3 & \dots & 4 \\ 560 & 220 & \dots & 150 \end{pmatrix} \begin{pmatrix} 1 & 7 & 560 \\ 1 & 3 & 220 \\ \vdots & \vdots & \vdots \\ 1 & 4 & 150 \end{pmatrix}$$

$$= \begin{pmatrix} 25 & 219 & 10\,232 \\ 219 & 3.05 & 133\,899 \\ 10\,232 & 133\,899 & 6\,725\,688 \end{pmatrix}$$

und der $X^T y$ Vektor ist dann:

$$\mathbf{X^T y} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 7 & 3 & \dots & 4 \\ 560 & 220 & \dots & 150 \end{pmatrix} \begin{pmatrix} 16{,}68 \\ 11{,}50 \\ \vdots \\ 10{,}75 \end{pmatrix} = \begin{pmatrix} 559{,}6 \\ 7\,375{,}44 \\ 337\,072 \end{pmatrix}$$

Der kleinste Quadrate Schätzer von β ist gegeben durch $\widehat{\boldsymbol{\beta}} = (\mathbf{X^T X})^{-1}\mathbf{X^T y} = \begin{pmatrix} 2{,}341 \\ 1{,}616 \\ 0{,}014 \end{pmatrix}$

Und zudem erhält man durch Bestimmung von $\hat{y}_i$ und e_i für jede beobachtung y_i die angepasste Gerade $\hat{y} = 2{,}34 + 1{,}62x_1 + 0{,}014x_2$

Nun erfolgt der Test auf Signifikanz der Regression anhand dieses Beispiels. Angenommen wir wollen den Wert der gegebenen Regressorvariable „Distanz“ (x2) dahingehend beurteilen, ob die Regressor- „Fälle“ (x1) im Modell auftreten oder nicht.

$$\begin{aligned} \sigma_{yy} &= \mathbf{y^T y} - \frac{(\sum_{i=1}^{n} y_i)^2}{n} \\ &= 18\,310{,}63 - \frac{559{,}6^2}{25} \\ &= 5\,784{,}547 \end{aligned}$$

$$\begin{aligned} \sigma_R &= \widehat{\boldsymbol{\beta}}^{\mathbf{T}}\mathbf{X^T y} - \frac{(\sum_{i=1}^{n} y_i)^2}{n} \\ &= 18\,076{,}90 - \frac{(559{,}6)^2}{25} \\ &= 5\,550{,}6177 \end{aligned}$$

und dadurch

$$\begin{aligned} \sigma_E &= \sigma_{yy} - \sigma_R \\ &= \boldsymbol{y^T y} - \widehat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{X^T} \\ &= 233{,}726 \end{aligned}$$

Um nun H_0: $\beta_1 = \beta_2 = 0$ zu testenberechnet man die Statisitk:

$$F_0 = \frac{\bar{\sigma}_R}{\bar{\sigma}_E} = \frac{2\,775{,}41}{10{,}63} = 261{,}24$$

Nachdem $F_0 > F_{0{,}05,\,2,\,22} = 3{,}44$ ist, schließt man, dass sich die Auslieferungszeit auf das Auslieferungsvolumen und/ oder die –distanz bezieht. Allerdings impliziert das nicht zwingend, dass die gefundene Beziehung eine geeignete ist, um die Auslieferungszeit als Funktion von Volumen und Distanz anzugeben. Hier sind weitere Tests für die Aussagekraft des Modells nötig.

4.3.2 Tests an einzelnen Regressionskoeffizienten

Wir sind hauptsächlich daran interessiert, Hypothesentests bei einzelnen Regressionskoeffizienten anzuwenden. Diese Tests sind hilfreich um den Wert jedes Regressors im Modell zu ermitteln. So könnte das Modell zum Beispiel durch Inklusion zusätzlicher Regressoren oder dem Streichen eines/mehrerer Regressor/en effektiver sein. Fügt man eine Variable hinzu, so bewirkt das ein Ansteigen der Quadratsumme für die Regression und ein Sinken der Quadratsumme des Residuums.
Deshalb muss entschieden werden, ob die Erhöhung der Regressions- Quadratsumme ausreichend ist, um die Verwendung eines zusätzlichen Regressors im Modell gewährleisten zu können. Das Hinzufügen eines Regressors bewirkt auch ein Ansteigen der Varianz der geschätzten Werte $\hat{y}$, weshalb man Acht geben muss nur solche Regressoren zu verwenden die einen tatsächlichen Wert haben zur Erklärung der unabhängigen Variable y. Außerdem kann das Hinzufügen eines unwichtigen Regressors x_i den Mittelwert der Residuen erhöhen, was die Nützlichkeit des Modells senkt.

Die Hypothesen zum Testen des Signifikanz eines individuellen Regresssionskoeffizienten , wie β_j, sind:

$$H_0: \beta_j = 0$$
$$H_1: \beta_j \neq 0$$

Sofern H_0: $\beta_j = 0$ nicht verworfen wird, weist das darauf hin, dass der Regressor x_j aus dem Modell gelöscht werden kann. Die Test-Statistik für diese Hypothese ist

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{\hat{\beta}_j}{s_e(\hat{\beta}_j)} \qquad (64)$$

wo das Diagonalelement von $(X^T X)$, also C_{jj}, mit $\hat{\beta}_j$ korreliert.
Die Nullhypothese H_0: $\beta_j = 0$ wird verworfen, wenn $|t_0| > t_{\frac{\alpha}{2},\ n-k-1}$.

Anzumerken ist die Tatsache, dass hier nur von einem partiellen bzw. marginalen Test die Rede ist, weil der Regressionskoeffizient $\hat{\beta}_j$ von allen anderen Regressorvariablen x_i ($i \neq j$) abhängt, die im Modell auftreten. Deshalb ist das ein Test von der Verteilung x_j, die durch die anderen Regressoren gegeben ist.

Beispiel:
Wiederum betrachten wir die Zeitdaten für die Limonadenauslieferung (Tabelle 11 im Anhang). Angenommen man möchte die Verteilung der Variable „Distanz (x_2)“ zum Modell untersuchen.
Dafür geeignete Hypothesen wären:

$$H_0: \beta_2 = 0$$
$$H_1: \beta_2 \neq 0$$

Um diese Hypothesen zu testen, brauchen wir eine extra Quadratsumme bedingt durch β_2, oder $\sigma_R(\beta_2|\beta_1, \beta_0) = \sigma_R(\beta_1, \beta_2, \beta_0) - \sigma_R(\beta_1, \beta_0) =$

$$= \quad \sigma_R\,(\beta_1, \beta_2|\beta_0) - \sigma_R(\beta_1|\beta_0)$$

Aus dem Beispiel in **4.3.1** weiß man:

$$\sigma_R(\beta_1, \beta_2|\beta_0) = \hat{\beta}^T X^T y - \frac{(\sum_{i=1}^{n} y_i)^2}{n} = 5\,550{,}82 \qquad (2 \text{ Freiheitsgrade})$$

Das reduzierte Modell $y = \beta_0 + \beta_1 x_1 + \varepsilon$ wurde in einem Beispiel im Kapitel zur Einfachen Linearen Regression berechnet und ist gegeben durch $\hat{y} = 3{,}32 + 2{,}18x_1$.
Die Quadratsumme der Regression für dieses Modell ist:

$$\sigma_R(\beta_1|\beta_0) = \hat{\beta}_1 \sigma_{xy} = 2{,}18 \cdot 2\,473{,}34 = 5\,382{,}41 \qquad (1 \text{ Freiheitsgrad})$$

Darum haben wir:

$$\sigma_R(\beta_2|\beta_1, \beta_0) = 5\,550.82 - 5\,382{,}41 = 168{,}41 \qquad (1 \text{ Freiheitsgrad})$$

Das ist genau jene Zunahme der Regressions-Quadratsumme, welche durch Hinzufügen von x_2 zu einem bereits bestehenden Modell (das x_1 beinhaltet), entsteht.
Um H_0: $\beta_2 = 0$ zu testen, formulieren wir den Test:

$$F_0 = \frac{\sigma_R(\beta_2|\beta_1, \beta_0)/1}{\overline{\sigma}_E} = \frac{168{,}41/1}{10{,}62} = 15{,}85 \qquad (65)$$

An dieser Stelle muss angemerkt werden, dass das $\overline{\sigma}_E$, sowohl x_1 als auch x_2 im Nenner der Teststatistik verwendet. Nachdem $F_{0{,}05,\ 1,22} = 4{,}30$ ist, wird H_0: $\beta_2 = 0$ verworfen und man schließt daraus, dass die Distanz (x_2) einen bedeutenden Beitrag zum Modell leistet. Seit dieser partielle F-Test eine einzige Variable umfasst, ist er äquivalent zum T-Test. Diese Erkenntnis erhält man, weil der T-Test auf H_0: $\beta_2 = 0$ aus der Teststatistik t_0 =3,98 resultiert und seitdem das Quadrat eine t-verteilte Zufallsvariable mit υ – Freiheitsgraden ist. Somit haben wir: $t_0^2 = (3{,}98)^2 = 15{,}84 = F_0$

4.3.3 Spezialfall von orthogonalen Spalten in X

Wir betrachten das Modell (aus Abschnitt 4.3.2):

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$$
$$= \mathbf{X_1\beta_1} + \mathbf{X_2\beta_2} + \boldsymbol{\varepsilon}$$

Die Extra-Quadratsummen-Methode lässt es zu, die Auswirkungen der Regressoren in $\mathbf{X_2}$ bedingt durch jene in $\mathbf{X_1}$ zu messen, indem man $\sigma_R(\boldsymbol{\beta_2}|\boldsymbol{\beta_1})$ berechnet. Im Allgemeinen können wir nicht darüber sprechen, die Quadratsummen bedingt durch $\boldsymbol{\beta_2}$, $\sigma_R(\boldsymbol{\beta_2})$ zu finden, ohne Zugang zur Abhängigkeit dieser Mengenmäßigkeit auf die Regressoren in $\mathbf{X_1}$ zu haben. Dennoch können wir eine Summe von Quadraten bedingt durch $\boldsymbol{\beta_2}$ bestimmen, die keinerlei Abhängigkeit von den Regressoren in $\mathbf{X_2}$ aufweist, aber nur dann, wenn die Spalten in $\mathbf{X_1}$ **orthogonal** zu jenen in $\mathbf{X_2}$ sind.
Um das zu demonstrieren, bilden wir die Normalgleichungen $(X^T X)\hat{\beta} = X^T y$ wiederum für das Modell aus Abschnitt 4.3.2. Die Normalgleichungen sind nun:

$$\begin{bmatrix} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} X_1^T y \\ X_2^T y \end{bmatrix} \qquad (66)$$

Wenn die Spalten von $\mathbf{X_1}$ jetzt orthogonal zu den Spalten in $\mathbf{X_2}$ sind, so ist $\boldsymbol{X_1^T X_2 = 0}$ und $\boldsymbol{X_2^T X_1 = 0}$. Die Normalgleichungen ergeben darum:

$$\mathbf{X_1^T X_1 \hat{\beta}_1 = X_1^T y}$$
$$\mathbf{X_2^T X_2 \hat{\beta}_2 = X_2^T y}$$

mit den Lösungen:

$$\boldsymbol{\hat{\beta}_1 = (X_1^T X_1)^{-1} X_1^T y}$$
$$\boldsymbol{\hat{\beta}_2 = (X_2^T X_2)^{-1} X_2^T y}$$

Anzumerken ist, dass der kleinste Quadrate Schätzer von β_1 die Variable $\hat{\beta}_1$ ist, gleichgültig ob X_2 im Modell enthalten ist oder nicht und der kleinste Quadrate Schätzer von β_2 ist $\hat{\beta}_2$ ungeachtet davon ob X_1 im Modell ist.

Als Regressionsquadratsumme für das komplette Modell folgert man:

$$\begin{aligned} \sigma_R(\beta) &= \hat{\beta}^T X^T y \\ &= \left[\hat{\beta}_1, \hat{\beta}_2\right] \begin{bmatrix} X_1^T y \\ X_2^T y \end{bmatrix} \\ &= \hat{\beta}_1^T X_1^T y + \beta_2^T X_2^T y \\ &= y^T X_1 (X_1^T X_1)^{-1} X_1^T y + y^T X_2 (X_2^T X_2)^{-1} X_2^T y \end{aligned}$$

Allerdings gibt es zwei Arten von Normalgleichungen, für die wir jeweils anmerken:

$$\sigma_R(\beta_1) = \hat{\beta}_1^T X_1^T y = y^T X_1 (X_1^T X_1)^{-1} X_1^T y$$
$$\sigma_R(\beta_2) = \beta_2^T X_2^T y = y^T X_2 (X_2^T X_2)^{-1} X_2^T y$$

Vergleicht man die beiden Gleichungen für $\sigma_R(\beta_1)$ bzw. $\sigma_R(\beta_2)$ mit $\sigma_R(\beta)$ so sehen wir, dass:

$$\sigma_R(\boldsymbol{\beta}) = \sigma_R(\boldsymbol{\beta_1}) + \sigma_R(\boldsymbol{\beta_2})$$

Deshalb ist:

$$\sigma_R(\boldsymbol{\beta_1}|\boldsymbol{\beta_2}) = \sigma_R(\boldsymbol{\beta}) - \sigma_R(\boldsymbol{\beta_2}) \equiv \sigma_R(\boldsymbol{\beta_1})$$

und

$$\sigma_R(\boldsymbol{\beta_2}|\boldsymbol{\beta_1}) = \sigma_R(\boldsymbol{\beta}) - \sigma_R(\boldsymbol{\beta_1}) \equiv \sigma_R(\boldsymbol{\beta_2})$$

Folglich misst $\sigma_R(\beta_1)$ die Verteilung des Regressors in X_1 zum Modell *ohne Vorbehalt*, ebenso wie $\sigma_R(\beta_2)$ die Verteilung des Regressors in X_2 zum Modell bedingungslos misst. Nachdem man den Effekt eines jeden Regressors eindeutig bestimmen kann, wenn die Regressoren orthogonal sind, macht es Sinn, die Daten für Berechnungen durch orthogonale Variablen auszudrücken.

Beispiel:
Als Regressionsmodell mit orthogonalen Regressoren betrachten wir das Modell $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$, mit gegebener $\mathbf{X}$ – Matrix:

$$X = \begin{pmatrix} +1 & -1 & -1 & -1 \\ +1 & +1 & -1 & -1 \\ +1 & -1 & +1 & -1 \\ +1 & -1 & -1 & +1 \\ +1 & +1 & +1 & -1 \\ +1 & +1 & -1 & +1 \\ +1 & -1 & +1 & +1 \\ +1 & +1 & +1 & +1 \end{pmatrix}$$

Daraus ist ersichtlich, dass die Spalten von X orthogonal sind. Deshalb misst $\sigma_R(\beta_j)$ für $j = 1,2,3$ die Verteilung des Regressors x_j, unabhängig davon, ob irgendwelche anderen Regressoren dieser Anpassung entsprechen.

4.3.4 Test der allgem. linearen Hypothese Tβ= 0

Viele Hypothesen über Regressionskoeffizienten können durch beinahe einheitliches Vorgehen getestet werden. Die Extra-Quadratsummen-Methode ist ein Spezialfall dieses Verfahrens, denn im allgemeinen Fall wird die verwendete Quadratsumme bei Hypothesentests für gewöhnlich als Differenz zwischen zwei Residualsummen berechnet. In diesem Abschnitt wird nur das Verfahren an sich umrissen und zugehörige Beweise die den Umfang der Arbeit sprengen würden, werden weggelassen.

Wir nehmen an, dass die Hypothese die uns interessiert, durch H_0: $\mathbf{T\beta = 0}$ ausgedrückt werden kann, wobei **T** eine *m x p* Matrix von Konstanten ist, so dass nur *r* von den *m* Gleichungen bei $\mathbf{T\beta = 0}$ unabhängig sind. Das vollständige Modell ist $\mathbf{y = X\beta + \varepsilon}$, mit $\widehat{\boldsymbol{\beta}} = (\mathbf{X^T X})^{-1}\mathbf{X^T y}$ und die Quadratsumme der Residuen ist

$$\sigma_E(\mathrm{FM}) = \mathbf{y^T y} - \widehat{\boldsymbol{\beta}}^{\mathbf{T}}\mathbf{X^T y}. \qquad (\mathrm{n - p\ Freiheitsgrade})$$

Um das reduzierte Modell zu erhalten, werden die *r* unabhängigen Gleichungen in $\mathbf{T\beta = 0}$ verwendet und für *r* Parameter der Regressionskoeffizienten im vollständigen Modell bezüglich der verbleibenden *p - r* Regressionskoeffizienten aufgelöst. Das führt uns zu dem reduzierten Modell $\mathbf{y = Z\gamma + \varepsilon}$, wo zum Beispiel Z eine *n x (p - r)* Matrix und γ ein *(p - r)* x 1 Vektor von unbekannten Regressionskoeffizienten ist. Die Schätzung von γ ist:

$$\hat{\boldsymbol{\gamma}} = (\mathbf{Z^T Z})^{-1}\mathbf{Z^T y} \qquad (67)$$

und die Quadratsumme der Residuen für das reduzierte Modell

$$\sigma_E(\mathrm{RM}) = \mathbf{y^T y} - \hat{\mathbf{y}}^{\mathbf{T}}\mathbf{Z^T y} \quad (\mathrm{n - p + r\ Freiheitsgrade})$$

Das reduzierte Modell (RM) beinhaltet weniger Parameter als das vollständige Modell (VM), weshalb $\sigma_E(RM) \geq \sigma_E(VM)$ ist. Zum Testen der Hypothese H_0:$T\beta = 0$, werwendet man die Differenz der Quadratsummen der Residuen:

$$\sigma_H = \sigma_E(RM) - \sigma_E(VM) \qquad (n - p + r - (n - p) = r - \text{Freiheitsgrade})$$

Hier wird σ_H als jene Quadratsumme bezeichnet, die sich auf die Hypothese **H_0: Tβ = 0** bezieht. Die Teststatistik dieser Hypothese ist:

$$F_0 = \frac{\sigma_H / r}{\sigma_E(VM)/(n - p)}$$

Wir verwerfen H_0: $T\beta = 0$ wenn $F_0 > F_{\alpha, r, n-pn-p}$

Beispiel 1: (Test auf Gleichheit zweier Regressionskoeffizienten)

Der allgemeine Hypothesenansatz kann verwendet werden, um die Gleichheit zweier Regressionskoeffizienten zu testen. Dazu wird folgendes Modell gewählt:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

Im vollständigen Modell hat $\sigma_E(VM)$ $n - p = n - 4$ Freiheitsgrade und wir wollen **H_0:** $\beta_1 = \beta_3$ testen. Diese Hypothese kann ausgewiesen werden als H_0: **Tβ = 0,** wobei

$$\mathbf{T} = [0, 1, 0, -1]$$

ein *1 x 4* -Zeilenvektor ist. Es existiert nur eine Gleichung in **Tβ = 0**, nämlich $\beta_1 - \beta_3 = 0$, die eingesetzt in das Vollständige Modell das folgende reduzierte Modell ergibt:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_1 x_3 + \varepsilon \\ &= \beta_0 + \beta_1 (x_1 + x_3) + \beta_2 x_2 + \varepsilon \\ &= \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + \varepsilon \end{aligned}$$

mit $\gamma_0 = \beta_0,\ \gamma_1 = \beta_1\ (= \beta_3),\ z_1 = x_1 + x_3, \gamma_2 = \beta_2$, und $z_2 = x_2$. Die Quadratsumme bedingt durch die hypothese $\sigma_H = \sigma_E(RM) - \sigma_E(VM)$ hat einen Freiheitsgrad. Das F-Verhältnis ist $F_0 = (\sigma_H/1)\ /\ [\sigma_E(VM)/(n-4)]$. Diese Hypothese könnte jedoch auch mit einer T-Statistik mit 3 Freiheitsgraden getestet werden.

Beispiel 2: Wir nehmen das folgende Modell an:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

und möchten H_0: $\beta_0 = \beta_3,\ \beta_2 = 0$ testen. Um das in Form einer allgemeinen Hypothese zu erklären, legen wir zuerst **T** fest

$$\mathbf{T} = \begin{bmatrix} 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Jetzt ergeben sich aus **Tβ = 0** zwei Gleichungen**,** nämlich $\beta_1 - \beta_3 = 0$ und $\beta_2 = 0$. Diese Gleichungen erzeugen ein reduziertes Modell

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_1 x_3 + \varepsilon \\ &= \beta_0 + \beta_1 (x_1 + x_3) + \varepsilon \\ &= \gamma_0 + \gamma_1 z_1 + \varepsilon \end{aligned}$$

In diesem Beispiel hat $\sigma_E(RM)$ *n-2*-Freiheitsgrade und darum hat σ_H n-2-(n-4)=2 Freiheitsgrade. Das F-Verhältnis ist $F_0 = (\sigma_H/2) \,/\, [\sigma_E(VM)/(n-4)]$.

Zudem kann die Teststatistik für die allgemeine lineare Form in einer anderen Form geschrieben werden, nämlich

$$F_0 = \frac{\widehat{\boldsymbol{\beta}}^{\mathbf{T}} \mathbf{T}^{\mathbf{T}} [\mathbf{T}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{T}^{\mathbf{T}}]^{-1} \mathbf{T}\widehat{\boldsymbol{\beta}} \,/\, r}{\sigma_E(\mathrm{VM}) \,/\, (n-p)}$$

Wiederum werden die Hypothesen gebildet:

$$H_0: \mathbf{T}\boldsymbol{\beta} = \mathbf{c}$$
$$H_1: \mathbf{T}\boldsymbol{\beta} \neq \mathbf{c}$$

Die Teststatistik dafür ist

$$F_0 = \frac{(\mathbf{T}\widehat{\boldsymbol{\beta}} - \mathbf{c})^{\mathbf{T}} [\mathbf{T}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{T}^{\mathbf{T}}]^{-1} (\mathbf{T}\widehat{\boldsymbol{\beta}} - \mathbf{c}) \,/\, r}{\sigma_E(\mathrm{VM})/(n-p)}$$

Wir verwerfen die Nullhypothese $H_0: \mathbf{T}\boldsymbol{\beta} = \mathbf{c}$ wenn $F_0 > F_{\alpha,\ r,\ n-p}$. Das ist das Testverfahren eines einseitigen F-Tests. Anzumerken ist, dass der Zähler von F_0 eine Maßzahl ausdrückt, welche die quadratische Distanz zwischen **Tβ** und c standardisiert durch die Kovarianzmatrix von $T\hat{\beta}$ angibt.
Um die Anwendung dieses erweiterten Verfahrens zu demonstrieren, greifen wir die beschriebene Situation aus Beispiel 1 aufund nehmen an, dass wir $H_0: \beta_1 - \beta_3 = 2$ testen wollen. Offensichtlich ist dabei T =[0,1,0, -1] und c =[2]

Sofern die Hypothese H_0: Tβ = 0 nicht verworfen werden kann, mag es außerdem vernünftig sein, β zu schätzen und zwar durch die von der Nullhypothese vorgegebene Bedingung.

