

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

FAKULTÄT FÜR MATHEMATIK UND INFORMATIK

INSTITUT FÜR MATHEMATISCHE STOCHASTIK

BACHELORARBEIT

Vergleich von Regressionsmodellen mittels gleichmäßiger Konfidenzbänder

Comparison of regression models using simultaneous confidence
bands

Autor:

Rolf Tobias Hajo Henrik
Henning Hause

Erstgutachter:

Prof. Dr. Tatyana
Krivobokova

Abgabetermin:

11.07.2017

Zweitgutachter:

Jun.-Prof. Dr. Andrea
Krajina

Studiengang:

B. Sc. Mathematik

Inhaltsverzeichnis

1	Regression und Konfidenzbänder	4
1.1	Regression	4
1.2	Konfidenzbänder	8
1.3	Konfidenzbänder auf \mathbb{R}^p für ein multiples lineares Regressionsmodell	9
1.4	Konfidenzbänder auf einem Intervall für ein einfaches lineares Regressionsmodell	15
1.5	Konfidenzbänder auf einem Intervall für ein multiples lineares Regressionsmodell	16
1.6	Konfidenzbänder auf auf einem Intervall für ein Regressionsmodell mit Polynomgestalt	21
2	Vergleich von zwei Regressionsmodellen	26
2.1	F-Test zum Vergleich von zwei Regressionsmodellen	27
2.2	Vergleich von Regressionsmodellen mit Konfidenzbändern	30
3	Einen Teil eines Regressionsmodells überprüfen	33
3.1	F-Test zum Überprüfen eines Teils eines Regressionsmodells	33
3.2	Einen Teil eines Regressionsmodells auf einem Intervall überprüfen	34
3.3	Einen Teil eines Regressionsmodells überprüfen, wenn das Modell Polynomgestalt hat	35
4	Regression und Konfidenzbänder für abhängige Daten	37
4.1	Autoregressive Modelle und AR(1)	37
4.2	AR(1) und das <i>nlme</i> Paket	39
4.3	Konfidenzbänder für AR(1)	42
5	Datenbeschreibung und Resultate	45
5.1	Datenbeschreibung	45
5.2	Vergleich verschiedener Konfidenzbänder	47
5.3	Vergleich von polynomen Regressionsmodellen mit Grad Vier, Fünf und Sechs	49
5.4	Vergleich von 10 kPa und 30 kPa Daten	54
5.5	Einen Teil eines Regressionsmodells überprüfen	56
	Selbstständigkeitserklärung	63

Einleitung

Diese Bachelorarbeit behandelt Methoden, um Regressionsmodelle zu vergleichen. Seien zwei Regressionsmodelle

$$Y_i = X_i \beta_i + e_i, \quad i \in \{1, 2\}$$

gegeben. Dabei seien $Y_i = (Y_{i,1}, \dots, Y_{i,n_i})'$ zwei Vektoren mit abhängigen Variablen.

Weiterhin seien X_i zwei $n_i \times (p+1)$ Designmatrix mit festem Design. Das bedeutet, für $l \in \{1, \dots, \min(n_1, n_2)\}$ ist die l -te Zeile von X_1 gleich der l -ten Zeile von X_2 .

Weiterhin seien $\beta_i = (\beta_{i,0}, \dots, \beta_{i,p})'$ zwei Koeffizientenvektor und $e_i = (e_{i,1}, \dots, e_{i,n_i})'$ zwei Vektoren mit Zufallsfehlern. Also sind Y_1 und Y_2 zwei Gruppen von Beobachtungen, die von den selben unabhängigen Variablen abhängen.

Ein Beispiel, entnommen [Liu64, S. 116], ist zu vergleichen, wie das Alter den Blutdruck von Männern und Frauen beeinflusst. Man hat zwei verschiedene Beobachtungen Y_1 und Y_2 , den Blutdruck der Männer und den Blutdruck der Frauen, und versucht diese Beobachtungen mit dem selben Modell X_i zu erklären. Da man eventuell verschieden viele Beobachtungen hat, haben die Y_i dann verschiedene Dimensionen.

Im Modell der gewöhnlichen linearen Regression ist $e_i \sim \mathcal{N}_{n_i}(0, \sigma^2 I_{n_i})$. Das Testproblem für den Vergleich von Regressionsmodellen ist

$$H_0 : \beta_1 = \beta_2 \text{ vs. } H_1 : \beta_1 \neq \beta_2$$

Dabei gleicht man die Dimensionen der β_i an, indem man Nullen einfügt.

Dieser Test wird normalerweise mit einem F-Test durchgeführt. In dieser Ausarbeitung werden alternative, auf Konfidenzbändern basierende Methoden nach [Liu64] betrachtet. Insbesondere die folgenden Spezialfälle sind von Interesse:

- Das Modell hat Polynomgestalt, das heißt die l -te ($1 \leq l \leq n_i$) Zeile der Designmatrix X ist von der Form $(1, x_{i,l}, x_{i,l}^2, \dots, x_{i,l}^p)$.
- Den Fehlern liegt ein AR(1)-Prozess zugrunde, das heißt $e_i \sim \mathcal{N}_{n_i}(0, \sigma^2 \Upsilon)$ mit $\Upsilon \neq I_{n_i}$.
- Man vergleicht die Regressionsmodelle nicht direkt sondern prüft, ob ein Teil des Regressionsmodells nicht signifikant ist. Das heißt man unterteilt $\beta_1 = (\beta_{1,1}, \beta_{1,2})$ und prüft, ob $\beta_{1,2}$ signifikant gleich Null ist.

Die Motivation für diese Betrachtungen ist der Vergleich von verschiedenen Regressionsmodellen für Stammzellen.

Das Kapitel 1 stellt die Konzepte der Regression und der Konfidenzbänder vor. Außerdem werden Methoden angegeben, wie für diese Regressionsgraphen Konfidenzbänder konkret zu berechnen sind.

Das Kapitel 2 führt aus, wie man Regressionsmodelle mittels des F-Tests vergleicht. Danach werden, mit Hilfe der Methoden aus dem ersten Kapitel, bessere Möglichkeiten zum Vergleich von Regressionsmodellen angegeben.

Das Kapitel 3 erläutert, wie man überprüfen kann, ob ein Teil des Regressionsmodells keinen Einfluss auf die Daten hat. Dazu wird zuerst die Methode des F-Tests angegeben. Dann wird mit Hilfe der Methoden aus Kapitel 1 bessere Möglichkeiten beschrieben, einen Teil des Konfidenzbandes zu überprüfen.

Das Kapitel 4 beschäftigt sich mit der Fragestellung, wie Schätzer für die Koeffizientenvektoren und für die Varianz der Fehler und Konfidenzbänder bei abhängigen Daten berechnet werden können. Konkret wird der Fall betrachtet, dass dem Regressionsmodell ein AR(1)-Prozess zugrunde liegt.

Das letzte Kapitel stellt die Daten der Stammzellen, die diese Ausarbeitung motivierten, vor. Außerdem werden die Resultate, die man erhält, wenn man die Methoden aus den ersten drei Kapiteln auf die Stammzelldaten anwendet, vorgestellt.

Am Ende jedes Abschnittes in dem ersten Kapitel folgt auf die dort behandelte Theorie oder Methode ein Beispiel. Dieses Beispiel zieht sich durch die ersten Vier Kapitel und wird jeweils ergänzt oder modifiziert.

Außerdem wird, nachdem eine Methode eingeführt wurde, die tatsächliche Überdeckungswahrscheinlichkeit mittels einer Simulation näherungsweise Bestimmt. Bei den Simulationen erhält man folgende Ergebnisse:

	unabhängig	AR bekannt	AR
Konfidenzband auf ganz \mathbb{R}^p	0.953	0.94	1.00
Konfidenzband auf einem Polyeder	.956	0.93	1.00
Konfidenzband auf einem Polyeder für Polynome	0.951	0.90	1.00

In den Spalten stehen die verschiedenen zugrunde liegenden Modelle. Dabei steht *unabhängig* für ein multilineares, homoskedastisches Modell mit unabhängigen Fehlern. *AR bekannt* steht für ein Modell dem ein AR(1)-Prozess zugrunde liegt, bei dem der Korrelationsparameter bekannt ist. Bei dem *AR* Modell ist der Korrelationsparameter unbekannt und muss zusätzlich geschätzt werden.

In den Zeilen stehen die verschiedenen Methoden. Von oben nach unten sind dies Konfidenzbänder auf ganz \mathbb{R}^p , Konfidenzbänder auf einem Polyeder $A \subset \mathbb{R}^p$ sowie Konfidenzbänder auf einem Polyeder A , wenn das Modell Polynomgestalt hat.

Die in dieser Ausarbeitung verwendeten R und C++ Codes sind in dem Paket KBminmaxpoly auf der Seite

<https://github.com/fake1884/KBminmaxpoly>

gefunden werden. Außerdem liegt das R-Paket auch bei. Eine Beschreibung ist in der beiliegenden README Datei zu finden.

Genauere Beschreibungen zum Code sind in der Arbeit an der Stelle zu finden, an dem der Code verwendet, beziehungsweise eingeführt wird. Ansonsten ist der Code durch kommentiert.

1 Regression und Konfidenzbänder

Ziel des ersten Kapitels ist es, gleichmäßige Konfidenzbänder für Regressionsmodelle auf einem Intervall zu konstruieren, wobei die Regressionsmodelle Polynomgestalt haben.

Ein Konfidenzband auf einer Menge A zu bestimmen, meint, dass die unabhängige Variable x aus dieser Menge kommt.

Im ersten Abschnitt dieses Kapitels wird das grundlegende Konzept der Regression eingeführt. Im darauf aufbauenden Abschnitt über Konfidenzbänder werden Konfidenzbänder definiert.

In den folgenden drei Abschnitten werden konkrete Methoden zum Bestimmen von Konfidenzbändern angegeben. Die einfachste Möglichkeit ist Konfidenzbänder auf \mathbb{R}^p zu konstruieren. Deswegen betrachten wir diese Möglichkeit zuerst.

Danach verallgemeinern wir diese Methode auf Konfidenzbänder auf einem Polyeder

$$A = \{(x_1, \dots, x_p)' : a_i \leq x_i \leq b_i, i = 1, \dots, p\} \subset \mathbb{R}^p \text{ für } -\infty \leq a_i < b_i \leq \infty, i = 1, \dots, p$$

Dabei bezeichnet x' die Transponierte von x während die Ableitung von x mit $\frac{d}{dx}x$ bezeichnet wird.

Weiterhin ist $x_i = (1, x_{(0)}) \in \mathbb{R}^{p+1}$ die i -te Zeile der Designmatrix X . Das heißt, das Konfidenzband wird für $x_{(0)} \in A \subset \mathbb{R}^p$ konstruiert.

Zuerst betrachten wir den einfacheren Fall, dass es nur eine unabhängige Variable gibt. Danach verallgemeinern wir auf den Fall von p unabhängigen Variablen mit $p \in \mathbb{N}$.

Im letzten Abschnitt betrachten wir, wie man Konfidenzbänder auf $A \subset \mathbb{R}^p$ konstruiert, wenn das Modell Polynomgestalt hat.

Das erste Kapitel orientiert sich stark an dem Buch [Liu64]. Außerdem werden in Abschnitt 1.1 das Buch [Geo09] und das Skript [Kri15] verwendet.

1.1 Regression

Bei einer Regression geht es darum, den Zusammenhang zwischen einer Zufallsvariablen $Y \in \mathbb{R}^n$ und einer Zufallsmatrix $X \in \mathbb{R}^{n \times (p+1)}$ mit $n, p \in \mathbb{N}$ zu beschreiben. Dazu versucht man eine Funktion f zu finden, sodass

$$\|f(X) - Y\|$$

minimiert wird. Dabei ist $\|\cdot\|$ eine beliebige, aber feste Abstandsfunktion. Außerdem geht man von dem Modell

$$Y = f(X) + e$$

aus. Dabei ist $e \in \mathbb{R}^n$ ein nicht beobachtbarer zufälliger Fehler. Wie man an der Gleichung erkennt, liegt die Annahme zugrunde, dass die Messung der Y -Werte fehlerbehaftet ist. Dagegen ist die Messung der $f(X)$ -Werte fehlerfrei.

In dieser Ausarbeitung wird immer ein festes Design verwendet. Beim festen Design sind die Einträge in der Matrix X fest gewählte Zahlen.

Da dann Y mittels f von X abhängt, nennt man Y die abhängige Variable und X die unabhängige Variable.

Andere übliche Bezeichnungen sind Zielvariable für Y und Kovariable oder Einflussgröße für X .

Weiterhin wird in dieser Ausarbeitung immer ein multiples, lineares Regressionsmodell benutzt. Das heißt, es liegt die Annahme zugrunde, dass $f(X) = X \cdot \beta$ ist. Dabei bezeichnet $\beta = (\beta_0, \dots, \beta_p)$ unbekannte Konstanten.

Ein besonderes Modell, das im Abschnitt 1.6 und Kapitel 5 wichtig wird, ist das Polynom-Modell. Bei diesem Modell ist die l -te Zeile von X gegeben durch $\tilde{x}_l = (1, x_l^1, x_l^2, \dots, x_l^p)$. Damit wird das Regressionsmodell zu

$$Y = X\beta + e \quad (1)$$

Zusätzlich ist in dieser Ausarbeitung für dieses Kapitel und die Kapitel 2 und 3 das Modell homoskedastisch. Das heißt, e folgt einer n -dimensionalen Normalverteilung mit Erwartungswert Null und unbekannter Varianz $\sigma^2 \cdot I_n$. Dabei ist I_n die n -dimensionale Einheitsmatrix.

In dieser Ausarbeitung werden Matrizen immer mit großen lateinischen Buchstaben bezeichnen, wenn ihre Einträge bekannt sind und mit großen griechischen Buchstaben, wenn ihre Einträge unbekannt sind. Diese Notation für bekannte und unbekannte Werte gilt auch für Skalare.

Außerdem bezeichnet $E(W)$ den Erwartungswert von W .

In Kapitel 4 wird geklärt, was passiert, wenn $e \sim \mathcal{N}_n(0, \sigma^2 \Upsilon)$. Dabei ist Υ eine Matrix mit unbekannten Werten, die die Abhängigkeitsstruktur der Y -Werte modelliert. Man nennt eine Matrix wie Υ Kovarianzmatrix.

Die Parameter β und σ^2 charakterisieren das Modell und werden aus den beobachteten Daten geschätzt. Ihre Schätzer werden mit $\hat{\beta}$ und $\hat{\sigma}^2$ bezeichnet.

Der folgende Satz gibt eine Möglichkeit die Schätzer $\hat{\beta}$ und \hat{e} beziehungsweise $\hat{\sigma}^2$ konkret zu berechnen.