4.4 Beispiel: „Arbeitsmotivation mit mehreren Prädiktoren“

Beispiel: y ist die Motivation (bzw. die Einschätzung der Arbeitsmotivation durch Experten) und die folgenden Prädiktoren sind beliebige Fragebogenwerte:
(siehe Tabelle 11 im Anhang: Arbeitsmotivation mit mehreren Prädiktoren)

Untersucht werden in der Tabelle folgende Prädiktoren (Tabelle 12 im Anhang):

Prädiktoren: *Eigenschaften*
x_1: Ehrgeiz (Fragebogen)
x_2: Kreativität (Fragebogen)
x_3: Leistungsstreben (Fragebogen)
Prädiktoren: *Rahmenbedingungen*
x_4: Hierarchie (Position in der Hierarchie des Unternehmens)
x_5: Lohn (Bruttolohn pro Monat)
x_6: Arbeitsbedingungen (Zeitsouveränität, Kommunikationsstruktur usw.)
Prädiktoren: *Inhalte der Tätigkeit*
x_7: Lernpotential (Lernpotential der Tätigkeit)
x_8: Vielfalt (Vielfalt an Teiltätigkeiten)
x_9: Anspruch (Komplexität der Tätigkeit)

Die Tabelle liefert Daten (x_1, y_1) ... (x_n, y_n), wobei es k unabhängige Variablen x_i = (x_{1i} ... x_{ki}) gibt und y_i die Realisation einer Zufallsvariablen (unter der Bedingung x_i) ist. Folgender Zusammenhang zwischen der Variablen Y und dem Vektor x_i wird nun angenommen (im Beispiel ist *k = 9*):

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_{ki} + \varepsilon_i$$

$$= \beta_0 + \sum_{j=1}^{k} \beta_j \; x_{ji} + \varepsilon_i$$

ε_i charakterisiert eine zufällige „Störung“ und es wird angenommen, dass die Störungen $\varepsilon_1 \ldots \varepsilon_n$ unabhängig und normalverteilt sind mit EW null und Varianz $\sigma^2 > 0$.
Somit kann ein linearer Zusammenhang zwischen x und Y postuliert werden, welcher noch zufälligen Störungen unterliegt.

Schätzung bei multipler linearer Regression

Bezüglich der Wahl von $\beta_0 \ldots \beta_k$ wird folgender Ausdruck mit der *Methode der kleinsten Quadrate* (analog zur *einfachen* linearen Regression) minimiert:

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_k x_{ki})^2$$

und die mathematische Statistik bzw. das allgemeine lineare Modell liefert die Schätzwerte $\hat{\beta}_0, \hat{\beta}_1 \ldots . \hat{\beta}_k$ für die Parameter $\beta_0 \ldots \beta_k$

$\Longrightarrow$ *Schätzer für die Varianz der Messfehler:*

$$S^2_{y|x} = \frac{1}{n-k-1} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \; \ldots - \hat{\beta}_k x_{ki})^2$$

Für das Beispiel ergeben sich laut *SPSS* folgende Ergebnisse für die Schätzwerte:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{0}} = -2{,}286 \quad \hat{\boldsymbol{\beta}}_{\boldsymbol{1}} = 0{,}18 \quad \hat{\boldsymbol{\beta}}_{\boldsymbol{2}} = 0{,}153 \quad \hat{\boldsymbol{\beta}}_{\boldsymbol{3}} = 0{,}46 \quad \hat{\boldsymbol{\beta}}_{\boldsymbol{4}} = 0{,}291$$

$$\widehat{\boldsymbol{\beta}}_{\mathbf{5}} = -0{,}001 \quad \widehat{\boldsymbol{\beta}}_{\mathbf{6}} = -0{,}26 \quad \widehat{\boldsymbol{\beta}}_{\mathbf{7}} = 0{,}19 \quad \widehat{\boldsymbol{\beta}}_{\mathbf{8}} = 0{,}213 \quad \widehat{\boldsymbol{\beta}}_{\mathbf{9}} = 0{,}029$$

Daraus resultieren einige mögliche Fragen:

A) Wie exakt sind diese Schätzungen tatsächlich?

B) Inwiefern besteht ein (signifikanter) Einfluss der unabhängigen Merkmale auf die Motivation?

H_0: $\beta_1 = 0$

H_1: $\beta_2 = 0$

⋮ ⋮

C) Wie zufriedenstellend ist das Ausmaß in dem das multiple Regressionsmodell die Situation beschreibt?

Zu A) Genauigkeit der Schätzung bei multipler linearer Regression

➔ Die Schätzer $\hat{e}_{\beta_0}$ ... $\hat{e}_{\beta_k}$ für die Standardfehler von $\hat{\beta}_0$ ... $\hat{\beta}_k$ sind aus dem allgemeinen linearen Modell verfügbar.

➔ Wenn der Stichprobenumfang wächst, konvergieren die Schätzer $\hat{e}_{\beta_j}$ gegen 0 nach dem Motto „je größer der Stichprobenumfang, desto genauer die Schätzungen“ (-> Konfidenzintervalle werden kleiner).

➔ Somit kann man Konfidenzintervalle für β_0 ... β_k bilden, so ist in etwa:

$$(\hat{\beta}_0 - t_{n-k-1,1-\frac{\alpha}{2}}\ \hat{e}_{\beta_o},\qquad \hat{\beta}_0 + t_{n-k-1,1-\frac{\alpha}{2}}\ \hat{e}_{\beta_o})$$

Ein $(1-\alpha)$ - Konfidenzintervall für β_0 und $t_{n-k-1,1-\frac{\alpha}{2}}$ ist ein $(1-\alpha/2)$ –Quantil der T-Verteilung mit n – k – 1 -Freiheitsgraden (siehe T-Verteilung)

➔ Für den Standardfehler der Schätzer im multiplen linearen Regressionsmodell ergeben sich die genannten Werte:

$$\hat{\mathbf{e}}_{\boldsymbol{\beta_0}} = 4{,}672 \quad \hat{\mathbf{e}}_{\boldsymbol{\beta_1}} = 0{,}082 \quad \hat{\mathbf{e}}_{\boldsymbol{\beta_2}} = 0{,}05 \quad \hat{\mathbf{e}}_{\boldsymbol{\beta_3}} = 0{,}065 \quad \hat{\mathbf{e}}_{\boldsymbol{\beta_4}} = 0{,}132$$

$$\hat{\mathbf{e}}_{\boldsymbol{\beta_5}} = 0{,}001 \quad \hat{\mathbf{e}}_{\boldsymbol{\beta_6}} = 0{,}055 \quad \hat{\mathbf{e}}_{\boldsymbol{\beta_7}} = 0{,}09 \quad \hat{\mathbf{e}}_{\boldsymbol{\beta_8}} = 0{,}051 \quad \hat{\mathbf{e}}_{\boldsymbol{\beta_9}} = 0{,}042$$

➔ Wegen $\mathbf{t_{15,\,0{,}975}} = \mathbf{2{,}1314}$ ist

[-0,093; 0,186]

zum Beispiel ein 95 % - Konfidenzintervall für den Parameter β_3.

➔ $0{,}05 + 2{,}1314 \cdot 0{,}065 \approx 0{,}186$

n = 25, k = 9 ⇒ n – k – 1 = 15

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	t	Sig.	Konfidenzintervall für B (95,0%)	
		B	Standardfehler	Beta			Untergrenze	Obergrenze
1	(Konstante)	-2,286	4,672		-,489	,632	-12,244	7,673
	x1	,180	,082	,316	2,205	,043	,006	,355
	x2	,153	,050	,234	3,050	,008	,046	,260
	x3	,046	,065	,089	,706	,491	-,093	,186
	x4	,291	,132	,278	2,200	,044	,009	,574
	x5	-,001	,001	-,121	-1,031	,319	-,004	,001
	x6	-,026	,055	-,037	-,473	,643	-,144	,092
	x7	,190	,090	,230	2,118	,051	-,001	,381
	x8	,213	,051	,366	4,166	,001	,104	,321
	x9	,029	,042	,077	,683	,505	-,061	,119

a. Abhängige Variable: Y

Tabelle 13: 95 % - Konfidenzintervall , Standardfehler etc

:

Mit diesem linearen Regressionsmodell können nun auch Vorhersagen für Werte (bzw. weitere Arbeiter) an der Stelle $\mathbf{x} = (x_1, \dots, x_k)$ (mit k = 9 im Beispiel) gemacht werden:

$$\hat{y}(\mathbf{x}) = \hat{\beta}_0 + \sum_{j=1}^{n} \hat{\beta}_j x_j$$

Somit ergibt sich z.B als Vorhersage der multiplen Regression an der Stelle:

$x_1 = 21,\ x_2 = 45,\ x_3 = 18,\ x_4 = 13,\ x_5 = 3000,\ x_6 = 39,\ x_7 = 27,\ x_8 = 55,\ x_9 = 53$

der Wert: $\hat{y}(\mathbf{x}) = 25{,}43$

Trotzdem muss man unterscheiden zwischen der Vorhersage für den Wert der multiplen Regression an der Stelle $x = (x_1, \dots, x_k)$ (im Beispiel ist k =9) und der Vorhersage für den Wert einer neuen Beobachtung an der Stelle x.

Für beide Vorhersagen können außerdem wieder Standardfehler bestimmt und Konfidenzintervalle angegeben werden.

Das Bestimmtheitsmaß bei multipler linearer Regression

Die Werte der abhängigen Variable zerfallen in Modellvorhersage($\hat{y}$) und Residuum ($\hat{\varepsilon}$) d.h: $y_i = \hat{y}_i + \hat{\varepsilon}_i$

Modellvorhersage:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_{ki} = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ji}$$

Residuum: $\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j\, x_{ji})$

Daraus kann das Bestimmtheitsmaß R^2 bzw. die Güte der Modellanpassung gefolgert werden (Anteil der erklärten Varianz):

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

In unserem Beispiel ist n = 25 und k = 9

➔ $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 53{,}651$

➔ $\sum_{i=1}^{n}(y_i - \bar{y}) = 790{,}96$

➔ $R^2 = 1 - \frac{53{,}65}{790{,}96} = 92{,}95$

Somit werden 92,95 % der Varianz der Variablen „Motivation“ durch das lineare Regressionsmodell erklärt.

Statistische Tests bei der multiplen linearen Regression

B) Inwiefern besteht ein (signifikanter) Einfluss der unabhängigen Merkmale auf die Motivation?

H_0: $\beta_1 = 0$

H_1: $\beta_2 = 0$

⋮ ⋮

Zuerst soll ein Gesamttest auf Signifikanz durchgeführt werden. Hierzu überprüft man ob mindestens eine der Prädiktorvariablen $x_1 \dots x_k$ einen Einfluss auf die abhängige Variable y hat und formuliert die Hypothese:

Nullhypothese: **H_0:** $\beta_j = 0$ *für alle* $j \in \{1 \dots k\}$
Alternative: **H_1:** $\beta_j \neq 0$ für mindestens ein $j \in \{1 \dots k\}$

Im Anschluss daran könnte die Frage auftreten, ob die Prädiktorvariable x_j (z.B Ehrgeiz) einen Einfluss auf die abhängige Variable y hat. Dann würde sich diese mathematische Formulierung der Hypothese ergeben:

Nullhypothese: **H_0:** $\beta_j = 0$
Alternative: **H_1:** $\beta_j \neq 0$

Schritt 1: Gesamttest auf Signifikanz

Mit **H_0:** $\beta_j = 0$ für alle j $\in$ {1,2…k}
H_1: $\beta_j \neq 0$ für min. ein j $\in$ {1,2…k}

➔ Man bestimmt die Varianz der Regression ($\sigma^2 = \frac{1}{k}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$) und die Residualvarianz $S_{xy}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, wobei genau wie im linearen Regressionsmodell vorgegangen wird.

➔ H_0 wird zugunsten der Alternative verworfen, falls gilt:

$$F_n = \frac{\sigma^2}{\sigma_{xy}^2} > F_{k;n-k-1;1-\alpha}$$

➔ Wenn H_0 durch diesen Test verworfen wird, so bleibt noch unklar, welches dieser Merkmale signifikant ist

Schritt 2: Tests für die Signifikanz einzelner Merkmale

Mit H_0: $\beta_j = 0$
H_1: $\beta_j \neq 0$

➔ Man verwirft die Nullhypothese H_0 zugunsten der Alternative H_1 falls gilt:

$$T_n = \left|\frac{\hat{\beta}_j}{\hat{e}_{\beta_j}}\right| > t_{n-k-1;\,1-\alpha/2}$$

(oder der entsprechende p-Wert kleiner als α ist)

➔ $t_{n-k-1;\,1-\alpha/2}$ ist das (1 - α/2) – Quantil der T-Verteilung mit n-k-1 Freiheitsgraden und $\hat{e}_{\beta_j}$ der Standardfehler von $\hat{\beta}_j$

Nun wird diese Theorie auf unser Beispiel angewendet:

„Besteht ein Einfluss von einer der 9 Prädiktorvariablen auf die abhängige Variable?“
Mathematische Hypothesen:

H_0: $\beta_j = 0$ für alle j = 1 … 9
H_1:$\beta_j \neq 0$ für mindestens ein j ϵ {1, … ,9}

➔ F_n = 21,404, $F_{9,\,15,\,0,95}$ = 2,59

➔ Die Nullhypothese wird mit Irrtumswahrscheinlichkeit von 5% verworfen, weil F_n > 21,404 > 2,59

ANOVA[a]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Sig.
1	Regression	705,988	9	78,443	21,404	,000[b]
	Residuum	54,972	15	3,665		
	Gesamtsumme	760,960	24			

a. Abhängige Variable: Y

b. Prädiktoren: (Konstante), x9, x2, x5, x6, x3, x8, x7, x4, x1

Tabelle 14: ANOVA

„Besteht ein Einfluss der Prädiktorvariable Ehrgeiz (x_1) auf die abhängige Variable Motivation (Signifikanz des Regressionskoeffizienten β_1)?“

Mathematische Hypothesen:
H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$

➔ $\hat{\beta}_1 = 0{,}18$, $\hat{e}_{\beta_1} = 0{,}082$, $t_{25-10,0,975} = 2{,}1314$ $\Rightarrow T_{25} = 2{,}19$

➔ Nachdem 2,19 > 2,1314 ist, wird die Nullhypothese H_0 zu Gunsten der Alternative H_1: $\beta_1 \neq 0$ verworfen (mit α =5%).
(vgl. mit den Signifikanzwerten in der Tabelle oben)

Kapitel 5

5. DIE VARIANZANALYSE

Zum Einstieg in dieses Kapitel betrachten wir ein Beispiel, welches sich auf die Variabilität der Gewichtszunahme einer fest vorgegebenen Anzahl von Tieren z.B Zuchtrindern etc. bezieht, auch wenn die Futterart und –menge bzw. die Lebensbedingungen völlig gleich sind. Diese Tatsache muss als Zufallsveränderliche angesehen werden, die durch Umstände hervorgerufen wird, die sich unserer Kenntnis und Kontrolle entziehen.
Wenn die Tiere hingegen unterschiedlich lange gefüttert werden, so wird die Zufallsvariabilität möglicherweise durch eine Variabilität überlagert, die sich durch Futterunterschiede ergeben. Nun stellt sich die Frage, ob die Futterart einen Einfluss auf die Gewichtszunahme hat und dazu muss man diesen Einfluss vom Zufallseinfluss zu trennen versuchen. Dies ist eine typische Aufgabe der *einfachen Varianzanalyse.* Um zwei Einflüsse gleichzeitig zu untersuchen erfordert dies die Anwendung einer *doppelten Varianzanalyse*, mit der versucht wird, die beiden Einflüsse voneinander und außerdem vom Zufallseinfluss, zu trennen.

Die Varianzanalyse als solche beruht auf einer rein arithmetischen Zerlegung der „**Quadratsumme**“ (=Summe der Quadrate der Abweichungen der Stichprobenwerte vom Mittelwert), wobei man in eine Summe von Bestandteilen zerlegt, die jede für sich einer bestimmten Variationsursache entspricht (z.B Futterart bzw. zu untersuchende Variable) und deren anderer Bestandteil der Zufallseinfluss ist.

Demnach wird bei der Varianzanalyse die vorliegende Stichprobe in Teilstichproben zerlegt, deren Mittelwerte verglichen werden,

5.1 Mittelwertvergleich von Normalverteilungen bei einfacher Varianzanalyse

n-Versuchstiere werden abgewogen und anschließend nach dem Zufallsprinzip in u - Gruppen eingeteilt. Diesen u- Gruppen werden jeweils u- verschiedene Futtermittel verabreicht. Nach einer bestimmten Zeit werden die Tiere wieder abgewogen und die Gewichtszunahme wird notiert. Somit erhält man eine Stichprobe von insgesamt n Werten, welche sich in **u - Gruppen** untergliedert, etwa:

$x_{11}, x_{12},...,x_{1n_1}$ (1. Zuchtrindgruppe)

$x_{21}, x_{22},...,x_{2n_2}$ (2. Zuchtrindgruppe)

Hier bezeichnet der erste Index die Gruppe und der zweite die Nummer des Tieres in der Gruppe, wobei die 1. Gruppe aus n_1 Tieren und die 2. Gruppe aus n_2 Tieren besteht, mit $n_1 + n_2 + ... + n_u = n$.
Nun soll mit der Varianzanalyse geprüft werden, ob hinsichtlich der mittleren Gewichtszunahme bei den auftretenden Gruppen signifikante, durch unterschiedliches Futter hervorgerufene, Unterschiede bestehen oder diese eine zufallsbedingte Ursache haben. Bestehen bloß zufallsbedingte Unterschiede, wäre es egal mit welchem der genannten Futtermittel man mästet. Unter der Annahme, dass die u- Gruppen von Zahlen aus u-normalverteilten Grundgesamtheiten entstammen, die alle dieselbe Varianz haben (σ^2 muss nicht bekannt sein) soll geprüft werden, ob die Mittelwerte $\mu_1, ..., \mu_u$ der genannten Grundgesamtheiten ebenfalls übereinstimmen.
Dementsprechend testet man die Hypothese, dass alle diese u Mittelwerte gleich sind und zerlegt die „Quadratsumme“ q in zwei Bestandteile q_1 und q_2:

$$q = \sum_{i=1}^{r} \sum_{k=1}^{ni} (x_{ik} - \bar{x})^2 \quad \Longrightarrow \qquad q = q_1 + q_2 \tag{68}$$

Dabei beschreibt der erste Bestandteil q_1 die Streuung zwischen den Gruppen und der zweite q_2 die Streuung innerhalb jeder Gruppe. Im Anschluss werden diese beiden Bestandteile dann miteinander verglichen.
Nun erfolgt der schrittweise Test der Hypothese, dass die normalverteilten Grundgesamtheiten gleicher Varianz, aus denen die u- Gruppen stammen, alle denselben Mittelwert haben.

1.Schritt: Berechnung der u-Mittelwerte $\bar{x}_1, \dots, \bar{x}_u$ der Gruppen:

$$\bar{x}_i = \frac{1}{n_i}(x_{i1} + x_{i2} + \dots + x_{in_i})$$

und Berechnung des Mittelwertes der gesamten Stichprobe:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{u}\sum_{k=1}^{n_i} x_{ik} = \frac{1}{n}\sum_{i=1}^{u} n_i \bar{x}_i.$$

2.Schritt: Berechnung der *„Quadratsumme zwischen den Mittelwerten der Gruppen“*:

$$q_1 = \sum_{i=1}^{u} n_i(\bar{x}_i - \bar{x})^2, \tag{69}$$

und der *„Quadratsumme innerhalb der Gruppen“*:

$$q_2 = \sum_{i=1}^{u}\sum_{k=1}^{n_i} (x_{ik} - \bar{x}_i)^2. \tag{70}$$

Daraus bildet man einen Quotienten und legt eine Signifikanzzahl α (5% oder 1%) fest:

$$v_0 = \frac{\frac{q_1}{u-1}}{\frac{q_2}{(n-u)}}$$

3.Schritt: Bestimmung der Lösung c der Gleichung aus der Tafel der F-Verteilung im Anhang mit (u - 1, n - u) Freiheitsgraden:

$$P(V \leq c) = 1 - \alpha$$

Ist nun $v_0 \leq c$, so wird die Hypothese $\mu_1 = \mu_2 = ... = \mu_u$ angenommen und wenn $v_0 > c$ ist, dann wird sie verworfen (d.h man nimmt an, dass die Mittelwerte nicht alle gleich sind)

Variation	Freiheitsgrade	Quadratsumme	Durchschnittsquadrat
Zwischen den Gruppen	u – 1	q_1	$q_1/(u-1)$
Innerhalb der Gruppen	n – u	q_2	$q_2/(n-u)$
Insgesamt	n – 1	q	

Tabelle 15: Ein Beispiel für die einfache Varianzanalyse

5.2 Ein Beispiel für die einfache Varianzanalyse

In diesem Beispiel soll untersucht werden, ob die Zugfestigkeit von Alufolien an allen Stellen dieselbe ist. Zu diesem Zweck wurden 4 Alufolien untersucht, und es ergeben sich die Werte in der Tabelle).

Messstelle	Messwerte			
1. Gruppe (Ecke)	137	142	128	137
2. Gruppe (Mitte)	140	139	117	137
3. Gruppe (Kante)	142	140	133	141

Tabelle 16: Stichprobenwerte für die Zugfestigkeit von Folien

1. Schritt: Als Gruppenmittelwerte ergeben sich:

$$\bar{x}_1 = 136, \quad \bar{x}_2 = 133{,}25, \quad \bar{x}_3 = 139$$

Als Mittelwert der gesamten Stichprobe berechnet man:

$$\bar{x} = \frac{1}{12}(4\bar{x}_1 + 4\bar{x}_2 + 4\bar{x}_3) = \frac{1}{3}(\bar{x}_1 + \bar{x}_2 + \bar{x}_3) = 136{,}08$$

2. Schritt: Berechnen der Quadratsumme zwischen den Gruppen:

$$q_1 = 4\,[(\bar{x}_1 - \bar{x})^2 + (\bar{x}_2 - \bar{x})^2 + (\bar{x}_3 - \bar{x})^2]$$

$$= 4\,[0{,}083^2 + 2{,}833^2 + 2{,}917^2] = 66{,}167.$$

Berechnen die Quadratsumme innerhalb der Gruppen:

$$q_2 = \sum_{i=1}^{3}\sum_{k=1}^{4}(x_{ik} - \bar{x}_i)^2 =$$

$$= (137 - 136)^2 + (142 - 136)^2 + \cdots + (141 - 139)^2 = 508{,}75.$$

$$\Longrightarrow v_0 = \frac{q_1/2}{q_2/9} = \frac{33{,}08}{56{,}52} = 0{,}585.$$ Zusätzlich wählen wir die Signifikanzzahl α=0,05.

3.Schritt: Nachdem u = 3, n = 12 ist, kann man u-1=2, n – u = 9 annehmen.

Aus der Tabelle im Anhang ergibt sich als Lösung der Gleichung:

$$P(V \leq c) = 0{,}95$$

der Wert **c = 4,26** mit $v_0 < c$, weshalb die Hypothese $\mu_1 = \mu_2 = \mu_3$ angenommen wird.
Somit kann man durch die Stichprobe aussagen, dass die Zugfestigkeit der Folien zwischen den verschiedenen Stellen der Messung nur zufallsbedingt schwankt, der Unterschied der Messwerte also nicht signifikant ist.

Variation	Freiheitsgrade	Quadratsumme	Quadratmittel
Zwischen den Gruppen Innerhalb der Gruppen	2 (k-1) 9 (n-k)	66,16 (L) 508,75 (F)	33,08 ($\frac{1}{k-1}L$) 56,53($\frac{1}{n-k}L$)
Insgesamt	11 (n-1)	574,92 (T)	

Tabelle 17: Vergleich der Variation zwischen und innerhalb der Gruppen

5.3 Die doppelte Varianzanalyse

Bislang wurden derartige Stichproben analysiert, welche sich nach einem Merkmal in Gruppen unterteilen ließen, so dass dies einer *einzelnen* Variationsursache (bzw. der einfachen Varianzanalyse) entsprach. Nun lassen sich die Gruppen nach einem zusätzlichen Merkmal untergliedern, weshalb man die doppelte Varianzanalyse zum Untersuchen des Einflusses zweier Variationsmerkmale verwendet. Jene Teile, die man bei der genannten Unterteilung der Gruppen einer Stichprobe erhält, werden als Klassen bezeichnet. Es wird eine Stichprobe von insgesamt n-Werten vorgegeben und in r Gruppen geteilt, wobei sich jede Gruppe in p-Klassen untergliedert. Danach behandeln wir den einfachsten Fall, dass jede Klasse nur einen einzelnen Fall enthält ($\Longrightarrow n = r \cdot p$).
Die Stichprobenwerte werden wiederum mit x_{ik} bezeichnet, wobei der erste Index die Gruppennummer und der zweite die Nummer der Klasse ist. Die Stichprobe lässt sich nun wie folgt anordnen: *p-Spalten (Klassen)*

r-Gruppen (Zeilen)

$$\begin{pmatrix} x_{11} & x_{11} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{r1} & x_{r2} & \dots & x_{rp} \end{pmatrix}$$

Dabei wird die Voraussetzung angenommen, dass die n-Beobachtungen aus n-unabhängigen normalverteilten Grundgesamtheiten mit derselben Varianz σ^2 und den möglicherweise verschiedenen Mittelwerten $\mu_{11}, \dots, \mu_{rp}$ entstammen (σ muss nicht bekannt sein). Jetzt ist die Hypothese auf Gleichheit der Mittelwerte zu testen, wodurch beurteilt werden kann, ob alle n - genannten Grundgesamtheiten völlig gleich verteilt sind. Somit wird der Mittelwert der i-ten Zeile mit $\bar{x}_{i\cdot}$ definiert und jener der k-ten Spalte mit $\bar{x}_{\cdot k}$.

$$\bar{x}_{i\cdot} = \frac{1}{p}\sum_{k=1}^{p} x_{ik} = \frac{\text{Zeilensumme}}{\text{Anzahl der Werte pro Zeile}} \quad \text{mit i} = 1, \dots, r \qquad (71)$$

$$\bar{x}_{\cdot k} = \frac{1}{r}\sum_{i=1}^{r} x_{ik} = \frac{\text{Spaltensumme}}{\text{Anzahl der Werte pro Spalte}} \quad \text{mit k} = 1, \dots, p$$

Dem Hinzuzufügen ist, dass die Punkte bei den Indizes zur Unterscheidung der beiden Arten von Mittelwerten stehen und zwar bei jenem Index, über den man summiert.
Der Mittelwert $\bar{x}$ ist gleich definiert wie oben und die Quadratsumme q

$$q = \sum_{i=1}^{r}\sum_{k=1}^{ni}(x_{ik} - \bar{x})^2$$

wird anstatt in zwei Bestandteile, nun in drei (q_1, q_2, q_3) zerlegt: (72)

$$q = \sum_i \sum_k (\bar{x}_{i\cdot} - \bar{x})^2 + \sum_i \sum_k (\bar{x}_{\cdot k} - \bar{x})^2 + \sum_i \sum_k (x_{ik} - \bar{x}_{i\cdot} - \bar{x}_{\cdot k} + \bar{x})^2 .$$

Nachdem in den ersten beiden Doppelsummen jeweils nur ein Index auftritt, lassen sie sich auf einfache Summen reduzieren.

$$q = q_1 + q_2 + q_3 \quad \text{mit:}$$

$$q_1 = p\sum_{i=1}^{r}(\bar{x}_{i\cdot} - \bar{x})^2 \qquad p \text{ ist die Anzahl der Klassen}$$

$$q_2 = r\sum_{k=1}^{n}(\bar{x}_{\cdot k} - \bar{x})^2 \qquad r \text{ ist die Anzahl der Gruppen}$$

$$q_3 = \sum_{i=1}^{r}\sum_{k=1}^{p}(x_{ik} - \bar{x}_{i\cdot} - \bar{x}_{\cdot k} + \bar{x})^2$$

Hier erklärt q_1 **die Quadratsumme zwischen den Mittelwerten der Zeilen**, q_2 die **Quadratsumme zwischen den Mittelwerten der Spalten** und q_3 die **Restsumme**.
Wiederum werden auch hier die n Stichprobenwerte x_{ik} als einzelne Beobachtungen von n Zufallsvariablen X_{ik} aufgefasst. Wenn man nun die x_{ik} im einfachen Varianzanalysemodell durch X_{ik} ersetzt, folgen daraus Zufallsvariable die mit den jeweils entsprechenden Großbuchstaben $\bar{X}_{i\cdot}$, $\bar{X}_{\cdot k}$, $\bar{X}$, Q, Q_1, Q_2, Q_3 benannt werden.

Sofern die Hypothese richtig ist haben Q/σ^2, Q_1/σ^2, Q_2/σ^2, Q_3/σ^2 jeweils eine χ^2-Verteilung mit n – 1, r – 1, p – 1 bzw. (r – 1) (p – 1) Freiheitsgraden und der Mittelwert dieser Variablen ist gleich σ^2:

$$\sigma_1^2 = \frac{1}{r-1}Q_1, \qquad \sigma_2^2 = \frac{1}{p-1}Q_2, \quad \sigma_3^2 = \frac{1}{(r-1)(p-1)}Q_3$$

Ein Vergleich der Varianzen von Normalverteilungen führt zu folgenden Quotienten

$$\mathbf{V_1} = \sigma_1^2/\sigma_2^2 \quad \text{und} \quad \mathbf{V_2} = \sigma_2^2/\sigma_3^2$$

die eine F-Verteilung mit $[r-1, (r-1)(p-1)]$ bzw. $[p-1, (r-1)(p-1)]$ Freiheitsgraden besitzen.

5.4 Ein Beispiel für die doppelte Varianzanalyse

Nun wird die doppelte Varianzanalyse in einem Beispiel angewendet und zwar beziehen wir uns wieder auf das Eingangsbeispiel mit den Zuchtrindern.

Zwölf Rinder werden aufgrund des Gewichtes in r=4 Gruppen zu p=3 Rindern unterteilt. Eine gewisse Zeit später stellt man die Gewichtszunahme fest und testet die Hypothese, dass die Unterschiede zwischen den beobachteten und gemessenen Werten rein zufällig sind. Es haben also weder die verwendeten Futterarten noch die Anfangsgewichte Einfluss auf die Gewichtszunahme und die Zufallsvariable ist normalverteilt.

	Futterart		
	A	**B**	**C**
Gruppe 1	7,0	14,0	8,5
... 2	16,0	15,5	16,5
... 3	10,5	15,0	9,5
... 4	13,5	21,0	13,5

Tabelle 18: Stichprobe der Gewichtszunahme durch Futterarten

1. Schritt: Hier wird die Variation zwischen den Zeilen, Zwischen den Spalten bzw. der Rest berechnet.

Variation	**Freiheitsgrade**	**Quadratsumme**	**Durchschnittsquad.**
Zwischen den Zeilen bzw. den Gruppen	r – 1 = 3	$q_1 = 87{,}73$	$\sigma_1^2 = \frac{q_1}{3} = 29{,}24$
Zwischen den Spalten bzw. Futterarten	p – 1 = 2	$q_2 = 54{,}12$	$\sigma_2^2 = \frac{q_2}{2} = 27{,}06$
Rest	(r-1) (p-1) = 6	$q_3 = 28{,}21$	$\sigma_3^2 = \frac{q_3}{6} = 4{,}70$
Insgesamt	n – 1 = 11	q = 170,06	

Tabelle 19: Quadratsummen zwischen Zeilen bzw. Gruppen und Spalten bzw. Futterarten

2. Schritt: Es ist

$$v_1 = \frac{29{,}24}{4{,}702} = 6{,}219 \quad \text{und} \quad v_2 = \frac{27{,}06}{4{,}702} = 5{,}755.$$

3. Schritt: Die Signifikanzzahl $\alpha = 0{,}05$ wird vorgegeben.

4. Schritt: Somit hat die Gleichung $P(V \leq c_1) = 1 - \alpha = 0{,}95$ für die F-Verteilung mit (3, 6) – Freiheitsgraden die Lösung

$$c_1 = 4{,}76$$

$$\Longrightarrow \quad v_1 = 6{,}219 \quad > \quad c_1 = 4{,}76.$$

Demnach kann angenommen werden, dass zwischen den Gruppen ein signifikanter Unterschied besteht, d. h dass das Anfangsgewicht das Endgewicht beeinflusst. Die zu testende Hypothese wird dadurch bereits verworfen.

5. Schritt: Zudem wird vorausgesetzt, dass Additivität vorliegt und für die F-Verteilung mit (2, 6) Freiheitsgraden hat die Gleichung $P(V \leq c_2) = 1 - \alpha = 0{,}95$ nun die Lösung:

$$c_2 = 5{,}14.$$

Damit ist $v_2 = 5{,}755 > c_2 = 5{,}14$. Darum darf man annehmen, dass zwischen den Spalten ebenfalls ein signifikanter Unterschied besteht, das heißt, dass auch die Futterart die Gewichtszunahme beeinflusst.

5.5 Das Schließen auf die allgemeine Form

Man kann derartige Fragestellungen auch explizit als lineare Regression formulieren, indem die unabhängigen Variablen nur durch die Werte 0 und 1 ausgedrückt werden und so eine Zerlegung in Teilstichproben erfolgt.

Demnach wird die j-te Indikatorvariable $v_j = \begin{cases} 1 & \text{z. B für Rindergruppe j} \\ 0 & \text{sonst} \end{cases}$ gesetzt und die Stichprobe $y_1, y_2, \ldots, y_n$ spaltet sich in k Teilstichproben mit entsprechenden Mittelwerten $M_1, M_2, \ldots, M_k$ auf.

$y_1, y_2, \ldots, y_{n_1}$ vom Umfang n_1 (n_1 –Rinder denen Futtermittel x_1 verabreicht wird)
$y_{n_1+1}, \ldots, y_{n_1+n_2}$ vom Umfang n_2
$\vdots$ $\vdots$
$y_{n_1+n_2+\cdots n_{k-1}+1}, \ldots, y_{n_1+n_2+\cdots+n_k}$ vom Umfang n_k

Im günstigsten Fall gilt $y = A\beta$ wobei $A = (v_1, v_2, \ldots, v_k)$ ist und β_i (für i = 1,...,w) die durchschnittliche Gewichtszunahme jener Rindergruppe ausdrückt, die mit Futtermittel x_i gefüttert wurde. (73)

Aus $A = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 1 & 0 & \ldots & \vdots \\ 0 & 1 & \ldots & \vdots \\ \vdots & \vdots & \ldots & 0 \\ \vdots & 1 & \ldots & \vdots \\ \vdots & 0 & \ldots & 1 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix}$, $\beta_i = \begin{pmatrix} \beta_1 \\ \vdots \\ \vdots \\ \beta_w \end{pmatrix}$ folgert man

$$\begin{array}{rcccc} y_1 = & \beta_1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{n_1} = & \beta_1 & 0 & 0 & 0 \\ y_{n_1+1} = & 0 & \beta_2 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{n_1+n_2} = & 0 & \beta_2 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_{\ldots} = & 0 & 0 & 0 & \beta_w \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{\ldots} = & 0 & 0 & 0 & \beta_w \end{array}$$

Allerdings sind diese Gleichungen aufgrund störender Einflüsse nicht korrekt, weshalb wir jene β_i suchen, die diese Gleichungen möglichst optimal erfüllen.

Der Schätzer $\hat{\beta}$ für β ist nach Kapitel 5.1 gegeben durch $\hat{\beta} = (A^T A)^{-1} A^T y$.
Daraus lässt sich folgendes ableiten:

$$A^T A = \text{diag}(n_1, \dots, n_k) = \begin{pmatrix} y_{n_1} & 0 & \dots & 0 \\ 0 & y_{n_2} & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & y_{n_k} \end{pmatrix} \Longrightarrow (A^T A)^{-1} = \begin{pmatrix} \frac{1}{y_{n_1}} & 0 & \dots & 0 \\ 0 & \frac{1}{y_{n_2}} & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \frac{1}{y_{n_k}} \end{pmatrix}$$

$$A^T y = \begin{pmatrix} y_1 + \dots + y_{n_1} \\ y_{n_1+1} + \dots + y_{n_1+n_2} \\ \vdots \\ y_{n_1+n_2+\cdots n_{k-1}+1}, \dots, y_{n_1+n_2+\cdots+n_k} \end{pmatrix} \quad \Longrightarrow$$

$$(A^T A)^{-1} A^T y = \begin{pmatrix} \frac{1}{n_1}(y_1 + \dots + y_{n_1}) \\ \frac{1}{n_1+n_2}(y_{n_1+1} + \dots + y_{n_1+n_2}) \\ \vdots \\ \frac{1}{y_{n_1+n_2+\cdots+n_k}}(y_{n_1+n_2+\cdots n_{k-1}+1}, \dots, y_{n_1+n_2+\cdots+n_k}) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix}$$

Somit erhält man also $\hat{\beta}_1 = \mu_1,\ \hat{\beta}_2 = \mu_2, \dots,\ \hat{\beta}_k = \mu_k$ (74)

Bei Varianzanalysen werden also die Mittelwerte von z - unabhängigen Stichproben verglichen. Seien μ_1, μ_2 ... μ_z die Mittelwerte, so wird die Hypothese H_0: $\mu_1, \mu_1, \dots, \mu_k$ getestet. Nun wird Y =(Y_1, ..., Y_n) verwendet und die folgenden Annahmen angenommen:

1. $Y_1, Y_2, \dots, Y_n$ sind voneinander unabhängig
2. $Y_{n_1+n_2+\cdots n_{j-1}+1}, \dots, Y_{n_1+n_2+\cdots+n_j}$ sind N (μ_j, σ) - verteilt

5.5.1 Test der Hypothese H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$

Man testet die Hypothese H_0: $\mu_1 = \mu_2 = \dots = \mu_k$ und sucht eine Testvariable.
Ferner bezeichnet man die Menge der Indizes der j-ten Teilstichprobe z.B mit
$I_j = \{n_1 + \dots + n_{j-1} + 1, \dots, n_1 + \dots + n_j\}$.

Zudem sei

$\hat{\mu}_j = \frac{1}{n_j} \sum_{l \epsilon I_j} Y_l$ ein Schätzer für μ_j

$\hat{\sigma}_j^2 = \frac{1}{n_j - 1} \sum_{l \epsilon I_j} (Y_l - \hat{\mu}_j)^2$ ein Schätzer für $\sigma_j^2 = \sigma^2$

$\hat{\mu} = \frac{1}{n} \sum_{l=1}^{n} Y_l = \frac{1}{n} \sum_{j=1}^{k} n_j \hat{\mu}_j$ ein Schätzer für den Gesamtmittelwert

Die Abweichungen zwischen den Mittelwerten werden geschätzt durch:

$$L = \sum_{j=1}^{k} n_j \left(\hat{\mu}_j - \hat{\mu}\right)^2 \quad (75)$$

Daraus folgert man, dass wegen $\hat{\mu}_j \approx \mu_j$ kleine Werte von L für H_0 und große Werte von L gegen H_0 sprechen. Außerdem muss L richtig normiert werden.

Nachdem $\frac{1}{n-k}F$ ein Mittelwert der $\hat{\sigma}_j^2$ ist, kann $\frac{1}{n-k}F$ als Schätzer für σ^2 gewählt werden, mit:

$$F = \sum_{j=1}^{k}(n_j - 1)\,\hat{\sigma}_j^2 = \sum_{j=1}^{k}\sum_{l \epsilon I_j}(Y_l - \hat{\mu}_j)^2 \qquad (76)$$

F misst z.B Messfehler und zufällig auftretende Fehler sozusagen die Abweichungen der Y_l von ihrem jeweiligen Mittelwert $\hat{\mu}_j = \mu_j$. Darum gibt $\sigma \approx \frac{1}{n-k}\,F$ an, in welchem Ausmaß die Teilstichproben um den jeweiligen Mittelwert μ_j schwanken.
Abschließend muss man noch die Abweichungen der $\hat{\mu}_j = \mu_j$ ($\approx \frac{1}{k-1}L$) in Relation zu σ ($\approx \frac{1}{n-k}\,F$) betrachten und die Verteilung dieses Quotienten bei Gültigkeit von H_0 berechnen.

$$H = \frac{\frac{1}{k-1}\,L}{\frac{1}{n-k}\,F} \qquad H \epsilon R^+ \qquad (77)$$

Diesbezüglich kann ein Satz aus der Wahrscheinlichkeitsrechnung herangezogen werden:

Satz: Seien die Zufallsvariablen X_1 und X_2 unabhängig, X_1 C(p) – verteilt und X_2 C(q) – verteilt, so hat $\frac{\frac{1}{p}X_1}{\frac{1}{q}X_2}$ die F(l, m) – Verteilung.

5.5.2 Quadratsummenzerlegung

Zusätzlich zu diesen Angaben gibt T die Abweichung vom Gesamtmittel $\hat{\mu} = \mu$ an:

$$T = \sum_{l=1}^{n}(Y_l - \mu)^2 \qquad (78)$$

Satz: (79)
Gilt $\mu_1 = \mu_2 = ... = \mu_k$, so hat H die F(k - 1, n - k) – Verteilung und es gilt L + F = T.

Beweis:
Seien $V_0 = \{0\} \subset V_1 \subset V_2 \subset .... \subset V_{r-1} \subset V_{r-1} \subset V_r = \mathbf{R^n}$ Teilräume mit Dimensionen $d_0 = 0 < d_1 < ... < d_{r-1} < d_r = n$ und sei P_j die orthogonale Projektion auf V_j so dass insbesondere $P_0x = 0$ und $P_rx = x$ für alle x gilt.

Dann wird V_1 vom Vektor $e = \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}$ und V_2 von den Vektoren $v_1, v_2, ..., v_k$ aufgespannt und v_j ist genau jener Vektor der in den I_j , also der Menge der j-ten Teilstichprobe Einsen hat und sonst lauter Nullen. Somit gilt $d_1 = 1$, $d_2 = k$ und $d_3 = n$ und wegen $v_1 + ... + v_k = e$ folgt $V_1 \subset V_2$.

Die orthogonale Projektion auf V_1 kann berechnet werden indem für A = e die 1 x 1-Matrix (1/n) durch $(A^TA)^{-1}$ angegeben wird und $P_1 = A(A^TA)^{-1}A^T$ die n x n –Matrix

$$\begin{pmatrix} \frac{1}{n} & \dots & \frac{1}{n} \\ \vdots & \ddots & \vdots \\ \frac{1}{n} & \dots & \frac{1}{n} \end{pmatrix} \text{ ist.}$$

$\Longrightarrow$ P_1x ist für alle $x \in \mathbf{R^n}$ die orthogonale Projektion von x auf den Teilraum V_1, insbesondere gilt $P_1Y = \hat{\mu}\, e$.
Zur Bestimmung von P_2 wird ähnlich vorgegangen wie bei P_1 und A als die Matrix mit den Spalten $v_1, \dots, v_k$ gewählt:

$$(A^TA)^{-1} = \begin{pmatrix} 1/n_1 & 0 & \dots & 0 \\ 0 & 1/n_2 & \dots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \dots & \dots & 1/n_k \end{pmatrix} \Longrightarrow (A^TA)^{-1}A^TY = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix} \Longrightarrow$$

$$P_2Y = A\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix} = \mu_1 v_1 + \mu_2 v_2 + \dots + \mu_k v_k =$$
$$= (\mu_1, \dots, \mu_1, \mu_2, \dots, \mu_2, \dots \mu_k, \dots, \mu_k)^T.$$

$\Longrightarrow$ $P_3\, Y = Y$. Somit kann gefolgert werden, dass:

$$||P_2Y - P_1Y||^2 = \sum_{j=1}^{k} n_j(\hat{\mu}_j - \hat{\mu})^2 = L$$

$$||Y - P_2Y||^2 = \sum_{j=1}^{k}\sum_{l \in I_j} (Y_l - \hat{\mu}_j)^2 = F$$

$$||Y - P_1Y||^2 = \sum_{l=1}^{n} (Y_l - \hat{\mu}) = T$$

Nachdem $\langle y - P_2Y, P_2Y - P_1Y\rangle = 0$ ist und $P_2Y - P_1Y \in V_2$, gilt:

$$||Y - P_2Y||^2 = ||Y - P_2Y + P_2Y - P_1Y|| =$$
$$= ||Y - P_2Y||^2 + ||P_2Y - P_1Y||^2 + 2\langle y - P_2Y, P_2Y - P_1Y\rangle =$$
$$= ||Y - P_2Y||^2 + ||P_2Y - P_1Y||^2$$

Somit ist die Behauptung T = L + F bewiesen.

Kapitel 6

6. TESTVERTEILUNGEN & TESTS FÜR VERTEILUNGSFUNKTIONEN

Jene Verteilungen die in der Statistik auftreten, kann man nach ihrem Verwendungszweck in zwei Klassen einteilen:

1. Solche Verteilungen, die in Beziehung zu mathematischen Modellen von Zufallsexperimenten auftreten.
2. Und in Prüf- bzw. Testverteilungen , welche die Basis statistischer Tests bilden.

Im ersten Teilkapitel werden nun zwei besonders wichtige Verteilungen betrachtet, um im Anschluss daran im zweiten Teil des Kapitels die zugehörigen Tests charakterisieren zu können.

6.1 Testverteilungen

6.1.1 Chi-Quadrat-Verteilung. Gammafunktion

Wir betrachten die unabhängigen Zufallsvariablen X_1, X_2, ..., X_n, wobei jede eine Normalverteilung mit Mittelwert 0 und Varianz 1 hat. Die daraus gebildete Summe der Quadrate dieser Variablen bezeichnet man allgemein mit $\chi^2 = \chi_1^2 + \chi_2^2 + \ldots + \chi_n^2$.
Die dazugehörige Verteilung nennt sich die Chi-Quadrat-Verteilung, wobei diese die folgende Wahrscheinlichkeitsdichte hat

$$\mathbf{f(x) = K_n x^{(n-2)/2} e^{-x/2}} \qquad \text{für } x > 0 \qquad (80)$$

und für negative x, f (x) = 0 gilt. Diese Bedingung ist deshalb zulässig, weil in dieser Dichtefunktion x für χ^2 steht. Die Anzahl der Freiheitsgrade wird durch n wiedergegeben und K_n ist eine Konstante.