Satz 1.1.1. In einem homoskedastischen, multiplen linearen Regressionsmodell sind der *ordinary least squares Schätzer* (OLS) und der *maximum likelihood Schätzer* (MLE) für $\hat{\beta}$, \hat{e} und $\hat{\sigma}^2$ identisch und wie im Folgenden gegeben:

$$\hat{\beta} = (X'X)^{-1}X'Y \quad (2)$$

Dabei ist X^{-1} die inverse Matrix zu X .

$$\hat{e} = (Y - X\hat{\beta}) = (I - H)Y \quad (3)$$

mit $H = X(X'X)^{-1}X'$

$$\hat{\sigma}^2 = \|\hat{e}\|^2/(n - p - 1) = \|Y - X\hat{\beta}\|^2/(n - p - 1) \quad (4)$$

Beweis. Das *least squares criterion* besagt, dass

$$L(\beta) = \|Y - X\beta\|^2 = (Y - X\beta)'(Y - X\beta)$$

zu minimieren ist. Da $Y'X\beta$ eine skalare Größe ist, stimmt sie mit ihrer Transponierten $\beta'X'Y$ überein. Benutzt man dies erhält man aus der obigen Gleichung

$$L(\beta) = Y'Y - 2\beta'X'Y + \beta'X'X\beta$$

Also muss der OLS folgende Gleichung erfüllen

$$\frac{d}{d\beta}L(\beta)|_{\beta=\hat{\beta}} = -2X'Y + 2X'X\hat{\beta} = 0 \quad (5)$$

Also muss gelten

$$X'X\hat{\beta} = X'Y \quad (6)$$

Da X vollen Zeilenrang hat, ist $X'X$ invertierbar und man findet

$$\hat{\beta} = (X'X)^{-1}X'Y$$

für die anderen beiden Schätzer siehe [Liu64, S. 4].

Für die Aussage, dass OLS und ML-Schätzer übereinstimmen siehe [Mun13, S. 74] oder [Liu64, S. 3] \square

Beispiel 1.1.2. In diesem einfachen Beispiel werden die Daten Y auf die Zeit regressiert. Dabei wird als Modell ein Polynom vom Grad Drei benutzt. Das heißt, es wird von dem homoskedastischen Modell

$$Y = X \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + e$$

ausgegangen.

Bei diesem Beispiel sind die Y -Werte zufällige Zahlen, die am Anfang der Datei *beispiele.R* erzeugt werden. Die Motivation für diese Daten ist, einen Wachstumsprozess zu beschreiben.

Um die Daten zu erzeugen, wird zuerst die seed auf 100 gesetzt. Danach wird auf dem Intervall $[0,1]$ ein äquidistantes Gitter G mit 50 Punkten erzeugt. Aus diesen wird die Designmatrix auf Grundlage eines polynomiellen Regressionsmodells vom Grad Drei erzeugt. Das heißt, die l -te Zeile von X ist gegeben durch $\tilde{x}_l = (1, x_l, x_l^2, x_l^3)$ für $\tilde{x}_l \in G, l \in \{1, \dots, 50\}$. Als nächstes wird $\beta = (10, 5, -4, 7)$ und $\sigma = 1$ gewählt.

Dann wird $e \sim \mathcal{N}_n(0, \sigma^2 I_n)$ generiert und Y nach dem Modell (1) berechnet.

In der Abbildung 1 sind die Datentupel (Y, G) eingezeichnet:

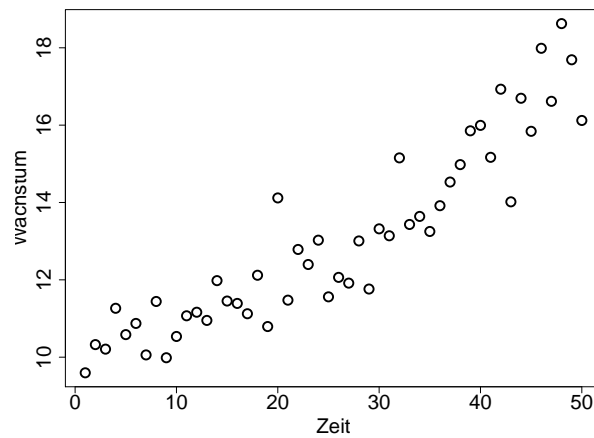


Abbildung 1: Plot der Beispieldaten

In den folgenden Kapiteln wird mit einem normierten Y gearbeitet, indem Y durch $\max(Y)$ geteilt wird. Dadurch erhält man Werte zwischen Null und Eins, wodurch sich die Berechnungen vereinfachen.

Man kann die Daten normieren, da von

$$Y = X\beta + e$$

ausgehend, man zu

$$\tilde{Y} := \max(Y)^{-1}Y = X \max(Y)^{-1}\beta + \max(Y)^{-1}e =: X\tilde{\beta} + \tilde{e}$$

gelangt und dieses Modell immer noch multipel linear und homoskedastisch ist.

Man kann ohne Einschränkung der Allgemeinheit davon ausgehen, dass die Elemente von X Werte in $[0, 1]$ annehmen. Sind die Werte nicht in $[0, 1]$ kann man mit $\max(X)$ normieren. Insgesamt erhält man aus $\tilde{\beta}$ wieder β indem man mit $\max(Y)$ multipliziert.

Benutzt man die Formeln (2) und (4) von oben erhält man $\hat{\beta} = (0.54780579 ; 0.19037112 ; 0.03639814 ; 0.18395970)$ und $\hat{\sigma}^2 = 0.04515636$. In der Abbildung 2 sind die Daten mit der Regressionsgerade eingezeichnet.

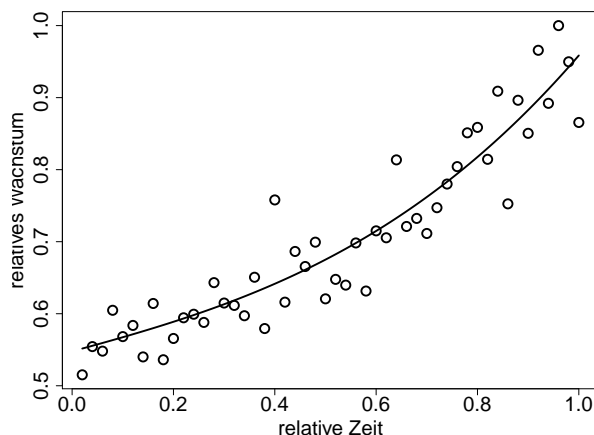


Abbildung 2: Plot der Beispieldaten mit Regressionsgerade

Transferiert man die Schätzwerte mittels Multiplikation mit $\max(Y) = 18.62582$ wieder zurück erhält man $\hat{\beta} = (10.2033308 ; 3.5458178 ; 0.6779452 ; 3.4263999)$ als Schätzer für $\beta = (10, 5, -4, 7)$ und $\hat{\sigma} = 0.8410741$ als Schätzer für $\sigma = 1$.

1.2 Konfidenzbänder

In diesem Abschnitt werden Konfidenzbänder definiert und der Unterschied zwischen punkweisen und gleichmäßigen Konfidenzbändern wird erklärt. Die folgende Definition orientiert sich an [Geo09, S. 229].

Definition 1.2.1. Sei $(\chi, \mathcal{F}, \mathbb{P}_\vartheta : \vartheta \in \Theta)$ ein stochastisches Modell, Σ eine beliebige Menge, $\tau : \Theta \rightarrow \Sigma$ eine zu ermittelnde Kenngröße für den Parameter ϑ , und $0 < \alpha < 1$. Dabei bezeichnet $\mathbb{P}(\cdot)$ ein Wahrscheinlichkeitsmaß.

Eine Abbildung $C : \chi \rightarrow \mathcal{P}(\Sigma)$, die jedem möglichen Beobachtungsergebnis $x \in \chi$ eine Menge $C(x) \subset \Sigma$ zuordnet, heißt ein Konfidenz- oder Vertrauensbereich für τ zum Irrtumsniveau α (beziehungsweise Sicherheitsniveau $1 - \alpha$), wenn

$$\inf_{\vartheta \in \Theta} \mathbb{P}_\vartheta(x \in \chi : C(x) \ni \tau(\vartheta)) \geq 1 - \alpha.$$

In dieser Ausarbeitung wird immer davon ausgegangen, dass $C(x)$ von der Form

$$C(x) = \left\{ y \in \mathbb{R}^n : x' \hat{\beta} - c \hat{\sigma} \sqrt{x'(X'X)^{-1}x} \leq y \leq x' \hat{\beta} + c \hat{\sigma} \sqrt{x'(X'X)^{-1}x} \right\}$$

ist. Dabei ist c der sogenannte kritische Wert. Für den Rest der Ausarbeitung wird dies durch

$$\mathbb{P} \left(x' \beta \in x' \hat{\beta} \pm c \hat{\sigma} \sqrt{x'(X'X)^{-1}x} \right) = 1 - \alpha$$

abgekürzt. Eine andere mögliche Schreibweise ist

$$\mathbb{P}(x'\beta \in M_n(x; Y, \alpha)) = 1 - \alpha$$

mit

$$M_n(x) = (x'\hat{\beta} - c, x'\hat{\beta} + c)$$

und c passend.