Dichtefunktion

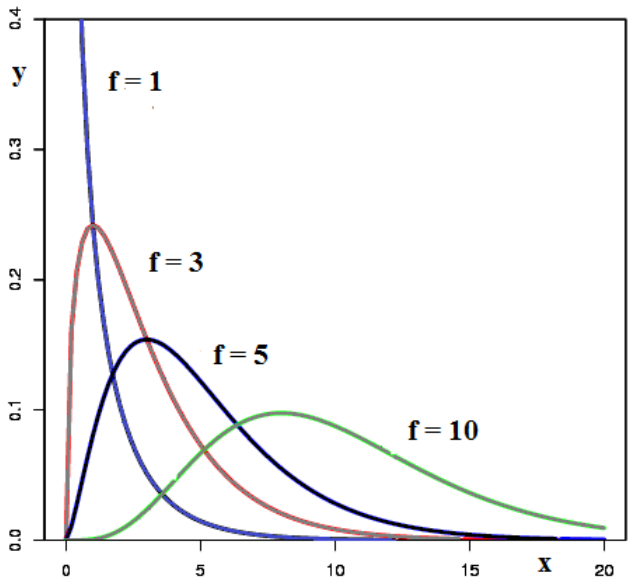

Abbildung 17: Dichtefunktion – Chi-Quadrat-Verteilung

Setzt man für n die Werte n = 1 und 2 ein, so sind die Kurven monoton fallend, während sie für n > 2 ein Maximum bei x = n – 2 haben, ersichtlich aus der f ´ (x) = 0

Aus der Dichtefunktion erhält man die Verteilungsfunktion

$$F(x) = K_n \int_0^x u^{(n-2)/2} e^{-u/2} du \qquad \text{für x } \geq 0$$

Dabei muss die auftretende Konstante K_n so gewählt werden, dass $F(\infty) = 1$ wird, woraus man folgendes erhält:

$$K_n = \frac{1}{2^{\frac{n}{2}}\,\Gamma(\frac{n}{2})} \tag{81}$$

Dabei ist Γ (α) die sogenannte **Gammafunktion,** die definiert ist durch das Integral

$$\Gamma(\alpha) = \int_0^\infty e^{-t} t^{\alpha-1} dt \qquad \alpha > 0 \tag{82}$$

Schreiben wir nun α + 1 statt α und integrieren partiell, so folgt unmittelbar

$$\Gamma(\alpha + 1) = \alpha\, \Gamma\,(\alpha) \tag{83}$$

Beispiel: 3,2 ! = Γ(4,2) = $\int_0^\infty e^{-t} t^{3,2}\, dt$

Nun können wir α = 1 in die Gleichung (6.1.2) einsetzen und integrieren

$$\Gamma(1) = \int_0^\infty e^{-t} dt = 1$$

und dadurch ergibt sich wegen (6.1.3) nacheinander

$$\Gamma(2) = 1 \cdot \Gamma(1) = 1\,!, \qquad \Gamma(3) = 2 \cdot \Gamma(2) = 2\,!$$

bzw. allgemein $\Gamma(\mathbf{n} + \mathbf{1}) = \mathbf{n}!$

Die Gammafunktion ist deshalb eine Verallgemeinerung der elementaren Fakultät. Ist n gerade, so ist demnach in (81)

$$\Gamma\left(\frac{n}{2}\right) = \left(\frac{n}{2} - 1\right)!$$

Übrig bleibt dann der Fall ungerader n

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

Wird nun (83) angewendet, so erhält man der Reihe nach: (siehe Tabelle im Anhang)

$$\Gamma\left(\frac{3}{2}\right) = \frac{1}{2}\Gamma\left(\frac{1}{2}\right) = \frac{1}{2}\sqrt{\pi}, \qquad \Gamma\left(\frac{5}{2}\right) = \frac{3}{2}\Gamma\left(\frac{3}{2}\right) = \frac{3}{4}\sqrt{\pi} \qquad \text{usw.}$$

6.1.2 T – Verteilung von Student

Eine weitere Grundlage wichtiger Tests, ist die sogenannte studentsche T-Verteilung (wurde unter dem Pseudonym „Student“ veröffentlicht). Darunter versteht man die Verteilung der Zufallsvariablen

$$T = \frac{X}{\sqrt{Y/n}}$$

mit n-Freiheitsgraden und X bzw. Y unabhängigen Zufallsvariablen.

Definition: Die Verteilung der Zufallsvariable T_n heißt t-Verteilung mit n-Freiheitsgraden und hat die Wahrscheinlichkeitsdichte (84)

$$f(z) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\ \Gamma\left(\frac{n}{2}\right)} \cdot \frac{1}{\left(1+\frac{z^2}{n}\right)^{(n+1)/2}}$$

(Herleitung siehe Anhang) und die zugehörige Verteilungsfunktion

$$F(z) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\ \Gamma\left(\frac{n}{2}\right)} \cdot \int_{-\infty}^{z} \frac{du}{\left(1+\frac{u^2}{n}\right)^{(n+1)/2}}.$$

Für die Freiheitsgrade n =1 und n =2 hat die T-Verteilung keine Varianz. Für n = 3.4--- ergibt sich aber

$$\sigma^2 = \frac{n}{n-2}.$$

Aus der Abbildung ist nun erkenntlich, dass mit wachsendem n die Verteilungsfunktion der t-Verteilung gegen die Verteilungsfunktion der Normalverteilung mit μ = 0 und σ = 1 strebt.

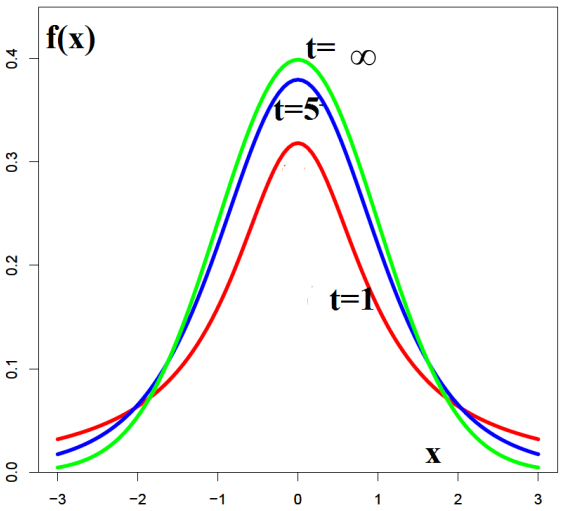

Abbildung 18: Dichte der T – Verteilung

6.1.3 F – Verteilung von Fischer

Definition: V_m und V_n seien zwei stochastisch unabhängige Zufallsvariable, die jeweils Chi-Quadrat verteilt sind mit m bzw. n - Freiheitsgraden. Dann heißt die Zufallsvariable

$$V = \frac{\frac{V_m}{m}}{\frac{V_n}{n}}$$

Fisher verteilt mit (m,n) Freiheitsgraden oder kurz F(m,n)-verteilt.

Satz: Die F(m,n)-Verteilung besitzt die Dichte (85)

$$g_{m,n} = \frac{\Gamma(\frac{m+n}{2})}{\Gamma\left(\frac{m}{2}\right)\Gamma(\frac{n}{2})} \cdot (\frac{m}{n})^{m/2} \cdot \frac{x^{\frac{m}{2}-1}}{(1+\frac{m}{n}x)^{\frac{m+n}{2}}} \qquad \text{für } \mathbf{x \geq 0}$$

Beweis: siehe Anhang

Für x > 0 gilt zum Beispiel für (m,n) = 6,4 bzw. (m, n) = (6, 10)

$g_{6,4}(x) = 12 \cdot 1{,}5^3 \frac{x^2}{(1+1{,}5x)^5}$

$g_{6,10}(x) = 105 \cdot 0{,}6^3 \frac{x^2}{(1+0{,}6x)^5}$

6.2 Tests für Verteilungsfunktionen

Nachdem in den bisherigen Kapiteln diverse Verteilungen bzw. Tests für Verteilungsfunktionen als bekannt vorausgesetzt wurden, sollen diese nun in diesem Abschnitt explizit charakterisiert werden. Es soll hier gezeigt werden, wie man von der Stichprobenverteilung auf die Verteilung der Grundgesamtheit schließen kann. In der Praxis hat man dabei oft eine Vermutung über die Art der Verteilung der Grundgesamtheit, die mit Hilfe einer Stichprobe entweder bestätigt oder widerlegt wird.
Somit testen wir ähnlich wie im Kapitel 4 die Hypothese, dass eine Zufallsvariable X, eine gewisse Verteilungsfunktion F(x) hat. Das empirische Analogon der Funktion F(x) ist also offenbar die Verteilungsfunktion $\tilde{F}(x)$ (Näherungsfunktion von F(x)) einer Stichprobe aus der betreffenden Grundgesamtheit. Um etwas über die Abweichung zwischen F(x) und $\tilde{F}(x)$ aussagen zu können, bedarf es einem Maß für die Abweichung zwischen diesen beiden Variablen. Außerdem muss man die Wahrscheinlichkeitsverteilung des Abweichungsmaßes unter der Annahme, die Hypothese sei richtig kennen, um eine Entscheidung darüber treffen zu können, ob man die Hypothese verwirft oder beibehält.
Im Folgenden werden nun zwei wichtige Testverfahren diskutiert, zum einen der Chi-Quadrat-Test für diskrete als auch stetige Verteilungen und zum anderen der Kolmogoroff-Smirnov-Test für stetige Verteilungen.

6.2.1 Chi-Quadrat-Test

Der Grundgedanke des Chi-Quadrat-Tests besteht darin, die x-Achse in Teilintervalle zu unterteilen, sich anschließend aus der hypothetischen Verteilungsfunktion F(x) die zu diesen Intervallen gehörenden Wahrscheinlichkeiten der betreffenden Zufallsvariablen X auszurechnen und diese dann mit den relativen Klassenhäufigkeiten einer gegebenen Stichprobe zu vergleichen. Sofern die Diskrepanz zu groß ist, wird die Hypothese, F(x) sei die Verteilungsfunktion von X, verworfen.

1. Schritt: Es wird eine Unterteilung der x-Achse in K Intervalle I_1, I_2, ... , I_k derart vorgenommen, dass jedes Intervall zumindest 5 Werte der gegebenen Stichprobe $x_1,...x_n$ enthält. Dann wird für jedes Intervall I_j die Anzahl b_j der Stichprobenwerte bestimmt, die in I_j liegen. Liegen Werte auf den Intervallgrenzen, so zählen sie jeweils zur Hälfte zum einen als auch zum anderen Intervall.

2. Schritt: Es folgt die Berechnung der Wahrscheinlichkeit p_j für jedes Intervall I_j aus F(x), mit der die betreffende Zufallsvariable irgendeinen Wert aus I_j annimmt. Daraus kann die Anzahl der theoretisch in I_j zu erwartenden Stichprobenwerte

$$e_j = np_j$$

berechnet werden.

3.Schritt: Berechnung der Abweichung

$$\chi_0^2 = \sum_{j=1}^{k} \frac{(b_j - e_j)^2}{e_j}$$

4. Schritt: Man wählt eine Signifikanzzahl α und bestimmt die Lösung c der Gleichung

$$P(\chi^2 \leq c) = 1 - \alpha$$

durch Ablesen aus der Tafel der Chi-Quadrat Verteilung mit K – 1 Freiheitsgraden. Wenn $\chi_0^2 \leq c$ ist, wird die Hypothese angenommen, andernfalls verwirft man sie.

Beispiel zum Chi-Quadrat-Test (Mendelsche Gesetze)

G. MENDEL erhielt bei seinen allseits bekannten Kreuzungsversuchen an zehn Erbsenpflanzen insgesamt 355 gelbe und 123 grüne Erbsen. Zu Testen ist nun, ob das für oder gegen die Mendelsche Theorie spricht, nach welcher sich gelb : grün wie 3 : 1 verhalten sollte.

1. Schritt: Für die beiden möglichen Ereignisse legen wir zum Beispiel fest

X = 0 (*gelbe Erbse*) und X = 1 (*grüne Erbse*)

Anschließend bestimmt man K = 2 Intervalle so, dass jedes Intervall eines von beiden Ereignissen enthält. Somit ist dann $b_1 = 355$ und $b_2 = 123$.

2. Schritt: Es ist n = 355 + 123 = 478 und wir erhalten

$$e_1 = 478 \cdot \frac{3}{4} = 358{,}5 \qquad \text{bzw.} \quad e_2 = 478 \cdot \frac{1}{4} = 119{,}5$$

3. Schritt: Dann berechnet man die Abweichung

$$\chi_0^2 = \frac{(355 - 358{,}5)^2}{358{,}5} + \frac{(123 - 119{,}5)^2}{119{,}5} = 0{,}137$$

4. Schritt: Schließlich hat die Gleichung

$$P(\chi^2 \leq c) = 1 - \alpha = 0{,}95$$

für die Signifikanzzahl α = 5%, die Lösung c =3,84. Nachdem $\chi_0^2 < c$ wird die Hypothese angenommen.

6.2.2 Kolmogoroff-Smirnov-Test

Der Kolmogoroff-Smirnov Test eignet sich im Gegensatz zum Chi-Quadrat Test nur für stetige Verteilungen. Wiederum gibt es eine Funktion F(x), die Verteilungsfunktion einer Grundgesamtheit, aus der eine Stichprobe entnommen wurde und für die es darum geht, eine Hypothese zu testen ist.

1. Schritt: Berechnung der Werte der stückweise konstanten Verteilungsfunktion $\tilde{F}(x)$ der Stichprobe $x_1, ..., x_n$.

2. Schritt: Bestimmung der Maximalabweichung

$$a = \max |\tilde{F}(x) - F(x)| \quad \text{bzw. genauer } sup\, |\tilde{F}(x) - F(x)|$$

zwischen $\tilde{F}(x)$ und F(x)

3.Schritt: Bei vorgegebener Signifikanzzahl α bestimmt man die Lösung c der Gleichung

$$P(A \leq c) = 1 - \alpha$$

aus der dem Stichprobenumfang *n* entsprechenden Zeile der Tafel für den Kolmogoroff-Smitnov Test im Anhang. Die hypothese wird angenommen, falls $a \leq c$ zutrifft.

Beispiel zum Kolmogoroff-Smirnoff-Test (zugehörige Tabelle 12, siehe Anhang)

Es ist zu überprüfen, ob die Stichprobe der Tabelle unten einer Normalverteilung mit Mittelwert μ = 165,05 cm und der Varianz $\sigma^2 = 34{,}31\ \text{cm}^2$ entspricht.
($\bar{x} = 165{,}05\ und\ s = \sqrt{34{,}31} = 5{,}86$)

1. Schritt: Die Werte der Verteilungsfunktion $\tilde{F}(x)$ der Stichprobe in der Tabelle erhält man durch Summenbildung der Spalte mit den relativen Häufigkeiten

2. Schritt: Nun muss man testen, ob die Grundgesamtheit die Verteilungsfunktion

$$F(x) = \Phi(\frac{x - 165{,}05}{5{,}86})$$

hat, deren Werte aus der Tafel im Anhang stammen. Anschließend werden a_1 und a_2 berechnet. Zum Beispiel für die zweite Zeile:

$$a_1 = F(154) - \tilde{F}(153) = 0{,}03 - 0{,}01 = 0{,}02$$
$$a_2 = F(154) - \tilde{F}(154) = 0{,}03 - 0{,}02 = 0{,}01$$

3. Schritt: Bei gegebener Signifikanzzahl α = 5 % und einem Stichprobenumfang von n = 100, entnimmt man als Lösung der Gleichung aus der Tafel im Anhang

$P(A \leq c) = 1 - \alpha = 0{,}95$ die Zahl **c = 0,134** **(→ Hypothese wird angenommen)**

Kapitel 7

7. EIGENWERTPROBLEM & ORTHOGONALITÄT

Dieses Kapitel verfolgt das Ziel, lineare Transformationen x → Ax in einzelne, leicht visualisierbare Elemente aufzugliedern. Die Hauptanwendungen der hier aufgegriffen Basiskonzepte – Eigenwerte und Eigenvektoren – beziehen sich auf diskrete dynamische Modelle, deren Anwendung sogar in Situationen fernab der Mathematik nützlich erscheint.

7.1 Eigenwerte & Eigenvektoren

Obwohl durch Transformationen **x** → **Ax** Vektoren in beliebige Richtungen gedreht werden können, ist es oft der Fall, dass spezielle Vektoren existieren, für die Transformationen durch A besonders günstig sind.

Beispiel 1

$$A = \begin{pmatrix} 3 & -2 \\ 1 & 0 \end{pmatrix}, u = \begin{pmatrix} -1 \\ 1 \end{pmatrix} \text{ und } v = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

Die Bilder von v und u durch Multiplikation von A werden in der Abbildung darunter gezeigt. Tatsache ist, dass Av gerade 2v ist und A dadurch nur v „streckt".

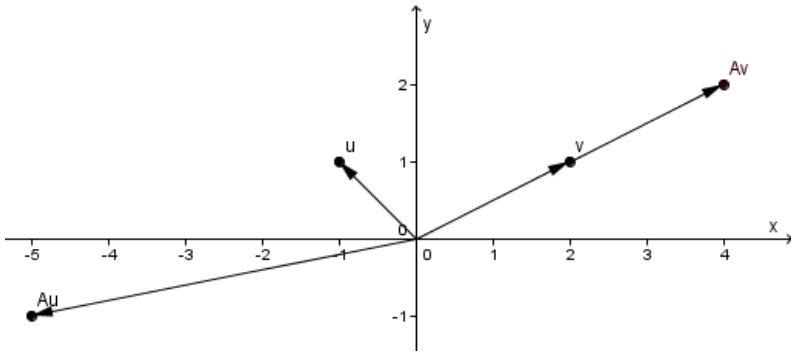

Abbildung 19: Bilder von v und u durch Multiplikation von A

Sofern nun A eine stochastische Matrix ist, erfüllt der stationäre Vektor **q** für A die Gleichung Ax = x. Dieser ist, Aq = 1· q.
In diesem Abschnitt werden folglich derartige Gleichungen wie in etwa

$$A\mathbf{x} = 2\mathbf{x} \quad \text{oder} \quad A\mathbf{x} = -4\mathbf{x}$$

diskutiert und wir werden nach solchen Vektoren Ausschau halten, die durch A in ein Skalarprodukt von sich selbst transformiert werden.

Definition: Ein Eigenvektor einer n x n Matrix A ist ein von null verschiedener Vektor x so dass Ax = λx für irgendein Skalar λ. Ein Skalar λ wird als Eigenwert von A bezeichnet, wenn eine nichttriviale Lösung x von A**x** = λ**x** vorliegt; ein derartiges x ist ein *Eigenvektor der mit λ korrespondiert*. (86)

Beispiel 2

Seien A = $\begin{pmatrix}1 & 6\\ 5 & 2\end{pmatrix}, u = \begin{pmatrix}6\\ -5\end{pmatrix} und\ v = \begin{pmatrix}3\\ -2\end{pmatrix}$. Sind u und v Eigenvektoren von A ?
→

$$\mathbf{Au} = \begin{pmatrix}1 & 6\\ 5 & 2\end{pmatrix}\begin{pmatrix}6\\ -5\end{pmatrix} = \begin{pmatrix}-24\\ 20\end{pmatrix} = -4\begin{pmatrix}6\\ -5\end{pmatrix} = -4\,\mathbf{u}$$

$$\mathbf{Av} = \begin{pmatrix}1 & 6\\ 5 & 2\end{pmatrix}\begin{pmatrix}3\\ -2\end{pmatrix} = \begin{pmatrix}-9\\ 11\end{pmatrix} \neq \lambda\begin{pmatrix}3\\ -2\end{pmatrix}$$

Deshalb ist u ein Eigenvektor, der mit dem Eigenwert -4 korrespondiert, während v kein Eigenvektor von A ist.

Beispiel 3
Hier soll gezeigt werden, dass 7 ein Eigenwert der Matrix A ist, um anschließend die korrespondierenden Eigenvektoren bestimmen zu können.

Damit 7 ein Eigenwert ist, muss A**x** = 7**x** eine nichttriviale Lösung haben.

Diese Gleichung ist allerdings äquivalent zu (A – 7 I) **x = 0**

$$\text{A} - 7\text{I} = \begin{pmatrix}1 & 6\\ 5 & 2\end{pmatrix} - \begin{pmatrix}7 & 0\\ 0 & 7\end{pmatrix} = \begin{pmatrix}-6 & 6\\ 5 & -5\end{pmatrix}$$

Die Spalten von A – 7I sind offensichtlich linear abhängig, somit hat (A – 7I) **x = 0** nichttriviale Lösungen und 7 ist ein Eigenwert von A. Um die dazu korrespondierenden Eigenvektoren zu finden, werden nun Zeilenoperationen angewendet:

$$\begin{pmatrix}-6 & 6 & 0\\ 5 & -5 & 0\end{pmatrix} \sim \begin{pmatrix}1 & -1 & 0\\ 0 & 0 & 0\end{pmatrix}$$

Die allgemeine Lösung hat die Form $y \cdot \begin{pmatrix}1\\ 1\end{pmatrix}$. Jeder Vektor dieser Form, mit $y \neq 0$ ist ein Eigenvektor der mit λ = 7 korrespondiert.

Insofern ist λ ein Eigenwert von A, dann und nur dann, wenn die Gleichung $(A-\lambda I)x = 0$ eine nichttriviale Lösung hat. Das Set mit allen Lösungen dieser Gleichung ist gerade der Nullraum der Matrix A –λI, welches ein Unterraum von R^n ist, den man *Eigenraum von A bezüglich λ* nennt. Der Eigenraum besteht aus dem Nullvektor und allen Eigenvektoren die mit λ korrespondieren.

Beispiel 3 zeigt dies für die Matrix A im Beispiel 2, wobei der Eigenraum der sich auf λ = 7 bezieht, aus allen Vielfachen vom Vektor (1,1) besteht (Linie durch (1,1) und Ursprung). Aus Beispiel 2 ergibt sich, dass der Eigenraum der mit λ = -4 korrespondiert

die Linie durch (6, -5) ist. Diese Eigenräume werden in der Abbildung 20 gezeigt und zwar zusammen mit den Eigenvektoren (1, 1) und (3/2, - 5/4) und der geometrischen Handlung der Transformation x → A x auf jeden Eigenraum.

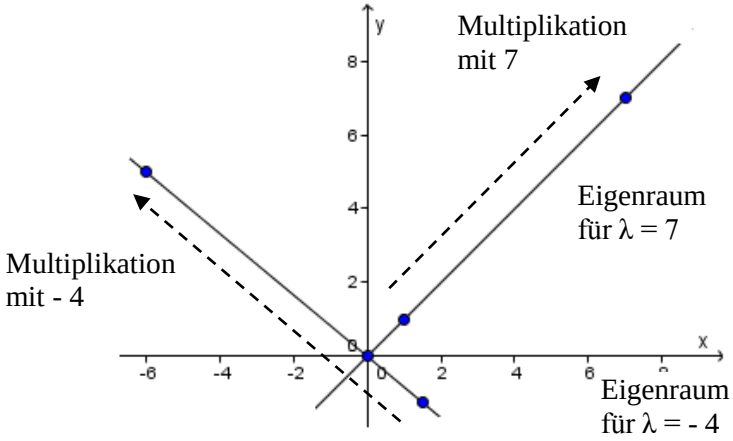

Abbildung 20: Eigenräume zum Beispiel 3

Beispiel 4

Wir wählen A = $\begin{pmatrix} 4 & -1 & 6 \\ 2 & 1 & 6 \\ 2 & -1 & 8 \end{pmatrix}$ und nehmen einen Eigenwert von A mit 2 an. Ziel ist es nun eine Basis für den korrespondierenden Eigenraum zu finden:

$$A - 2I = \begin{pmatrix} 4 & -1 & 6 \\ 2 & 1 & 6 \\ 2 & -1 & 8 \end{pmatrix} - \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 2 & -1 & 6 \\ 2 & -1 & 6 \\ 2 & -1 & 6 \end{pmatrix}$$

$$\rightarrow \quad \begin{pmatrix} 2 & -1 & 6 & 0 \\ 2 & -1 & 6 & 0 \\ 2 & -1 & 6 & 0 \end{pmatrix} \sim \begin{pmatrix} 2 & -1 & 6 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Somit haben wir uns davon überzeugt, dass 2 tatsächlich ein Eigenwert von A ist, weil die Gleichung (A-2I)x = 0 frei Variablen hat. Die allgemeine Lösung ist

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = y \begin{pmatrix} 0{,}5 \\ 1 \\ 0 \end{pmatrix} + z \begin{pmatrix} -3 \\ 0 \\ 1 \end{pmatrix} \quad \text{, y und z sind frei wählbar}$$

Der Eigenraum ist ein zweidimensionaler Teilraum des R^3. Eine Basis ist dadurch gegeben mit:

$$\left\{ \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} -3 \\ 0 \\ 1 \end{pmatrix} \right\}$$

Theorem 1: Die Eigenwerte einer Dreiecksmatrix sind die Einträge auf der Hauptdiagonalen.