Weiterhin unterscheidet man punktweise und gleichmäßige Konfidenzbänder. Für punktweise Konfidenzbänder gilt für alle $x_{(0)} = (x_1, \dots, x_p) \in \mathbb{R}^p$

$$\mathbb{P}(x'\beta \in x'\hat{\beta} \pm c) = 1 - \alpha$$

während für gleichmäßige

$$\mathbb{P}(x'\beta \in x'\hat{\beta} \pm c \text{ für alle } x_{(0)} \in \mathbb{R}^p) = 1 - \alpha$$

gilt. Dabei ist $\beta = (\beta_0, \dots, \beta_p)$ der Koeffizientenvektor, der das Modell festlegt, und $\hat{\beta}$ ein Schätzer für β .

In dieser Ausarbeitung werden nur gleichmäßige Konfidenzbänder betrachtet.

1.3 Konfidenzbänder auf \mathbb{R}^p für ein multiples lineares Regressionsmodell

Seien $x, \beta, \hat{\beta} \in \mathbb{R}^{p+1}$, $X \in \mathbb{R}^{n \times (p+1)}$. Wir gehen davon aus, dass $x = (1, x_{(0)}) \in \mathbb{R}^{p+1}$. In dem verbleibenden Teil dieses Kapitels werden Konfidenzbänder der Form

$$x'\beta \in x'\hat{\beta} \pm c \hat{\sigma} \sqrt{x'(X'X)^{-1}x}$$

bestimmt. Dabei wird der kritische Parameter c so gewählt, dass

$$\mathbb{P}\left(x'\beta \in x'\hat{\beta} \pm c \hat{\sigma} \sqrt{x'(X'X)^{-1}x} \text{ für alle } x_{(0)} \in A\right) = 1 - \alpha$$

gilt. Dabei ist A ein Bereich, dem aus bestimmten Gründen besonderes Interesse gilt. Mehr zur Wahl von A steht im nächsten Abschnitt. In diesem Abschnitt ist $A = \mathbb{R}^p$.

Wegen des folgenden Resultates aus [Liu64, S. 66] benutzen wir diese Form von Konfidenzbändern.

Satz 1.3.1. Für beliebiges $\beta \in \mathbb{R}^{p+1}$ und $\sigma > 0$ gilt

$$\mathbb{P}\left(x'\beta \in x'\hat{\beta} \pm \sqrt{(p+1)f_{p+1, n-p-1}^\alpha} \hat{\sigma} \sqrt{x'(X'X)^{-1}x} \text{ für alle } x_{(0)} \in \mathbb{R}^p\right) = 1 - \alpha$$

Dabei ist $f_{p+1, n-p-1}^\alpha$ das α -Quantil der F-Verteilung mit Freiheitsgraden $p+1$ und $n-p-1$.

Dieser Satz er6ffnet uns eine einfache M6glichkeit, Konfidenzbänder auf \mathbb{R}^p zu bestimmen. Da das Konfidenzband in Satz 1.3.1 von hyperbolischer Form ist, betrachten wir in den folgenden Abschnitten immer hyperbolische Konfidenzbänder.

Um den Satz 1.3.1 beweisen zu können, brauchen wir die folgenden Resultate:

Lemma 1.3.2. Es ist

$$\begin{aligned} & \mathbb{P} \left(x' \beta \in x' \hat{\beta} \pm c \hat{\sigma} \sqrt{x'(X'X)^{-1}x} \text{ für alle } x_{(0)} \in A \right) \\ &= \mathbb{P} \left(\sup_{x_0 \in A} \left| \frac{x'(\beta - \hat{\beta})}{\hat{\sigma} \sqrt{x'(X'X)^{-1}x}} \right| \leq c \right) \\ &= \mathbb{P}(S \leq c) \end{aligned}$$

mit

$$\sup_{x_0 \in A} \frac{|x'(\beta - \hat{\beta})|}{\hat{\sigma} \sqrt{x'(X'X)^{-1}x}} =: S$$

Beweis.

$$\begin{aligned} & \mathbb{P} \left(x' \beta \in x' \hat{\beta} \pm c \hat{\sigma} \sqrt{x'(X'X)^{-1}x} \text{ für alle } x_{(0)} \in A \right) \\ &= \mathbb{P} \left(x' \hat{\beta} - c \hat{\sigma} \sqrt{x'(X'X)^{-1}x} \leq x' \beta \leq x' \hat{\beta} + c \hat{\sigma} \sqrt{x'(X'X)^{-1}x} \text{ für alle } x_{(0)} \in A \right) \\ &= \mathbb{P} \left(-c \hat{\sigma} \sqrt{x'(X'X)^{-1}x} \leq x'(\beta - \hat{\beta}) \leq c \hat{\sigma} \sqrt{x'(X'X)^{-1}x} \text{ für alle } x_{(0)} \in A \right) \\ &= \mathbb{P} \left(-c \leq \frac{x'(\beta - \hat{\beta})}{\hat{\sigma} \sqrt{x'(X'X)^{-1}x}} \leq c \text{ für alle } x_{(0)} \in A \right) \\ &= \mathbb{P} \left(\sup_{x_0 \in A} \frac{|x'(\beta - \hat{\beta})|}{\hat{\sigma} \sqrt{x'(X'X)^{-1}x}} \leq c \right) \end{aligned}$$

□

Es geht in den folgenden Abschnitten immer darum, die Verteilung von S zu bestimmen. Um diese Verteilung auf ganz \mathbb{R}^p zu bestimmen, wird das folgende Ergebnis benutzt, welches sich an [Liu64, S. 6] orientiert.

Satz 1.3.3. Wenn Modell (1) homoskedastisch ist, gilt mit Satz 1.1.1

1. $\hat{\beta} \sim \mathcal{N}_{p+1}(\beta, \sigma^2(X'X)^{-1})$. Dabei ist $\mathcal{N}_{p+1}(\beta, \sigma^2(X'X)^{-1})$ die $(p+1)$ -dimensionale Normalverteilung mit Erwartungswert β und Kovarianzmatrix $\sigma^2(X'X)^{-1}$.
2. $\hat{e} \sim \mathcal{N}_n(0, \sigma^2(I - H))$ mit $H = X(X'X)^{-1}X'$

3. $\hat{\sigma}^2 \sim \frac{\sigma^2}{n-p-1} \chi_{n-p-1}^2 = \frac{\sigma^2}{v} \chi_v^2$. Dabei ist $n - p - 1 := v$, was auch im Verlauf dieser Ausarbeitung so beibehalten wird. Weiterhin ist χ_v^2 die Chi-Quadrat-Verteilung mit v Freiheitsgraden.
4. $\hat{\beta}$ und $\widehat{\sigma}^2$ sind unabhängig.

Beweis. 1. Benutzt man Modell (1) und die Voraussetzung $e \sim \mathcal{N}_n(0, \sigma^2 I_n)$, so folgt $Y = X\beta + e \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$. Da aus (2) folgt, dass $\hat{\beta} = (X'X)^{-1}X'Y$ gilt, ist $\hat{\beta}$ eine Linearkombination von Y . Daraus folgt, dass auch $\hat{\beta}$ normalverteilt ist. Es genügt also den Erwartungswert und die Kovarianz zu bestimmen.

$$\mathbb{E}(\hat{\beta}) = \mathbb{E}((X'X)^{-1}X'Y) = (X'X)^{-1}X'\mathbb{E}(Y) = (X'X)^{-1}X'X\beta = \beta$$

und

$$\begin{aligned} \text{Cov}(\hat{\beta}) &= \text{Cov}((X'X)^{-1}X'Y) = ((X'X)^{-1}X')\text{Cov}(Y)(X(X'X)^{-1}) \\ &= ((X'X)^{-1}X')\sigma^2 I(X(X'X)^{-1}) = \sigma^2(X'X)^{-1} \end{aligned}$$

2. Da nach (3) gilt, dass $\hat{e} = (I - H)Y$, ist \hat{e} auch eine Linearkombination von Elementen aus Y . Da Y normalverteilt ist, ist es also auch \hat{e} . Berechnet man wieder Erwartungswert und Kovarianz, erhält man:

$$\begin{aligned} \mathbb{E}(\hat{e}) &= \mathbb{E}((I - H)Y) = (I - H)\mathbb{E}(Y) = (I - H)X\beta \\ &= (I - X(X'X)^{-1}X')X = X - X(X'X)^{-1}X'X = X - X = 0 \end{aligned}$$

und

$$\text{Cov}(\hat{e}) = \text{Cov}((I - H)Y) = (I - H)\text{Cov}(Y)(I - H) = \sigma^2(I - H)$$

Bei dieser Berechnung wird benutzt, dass die Matrix $I - H$ sowohl symmetrisch als auch idempotent ist.

Die Matrix $I - H$ ist symmetrisch, da $H = X'(X'X)^{-1}X$ symmetrisch ist. Weiterhin ist die Matrix idempotent, da

$$\begin{aligned} (I - H)^2 &= (I - X(X'X)^{-1}X')(I - X(X'X)^{-1}X') \\ &= I + X(X'X)^{-1}X'X(X'X)^{-1}X' - 2X(X'X)^{-1}X' \\ &= I - X(X'X)^{-1}X' - X(X'X)^{-1}X' + X(X'X)^{-1}X' \\ &= (I - H) \end{aligned}$$

3. Sei $Q = I - H$ mit $H = X(X'X)^{-1}X'$, dann ist Q symmetrisch und idempotent. Daraus folgt

$$\begin{aligned}\text{Rang}(Q) &= \text{Spur}(Q) = \text{Spur}(I) - \text{Spur}(H) \\ &= n - \text{Spur}(X(X'X)^{-1}X') = n - \text{Spur}((X'X)^{-1}X'X) \\ &= n - (p + 1)\end{aligned}$$

Dabei benutzt man $\text{Spur}(AB) = \text{Spur}(BA)$.

Also kann man Q ausdrücken als $Q = T'LT$. Dabei ist T eine orthogonale Matrix und L eine passende Diagonalmatrix mit den ersten $n - (p + 1)$ Diagonalelementen Eins und den verbleibenden Diagonalelementen gleich Null.

Es ist $e = Y - X\beta \sim \mathcal{N}_n(0, \sigma^2 I_n)$. Sei $z = Te$. Dann ist $z \sim \mathcal{N}_n(0, \sigma^2 I_n)$, da aus der Orthogonalität von T folgt, dass

$$\text{Var}(z) = \text{Var}(Te) = T\text{Var}(e)T' = TT'\text{Var}(\sigma^2 I_n) = I_n \sigma^2$$

.

Weiterhin ist

$$Qe = (I - H)(Y - X\beta) = (I - H)Y - (I - H)X\beta = (I - H)Y = \hat{e}$$

da $(I - H)X = 0$, wie oben bereits berechnet. Es folgt

$$\begin{aligned}\|\hat{e}\|^2 &= \|Qe\|^2 = e'Q'Qe = e'Qe = (Te)'L(Te) \\ &= z_1^2 + \dots + z_{n-p-1}^2 \sim \sigma^2 \chi_{n-p-1}^2\end{aligned}$$

Setzt man dies nun in Gleichung (4), das heißt $\hat{\sigma}^2 = \|\hat{e}\|^2/(n - p - 1)$, ein, erhält man

$$\hat{\sigma}^2 = \|\hat{e}\|^2/(n - p - 1) \sim \frac{\sigma^2}{n - p - 1} \chi_{n-p-1}^2$$

4. Da $\hat{\beta}$ und \hat{e} normalverteilt sind, reicht es für die Behauptung der Unabhängigkeit zu zeigen, dass die Kovarianz zwischen den Beiden Null ist.

$$\begin{aligned}\text{Cov}(\hat{\beta}, \hat{e}) &= \text{Cov}((X'X)^{-1}X'Y, (I - H)Y) \\ &= (X'X)^{-1}X'\text{Cov}(Y, Y)(I - H) \\ &= \sigma^2(X'X)^{-1}X'(I - H) \\ &= \sigma^2(X'X)^{-1}((I - H)'X'X)' \\ &= \sigma^2(X'X)^{-1}((I - H)X)' = \sigma^2(X'X)^{-1}0 = 0\end{aligned}$$

□

Da wir nun von allen Bausteinen von S die Verteilung kennen, können wir den Satz 1.3.1 beweisen. Satz und Beweis orientiert sich an [Liu64, S. 66].

Beweis. Zu zeigen war, dass für beliebiges $\beta \in \mathbb{R}^{p+1}$ und $\sigma > 0$ im homoskedastischen Fall

$$\mathbb{P} \left(x' \beta \in x' \hat{\beta} \pm \sqrt{(p+1)f_{p+1,n-p-1}^\alpha} \hat{\sigma} \sqrt{x'(X'X)^{-1}x} \text{ für alle } x_{(0)} \in \mathbb{R}^p \right) = 1 - \alpha$$

gilt.

Sei P die eindeutige Wurzel aus $(X'X)^{-1}$ und somit $(X'X)^{-1} = P^2$. Definiere außerdem, dass $N := P^{-1}(\hat{\beta} - \beta)/\sigma$. Dann folgt N einer $(p+1)$ -dimensionalen Normalverteilung. Man hat also $N \sim \mathcal{N}_{p+1}(0, I_n)$.

Damit gilt

$$\begin{aligned} & \mathbb{P} \left(x' \beta \in \sqrt{(p+1)f_{p+1,n-p-1}^\alpha} \hat{\sigma} \sqrt{x'(X'X)^{-1}x} \text{ für alle } x_{(0)} \in \mathbb{R}^p \right) \\ &= \mathbb{P} \left(\sup_{x_{(0)} \in \mathbb{R}^p} \frac{|x'(\hat{\beta} - \beta)|}{\hat{\sigma} \sqrt{x'(X'X)^{-1}x}} \leq \sqrt{(p+1)f_{p+1,n-p-1}^\alpha} \right) \\ &= \mathbb{P} \left(\sup_{x_{(0)} \in \mathbb{R}^p} \frac{|(Px)'N|}{(\hat{\sigma}/\sigma) \sqrt{(Px)'(Px)}} \leq \sqrt{(p+1)f_{p+1,n-p-1}^\alpha} \right) \\ &= \mathbb{P} \left(\frac{\|N\|}{(\hat{\sigma}/\sigma)} \left(\sup_{x_{(0)} \in \mathbb{R}^p} \frac{|(Px)'N|}{\|Px\| \|N\|} \right) \leq \sqrt{(p+1)f_{p+1,n-p-1}^\alpha} \right) \\ &= \mathbb{P} \left(\frac{\|N\|}{(\hat{\sigma}/\sigma)} \leq \sqrt{(p+1)f_{p+1,n-p-1}^\alpha} \right) \\ &= \mathbb{P} \left(\frac{(\hat{\beta} - \beta)'(X'X)(\hat{\beta} - \beta)}{(p+1)\hat{\sigma}^2} \leq f_{p+1,n-p-1}^\alpha \right) \\ &= 1 - \alpha \end{aligned}$$

Dabei wird im vorletzten Schritt die Cauchy-Schwarz-Ungleichung $|(Px)'N| \leq \|Px\| \|N\|$ verwendet. Gleichheit gilt, falls $Px = \lambda N$ für ein N .