Beweis: Zur Vereinfachung betrachten wir den 3 x 3 Fall. Wenn A eine obere Dreiecksmatrix ist, so hat A – λI die Form

$$A - \lambda I = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{pmatrix} - \begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix}$$

$$= \begin{pmatrix} a_{11} - \lambda & a_{12} & a_{13} \\ 0 & a_{22} - \lambda & a_{23} \\ 0 & 0 & a_{33} - \lambda \end{pmatrix}$$

Das Skalar λ ist ein Eigenwert von A, wenn die Gleichung $(A - \lambda I)\, x = 0$ eine nichttriviale Lösung hat. Dieser Fall tritt auf, wenn die Gleichung freie Variable aufweist. Nachdem A – λI Nulleinträge hat, sieht man, dass $(A - \lambda I)x = 0$ freie Variable hat, wenn zumindest ein Diagonaleintrag gleich null ist. Dieser Fall tritt nur dann auf, wenn λ gleich einem der Einträge a_{11}, a_{22} oder a_{33} ist.

Theorem 2: Wenn v_1, ..., v_r Eigenvektoren sind, die mit individuellen Eigenwerten $\lambda_1,...,\lambda_r$ einer n x n Matrix A korrespondieren, dann ist das Set $\{v_1,...,v_r\}$ linear unabhängig.

Beweis: Wenn $\{v_1, ..., v_r\}$ linear abhängig ist, so gibt es einen minimalen Index p so dass v_{p+1} eine Linearkombination der vorangehenden (linear unabhängigen) Vektoren ist, und es existieren Skalare $c_1, ..., c_p$ so dass

$$\text{(a)} \quad c_1 v_1 + \cdots + c_p v_p = v_{p+1} \quad \text{gilt.}$$

Multipliziert man nun beide Seiten von (a) mit A und verwendet die Tatsache dass $Av_k = \lambda_k v_k$ für jedes k ist, so erhalten wir

$$\text{(b)} \quad \begin{aligned} c_1 A v_1 + \ldots + c_p A v_p &= A v_{p+1} \\ c_1 \lambda_1 v_1 + \ldots + c_p \lambda_p v_p &= \lambda_{p+1} v_{p+1} \end{aligned}$$

Multipliziert man beide Seiten von (a) mit λ_{p+1} und zieht das Ergebnis von (b) ab, so haben wir

$$\text{(c)} \quad c_1(\lambda_1 - \lambda_{p+1})v_1 + \ \ldots + c_p(\lambda_p - \lambda_{p+1})v_p = 0$$

Nachdem $\{v_1,...,v_p\}$ linear unabhängig ist, sind alle Werte $c_i = 0$, aber keiner der Faktoren $\lambda_i - \lambda_{p+1}$ ist, aufgrund der unterschiedlichen Eigenwerte. Allerdings sagt (a) aus, dass $v_{p+1} = 0$ ist, was unmöglich ist. Darum kann v_1 ... v_r nicht linear abhängig sein und ist deshalb linear unabhängig.

7.2 Die charakteristische Gleichung

Nützliche Informationen über die Eigenwerte einer quadratischen Matrix A sind verschlüsselt in einer speziellen Skalargleichung, die man charakteristische Gleichung von A nennt

7.2.1 Determinanten

Wir betrachten A als eine n x n Matrix und erhalten eine geeignete Treppenform U dieser Matrix durch Gauß – Elimination (k ist die Anzahl der Zeilen - Vertauschungen). Bei auftretenden Zeilenvertauschungen im Eliminationsverfahren ist zusätzlich zur

Multiplikation der Diagonalelemente der Treppenform noch der Faktor $(-1)^k$ hinzu zu multiplizieren.

$$\rightarrow \det A = \begin{cases} (-1)^r \cdot (\text{Pivotprodukt von U}) & \text{wenn A invertierbar} \\ 0 & \text{wenn A nicht invertierbar} \end{cases}$$

Theorem 3: Eigenschaften von Determinanten

Seien A eine n x n Matrix, dann gilt:

a) A ist invertierbar, dann und nur dann wenn **det A ≠ 0**

b) det AB =(det A)(det B)

c) det A^T = det A

d) Ist A eine Dreiecksmatrix, dann ist det A das Produkt der Einträge auf der Hauptdiagonale

e) Ein Zeilentausch von A ändert die Determinante nicht. Multiplikation einer Zeiel mit einem Skalar führt jedoch dazu, das die Determinante mit dem selben faktor multipliziert wird.

Theorem 4: Wenn n x n Matrizen ähnlich sind, haben sie das selbe charakteristische Polynom und daher dieselben Eigenwerte (mit den selben Vielfachheiten).

Beweis
Wenn $B = P^{-1} A P$, so ist

$$B - \lambda I = P^{-1}AP - \lambda P^{-1}P = P^{-1}(AP - \lambda P) = P^{-1}(A - \lambda I)P$$

Wir verwenden Eigenschaft (b) aus Theorem 3 und berechnen

$$\begin{aligned} det(B - \lambda I) &= \det[P^{-1}(A - \lambda I)\, P\,] \\ &= \det(P^{-1}) \cdot \det(A - \lambda I) \cdot \det(P) \end{aligned}$$

Nachdem $\det(P^{-1}) \cdot \det(P) = \det(P^{-1}P) = \det I = I$ ist, ist tatsächlich
$\det(B - \lambda I) = \det(A - \lambda I)$

7.3 Diagonalisierung

In vielen Fällen kann die Eigenwert – Eigenvektor Information die in einer Matrix A enthalten ist, in einer nützlichen Faktorisierung der Form $A = PDP^{-1}$ gezeigt werden. Mit dieser Faktorisierung gelingt es, A^k für große Werte von k möglichst schnell zu berechnen und ist somit eine fundamentale Idee in einigen Anwendungen der linearen Algebra.

Beispiel: Gegeben ist $A = \begin{pmatrix} 7 & 2 \\ -4 & 1 \end{pmatrix}$ und es soll eine Formel A^k gefunden werden, so dass $A = PDP^{-1}$ gilt,

$$\text{mit} \quad P = \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix} \quad \text{und} \quad D = \begin{pmatrix} 5 & 0 \\ 0 & 3 \end{pmatrix}$$

→ Die Standardformel für die Inverse einer 2 x 2 Matrix ergibt

$$P^{-1} = \begin{pmatrix} 2 & 1 \\ -1 & -1 \end{pmatrix}$$

Anschließend verwenden wir die Assoziativität der Matrixmultiplikation:

$$A^2 = (PDP^{-1})(PDP^{-1}) = PD(P^{-1}P)DP^{-1} = PDDP^{-1}$$

$$= PD^2P^{-1} = \begin{pmatrix} 1 & 1 \\ -1 & -2 \end{pmatrix}\begin{pmatrix} 5^2 & 0 \\ 0 & 3^2 \end{pmatrix}\begin{pmatrix} 2 & 1 \\ -1 & -1 \end{pmatrix}$$

Im Allgemeinen gilt für $k \geq 1$

$$A^k = PD^kP^{-1} = \begin{pmatrix} 1 & 1 \\ -1 & -2 \end{pmatrix}\begin{pmatrix} 5^k & 0 \\ 0 & 3^k \end{pmatrix}\begin{pmatrix} 2 & 1 \\ -1 & -1 \end{pmatrix}$$

$$= \begin{pmatrix} 2 \cdot 5^k - 3^k & 5^k - 3^k \\ 2 \cdot 3^k - 2 \cdot 5^k & 2 \cdot 3^k - 5^k \end{pmatrix}$$

Theorem 5: Das Diagonalisierungstheorem
Eine n x n Matrix A ist diagonalisierbar, dann und nur dann, wenn A n linear unabhängige Eigenvektoren hat. Tatsächlich trifft $A = PDP^{-1}$, mit der Diagonalmatrix D, nur zu, wenn die Spalten von P n linear unabhängige Eigenvektoren von A sind. In diesem Fall sind die Diagonaleinträge von D Eigenwerte von A die zu mit den Eigenvektoren in P korrespondieren.

7.4 Orthogonalprojektionen und ihre Anwendung bei der Methode der kl. Quadrate

Ein Set von Vektoren $\{u_1, ..., u_p\}$ im R^n wird als *orthogonales Set* bezeichnet, wenn jedes Paar unterschiedlicher Vektoren des Sets orthogonal ist, sodass $u_i \cdot u_j = 0$ (mit $i \neq j$).

Theorem 6: Wenn $S = \{u_1, ..., u_p\}$ ein orthogonales Set von Vektoren ungleich 0 im R^n ist, so ist S linear unabhängig und daher eine Basis für den Unterraum der durch S aufgespannt iwrd.

Beweis: Wenn $0 = c_1u_1 + ... + c_pu_p$ für einige Skalare $c_1, ..., c_p$, dann ist

$$\begin{aligned} 0 = \mathbf{0} \cdot \mathbf{u_1} &= (c_1\mathbf{u}_1 + c_2\mathbf{u}_2 + ... + c_p\mathbf{u}_p) \cdot \mathbf{u_1} \\ &= (c_1\mathbf{u}_1) \cdot \mathbf{u}_1 + (c_2\mathbf{u}_2) \cdot \mathbf{u}_1 + ... + (c_p\mathbf{u}_p) \cdot \mathbf{u}_1 \\ &= c_1(\mathbf{u}_1 \cdot \mathbf{u}_1) + c_2(\mathbf{u}_2 \cdot \mathbf{u}_1) + ... + c_p(\mathbf{u}_p \cdot \mathbf{u}_1) \\ &= c_1(\mathbf{u}_1 \cdot \mathbf{u}_1) \end{aligned}$$

weil u_1 orthogonal zu $u_2, ..., u_p$ ist. Nachdem u_1 ungleich null, $u_1 \cdot u_1$ nicht null und $c_1 = 0$ ist. Ähnlich dazu muss $c_2, ..., c_p$ null sein, wodurch S linear unabhängig ist.

Definition: Eine Orthogonalbasis für einen Unterraum W des R^n ist eine Basis für W, die außerdem ein orthogonales Set bildet. (87)

Theorem 7: Sei $\{u_1, ..., u_p\}$eine Orthogonalbasis für einen Unterraum W des R^n. Dann wird jedes y in W eindeutig als Linearkombination der $u_1,..., u_p$ repräsentiert. Tatsächlich gilt, wenn

$$y = c_1u_1 + ... + c_pu_p$$

dass $c_j = \frac{y \cdot u_j}{u_j \cdot u_j}$ (mit j = 1, ..., p) ist.

Beweis: Ähnlich wie im vorhergehenden Beispiel, zeigt die Orthogonalität von $\{u_1, ..., u_p\}$ dass $\mathbf{y} \cdot \mathbf{u}_1 = (c_1\mathbf{u}_1 + c_2\mathbf{u}_2 + ... + c_p\mathbf{u}_p) \cdot \mathbf{u}_1 = c_1 \cdot (\mathbf{u}_1 \cdot \mathbf{u}_1)$
Nachdem $u_1 \cdot u_1$ ungleich null ist, kann die Gleichung für c_1 gelöst werden. Um c_j für j = 2, ..., p zu finden, berechnet man $y \cdot u_j$ und löst für c_j auf.

Die Orthogonalprojektion

Nun sei ein Vektor u (im R^n) ungleich null gegeben und wir betrachten das Problem, dass der Vektor y in eine Summe von zwei Vektoren, einer ein Vielfaches von u und der andere orthogonal zu u, so dass

$$y = \hat{y} + z \quad (88)$$

wobei $\hat{y} = \alpha u$ für ein Skalar α und z irgendein orthogonaler Vektor zu u ist.

$\rightarrow 0 = (\mathbf{y} - \alpha\mathbf{u}) \cdot \mathbf{u} = \mathbf{y} \cdot \mathbf{u} - (\alpha\,\mathbf{u}) \cdot \mathbf{u} = y \cdot u - \alpha\,(\mathbf{u} \cdot \mathbf{u})$

Deshalb ist $\alpha = \frac{y \cdot u}{u \cdot u}$ und somit $\hat{y} = \frac{y \cdot u}{u \cdot u}\, u$ die Orthogonalprojektion von y auf u
und $z = y - \frac{y \cdot u}{u \cdot u} u$ der orthogonale Bestandteil von y zu u.

Beispiel: Gegeben seien $u_1 = \begin{pmatrix} 2 \\ 5 \\ -1 \end{pmatrix}$, $u_2 = \begin{pmatrix} -2 \\ 1 \\ 1 \end{pmatrix}$ und $y = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$.

Wir beobachten, dass $\{u_1, u_2\}$ eine Orthogonalbasis für W = Span $\{u_1, u_2\}$ist und und schreiben y als die Summe eines Vektors in W und eines Vektors orthogonal zu W.

$$\rightarrow \quad \hat{y} = \frac{y \cdot u_1}{u_1 \cdot u_1} + \frac{y \cdot u_2}{u_2 \cdot u_2} u_2 = \frac{9}{30} \begin{pmatrix} 2 \\ 5 \\ -1 \end{pmatrix} + \frac{3}{6} \begin{pmatrix} -2 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -2/5 \\ 2 \\ 1/5 \end{pmatrix}$$

Also

$$y - \hat{y} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} - \begin{pmatrix} -2/5 \\ 2 \\ 1/5 \end{pmatrix} = \begin{pmatrix} 7/5 \\ 0 \\ 14/5 \end{pmatrix}$$

Nun können wir uns davon überzeugen, dass y - $\hat{y}$ tatsächlich orthogonal zu u_1 und u_2 ist. Die gewünschte Zerlegung von y ist somit $y = \begin{pmatrix}1\\2\\3\end{pmatrix} = \begin{pmatrix}-2/5\\2\\1/5\end{pmatrix} + \begin{pmatrix}7/5\\0\\14/5\end{pmatrix}$

Theorem 8: Eine m x n Matrix U hat orthonormale Spalten, dann und nur dann, wenn $U^T U = I$.

7.5 Das Gram Schmidt Verfahren

Das Gram Schmidt Verfahren ist ein einfacher Algorithmus, um eine Orthogonal- oder Orthonormalbasis für beliebige Unterräume ($\neq 0$) des R^n zu erzeugen.
Ist eine Basis $\{x_1, ..., x_p\}$ für einen Unterraum W des R^n gegeben, so definiert man das Gram Schmidt Verfahren wie folgt: (89)

$$
\begin{aligned}
v_1 &= x_1 \\
v_2 &= x_2 - \frac{x_2 \cdot v_2}{v_1 \cdot v_1} v_1 \\
v_3 &= x_3 - \frac{x_3 \cdot v_1}{v_1 \cdot v_1} v_1 - \frac{x_3 \cdot v_2}{v_2 \cdot v_2} v_2 \\
&\vdots \\
v_p &= x_p - \frac{x_p \cdot v_1}{v_1 \cdot v_1} v_1 - \frac{x_p \cdot v_2}{v_2 \cdot v_2} v_2 - \cdots - \frac{x_p \cdot v_{p-1}}{v_{p-1} \cdot v_{p-1}} v_{p-1}
\end{aligned}
$$

Beispiel: Gegeben sind die beiden Vektoren $x_1 = \begin{pmatrix}3\\6\\0\end{pmatrix}$ und $x_2 = \begin{pmatrix}1\\2\\2\end{pmatrix}$ mit W = Span $\{x_1, x_2\}$ und wir konstruieren nun eine Orthogonalbasis $\{v_1, v_2\}$ für W.

→ Die Komponente von x_2 orthogonal zu x_1 ist $x_2 - p$, (p ist die Projektion von x_2 auf x_1) und liegt in W, weil sie durch x_2 und einem Vielfachen von x_1 erzeugt wird. $x_1 = v_1$.

$$
\begin{aligned}
v_2 &= x_2 - p = x_2 - \frac{x_2 \cdot x_1}{x_1 \cdot x_1} \cdot x_1 \\
&= \begin{pmatrix}1\\2\\2\end{pmatrix} - \frac{15}{45}\begin{pmatrix}3\\6\\0\end{pmatrix} = \begin{pmatrix}0\\0\\2\end{pmatrix}
\end{aligned}
$$

Somit ist $\{v_1, v_2\}$ schließlich ein orthogonales Set von Vektoren ungleich null in W

Und eine Orthonormalbasis ergibt sich durch Division von $\{v_1, v_2\}$ durch die Norm:

$$
u_1 = \frac{1}{||v_1||} v_1 = \frac{1}{\sqrt{45}}\begin{pmatrix}3\\6\\0\end{pmatrix} = \begin{pmatrix}1/\sqrt{5}\\2/\sqrt{5}\\0\end{pmatrix}
$$

$$
u_2 = \frac{1}{||v_2||} v_2 = \begin{pmatrix}0\\0\\1\end{pmatrix}
$$

Diese Orthonormalbasen bilden im Wesentlichen die Voraussetzung für eine **QR – Faktorisierung** (sofern die Matrix A m x n unabhängige Spalten hat), welche die Matrix A in die Faktoren Q · R zerlegt. Zuerst bildet man bei diesem Verfahren die Orthonormalbasis, wie eben gezeigt und R ist eine invertierbare n x n Dreiecksmatrix mit positiven Einträgen auf der Diagonale (mit A = QR):

$$\mathbf{Q^T A = Q^T (QR) = IR = R}$$

Auf Anwendungen und Beispiele dieser Faktorisierung wird an dieser Stelle verzichtet, jedoch werden wir im nächsten Unterkapitel nochmal auf diese praktische Form der Faktorisierung zurück kommen.

7.6 Anwendungen auf Kleinste-Quadrate-Probleme

Dieses Unterkapitel greift teilweise bereits besprochene Inhalte auf und versucht nochmal einen Gesamtüberblick über „Lineare Algebra“ im Kontext der „Methode der kleinsten Quadrate“ zu geben, ehe in den letzten beiden Kapiteln explizit die Hauptfaktorenanalyse samt ihrem Kern der Singulärwertzerlegung näher erläutert wird. Ausgehend vom Problem, dass Ax = b keine Lösung hat, obwohl eine solche gesucht wird, ist es das Beste, ein solches x zu finden, welches Ax so gut wie möglich an b annähert. Je kleiner der Abstand zwischen b und Ax, gegeben durch || b – Ax||, desto besser die Approximation.

Definition: Sei A eine m x n Matrix und b im R^m, so ist eine kleinste-Quadrate Lösung von Ax = b ein $\hat{x}$ im R^n so dass für alle x im R^n gilt:

$$\left|\left|b - A\hat{x}\right|\right| \leq ||b - Ax|| \qquad (90)$$

Dabei spielt es keine Rolle welches x man wählt, der Vektor Ax liegt notwendigerweise im Spaltenraum von A und darum suchen wir ein x, so dass Ax der nahste Punkt des Spaltenraumes A zu **b** ist.

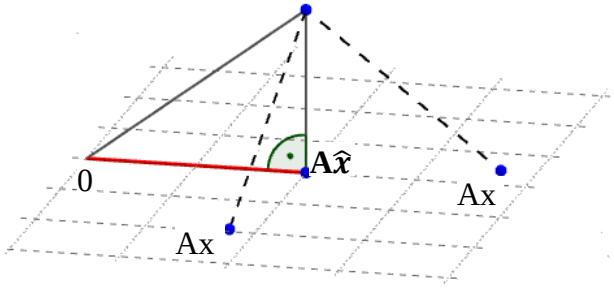

Abbildung 21: b ist näher zu A$\hat{x}$ als zu Ax für andere x

Bei gegebenem **A** und **b** wie oben, fügen wir nun die „Beste Näherung“ zum Unterraum der Spalten von **A** hinzu. Sei dazu:

$$\hat{\mathbf{b}} = \text{proj}_{\text{Spalten A}}\,\mathbf{b}$$

und weil $\hat{b}$ der nahste Punkt im Spaltenraum A ist, ist die Gleichung A$\hat{x}$ = $\hat{b}$ konsistent und es existeirt ein $\hat{x}$ im R^n so dass gilt: $A\hat{\mathbf{x}} = \hat{\mathbf{b}}$

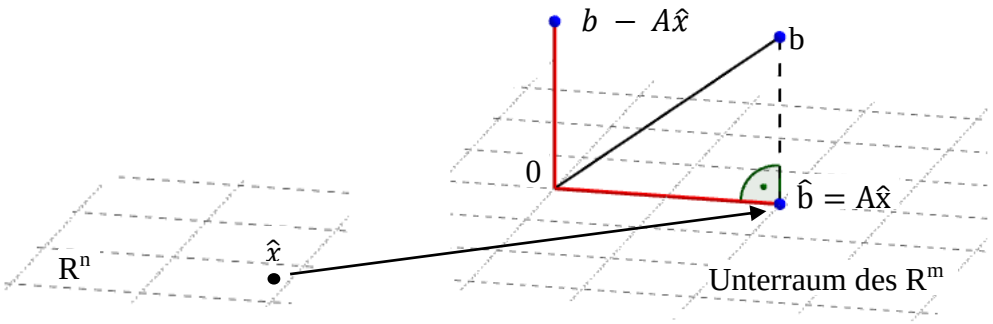

Abbildung 22: Die kleinste Quadrate Lösung $\hat{x}$ liegt im R^m

→
$$A^T(b - A\hat{x}) = 0$$
$$A^T b - A^T A\hat{x} = 0$$
$$A^T A\,\hat{x} = A^T b$$

Das Set von kleinste Quadrate Lösungen von Ax = b deckt sich mit dem „nicht leeren" Set von Normalgleichungen $A^T A\,\hat{x} = A^T b$.

Beispiel:
Gesucht ist eine kleinste – Quadrate Lösung des inkonsistenten Systems Ax = b

$$A = \begin{pmatrix} 4 & 0 \\ 0 & 2 \\ 1 & 1 \end{pmatrix},\quad b = \begin{pmatrix} 2 \\ 0 \\ 11 \end{pmatrix}$$

Nun berechnet man:

$$A^T A = \begin{pmatrix} 4 & 0 & 1 \\ 0 & 2 & 1 \end{pmatrix}\begin{pmatrix} 4 & 0 \\ 0 & 2 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 17 & 1 \\ 1 & 5 \end{pmatrix}$$

$$A^T b = \begin{pmatrix} 4 & 0 & 1 \\ 0 & 2 & 1 \end{pmatrix}\begin{pmatrix} 2 \\ 0 \\ 11 \end{pmatrix} = \begin{pmatrix} 19 \\ 11 \end{pmatrix}$$

Nun wird die Gleichung $A^T A\mathbf{\hat{x}} = A^T\mathbf{b}$ zu

$$\begin{pmatrix} 17 & 1 \\ 1 & 5 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 19 \\ 11 \end{pmatrix}$$

und $(A^T A)^{-1}$ ergibt: $(A^T A)^{-1} = \frac{1}{84}\begin{pmatrix} 5 & -1 \\ -1 & 17 \end{pmatrix}$ woraus $\mathbf{\hat{x}} = (A^T A)^{-1} A^T b$ folgt.

$$\mathbf{\hat{x}} = \frac{1}{84}\begin{pmatrix} 5 & -1 \\ -1 & 17 \end{pmatrix}\begin{pmatrix} 19 \\ 11 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

Alternativ können bei orthogonalen Spalten von A Berechnungen der kleinste Quadrate Lösungen natürlich auch mit Orthonormalbasen, dem Gram Schmidt Verfahren bzw. der A = QR Faktorisierung (Ax = b wird zu $\hat{x} = R^{-1}Q^T b$ mit $A = QR$) erfolgen.