Im letzten Schritt wird aus Satz 1.3.3 benutzt, dass sowohl $(\hat{\beta} - \beta)'(X'X)(\hat{\beta} - \beta)$ als auch $(p+1)\hat{\sigma}^2$ einer χ^2 -Verteilung folgen. Außerdem sind beide unabhängig und somit ist ihr Quotient F-verteilt.

□

Beispiel 1.3.4. In der Abbildung 3 sind die Daten aus dem letzten Beispiel wieder mit einer polynomiellen Regression vom Grad Drei eingezeichnet.

Diesmal wird dazu noch ein Konfidenzband auf ganz \mathbb{R}^p eingezeichnet. Die Berechnung ist in dem Verzeichnis `/man/Beispiel-R.R` zu finden. Der kritische Parameter für dieses Konfidenzband ist 3.20876.

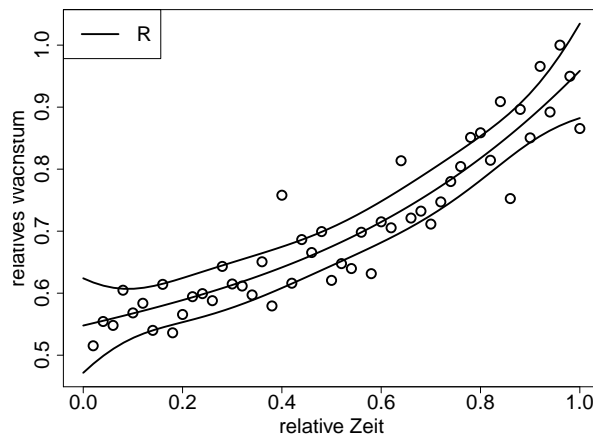


Abbildung 3: Weiterführung des Beispiels durch Konfidenzband auf ganz \mathbb{R}^p

Simulation 1.3.5. Außerdem wurde eine Simulation durchgeführt, um die Überdeckungswahrscheinlichkeit der Methode zu überprüfen.

Dazu wurden Daten auf zwei verschiedene Arten erzeugt. Zum Einen wurden die Daten auf Grundlage eines multiplen, linearen, homoskedastisches Regressionsmodell erzeugt. Zum anderen wurde ein multiples lineares Regressionsmodell mit einem AR(1)-Prozess zugrunde gelegt.

Der Grad des wahren Modelles ist 5 und es wird auch eine Regression mit Grad 5 durchgeführt. Dabei sind sowohl das wahre, als auch das geschätzte Modell von Polynomgestalt. Bei der Simulation wurden zuerst 100 Mal Daten mit den oben genannten Grundmodellen erzeugt. Danach wurden Konfidenzbänder um ein OLS-Schätzer beziehungsweise mit einem AR(1) transformierten OLS-Schätzer berechnet. Dann wurde gezählt, wie oft das wahre Modell in dem Konfidenzband lag.

Für das multiple lineare Regressionsmodell kamen keine guten Werte bei den Überdeckungswahrscheinlichkeiten heraus. Deswegen wurde ein vereinfachtes Modell verwendet. Dabei wurde der Grad des Modells verringert, der Fehler auf 0.5 fixiert und die Anzahl an Beobachtungen erhöht. Außerdem wurde die Überdeckungswahrscheinlichkeit auf drei Stellen nach dem Komma genau bestimmt. Für nähere Informationen zu diesem Modell siehe das R Skript *make-test-data-R.R*. Für das AR-Modell siehe das Skript *make-test-data-AR.R*.

Auf Grund von Laufzeitproblemen werden die Daten von den Funktionen *make-test-data-R* beziehungsweise *make-test-data-AR* erzeugt und gespeichert. Auch die jeweils berechneten Regressionsmodelle, Designmatritzen und weitere Werte werden abgespeichert. Diese Daten sind in */data* zu finden.

Bei der Berechnung der Überdeckungswahrscheinlichkeit müssen dann nur noch die Konfidenzbänder, die von der Anzahl an Simulationen abhängen, berechnet werden. Für eine Erklärung warum Simulationen benötigt werden, siehe die Kapitel 1.3 und 1.6.

Die Laufzeit für die Berechnung der Überdeckungswahrscheinlichkeiten von Konfidenzbän-

den bei denen Simulationen durchgeführt werden müssen beträgt ungefähr 15 Minuten bei den empfohlenen 5000 Wiederholungen.

Wie genau bei einem AR(1) Modell OLS-Schätzer bestimmt werden können wird in Kapitel 4 erläutert.

Die Ergebnisse sind in der folgenden Tabelle zusammengestellt:

	unabhängig	AR bekannt	AR
Konfidenzband auf ganz \mathbb{R}^p	0.953	0.94	1.00

Die erste Spalte gibt die Überdeckungswahrscheinlichkeit bei dem homoskedastischen Modell an und die zweite bei einem AR(1)-Grundmodell mit bekanntem Korrelationsparameter. Bei der dritten Spalte ist dieser Parameter unbekannt und muss zusätzlich geschätzt werden.

Man sieht, dass bei dieser Methode die Überdeckungswahrscheinlichkeit über der geforderten Wahrscheinlichkeit von 0.95 liegt. Der Grund dafür ist, dass das Konfidenzniveau auf ganz \mathbb{R}^4 eingehalten wird und nicht nur auf dem Intervall $[0,1]$. Für nähere Informationen dazu siehe Kapitel 1.6.

Diese Tabelle wird um weitere Zeilen mit den anderen Methoden Konfidenzbänder zu erzeugen erweitert, nachdem die neuen Methoden eingeführt worden sind.

1.4 Konfidenzbänder auf einem Intervall für ein einfaches lineares Regressionsmodell

Dieser Abschnitt orientiert sich an [Liu64, S. 17-23] und [LLP08] und [WB71].

Für den Fall, dass wir nur eine unabhängige Variable x_1 haben, vereinfacht sich das Konfidenzband aus Gleichung (1.3) zu

$$\mathbb{P} \left(\beta_0 + \beta_1 x_1 \in \hat{\beta}_0 + \hat{\beta}_1 x_1 \pm c\hat{\sigma} \sqrt{\nu(1, x_1)} \text{ für alle } x_1 \in (a, b) \right) \quad (7)$$

Dabei und für diesen Abschnitt gelten die folgenden Definitionen

$$\begin{aligned} \beta &:= (\beta_0, \beta_1)' \\ \hat{\beta} &:= (\hat{\beta}_0, \hat{\beta}_1)' \sim \mathcal{N}_2(\beta, \sigma^2(X'X)^{-1}) \\ P^2 &:= (X'X)^{-1} \\ N &:= P^{-1}(\hat{\beta} - \beta)/\sigma \sim \mathcal{N}_2(0, I_2) \\ T &:= N/(\hat{\sigma}/\sigma) = P^{-1}(\hat{\beta} - \beta)/\hat{\sigma} \\ v &:= n - 2 \\ x_1 &:= x \\ \nu(c, d) &:= (c, d)(X'X)^{-1}(c, d)' \\ A &:= (a, b) \end{aligned}$$

In [Liu64, S. 19-20] wird folgendes Ergebnis gezeigt

$$\begin{aligned}
& \mathbb{P} \left(\beta_0 + \beta_1 x \in \hat{\beta}_0 + \hat{\beta}_1 x \pm c\hat{\sigma}\sqrt{\nu(1, x)} \text{ für alle } x \in (a, b) \right) \\
&= \mathbb{P} \left(\sup_{x \in (a, b)} |(1, x)(\hat{\beta} - \beta)/\hat{\sigma}|/\sqrt{\nu(1, x)} < c \right) \\
&= \mathbb{P} \left(\sup_{x \in (a, b)} |(P \cdot (1, x)')' \cdot T|/\|P \cdot (1, x)'\| < c \right) \\
&= \mathbb{P}(T \in R_{h,2})
\end{aligned}$$

dabei ist

$$R_{h,2}(x) = \{T : |(P \cdot (1, x)')' \cdot T|/\|P \cdot (1, x)'\| < c\}$$

eine Region von der Form $(T : |v'T|/\|v\| < r) \subset \mathbb{R}^2$. Deswegen kann man den Winkel ϕ zwischen den Vektoren $P \cdot (1, a)'$ und $P \cdot (1, b)'$ mittels

$$\begin{aligned}
\cos \phi &= (P \cdot (1, a)')'(P \cdot (1, b)')/\|P \cdot (1, a)'\|\|P \cdot (1, b)'\| \\
&= (1, a)(X'X)^{-1}(1, b)'/\sqrt{\nu(1, a)\nu(1, b)}
\end{aligned}$$

berechnen. Jetzt folgt mit [WB71], dass gilt

$$\begin{aligned}
& \mathbb{P} \left(\beta_0 + \beta_1 x \in \hat{\beta}_0 + \hat{\beta}_1 x \pm c\hat{\sigma}\sqrt{\nu(1, x)} \text{ für alle } x \in (a, b) \right) \\
&= 1 - \frac{\phi}{\pi} \left(1 + \frac{c^2}{v} \right)^{-v/2} - \frac{2}{\pi} \int_0^{(\pi-\phi)/2} \left(1 + \frac{c^2}{v \sin^2(\theta + \phi/2)} \right)^{-v/2} d\theta
\end{aligned}$$

Man muss also

$$1 - \frac{\phi}{\pi} \left(1 + \frac{c^2}{v} \right)^{-v/2} - \frac{2}{\pi} \int_0^{(\pi-\phi)/2} \left(1 + \frac{c^2}{v \sin^2(\theta + \phi/2)} \right)^{-v/2} d\theta = 1 - \alpha$$

nach c lösen und kann dann das Konfidenzband (7) bestimmen.

1.5 Konfidenzbänder auf einem Intervall für ein multiples lineares Regressionsmodell

Im Abschnitt 1.3 wurden Konfidenzbänder auf ganz \mathbb{R}^p konstruiert. Allerdings kann es bei der Verwendung einiger unabhängigen Variablen nicht sinnvoll sein, das Konfidenzband auf ganz \mathbb{R}^p zu konstruieren. Ist die unabhängige Variable zum Beispiel ein Gewicht, reicht es, ein Konfidenzband auf \mathbb{R}_+^p zu konstruieren. Schließlich sind Gewichte immer positiv.

Nachfolgend wird das uns interessierende Gebiet mit A bezeichnet. Dabei ist

$$A = \{(x_1, \dots, x_p)' : a_i \leq x_i \leq b_i, i = 1, \dots, p\} \subset \mathbb{R}^p \text{ für } -\infty \leq a_i < b_i \leq \infty, i = 1, \dots, p$$

Außerdem definieren wir $x_{(0)} := (x_1, \dots, x_p)'$.

Eine Möglichkeit, ein Konfidenzband auf A zu bestimmen, ist, ein Konfidenzband auf ganz \mathbb{R}^p zu konstruieren und dann den Teil auf $\mathbb{R}^p \setminus A$ zu vernachlässigen.

Da dann die Bedingung $\mathbb{P}(x'\beta \in x'\hat{\beta} \pm d \text{ für alle } x_{(0)} \in A) = 1 - \alpha$ für mehr $x_{(0)}$ -Werte als nötig erfüllt sein muss, ist das Konfidenzband dann weiter, als es sein müsste. Dabei ist d eine beliebig, aber fest gewählte Konstante.

Besser ist die folgende Möglichkeit, die sich an [Liu64, S. 70] orientiert. Das Ziel ist wieder ein hyperbolisches Konfidenzband der Form

$$x'\beta \in x'\hat{\beta} \pm c \hat{\sigma} \sqrt{x'(X'X)^{-1}x} \quad (8)$$

zu erzeugen, um es mit dem Band aus Abschnitt 1.3 vergleichen zu können. Das Problem ist, ein geeignetes c zu finden, sodass

$$\mathbb{P}\left(x'\beta \in x'\hat{\beta} \pm c \hat{\sigma} \sqrt{x'(X'X)^{-1}x} \text{ für alle } x_{(0)} \in A\right) = 1 - \alpha$$

gilt.

Aus Satz 1.3.2 aus Abschnitt 1.3 ist bekannt, dass

$$\mathbb{P}\left(x'\beta \in x'\hat{\beta} \pm c \hat{\sigma} \sqrt{x'(X'X)^{-1}x} \text{ für alle } x_{(0)} \in A\right) = \mathbb{P}(S < c)$$

Dabei ist S gegeben durch

$$S = \sup_{x_{(0)} \in A} \frac{|x'(\beta - \hat{\beta})|}{\hat{\sigma} \sqrt{x'(X'X)^{-1}x}}$$

Da die Verteilung von S nicht von β und σ^2 abhängt, ist es in dieser Hinsicht ein Pivotelement. Dabei ist β wieder der Koeffizientenvektor und σ^2 die Varianz der Fehler.

Um zu sehen, dass die Verteilung von S nicht von β und σ^2 abhängt, betrachte Satz 1.3.3. Dort wurde gezeigt, dass $\hat{\beta} \sim \mathcal{N}_{p+1}(\beta, \sigma^2(X'X)^{-1})$ und $\hat{\sigma}^2 \sim \sigma^2/(n - p - 1) \cdot \chi_{n-p-1}^2$. Fügt man nun $1 = \sigma/\sigma$ ein, folgt die Behauptung.

Allerdings hängt die Verteilung von S immer noch in komplizierter Art und Weise von der Designmatrix X und den Grenzen a_i und b_i von A für $i = 1, \dots, p$ ab.

Seien P und N definiert wie im Beweis zu Satz 1.3.1. Das heißt,

$$\begin{aligned} P^2 &:= (X'X)^{-1} \\ N &:= \frac{P^{-1}(\hat{\beta} - \beta)}{\sigma} \end{aligned}$$

Dann hat

$$T = \frac{N}{(\hat{\sigma}/\sigma)} = \frac{P^{-1}(\hat{\beta} - \beta)}{\hat{\sigma}}$$

eine so genannte $\tau_{p+1,v}$ Verteilung. Damit kann man S ausdrücken als

$$\begin{aligned} S &= \sup_{x_0 \in A} \frac{|(Px)'(P^{-1}(\hat{\beta} - \beta))/\hat{\sigma}|}{\sqrt{(Px)'(Px)}} \\ &= \sup_{x_0 \in A} \frac{|(Px)'T|}{\|Px\|} \\ &= \sup_{v \in C(P,A)} \frac{|v'T|}{\|v\|} \end{aligned}$$

mit

$$\begin{aligned} C(P, A) &= \{\lambda Px : \lambda \geq 0, x \in A\} \\ &= \{\lambda(p_0 + x_1 p_1 + \dots + x_p p_p) : \lambda \geq 0, x_i \in [a_i, b_i] \text{ für alle } i = 1, \dots, p\} \end{aligned}$$

wobei $P = (p_0, \dots, p_p)$.