Kapitel 8

8. SYM. MATRIZEN & QUAD. FORMEN

Symmetrische Matrizen treten in Anwendungen öfter auf, als jede andere Hauptklasse von Matrizen. Die Diagonalisierung einer symmetrischen Matrix, die nun im Kapitel 7.1 diskutiert wird, bildet die Grundlage für weitergehende Diskussionen in den Abschnitten 7.2 und 7.3, die quadratischen Formen betreffend.

8.1 Diagonalisierung symmetrischer Matrizen

Eine symmetrische Matrix, ist eine Matrix A mit $A^T = A$, die notwendigerweise quadratisch ist. Die Einträge der Hauptdiagonale sind willkürlich, aber die anderen Einträge treten paarweise auf – und zwar auf entgegengesetzten Seiten der Hauptdiagonale.

Beispiel

Gegeben sei $A = \begin{pmatrix} 6 & -2 & -1 \\ -2 & 6 & -1 \\ -1 & -1 & 5 \end{pmatrix}$ mit den Eigenwerten und Eigenvektoren:

$\lambda = 8;\ v_1 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}; \quad \lambda = 6;\ v_2 = \begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix}; \quad \lambda = 3;\ v_3 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$

Diese 3 Vektoren formen eine Basis für R^3 und somit könnten wir sie als Spalten einer Matrix P verwenden, die A diagonalisiert (sind orthogonal aufeinander). Trotzdem wäre es sinnvoller, wenn die Spalten orthonormal wären:

$$u_1 = \begin{pmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{pmatrix}, \quad u_2 = \begin{pmatrix} -1/\sqrt{6} \\ -1/\sqrt{6} \\ 2/\sqrt{6} \end{pmatrix}, \quad u_3 = \begin{pmatrix} 1/\sqrt{3} \\ 1/\sqrt{3} \\ 1/\sqrt{3} \end{pmatrix}$$

$$\rightarrow \qquad P = \begin{pmatrix} -1/\sqrt{2} & -1/\sqrt{6} & 1/\sqrt{3} \\ 1/\sqrt{2} & -1/\sqrt{6} & 1/\sqrt{3} \\ 0 & 2/\sqrt{6} & 1/\sqrt{3} \end{pmatrix}$$

Dann ist für gewöhnlich $A = PDP^{-1}$, wobei dieses Mal P quadratisch ist, orthonormale Spalten hat, die orthogonal sind und $P^{-1} = P^T$ gilt.

Theorem 9: Wenn A symmetrisch ist, dann sind irgendwelche zwei Eigenvektoren von unterschiedlichen Eigenräumen orthogonal.

Theorem 10: Eine n x n Matrix A ist orthogonal diagonalisierbar, dann und nur dann, wenn A symmetrisch ist.

Beispiel

Gegeben ist die Matrix A = $\begin{pmatrix} 3 & -2 & 4 \\ -2 & 6 & 2 \\ 4 & 2 & 3 \end{pmatrix}$ und wir diagonalisieren diese Matrix orthogonal, ausgehend von der charakteristischen Gleichung

$$0 = -\lambda^3 + 12\lambda^2 - 21\lambda - 98 = -(\lambda - 7)^2(\lambda + 2)$$

→ $\lambda = 7;\ v_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix},\ v_2 = \begin{pmatrix} -1/2 \\ 1 \\ 0 \end{pmatrix};\quad \lambda = -2;\ v_3 = \begin{pmatrix} -1 \\ -1/2 \\ 1 \end{pmatrix}$

Obwohl v_1 und v_2 linear unabhängig sind, sind sie nicht orthogonal. Allerdings ist eine Projektion von v_2 auf v_1 gegeben durch $\frac{v_2 \cdot v_1}{v_1 \cdot v_1} v_1$ und die Komponente von v_2 orthogonal zu v_1 ist:

$$z_2 = v_2 - \frac{v_2 \cdot v_1}{v_1 \cdot v_1} v_1 = \begin{pmatrix} -1/2 \\ 1 \\ 0 \end{pmatrix} - \frac{-1/2}{2} \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -1/4 \\ 1 \\ 1/4 \end{pmatrix}$$

Dann ist $\{v_1, v_2\}$ ein orthogonales Set im Eigenraum für $\lambda = 7$ (z_2 ist eine Linearkombination der Eigenvektoren v_1 und v_2 und liegt somit im Eigenraum). Nachdem der Eigenraum 2- dimensional ist, haben wir mit dem orthogonalen Set $\{v_1, z_2\}$ eine Orthogonalbasis für den Eigenraum bzw. durch normieren die Orthonormalbasis für den Eigenraum (für $\lambda = 7$).

$$u_1 = \begin{pmatrix} 1/\sqrt{2} \\ 0 \\ 1/\sqrt{2} \end{pmatrix},\quad u_2 = \begin{pmatrix} -1/\sqrt{18} \\ 4/\sqrt{18} \\ 1/\sqrt{18} \end{pmatrix}$$

Eine Orthonormalbasis für den Eigenraum $\lambda = -2$ ist:

$$u_3 = \frac{1}{||2v_2||} 2v_3 = \frac{1}{3} \begin{pmatrix} -2 \\ -1 \\ 2 \end{pmatrix} = \begin{pmatrix} -2/3 \\ -1/3 \\ 2/3 \end{pmatrix}$$

8.2 Quadratische Formen

Bis jetzt erfolgte in diesem Kapitel eine Fokussierung auf lineare Gleichungen, außer für die Quadratsummen die bei der Berechnung von x^Tx auftraten, aber auch bei der Methode der kleinsten Quadrate. Solche Summen, sogenannte quadratische Formen, treten somit hauptsächlich in Anwendungen der linearen Algebra auf.

Eine **quadratische Form im R^n** ist eine Funktion Q, deren Wert bei einem Vektor x durch einen Ausdruck der Form $Q(x) = x^T Ax$ (mit A symmetrisch) berechnet werden kann. Das simplste Beispiel einer quadratischen Form ≠ 0 ist $Q(x) = x^T I\ x = ||x||^2$ bzw. kann statt I auch eine beliebige symmetrische Matrix A eingesetzt werden.

Beispiel

Gegeben: $A = \begin{pmatrix} 4 & 0 \\ 0 & 3 \end{pmatrix} \rightarrow x^T Ax = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 4 & 0 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 4x_1^2 + 3x_2^2$

Theorem 11 (Hauptachsentheorem)
Sei A eine symmetrische n x n Matrix. Dann gibt es einen orthogonalen Wechsel der Variable, x = Py, der die quadratische Form x^TAx in die quadratische Form y^TDy transformiert, ohne das Kreuzprodukt zu verwenden.

Die Spalten von P im Theorem sind die Hauptachsen der quadratischen Form x^TAx und der Vektor y ist der Koordinatenvektor von x bezogen auf die Orthonormalbasis von R^n, die durch diese Hauptachsen gegeben sind.

Beispiel

Gegeben sei die Matrix A= $\begin{pmatrix} 1 & -4 \\ -4 & -5 \end{pmatrix}$ aus der sich folgende Orthonormalbasis ergibt:

$$P = \begin{pmatrix} 2/\sqrt{5} & 1/\sqrt{5} \\ -1/\sqrt{5} & 2/\sqrt{5} \end{pmatrix}, \quad D = \begin{pmatrix} 3 & 0 \\ 0 & -7 \end{pmatrix}$$

Dann ist $A = PDP^{-1}$ und $D = P^{-1}AP = P^TAP$ und x kann wie folgt geändert werden:

$$\mathbf{x = Py}, \quad wo\ x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\ und\ y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

$$\rightarrow \quad x_1^2 - 8x_1x_2 - 5x_2^2 = x^TAx = (Py)^T A\,(Py)$$
$$= y^TP^TA\,Py = y^TDy$$
$$= 3y_1^2 - 7y_2^3$$

Um die Bedeutung der Gleichheit der quadratischen Formen in diesem Beispiel zu zeigen, können wir Q(x) für x =(2, -2) berechnen, indem wir die quadratische Form verwenden. Nachdem x = Py ist, ergibt sich $y = P^{-1}x = P^Tx$.

8.3 Singulärwertzerlegung

Die besprochenen Diagonalisierungstheoreme spielen in vielen interessanten Anwendungen eine Rolle. Doch leider können nicht alle faktorisiert werden durch $A = PDP^{-1}$ und der Diagonalmatrix D. Allerdings ist eine Faktorisierung $A = QDP^{-1}$ für irgendeine m x n Matrix A möglich. Eine spezielle Faktorisierung dieses Typs, die sogenannte Singulärwertzerlegung, ist eine der nützlichsten Matrixfaktorisierungen der angewandten linearen Algebra. Sie basiert auf der folgenden Eigenschaft der gewöhnlichen Diagonalisierung, die von Rechtecksmatrizen imitiert werden kann:
Die absoluten Werte der Eigenwerte einer symmetrischen Matrix A messen den Umfang, in dem A gewisse Vektoren (die Eigenvektoren) streckt oder staucht.
Wenn $Ax = \lambda x$ und $||x|| = 1$, dann ist

$$\big||Ax|\big| = \big||\lambda x|\big| = |\lambda| \cdot \big||x|\big| = |\lambda| \qquad (91a)$$

Wenn λ_1 jener Eigenwert mit der größten Magnitude ist, dann identifiziert der dazu korrespondierende Einheitseigenvektor v_1 eine Richtung, in welcher der Ausdehnungseffekt von A am größten ist. Die Länge von Ax wird also durch (91a) dann maximiert wenn $x = v_1$ und $||Av_1|| = |\lambda_1|$ ist. Diese Beschreibung von v_1 und $|\lambda_1|$ gilt analog für alle Rechtecksmatrizen, bei denen eine Singulärwertzerlegung durchgeführt wird.

Beispiel

Angenommen $A = \begin{pmatrix} 4 & 11 & 14 \\ 8 & 7 & -2 \end{pmatrix}$, so bildet die lineare Transformation $x \rightarrow Ax$ die Einheitskugel $\{x: ||x|| = 1\}$ im R^3 auf eine Ellipse im R^2 ab. Ziel ist es nun, einen Einheitsvektor x zu finden, bei welchem die Länge ||Ax|| maximiert wird.

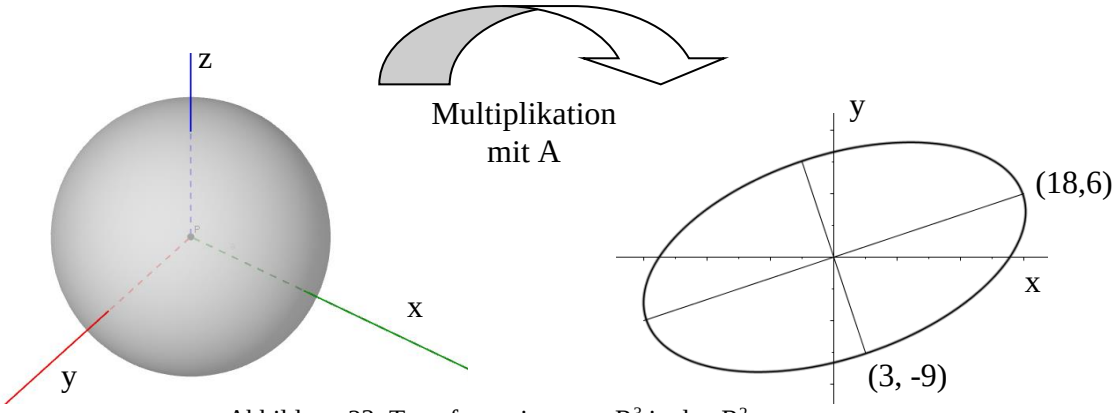

Abbildung 23: Transformation vom R^3 in den R^2

Lösung: Die Größe von $||Ax||^2$ ist beim selben x maximal, welches auch ||Ax|| maximiert, wobei $||Ax||^2$ leichter handhabbar ist. Wir beobachten, dass:

$$||Ax||^2 = (Ax)^T(Ax) = x^TA^TAx = x^T(A^TA)x \quad \text{ist.}$$

A^TA ist auch eine symmetrische Matrix, nachdem gilt: $(A^TA)^T = A^TA^{TT} = A^TA$. Also besteht das Problem jetzt darin, die quadratische Form $x^T(A^TA)x$ so zu maximieren, dass sie Gegenstand der Bedingung || x || = 1 wird. Ohne explizit darauf einzugehen ist der Maximalwert aber genau der größte Eigenwert λ_1 von A^TA. Außerdem gelangt man zum Maximalwert durch einen Einheitseigenvektor von A^TA der zu λ_1 gehört.

$$A^TA = \begin{pmatrix} 4 & 8 \\ 11 & 7 \\ 14 & -2 \end{pmatrix}\begin{pmatrix} 4 & 11 & 14 \\ 8 & 7 & -2 \end{pmatrix} = \begin{pmatrix} 80 & 100 & 40 \\ 100 & 170 & 140 \\ 40 & 140 & 200 \end{pmatrix}$$

Die Eigenwerte von A^TA sind $\lambda_1 = 360, \lambda_2 = 90$ und $\lambda_3 = 0$. Die dazu korrespondierenden Eigenvektoren entsprechen dann

$$v_1 = \begin{pmatrix} 1/3 \\ 2/3 \\ 2/3 \end{pmatrix} \qquad v_2 = \begin{pmatrix} -2/3 \\ -1/3 \\ 2/3 \end{pmatrix}, \qquad v_3 = \begin{pmatrix} 2/3 \\ -2/3 \\ 1/3 \end{pmatrix}$$

Der Maximalwert von $||Ax||^2$ ist 360 und wird erreicht, wenn x der Einheitsvektor v_1. Der Vektor Av_1 ist ein Punkt auf der Ellipse, der am weitesten entfernt ist vom Ursprung, nämlich

$$Av_1 = \begin{pmatrix} 4 & 11 & 14 \\ 8 & 7 & -2 \end{pmatrix} \begin{pmatrix} 1/3 \\ 2/3 \\ 2/3 \end{pmatrix} = \begin{pmatrix} 18 \\ 6 \end{pmatrix}$$

Für $||x|| = 1$ ist der Maximalwert von $||Ax||$ demnach $||Av_1|| = \sqrt{360} = 6\sqrt{10}$

Dieses Beispiel geht davon aus, dass der Effekt von A auf den Einheitskreis im R^3 sich auf die quadratische Form $x^T (A^TA) x$ bezieht. Tatsächlich kann also das gesamte geometrische Verhalten der Transformation $x \rightarrow Ax$ durch diese quadratische Form „eingefangen“ werden.

8.3.1 Singulärwerte einer m x n Matrix

Sei A eine m x n Matrix, so ist A^TA symmetrisch und kann orthogonal diagonalisiert werden. Sei $\{v_1, \ldots, v_2\}$eine Orthonormalbasis für R^n bestehend aus Eigenvektoren von A^TA und angenommen $\lambda 1, \ldots \lambda_n$ sind die mit A^TA assoziierten Eigenwerte, dann gilt für $1 \leq i \leq n$,

$$||Av||^2 = (Av_i)^T Av_i = v_i^T A^T A\, v_i \qquad \textbf{(91b)}$$

$$= v_i^T(\lambda_i v_i) \quad \text{nachdem } v_i \text{ ein Eigenvektor von } A^TA$$

$$= \lambda_i \quad \text{nachdem } v_i \text{ ein Einheitsvektor ist}$$

Die **Singulärwerte** von A sind die Quadratwurzeln der Eigenwerte von A^TA, die bezeichnet werden mit $\sigma_1, \ldots, \sigma_n$ und in absteigender Reihenfolge angegeben werden.
So ist $\sigma_i = \sqrt{\lambda_i}$ für $1 \leq i \leq n$

Nach (b) sind die Singulärwerte von A die Längen der Vektoren $Av_1, \ldots, Av_n$.

Beispiel
Sei A dieselbe Matrix wie im vorigen Beispiel. Nachdem die Eigenwerte von A^TA 360, 90 und 0 sind, sind die Singulärwerte von A:

$$\sigma_1 = \sqrt{360} = 6\sqrt{10}, \quad \sigma_2 = \sqrt{90} = 3\sqrt{10}, \quad \sigma_3 = 0$$

Aus dem vorigen Beispiel ergibt sich als erster Singulärwert von A das Maximum von $||Ax||$ über alle Einheitsvektoren, wobei das Maximum beim Einheitseigenvektor v_1 angenommen wird. Der zweite Singulärwert von A ist, ohne explizit die Ursache dafür anzugeben, gerade das Maximum von $||Ax||$ über alle Einheitsvektoren die orthogonal zu v_1 sind, und dieses Maximum wird beim zweiten Einheitseigenvektor v_2 erreicht.

$$Av_2 = \begin{pmatrix} 4 & 11 & 14 \\ 8 & 7 & -2 \end{pmatrix} \begin{pmatrix} -2/3 \\ -1/3 \\ 2/3 \end{pmatrix} = \begin{pmatrix} 3 \\ -9 \end{pmatrix}$$

Dieser Punkt Av_2 liegt auf der Nebenachse der Ellipse, so wie Av_1 auf der Hauptachse liegt. Die ersten beiden Singulärwerte von A sind somit die Längen der Haupt- und Nebenhalbachsen der Ellipse.
Außerdem ist die Tatsache, dass Av_1 und Av_2 orthogonal aufeinander sind, kein Zufall, wie der nächste *Satz* zeigt.

SATZ:
Angenommen $\{v_1, ..., v_n\}$ist eine Orthonormalbasis des R^n bestehend aus Eigenvektoren von A^TA, so dass die dazu korrespondierenden Eigenwerte von A^TA angeordnet sind durch $\lambda_1 \geq ... \geq \lambda_n$, wobei A zusätzlich r Singulärwerte ungleich 0 hat. Dann ist $\{Av_1..., Av_r\}$ eine Orthogonalbasis für den Spaltenraum A und Rang A = r.

8.3.2 Singulärwertzerlegung

Die Zerlegung von A involviert eine m x n „Diagonal-“ Matrix Σ der Form

$$\Sigma = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} \quad (91c)$$

m – r Zeilen

n – r Spalten

ist und D eine r x r Diagonalmatrix beschreibt.

SATZ (Singulärwertzerlegung):
Sei A eine m x n Matrix mit Rang r, so existiert eine *m x n* - Matrix Σ, wo die Diagonaleinträge in D die ersten r Singulärwerte von A sind, mit $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ und es existieren eine orthogonale *m x m* – Matrix U und eine orthogonale *n x n* – Matrix V, so dass gilt

$$\mathbf{A = U\Sigma V^T} \quad (92)$$

Die Spalten von U in einer solchen Zerlegung werden *linke Singulärvektoren* von A genannt und die Spalten von V sind die *rechten Singulärvektoren* von A.

Beispiel a): Gesucht ist eine Singulärwertzerlegung von A = $\begin{pmatrix} 4 & 11 & 14 \\ 8 & 7 & -2 \end{pmatrix}$
Aus den bisherigen zwei Beispielen in diesem Kapitel können wir zum einen v_1, v_2 und v_3 als die rechten Singulärvektoren von A und zum anderen Av_1 und Av_2 verwenden.

$$u_1 = \frac{1}{\sigma_1} Av_1 = \frac{1}{6\sqrt{10}} \begin{pmatrix} 18 \\ 6 \end{pmatrix} = \begin{pmatrix} 3/\sqrt{10} \\ 1/\sqrt{10} \end{pmatrix}$$

$$u_2 = \frac{1}{\sigma_2} Av_2 = \frac{1}{3\sqrt{10}} \begin{pmatrix} 3 \\ -9 \end{pmatrix} = \begin{pmatrix} 1/\sqrt{10} \\ -3/\sqrt{10} \end{pmatrix}$$

Dann ist $\{u_1, u_2\}$ eine Basis für R^2. Sei U = $(u_1 \quad u_2)$, V = $(v_1 \quad v_2 \quad v_3)$ und

$$D = \begin{pmatrix} 6\sqrt{10} & 0 \\ 0 & 3\sqrt{10} \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} 6\sqrt{10} & 0 & 0 \\ 0 & 3\sqrt{10} & 0 \end{pmatrix}$$

$$\rightarrow \quad \mathbf{A} = \underbrace{\begin{pmatrix} 3/\sqrt{10} & 1/\sqrt{10} \\ 1/\sqrt{10} & -3/\sqrt{10} \end{pmatrix}}_{\text{U}} \underbrace{\begin{pmatrix} 6\sqrt{10} & 0 & 0 \\ 0 & 3\sqrt{10} & 0 \end{pmatrix}}_{\Sigma} \underbrace{\begin{pmatrix} 1/3 & 2/3 & 2/3 \\ -2/3 & -1/3 & 2/3 \\ 2/3 & -2/3 & 1/3 \end{pmatrix}}_{V^T}$$

Beispiel b): Gesucht ist eine Singulärwertzerlegung von A = $\begin{pmatrix} 1 & -1 \\ -2 & 2 \\ 2 & -2 \end{pmatrix}$

→ Zuerst berechnet man $A^TA = \begin{pmatrix} 9 & -9 \\ -9 & 9 \end{pmatrix}$. Die Eigenwerte von A^TA sind 18 und 0, mit korrespondierenden Einheitseigenvektoren

$$\mathbf{v_1} = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}, \quad \mathbf{v_2} = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$$

→ Dann folgt $A\mathbf{v_1} = \begin{pmatrix} 2/\sqrt{2} \\ -4/\sqrt{2} \\ 4/\sqrt{2} \end{pmatrix}$, $\sigma_1 = ||A\mathbf{v_1}|| = \sqrt{18} = 3\sqrt{2}$

$$\text{und } \mathbf{u_1} = \frac{1}{3\sqrt{2}} A\mathbf{v_1} = \begin{pmatrix} 1/3 \\ -2/3 \\ 2/3 \end{pmatrix}$$

Außerdem ist $Av_2 = 0$ nachdem v_2 mit dem Null-Eigenwert von A^TA korrespondiert.

→ Im nächsten Schritt möchte man $\{u_1\}$so verändern, dass man eine Orthonormalbasis im R^3 erhält. Insofern braucht man zwei Orthonormalvektoren die orthogonal sind zu u_1, wobei jeder Vektor die Gleichung $u_1^T x = 0$ erfüllen muss (ist äuivalent zur Gleichung $x_1 - 2x_2 + 2x_3 = 0$). Eine Basis für das Lösungsset dieser Gleichung ist

$$w_1 = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}, \quad w_2 = \begin{pmatrix} -2 \\ 0 \\ 1 \end{pmatrix}$$

→ Die Anwendung des Gram Schmidt-Verfahrens auf $\{w_1, w_2\}$ führt zu

$$u_2 = \begin{pmatrix} 2/\sqrt{5} \\ 1/\sqrt{5} \\ 0 \end{pmatrix}, \quad u_3 = \begin{pmatrix} -2/\sqrt{45} \\ 4/\sqrt{45} \\ 5/\sqrt{45} \end{pmatrix}$$

Schlussendlich ist $U = (u_1 \quad u_2 \quad u_3)$, $V = (v_1 \quad v_2)$ und $\Sigma = \begin{pmatrix} 3\sqrt{2} & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$

$$\rightarrow A = \begin{pmatrix} 1 & -1 \\ -2 & 2 \\ 2 & -2 \end{pmatrix} = \begin{pmatrix} \frac{1}{3} & \frac{2}{\sqrt{5}} & -\frac{2}{\sqrt{45}} \\ -\frac{2}{3} & \frac{1}{\sqrt{5}} & \frac{4}{\sqrt{45}} \\ \frac{2}{3} & 0 & \frac{5}{\sqrt{45}} \end{pmatrix} \begin{pmatrix} 3\sqrt{2} & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$$

8.3.3 Anwendungen der Singulärwertzerlegung

Die Singulärwertzerlegung findet insbesondere in der numerischen Mathematik Anwendung, weil sich beispielsweise dadurch fast singuläre lineare Gleichungssysteme im Rahmen rechentechnischer Genauigkeiten passabel lösen lassen.
In der Statistik ist die Singulärwertzerlegung der rechnerische Kern der **Hauptkomponentenanalyse** (siehe nächstes Kapitel) und spielt somit auch eine entscheidende Rolle bei der Methode der kleinsten Quadrate.

Außerdem beruhen moderne Bildkompressionsverfahren auf einem Algorithmus, der das Bild (bzw. die Matrix aus Farbwerten) in eine Singulärwertzerleung überführt und anschließend die Matrix Σ reduziert, indem nur stark von null verschiedene Elemente berücksichtigt und gespeichert werden. Demnach führt das Weglassen von kleinen Singulärwerten also zu einem verlustbehafteten Modellreduktionsverfahren.

Beispiel (reduzierte Singulärwertzerlegung und die Pseudoinverse von A)

Wenn Σ Zeilen und Spalten mit Nullen beinhaltet, ist eine kompaktere Zerlegung von A möglich. Ausgehend von der Notation vorher, sei r = rang A, und es erfolgt eine Zerlegung von U und V in Teilmatrizen, wobei deren erster Block jeweils r Spalten beinhaltet:

$$U = (U_r \quad U_{m-r}), \quad \text{wo} \quad U_r = (u_1 \dots u_r) \text{ ist}$$
$$V = (V_r \quad V_{n-r}), \quad \text{wo} \quad V_r = (v_1 \dots v_r) \text{ ist}$$

Dann ist U_r eine m x r und V_r eine n x r Matrix und die unterteilte Matrixmultiplikation zeigt, dass:

$$A = (U_r \quad U_{m-r}) \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V_r^T \\ V_{n-r}^T \end{pmatrix} = U_r D V_r^T$$

Diese Faktorisierung von A nennt man eine **reduzierte Singulärwertzerlegung**. Nachdem die Diagonaleinträge in D ungleich Null sind, kann nun folgende Matrix geformt werden (die sogenannte Pseudoinverse von A):

$$A^+ = V_r D^{-1} U_r^T \tag{93}$$

Beispiel (kleinste Quadrate Lösung):
Gegeben ist die Gleichung Ax = b und wir verwenden die eben definierte Pseudoinverse von A um folgendes zu definieren:

$$\hat{\mathbf{x}} = A^+ \mathbf{b} = V_r D^{-1} U_r^T \mathbf{b}$$

Außerdem setzen wir auch den durch A definierten Ausdruck aus dem ersten Beispiel in $A\hat{x}$ ein, so dass

$$\begin{aligned} \mathbf{A}\hat{\mathbf{x}} &= (U_r D V_r^T)\,(V_r D^{-1} U_r^T \mathbf{b}) \\ &= U_r D D^{-1} U_r^T \mathbf{b} \\ &= U_r U_r^T \mathbf{b} \end{aligned}$$

Der Ausdruck $U_r U_r^T \mathbf{b}$ ist die Orthogonalprojektion $\hat{b}$ von b auf den Spaltenraum von A. Deshalb ist $\hat{x}$ eine kleinste Quadrate Lösung von **Ax = b**. Durch Nachprüfen ergibt sich

tatsächlich dieses $\hat{x}$ als kleinster Abstand unter allen kleinste-Quadrate Abständen/Lösungen.

Kapitel 9

9. HAUPTKOMPONENTENANALYSE

Die „Hauptkomponentenanalyse“ bzw. „Hauptachsentransformation“ ist eine Methode der multivariaten Statistik und wird angewendet um ausgedehnte Datensätze zu strukturieren bzw. zu vereinfachen. Zu diesem Zweck wird eine große Menge an statistischen Variablen durch eine geringere Zahl möglichst repräsentativer Linearkombinationen (sogenannten „Hauptkomponenten“) approximiert.
Der Unterschied zur Faktorenanalyse, mit der Ähnlichkeiten bestehen und in der die Hauptkomponentenanalyse auch als Näherungsmethode zur Faktorenextraktion zur Anwendung kommt, wird im Unterkapitel Faktorenanalyse diskutiert.

Ein Anwendungsbeispiel der HKA welches als Motivation herangezogen werden soll, ist das Problem der *Gesichtserkennung*, die mit der Bewältigung von enormen, hochdimensionalen Datenmengen verbunden ist. Allerdings sind oft einige Komponenten einer Datenmenge irrelevant oder weniger relevant als andere, da sie beinahe konstant sind. Die Gesichter unterscheiden sich in Nasen-, Augen und Mundpartie in etwa stärker voneinander als in Ausschnitten der Stirn oder der Wangen, so dass es zweckmäßig ist, nur diese signifikanten Partien als Unterscheidungsmerkmale zu speichern

9.1 Grundgedanken der Hauptkomponentenanalyse

Ausgehend von einem Datensatz mit Matrixstruktur werden an n-Objekten jeweils m-Merkmale gemessen. Dieser Datensatz kann somit als Menge von n Punkten im n-dimensionalen Raum veranschaulicht werden. Ausgewiesenes Ziel der Hauptkomponentenanalyse ist eine Dimensionsreduktion der Variablen durch eine Projektion der Datenpunkte in einen k- dimensionalen Teilraum R^k ($k < n$), so dass dabei nur ein Minimum an Information verloren geht und die auftretende Redundanz in Form von Korrelation in den Datenpunkten komprimiert wird.

Zur besseren Anschauung wird nun zunächst ein theoretisches Beispiel einer drei-dimensionalen Datenmenge Schritt für Schritt erklärt, bis schließlich weiter unten im Kapitel ein praktisches Anwendungsbeispiel der HKA folgt.

Gegeben sei zunächst eine Reihe mehrdimensionaler Messungen (Datenmenge), die eine Punktwolke bilden und im Sinne der HKA transformiert und anschließend in ihren Dimensionen reduziert wird.

1.Schritt: Ursprung des Koordinatensystems wird in den Schwerpunkt der Punktwolke gesetzt

2. Schritt: Das Koordinatensystem wird gedreht, so dass die erste Achse in Richtung der größten Abweichung bzw. der größten Varianz ausgerichtet ist

3.Schritt: Die zweite Achse wird in Richtung der größtmöglichen Varianz unkorreliert zur ersten Achse gedreht, wodurch die Drehung des Koordinatensystems in jene Richtung der größtmöglichen Varianz ausgerichtet wird, die möglich ist, ohne die Richtung der ersten Achse zu ändern (Drehung des Systems um x-Achse) .

4.Schritt: Fortsetzung des Verfahrens, bis die k-te Achse in Richtung der größten Varianz ausgerichtet ist, unkorreliert zu den ersten (k – 1)-Achsen. Dadurch bezeichnet die k-te Achse die k-te Hauptkomponente, die geometrisch als Hauptachsen eines Ellipsoiden (Punktwolke) gedeutet werden können.

9.2 Herleitung der Problemlösung

Ausgehend von einer Datenmenge aus n, p- elementigen Beobachtungen in Form einer (p x n) Matrix X wird der p-dimensionale Vektor a_1 gesucht für den gilt, dass Var(a_1^T) maximal wird. Die (p x p) Kovarianzmatrix zu X ist definiert durch S = Kov (X).
Diese Bedingung entspricht nach Definition von Varianz und Kovarianz dem Problem $a_1^T S a_1$ zu maximieren. Nachdem allerdings der Ausdruck für beliebige a_1 beliebig groß wird, braucht man eine Schrankenbedingung z.B $a_1^T = 1$

Problematisch ist nun die Maximierung eines Ausdrucks mit Nebenbedingung für deren Lösung der Lagrange-Multiplikator λ in der Gleichung $a_1^T S a_1 - \lambda(a_1^T a_1 - 1)$verwendet wird, die Ausdruck und Nebenbedingung in einer Gleichung zusammenfasst. Ziel ist es den Vektor a_1 zu suchen, der das Ergebnis der Gleichung maximiert.

Es wird nach a_1 differenziert, um einen Extremwert zu erhalten.
Die Ableitung liefert:

$$S a_1 - \lambda a_1 = 0 \Rightarrow (S - \lambda E) a_1 = 0$$

Offensichtlich ist dies nun ein Eigenwertproblem von S, wobei λ ein Eigenwert (EW) von a_1 ist. Aus $S a_1 - \lambda a_1 = 0$ folgt $S a_1 = \lambda a_1$. Diese Erkenntnis ergibt eingesetzt in das ursprüngliche Problem, das durch Maximierung von $a_1^T S$ gegeben war:

$= \max\{a_1' S a_1 | a_1' a_1 = 1\} = \max\{a_1' \lambda a_1 | a_1' a_1 = 1 \wedge \lambda \text{ ist EW von S}\}$
$= \max\{a_1' \lambda a_1 | a_1' a_1 = 1 \wedge \lambda \text{ ist EW von S}\} = \max\{\lambda | \lambda \text{ ist EW von S}\}$

Darum ist der größte EW von S gesucht.

Anschließend möchte man den q-dimensionalen Vektor a_2 ermitteln, für den gilt: Var (a_2X) wird maximal, $a_2^T a_2 = 1$ und a_1 ist unkorreliert zu a_2.

Somit muss folgendes zutreffen:

$$0= \text{Kov}(a_1^T X,\ a_2^T X) = a_1^T S\, a_2 = \lambda_1\, a_1^T a_2 = \lambda_1\, a_2^T a_1$$

$$\Rightarrow\ a_1 \text{ unkorreliert zu } a_2 \Leftrightarrow a_1^T S a_2 = a_2^T S a_1 = a_1^T a_2 = a_2^T a_1 = 0$$

Daraus ergibt sich eine ähnliche Situation wie in Schritt 1, wodurch eine erweiterte Lagrange-Multiplikatorgleichung angewendet wird, welche zwei Multiplikationen verwendet:

$$a_2' S a_2 - \lambda\left(a_2' a_2 - 1\right) - \phi a_2' a_1 = \max \qquad (94a)$$

Die Ableitung nach a_2 liefert

$$S a_2 - \lambda a_2 - \phi a_1 = 0.$$

Multiplizieren mit a_1 ergibt dann

$$a_1^T S a_2 - \lambda a_1 a_2 - \phi a_1 a_1 = 0$$
$$\Rightarrow \phi = 0 \qquad (94b)$$

(A) und (B) liefern dann

$$S a_2 - \lambda a_2 = 0\ \Rightarrow (S - \lambda E) a_2 = 0$$

Gesucht ist also der zweitgrößte EW

Die Fortsetzung bis q liefert die folgenden Werte:

- $\{a_1,\ldots,a_q\}$ als Hauptvektoren und somit $\{a_1 I,\ldots,a_q I\}$ als Hauptkomponenten mit I=Matrix aus den Basisvektoren des Ausgangssystems
- $\{\lambda_1,\ldots\lambda_m\}$ als deren Varianzen
- $\frac{\lambda_m}{\lambda_1+\ldots+\lambda_q}$ als ein Maß für den Anteil der m-ten Hauptkomponente an der Gesamtvarianz

9.3 Eigenschaften der Hauptkomponentenanalyse

Eine der wichtigsten Eigenschaften der Hauptkomponentenanalyse ist es, dass sie eine optimale Rekonstruktion im Sinne der kleinsten Fehlerquadrate erlaubt, wodurch eine Dimensionsreduktion ermöglicht wird, bei welcher der Informationsverlust minimal ist.
A bezeichnet eine n x p –Matrix und B eine p x q –Matrix.
Eigenschaft a) : (95a)

Sei $y = B' x$ eine orthonormale Abbildung (mit B = p x q, $1 \le q \le p$), also
$S_y = B' S_x B$, mit $S_y = Kov(y)\ und\ S_x = Kov(x)$

Dann trifft das Folgende zu:

Spur(S_y) wird maximal genau dann, wenn B = A_q ist,
wobei A_q aus den ersten q Hauptkomponenten besteht.

Beweis:

Sei B = AC (C ist eine p x q Transformationsmatrix)

Dann folgt:

$$\mathrm{Spur}(B'SB) = \mathrm{Spur}\big(C'\underbrace{A'S\,A}_{D=\mathrm{diag}(\lambda_1\ldots\lambda_q)}C\big) = \mathrm{Spur}\,(C'D\,C) = \sum_{j,q=1}^{p} \lambda_j\, c_{jk}^2. \qquad (1)$$

Nachdem die Spalten von A und B orthonormal sind, folgt Orthonormalität für die Spalten von C:

$$C'C = B'AA'B = B'B = E_q$$

$$\Rightarrow \quad \mathrm{Spur}(C'C) = \sum_{j,k=1}^{p} c_{jk}^2 = \mathrm{Spur}(E_q) = q$$

Für die Zeilen von C gilt:

$c_j' c_j \leq 1$ da C Teil einer Orthogonalmatrix ist

$$\Rightarrow \sum_{j,k=1}^{p} c_{jk}^2 \leq 1. \qquad (2)$$

Aus (1) und (2) folgt also dass $\sum \lambda_j c_{jk}^2$ maximal wird, falls gilt: $\sum_{k=1}^{p} c_{jk}^2 = \begin{cases} 1 & j=1...q \\ 0 & j=q+1\ldots p \end{cases}$

Dies wird erfüllt durch $C = E_q$, also $B = A_q$

Umgekehrt wird der Wert minimal, falls $B=\widetilde{A}_q$ gilt.
Dabei werden die Spalten von $\widetilde{A}_q$ aus EV zu den q kleinsten EW gebildet.

Eigenschaft b): Optimale Rekonstruktion (im Sinne der kleinsten Fehlerquadrate) (95b)

Gegeben sei eine Punktwolke $\{x_1,\ldots,x_n\}$in einem p-dimensionalen Raum und durchzuführen ist eine Projektion auf einen q-dimensionalen Unterraum $y_i = Bx_i$ mit kleinstmöglichem Datenverlust.

Eine Maßzahl die den Datenverlust ausdrückt ist jene der Summe der quadrierten Abstände der Punkte zum Unterraum.

Es gilt, dass die Summe minimal wird, wenn $B = A_q$ ist, sofern A_q die Matrix aus den EV der größten q EW von S ist.

$\Rightarrow y_i = A_q x_i$ ist jene Projektion mit dem geringsten Datenverlust

Beweis:

$\sum_{i=1}^{n} r_i' r_i$ ist also die Summe der quadrierten Fehler

Sowohl die Benennung als auch die Anschauung ist im Fall höherdimensionaler p und q, mit p > q, gleich. Der Beweis kann darum für den allgemeinen Fall durchgeführt werden.

Es gilt $x_i' x_i = (m_i + r_i)'(m_i + r_i)) = m_i' m_i + r_i' r_i + 2 r_i m_i = m_i' m_i + r_i' r_i$

Da m_i orthogonal zu r_i ist, also

$$\sum_{i=1}^{n} r_i' r_i = \sum_{i=1}^{n} x_i' x_i - \sum_{i=1}^{n} m_i' m_i.$$

Um $\sum_{i=1}^{n} r_i' r_i$ zu minimieren, muss man $\sum_{i=1}^{n} m_i' m_i$ maximieren, also $\sum_{i=1}^{n} y_i' y_i$ maximieren.

Einfache Umformungsschritte führen zu einem Maximierungsproblem, das mit von Hilfe Eigenschaft a) lösbar ist.

$$\sum_{i=1}^{n} y_i' y = \sum_{i=1}^{n} x_i' B B' x_i = \text{Spur}\left(\sum_{i=1}^{n} x_i' B B' x_i\right) = \sum_{i=1}^{n} \text{Spur}(x_i' B B' x_i) =$$
$$\sum_{i=1}^{n} \text{Spur}(B' x_i x_i' B) = \text{Spur}\left[B'\left(\sum_{i=1}^{n} x_i x_i'\right) B\right] = \text{Spur}[B' X' X B] =$$

$$(n-1)\text{Spur}(B' S B) = \max.$$

Nach Eigenschaft a) trifft das genau dann zu wenn gilt: B= A_q

9.4 Beispiel für die Hauptkomponentenanalyse

Beispiel 1: Ski - Weltcupabfahrt (Tabelle 20 im Anhang)

Bei einer Weltcupabfahrt wurden die Zeiten von sechs Teilstücken gemessen. Jene Fahrer die entweder disqualifiziert wurden bzw. deren Zeiten weit von jenen der anderen Fahrer abwichen wurden weggelassen.(hier 3 Fahrer, siehe Tabelle im Anhang)

Problemstellung
Es soll eine Faktorenanalyse auf den sechs Teilzeiten durchgeführt und als Zahl der Faktoren soll drei gewählt werden.

Die *Korrelationsmatrix* der Teilzeiten sieht dann wie folgt aus:

Korrelationsmatrix

		1.Teil	2.Teil	3.Teil	4.Teil	5.Teil	6.Teil
Korrelation	1.Teil	1,000	,815	,786	,528	,405	,316
	2.Teil	,815	1,000	,764	,545	,478	,413
	3.Teil	,786	,764	1,000	,829	,573	,364
	4.Teil	,528	,545	,829	1,000	,585	,366
	5.Teil	,405	,478	,573	,585	1,000	,420
	6.Teil	,316	,413	,364	,366	,420	1,000

Tabelle 21: Korrelationsmatrix (erstellt mit Spss)

In dieser Abbildung ist gleich erkennbar, dass nur positive Korrelationen auftreten, die alle im Bereich zwischen 0,31 und 0,83 liegen. Zudem kann daraus geschlossen werden, dass die Zeiten benachbarter Streckenabschnitte tendenziell stärker korrelieren als weiter voneinander entfernt liegende.

KMO und Bartlett-Test

Kaiser-Meyer-Olkin-Maß der Stichprobeneignung.		,785
Bartlett-Test auf Sphärizität	Näherungsweise Chi-Quadrat	277,289
	df	15
	Sig.	,000

Abbildung 18/Tabelle 22: KMO-Index und Bartlett-Test (erstellt mit Spss)

Mit **KMO-Index** und **Bartlett-Test** wird überprüft, ob ein nennenswerter Zusammenhang zwischen allen Variablen besteht. Ist dies nicht der Fall, macht die Faktorenanalyse keinen Sinn. Ein KMO-Index von 0.785 entspricht einem „halbwegs guten" Ausmaß an Interkorrelation zwischen allen Variablen Der Bartlett-Test prüft die Nullhypothese, dass in der Population kein Zusammenhang zwischen den Variablen besteht. Wird der Test signifikant, ist diese Hypothese mit einer Irrtumswahrscheinlichkeit von höchstens 5% widerlegt.

Um die Faktoren (oberer, mittlerer und unterer Streckenabschnitt) zu bestimmen wird nun die Methode der Hauptkomponenten angewendet und darauffolgend eine Orthogonalrotation der Faktoren anhand der Varimax-Methode durchgeführt.

	1.Hauptkomp.	2.Hauptkomp.	3.Hauptkomp.
zugehöriger Eigenwert	3,79	0,84	0,66
Anteil an Gesamtvarianz	0,63	0,14	0,11
Kumulativer Anteil an Varianz	0,63	0,77	0,88

Tabelle 23: bedeutende Kennzahlen der drei Hauptkomponenten

Dadurch können mit dieser Dimensionsreduktion durch die Hauptkomponentenanalyse mit den drei Faktoren 88% der gesamten Varianz beschrieben werden. Allerdings beschreibt allein in der unrotierten Lösung der erste Faktor 63 % der Varianz.
Vor einer genaueren Betrachtung der einzelnen Faktoren, wird die Varimax-Methode für eine Rotation verwendet, welche zu folgenden Faktorladungen in der *„rotierten Komponentenmatrix*" führt:

Komponentenmatrix[a]

	Komponente		
	1	2	3
1.	,833	-,378	,264
2.	,861	-,228	,287
3.	,931	-,187	-,126
4.	,823	,040	-,398
5.	,714	,383	-,378
6.	,557	,682	,444

Extraktionsmethode: Analyse der Hauptkomponente.

Rotierte Komponentenmatrix[a]

	Komponente		
	1	2	3
1.	**,921**	,214	,113
2.	**,865**	,256	,250
3.	**,731**	**,615**	,076
4.	,416	**,812**	,068
5.	,150	**,825**	,311
6.	,187	,205	**,946**

Extraktionsmethode: Analyse der Hauptkomponente.

Tabelle 24: (rotierte) Komponentenmatrix

Die Werte der „rotierten Komponentenmatrix“ entsprechen den Korrelationen zwischen den ursprünglichen Variablen und den Faktoren. Demnach sollen also die drei unabhängig voneinander wirkenden Faktoren die berechnet wurden, die sechs ursprünglichen Variablen möglichst ideal widerspiegeln. Die stärkeren Korrelationen wurden zu diesem Zweck fett markiert. Die Abbildung zeigt, dass Faktor 1 den oberen Streckenabschnitt (bzw. die ersten 3 Teilstücke), Faktor 2 eher den mittleren Streckenabschnitt (bzw. die Gleitpassage) und Faktor 3 den unteren Streckenabschnitt (bzw. der Steilhang bis ins Ziel) beschreibt.
An dieser Stelle ist ohne zusätzliche Information über die Abfahrtsstrecke keine weitere Interpretation möglich. Diverse Cheftrainer der Skinationen Schweiz und Österreich gaben zu den Resultaten sich deckende Kommentare ab:

- **Faktor 1:** Der leichte Wind zu Beginn des Rennens wurde mit Fortdauer des Wettkampfes immer stärker und führte fast zu einem Abbruch der Veranstaltung
- **Faktor 2:** Der Zwischenteil war ein typisches Gleitstück
- **Faktor 3:** Der Steilhang war mit Kunstdünger stark präpariert und wurde zu einer harten Eisunterlage

Somit könnte vermutet werden, dass das Rennen wesentlich durch diese drei Faktoren beeinflusst wurde. Diesen Umstand bestätigt auch die Abbildung 19, denn der erste Faktor, der hauptsächlich die Zeiten in den oberen Abschnitten charakterisiert, nimmt mit Fortdauer des Rennens und Höhe der Startnummer beständig zu. Ein derartiger Trend ist bei den Faktoren 2 und 3 nicht vorhanden.
Zusätzlich ergibt sich die Frage nach der Gewichtung der drei Faktoren bei der Beschreibung der sechs ursprünglichen Teilzeiten. Man weiß, dass die drei Faktoren der Dimensionsreduktion 88 % der Gesamtvarianz erklären und durch die Rotation nicht beeinflusst werden. Allerdings haben sich die Anteile der einzelnen Faktoren wie folgt verschoben:

	1.Faktor	2.Faktor	3.Faktor
Anteil an Gesamtvarianz	0,394	0,312	0,18
Kumulierter Anteil an Varianz	0,394	0,706	0,88

Tabelle 25: Anteil an der Gesamtvarianz/ Kum. Anteil an der Varianz

Daraus kann schlussgefolgert werden, dass die Verteilung der Anteile der drei Faktoren bezogen auf die 88% der Gesamtvariabilität nach der Rotation gleichmäßiger ist.

Eine andere Zerlegung der erklärbaren Gesamtvarianz durch Faktoren ist durch den Anteil der Varianzen der Teilzeiten, welche die drei Faktoren zu erklären imstande sind, gegeben:

Kommunalitäten

	Anfänglich	Extraktion
1.	1,000	,907
2.	1,000	,876
3.	1,000	,918
4.	1,000	,837
5.	1,000	,799
6.	1,000	,972

Extraktionsmethode: Analyse der Hauptkomponente.

Tabelle 26: Kommunalitäten

Diese Größen werden oft unter dem Begriff „Kommunalitäten“ zusammengefasst und kennzeichnen das Ausmaß der Varianz der Teilzeiten (Variablen), dass durch die Varianz erklärt wird. Hier erklären die Faktoren zumindest 80% der Varianz und die Kommunalitäten ergeben eine Summe von 5,3 (jene 88% der Gesamtvarianz 6, weil 6 standardisierte Variablen vorliegen)

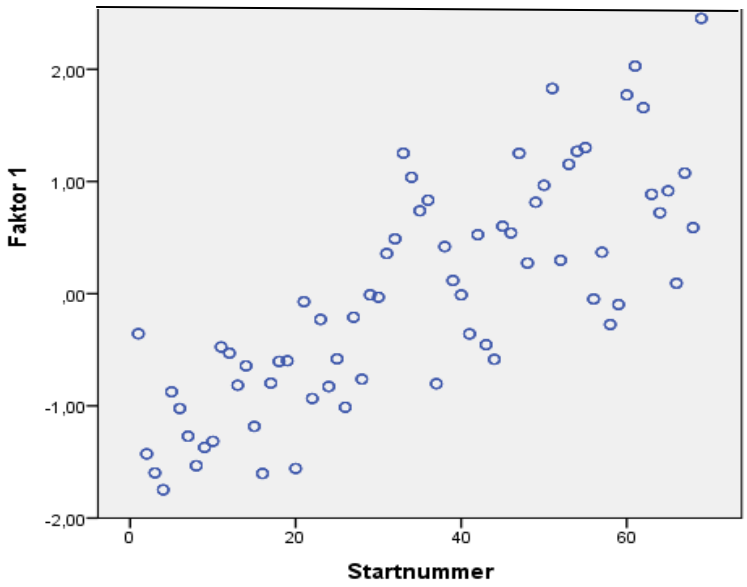

Abbildung 24: Punktwolke welche den Faktor 1 und die Startnummer gegenüber stellt

Beispiel 2: FIS Alpine Ski WM in Vail/Beaver Creek 2015 (Tabelle 28, Anhang)

Dass ein derartiger Zusammenhang zwischen allen Variablen wie im Beispiel oben nicht immer zutrifft, zeigt der WM-Riesentorlauf der Herren in Vail/Beaver Creek.

Die zugehörige Tabelle im Anhang enthält die Endergebnisse und Zwischenzeiten des alpinen WM-Riesentorlaufes der Herren vom 13. 2. 2015. Es werden jene Startnummern außer Acht gelassen die vom Rennkomitee des alpinen Skisports disqualifiziert wurden und im Endresultat einen Rang außerhalb der „Top-30 Athleten“ einnehmen.

Wiederum soll eine Faktoranalyse auf den acht Teilzeiten durchgeführt werden, wobei als Zahl der Faktoren vier gewählt werden soll.

Mit **KMO-Index** und **Bartlett-Test** wird wiederum überprüft, ob ein nennenswerter Zusammenhang zwischen allen Variablen besteht. Sofern dies nicht zutrifft macht die Faktorenanalyse keinen Sinn ist. Bei der Stichprobeneignung ergab die Berechnung mit SPSS den Wert 0,6 und eine Irrtumswahrscheinlichkeit von 0,005.

Die Korrelationsmatrix beinhaltet sogar einige negative Zahlen, weshalb dieses Beispiel mit einer Dimensionsreduktionsmethode wie der Hauptkomponentenanalysen nur unter speziellen Annahmen gelöst werden kann und im Prinzip keinen eindeutigen linearen Zusammenhang darstellt.

Im ersten Schritt wird darum wiederum mit SPSS eine Korrelationsmatrix für die Streckenabschnitte 1a bis 1d (für die Abschnitte a-d im 1. Durchgang) bzw. 2a bis 2d (für die Abschnitte a-d im 2. Durchgang) berechnet. Die restlichen Schritte erfolgen ähnlich zum Beispiel 2.

KMO und Bartlett-Test

Kaiser-Meyer-Olkin-Maß der Stichprobeneignung.		,600
Bartlett-Test auf Sphärizität	Näherungsweise Chi-Quadrat	50,746
	df	28
	Sig.	,005

Tabelle 27: KMO und Bartlett Test

ANHANG

Tabelle 11: Auslieferungszeit eines Getränkelieferanten

Beobachtung	Auslieferungszeit in Minuten	Anzahl der Produkte (x1)	Distanz in m (x2)
1	16,68	7	560
2	11,50	3	220
3	12,03	3	340
4	14,88	4	80
5	13,75	6	150
6	18,11	7	330
7	8,00	2	110
8	17,83	7	210
9	79,24	30	1460
10	21,50	5	605
11	40,33	16	688
12	21,00	10	215
13	13,50	4	255
14	19,75	6	462
15	24,00	9	448
16	29,00	10	776
17	15,35	6	200
18	19,00	7	132
19	9,50	3	36
20	35,10	17	770
21	17,90	10	140
22	52,32	26	810
23	18,75	9	450
24	19,83	8	635
25	10,75	4	150

Tabelle 12: Arbeitsmotivation mit mehreren Prädiktoren

i	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
1	32	36	30	20	20	3100	34	29	69	66
2	14	30	11	30	7	2600	39	16	47	36
3	12	19	15	15	8	3200	42	13	32	17
4	27	42	16	39	13	2500	43	15	63	49
5	20	14	22	5	22	3700	42	29	38	62
6	13	12	16	6	11	2600	36	17	39	51
7	17	17	20	12	11	2500	41	18	44	15
8	8	4	5	0	16	3800	23	9	31	33
9	22	32	20	35	20	3500	25	21	40	55
10	19	15	13	8	13	3100	29	21	57	56
11	25	38	5	34	21	3600	59	27	53	67
12	23	24	6	26	9	2600	45	31	54	62
13	17	28	11	32	10	2600	30	7	45	26
14	22	36	4	26	16	2500	52	23	56	64
15	19	18	26	12	6	2500	40	17	54	55
16	27	40	27	36	12	2500	42	29	44	62
17	26	30	28	27	18	3000	38	34	43	64
18	20	27	11	26	10	2600	35	19	46	55
19	11	18	23	13	11	2800	42	18	31	43
20	24	32	18	19	15	2700	48	23	51	53
21	19	33	9	25	6	2400	38	23	37	65
22	19	33	22	30	5	2600	36	30	39	39
23	22	27	28	18	17	4000	45	23	52	54
24	24	30	32	21	11	2700	44	20	41	47
25	17	37	8	11	2	2300	32	20	44	41

Tabelle 28: Körpergrößen

Körpergröße x in [cm]	Absolute Häufigkeit	Relative Häufigkeit	$\tilde{F}(x)$	$\Phi\left(\frac{x-165{,}05}{5{,}86}\right)$	a_1	a_2
153	1	0,01	0,01	0,02	0,02	0,01
154	1	0,01	0,02	0,03	0,02	0,01
155	2	0,02	0,04	0,04	0,02	0,00
156	3	0,03	0,07	0,06	0,02	0,01
157	3	0,03	0,10	0,09	0,02	0,01
158	5	0,05	0,15	0,12	0,02	0,03
159	6	0,06	0,21	0,15	0,00	0,06
160	4	0,04	0,25	0,19	0,02	0,06
161	5	0,05	0,30	0,25	0,00	0,05
162	7	0,07	0,37	0,30	0,00	**0,07**
163	5	0,05	0,42	0,36	0,01	0,06
164	5	0,05	0,47	0,43	0,01	0,04
165	6	0,06	0,53	0,50	0,03	0,03
166	7	0,07	0,60	0,56	0,03	0,04
167	5	0,05	0,65	0,63	0,03	0,02
168	4	0,04	0,69	0,69	0,04	0,00
169	5	0,05	0,74	0,75	0,06	0,01
170	5	0,05	0,79	0,80	0,06	0,01
171	6	0,06	0,85	0,85	0,06	0,00
172	4	0,04	0,89	0,88	0,03	0,01
173	3	0,03	0,92	0,91	0,02	0,01
174	2	0,02	0,94	0,94	0,02	0,00
175	3	0,03	0,97	0,96	0,02	0,01
176	1	0,01	0,98	0,97	0,00	0,01
177	1	0,01	0,99	0,98	0,00	0,01
178	1	0,01	1,00	0,99	0,00	0,01

Tabelle 20: Ski-Weltcupabfahrt

		Teilzeiten in Sekunden						
Startnr.	Skifahrer	1.	2.	3.	4.	5.	6.	total
1	Fahrer 1	17,79	32,47	19,73	21,65	14,91	12,41	118,96
2	Fahrer 2	17,52	32,67	19,75	21,8	15,46	12,3	119,5
3	Fahrer 3	17,48	32,25	19,92	22,01	15,15	12,64	119,45
4	Fahrer 4	17,61	32,07	19,59	21,67	15,35	12,4	118,69
5	Fahrer 5	17,71	32,38	20	22,1	15,18	12,34	119,71
6	Fahrer 6	17,79	32,03	19,87	21,64	15,2	12,42	118,95
7	Fahrer 7	17,79	32,74	20,21	22,56	15,81	12,74	121,85
8	Fahrer 8	17,74	32,1	20,09	22,25	15,56	12,4	120,14
9	Fahrer 9	17,76	32,62	20,17	22,32	15,78	12,75	121,4
10	Fahrer 10	17,69	32,41	20,17	22,19	15,66	12,13	120,25
11	Fahrer 11	17,81	32,91	20,28	22,37	15,46	12,54	121,37
12	Fahrer 12	17,86	32,68	19,9	22,04	15,29	12,57	120,34
13	Fahrer 13	17,67	32,46	20,19	22,33	15,23	12,12	120
14	Fahrer 14	17,73	32,58	20,18	22,31	15,17	12,52	120,49
15	Fahrer 15	17,74	32,73	19,93	21,77	15,58	13,08	120,83
16	Fahrer 16	17,61	31,84	19,8	21,74	15,23	12,13	118,35
17	Fahrer 17	17,78	32,61	19,65	21,54	15,4	12,22	119,2
18	Fahrer 18	17,89	32,63	20,38	22,3	15,66	12,23	121,09
19	Fahrer 19	17,91	32,42	20,16	22,33	15,38	12,33	120,53
20	Fahrer 20	17,68	32,24	19,96	22,06	15,44	12,79	120,17
21	Fahrer 21	17,86	32,72	19,98	21,9	15,19	11,98	119,63
22	Fahrer 22	17,74	32,54	19,83	21,98	15,34	12,35	119,73
23	Fahrer 23	17,84	32,56	20,42	22,29	15,24	12,2	120,55
24	Fahrer 24	17,71	32,41	20,09	22,37	15,18	12,23	119,99
25	Fahrer 25	17,8	32,23	19,77	21,93	14,95	12,19	118,87
26	Fahrer 26	17,62	32,37	19,95	21,93	15,06	12,6	119,53
27	Fahrer 27	17,69	32,77	20,03	21,95	15,03	12,04	119,51
28	Fahrer 28	17,67	32,4	19,78	21,69	14,94	12,58	119,06
29	Fahrer 29	17,83	32,91	19,76	21,56	15,12	12,23	119,41
30	Fahrer 30	17,75	32,93	19,75	21,61	14,94	12,4	119,38
31	Fahrer 31	17,94	33,12	20,53	22,22	15,32	12,6	121,73
32	Fahrer 32	18,06	33,54	20,33	22,24	15,62	12,66	122,45
33	Fahrer 33	18,28	33,29	20,91	22,99	15,32	12,7	123,49
34	Fahrer 34	18,21	32,99	20,3	21,75	15,09	12,78	121,12
35	Fahrer 35	18,1	33,1	20,49	21,94	15,38	12,42	121,43
36	Fahrer 36	18,18	33,29	20,31	22,24	15,41	12,47	121,9
37	Fahrer 37	17,78	32,42	19,84	21,88	15,17	12,6	119,69
38	Fahrer 38	18,11	33,27	20,4	22,08	15,61	12,76	122,23
39	Fahrer 39	17,92	32,82	20,23	21,76	15,26	12,41	120,4
40	Fahrer 40	17,89	33,02	20,32	21,98	15,33	12,85	121,39
41	Fahrer 41	17,91	32,78	20,18	22,31	15,33	12,81	121,32
42	Fahrer 42	17,98	32,8	20,2	21,93	14,97	12,38	120,26
43	Fahrer 43	17,83	33,04	20,07	22,02	15,39	13,09	121,44
44	Fahrer 44	17,84	32,97	20,4	22,37	15,66	12,73	121,97

45	Fahrer 45	17,97	33,32	20,38	22,03	15,39	12,26	121,35
46	Fahrer 46	18,3	33,32	20,65	22,29	16,09	12,37	123,02
47	Fahrer 47	18,33	33,18	20,4	21,95	15,41	12,34	121,61
48	Fahrer 48	18,1	32,84	20,27	22,23	15,25	12,87	121,56
49	Fahrer 49	18	33,28	20,56	22,56	15,1	12,63	122,13
50	Fahrer 50	18,04	32,9	20,39	22,08	14,93	12,13	120,47
51	Fahrer 51	18,63	33,99	21,1	22,86	16,12	12,78	125,48
52	Fahrer 52	18,14	32,96	20,79	22,41	15,65	12,79	122,74
53	Fahrer 53	18,25	33,27	20,17	22,28	15,16	12,52	121,65
54	Fahrer 54	18,21	33,35	20,52	22,19	15,21	12,69	122,17
55	Fahrer 55	18,36	33,23	20,39	22,42	15,37	12,25	122,02
56	Fahrer 56	17,93	33,51	20,16	21,91	15,78	12,7	121,99
57	Fahrer 57	18,08	33,33	20,18	22,19	15,47	12,86	122,11
58	Fahrer 58	17,99	32,99	20,06	21,96	15,6	12,76	121,36
59	Fahrer 59	18	33,27	20,78	23,06	15,74	12,98	123,55
60	Fahrer 60	18,23	34,37	20,89	22,91	15,7	12,45	124,55
61	Fahrer 61	18,5	33,48	20,55	22,25	15,27	12,49	122,54
62	Fahrer 62	18,34	33,49	20,42	22,03	15,29	12,44	122,01
63	Fahrer 63	18,11	33,53	20,63	22,47	15,44	12,8	122,98
64	Fahrer 64	18,07	33,16	20,5	22,11	15,35	12,42	121,61
65	Fahrer 65	18,3	33,39	20,41	22,49	15,48	12,98	123,05
66	Fahrer 66	18,18	33,08	20,39	22,43	15,76	12,81	122,65
67	Fahrer 67	18,43	33,34	20,96	22,83	15,83	12,74	124,13
68	Fahrer 68	18,22	33,64	21,09	23,13	16,05	12,82	124,95
69	Fahrer 69	18,49	34,76	20,92	22,61	15,83	13,01	125,62

Tabelle 1: Kraftstoffverbrauchsdaten von VW-Modellen

VW-Modell	Leistung in KW	Verbrauch in l/ 100 km	CO2 Emission in g/km
1-VW Polo	44	4,5	114
2-VW Golf 6	63	4,9	113
3-VW Golf Variant	77	6,8	120
4-VW Beetle	77	5,9	137
5-Golf Sportsvan	63	4,9	113
6-VW Scirocco	132	6	139
7-VW Mini	30	4,2	105
8-VW Touran	103	5,7	149
9-VW Tiguan	90	6,5	152
10-VW Sharan	120	7	167
11-VW Touareg	150	7,5	173
12-VW Phaeton	175	9,5	240
13-VW Caddy	75	5,1	134
14-VW T5 Multivan	110	6,3	195
15-VW Passat	125	6,3	130
16-VW Eos	60	5,2	112
17-VW up!	40	3,8	106
18-VW Pan Americana	160	8,5	219
19-VW T5 California	160	7,9	189
20-VW Crafter	140	7,3	155

Tabelle 26: FIS Alpine Ski WM in Vail/Beaver Creek 2015

Start	Name des	Zeiten für die Teilstücke im 1. Durchgang				Zeiten für die Teilstücke im 2. Durchgang				totale
nr.	Fahrers	1.	2.	3.	4.	5.	6.	7.	8.	Zeit (Sek)
2	Pinturault	16,57	15,15	24,61	19,08	15,87	17,33	26,17	20,26	**155,04**
3	Hirscher	16,57	15,16	24,50	18,95	15,70	17,97	25,37	20,39	**154,61**
5	Dopfer	16,99	14,98	25,16	19,56	15,83	17,25	26,63	21,41	**157,81**
6	Ligety	16,37	14,98	24,84	19,23	15,83	17,01	26,04	19,86	**154,16**
7	Neureuther	16,87	14,74	24,80	19,00	16,17	17,02	26,44	20,22	**155,26**
8	Muffat-Jeande	16,62	15,37	25,08	18,93	15,76	17,12	26,54	20,3	**155,72**
9	Jitloff	16,67	14,90	24,87	19,35	15,83	18,20	25,56	20,66	**156,04**
10	Sandell	16,49	15,18	24,80	19,71	16,27	17,19	26,57	20,68	**156,89**
12	Nani	16,48	14,81	24,72	19,35	15,83	17,39	26,53	20,46	**155,57**
14	Kristoffersen	16,68	15,03	24,97	19,30	15,85	17,49	26,75	20,82	**156,89**
15	Janka	16,89	15,03	25,16	19,28	15,89	17,33	26,24	20,57	**156,39**
16	Luitz	16,58	14,97	25,53	19,34	16,06	17,26	27,33	20,76	**157,83**
17	Olsson	16,59	14,74	24,92	19,41	15,78	17,2	26,38	20,37	**155,39**
18	Simoncelli	16,70	15,48	24,91	19,23	16,56	16,69	26,86	20,62	**157,05**
19	Schörghofer	16,59	14,73	25,27	19,26	15,89	17,14	26,69	20,71	**156,28**
21	Eisath	14,33	15,11	25,24	19,46	15,91	17,06	26,42	19,9	**155,93**
22	Caviezel	16,92	15,62	25,45	19,21	16,22	17,32	26,93	20,84	**158,51**
23	Borsotti	17,02	15,20	25,46	19,86	15,92	17,25	26,87	20,77	**158,35**
25	Zubcic	16,93	14,84	25,29	19,28	16,13	17,69	26,61	20,25	**157,02**
26	Torsti	16,64	14,69	26,55	19,52	16,07	17,14	26,79	20,49	**157,89**
27	Zurbriggen	17,23	15,32	25,57	19,41	15,56	17,69	27,35	20,52	**158,65**
28	Murisier	17,05	14,95	26,03	20,08	15,97	16,88	26,17	21,82	**158,95**
29	Cook	16,83	15,26	25,12	19,39	15,92	16,87	26,68	20,35	**156,42**
31	Brown	17,24	15,53	25,43	19,51	16,3	17,17	26,31	20,5	**157,99**
33	Rubie	16,88	15,51	25,29	19,80	15,79	17,64	27,04	20,52	**158,47**
34	Ford	17,04	15,44	25,24	19,69	16,05	17,15	26,49	20,52	**157,62**
36	Kryzl	16,99	15,59	25,48	19,44	15,82	17,48	26,68	20,53	**158,01**
38	Philp	17,20	15,50	25,49	19,43	16,08	16,89	26,42	20,4	**157,41**
40	Choudounsky	17,44	15,40	25,51	19,51	16,28	17,4	26,89	20,47	**158,9**
44	Read	17,03	15,02	26,08	20,13	15,75	16,87	26,09	21,76	**158,73**

ABBILDUNGSVERZEICHNIS

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

LITERATURVERZEICHNIS

BELLGARDT, E. (2004): „Statistik mit SPSS - Ausgewählte Verfahren für Wirtschaftswissenschaftler“ (Verlag Franz Vahlen Gmbh)

BLUME, J. (1970): „Statistische Methoden für Ingenieure und Naturwissenschaftler – Grundlagen, Beurteilung von Stichproben, einfache lineare Regression, Korrelation“ (VDI Verlag Gmbh - Düsseldorf)

HACKL,P. /KATZENBEISSER, W. (1994): „Statistik – für Sozial- und Wirtschaftswissenschaften“ (Oldenbourg Verlag)

HAUER, P. (1991): „Einführung in die lineare Regression: Theoretische und anwendungsorientierte Aspekte“ (Diplomarbeit der Uni Wien)

KREYSZIG, E. (1977): „Statistische Methoden und ihre Anwendungen“ (Verlag Vandenhoeck & Ruprecht in Göttingen)

KURZ, B. (2003): „Lineare Regressionsanalyse“ (Diplomarbeit der Uni Wien)

LAY, D. (1996): „Linear Algebra and it´s applications“ (Addison Wesley Longman Verlag)

MONTGOMERY, D./PECK, E. (1991): „Introduction to linear regression analysis“ (Verlag John Wiley and Sons)

POKROPP, F. (1994): „Lineare Regression und Varianzanalyse“ (Oldenbourg Verlag)

RIEDWYL, H. (1997): „Lineare Regression und Verwandtes“ (Birkhäuser Verlag)

STRANG, G. (2003): „Lineare Algebra“ (Springer Verlag)

YOUDEN (1957): Industrial and Engin. Chem. S. 49, Band 71

LEBENSLAUF

Der Verfasser Hofegger Manuel wurde am 22.04.1990 in Scheibbs (Niederösterreich) geboren.

Er besuchte klassisch jeweils 4 Jahre die Volksschule, die Hauptschule und das BORG in Scheibbs, welches er im Juni 2008 erfolgreich abgeschlossen hat.

Nach 9 – Monaten Zivildienst beim Roten Kreuz entschloss er sich für ein Studium nach Wien zu gehen und begann dort im Wintersemester 2009 mit Geodäsie & Geoinformation zunächst sein einjähriges Gastspiel an der TU Wien, welches von einigen Abtastversuchen auch in anderen Studiengängen geprägt war, ehe er sich dann im Wintersemester 2010 für das Lehramtsstudium Mathematik und Geographie/Wirtschaftskunde an der Universität Wien entschied.

ABSTRACT

Die *Regressionsanalyse* umfasst alle statistischen Verfahren die der statistischen Analyse von Zusammenhängen zwischen zwei oder mehreren Zufallsvariablen dienen.
Ausgehend von einer Stichprobenerhebung aus der komplexeren Grundgesamtheit versucht die lineare Regression, die Art der Beziehung zwischen zwei Variablen festzustellen und durch eine mathematische Funktion diesen Zusammenhang zu beschreiben, da sie sich naturgemäß anschaulich repräsentieren lässt und sich somit adäquat zur Vermittlung grundsätzlicher Überlegungen eignet.
Grundsätzlich wird in vielen Praxisbeispielen, als Standardinstrument für derartige Schätzungen, die *Methode der kleinsten Quadrate* heran gezogen.
In weiterer Folge spielt auch die Herleitung von statistischen Tests und Konfidenzintervallen eine Rolle und das Modell wird zusätzlich durch die Normalverteilungsannahme erweitert.
Ein sehr praxisnahes Bild ergibt sich dann durch die Varianzanalyse, indem in Form eines Beispiels ein Vergleich mehrerer VW - Automodelle in Hinblick auf eine quantitative Variable y durchgeführt wird.
Allerdings dürfen auch die Abweichungen der Modellvoraussetzungen nicht zu kurz kommen, indem die Ursachen, Gründe, bzw. eventuellen Lösungsmöglichkeiten thematisiert werden. Im Fokus stehen hier vor allem das Problem der Kollinearität der unabhängigen Variablen bei der linearen Mehrfachregression, ebenso wie mögliche Lösungen für Varianzinhomogenität. Außerdem wird darauf geachtet, dass durch Eigenwerte/Eigenvektoren, Diagonalisierbarkeit, Orthogonalitätsprojektionen und schließlich der Singulärwertzerlegung, der Bezug zur Linearen Algebra mit zunehmendem Lesefortschritt der Arbeit sich zusehends vernetzter repräsentiert, ehe abschließend zur Hauptkomponentenanalyse übergeleitet wird.