Allerdings ist es für ein allgemeines $p \geq 1$ sehr schwierig, eine Formel für die Verteilung von S explizit zu berechnen, um die kritische Konstante c zu bestimmen [Liu64, S.70, Z.18]. Deswegen wird eine Simulation benutzt.

Man beachte, dass $C(P, A)$ der Kegel ist, der von den Vektoren $p_0 + c_1 p_1 + \dots + c_p p_p$ aufgespannt wird, wobei c_i entweder a_i oder b_i für $i = 1, \dots, p$ ist.

Sei $\pi(t; P, A)$ die Projektion von $t \in \mathbb{R}^{p+1}$ auf den Kegel $C(P, A)$, d.h. $\pi(t, P, A)$ löst $\min_{v \in C(P,A)} \|v - t\|$. Jetzt folgt mit [Nai87, Theorem 2.1], dass S die Gleichung

$$S = \max(\|\pi(t, P, A)\|, \|\pi(-t, P, A)\|)$$

löst.

Man kann also die kritische Konstante c mit folgender Simulation finden:

1. Simuliere $N \sim \mathcal{N}_{p+1}(0, I_n)$ und $\hat{\sigma}/\sigma \sim \sqrt{\chi_v^2/v}$ mit $v = n - p - 1$ und p der Anzahl an unabhängigen Parametern.
2. Berechne $\pi(t, P, A)$ und $\pi(-t, P, A)$.
3. Berechne $S = \max(\|\pi(t, P, A)\|, \|\pi(-t, P, A)\|)$.

Wiederholt man diese Schritte k mal, so erhält man S_1, \dots, S_k . Dann ist c das $1 - \alpha$ Quantil von S_1, \dots, S_k .

Um $\pi(t, P, A)$ und $\pi(-t, P, A)$ zu bestimmen, wird folgendes Verfahren verwendet, das aus [Liu64, Appendix B] stammt:

Es soll das v in \mathbb{R}^{p+1} , das $\|v - t\|^2$ minimiert, gefunden werden. Dabei soll $v \in C(P, A_r)$ sein, mit $C(P, A_r)$ wie oben definiert.

Dazu betrachten wir

$$\|v - t\|^2 = v'v - 2t'v + t't$$

Man sieht, dass $t't$ unabhängig von v ist. Deshalb muss man es bei der Minimierung von $\|v - t\|^2$ nicht berücksichtigen.

Sei $e_j \in \mathbb{R}^{p+1}$ der j -te Einheitsvektor. Aus der Definition von $C(P, A_r)$ ist zu sehen, dass $v \in C(P, A_r)$ impliziert, dass $v = \lambda Px$ oder gleichwertig $P^{-1}v = \lambda x = (\lambda, \lambda x_1, \dots, \lambda x_p)'$ für ein $x \in A_r$ und $\lambda \geq 0$ gelten muss. Deshalb ist $e_1'P^{-1}v = \lambda \geq 0$ und $a_j \leq e_{j+1}'P^{-1}v/e_1'P^{-1}v = x_j \leq b_j$ für $j = 1, \dots, p$ oder gleichwertig

$$\begin{aligned} -e_1'P^{-1}v &\leq 0 \\ (e_{j+1}' - b_j e_1')P^{-1}v &\leq 0 \quad , \text{für } j = 1, \dots, p \\ (a_j e_1' - e_{j+1}')P^{-1}v &\leq 0 \quad , \text{für } j = 1, \dots, p \end{aligned}$$

Diese Beschränkungen kann man zusammenfassen zu

$$Av \leq 0$$

wobei A die $(2p+1) \times (p+1)$ Matrix

$$A = \begin{bmatrix} (e_2' - b_1 e_1')P^{-1} \\ (a_1 e_1' - e_2')P^{-1} \\ \vdots \\ (e_{p+1}' - b_p e_1')P^{-1} \\ (a_p e_1' - e_{p+1}')P^{-1} \\ -e_1'P^{-1} \end{bmatrix}$$

ist. Für Minimierungsprobleme dieser Art gibt es ein R-Paket namens *quadprog*, das zur Lösung dieser Minimierungsaufgabe benutzt wird.

Beispiel 1.5.1. Jetzt wird das Beispiel aus Abschnitt 1.3 fortgeführt, indem zu der Regression mit Grad Drei und dem Konfidenzband auf ganz \mathbb{R}^3 noch ein Konfidenzband auf A eingezeichnet wird. Dabei benutzt man $A = \{x_0 \in \mathbb{R}^3 : x_i \in [0, 1] \text{ für alle } i = 1, 2, 3\}$.

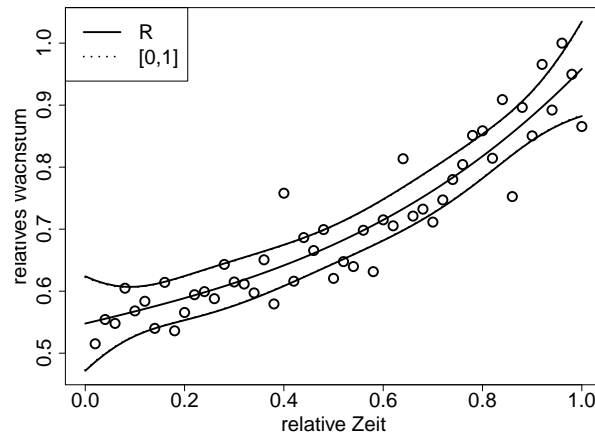


Abbildung 4: Beispieldaten aus Kapitel 1 mit einem Konfidenzband nach Kapitel 1.3 und einem nach Kapitel 1.5

Der kritische Wert für das Konfidenzband auf ganz \mathbb{R}^3 , der bereits in Abschnitt 1.3 berechnet wurde, ist 3.20876. Für das Konfidenzband auf A ergibt sich 3.143099 als kritischer Wert für das Konfidenzband. Dieser Wert ist kleiner als der bisherige, was auch zu erwarten war.

Dies sieht man auch daran, dass das Konfidenzband auf A schmäler ist, als das Konfidenzband auf \mathbb{R}^3 .

Da bei der Berechnung des kritischen Wertes für das Konfidenzband auf A eine Simulation verwendet wird und Reproduktivität gewährleisten sein soll, wird an dieser Stelle die Seed auf 4 gesetzt.

Simulation 1.5.2. Jetzt wird die Simulationstabelle aus dem Abschnitt 1.3 durch Überdeckungswahrscheinlichkeiten für die Konfidenzbänder auf A fortgesetzt.

In diesem Fall ist $A = \{x \in \mathbb{R}^4 : x_i \in [0, 1] \text{ für alle } i = 1, 2, 3, 4\}$. Dies liegt daran, dass wir die Matrix X so wählen, dass die unabhängigen Variablen im Intervall $[0, 1]$ liegen.

Die neuen Überdeckungswahrscheinlichkeiten für Konfidenzbänder auf Intervallen werden mit Minmax bezeichnet. Auch diese Überdeckungswahrscheinlichkeiten liegen über den geforderten 0.95.

	unabhängig	AR bekannt	AR
Konfidenzband auf ganz \mathbb{R}^p	0.953	0.94	1.00
Konfidenzband auf einem Polyeder	.956	0.93	1.00

1.6 Konfidenzbänder auf einem Intervall für ein Regressionsmodell mit Polynomgestalt

In diesem Abschnitt betrachten wir wieder das Regressionsmodell

$$Y = X\beta + e$$

mit Y einem $1 \times n$ Datenvektor, β ein $1 \times (p+1)$ Koeffizientenvektor und $e \sim \mathcal{N}_n(0, \sigma^2 I_n)$ ein $1 \times n$ Zufallsfehlervektor.

In den ersten Abschnitten lag die Annahme zugrunde, dass X irgendeine feste, aber beliebige $n \times (p+1)$ Matrix ist. Im letzten Abschnitt sind wir zusätzlich implizit davon ausgegangen, dass die erste Spalte von X nur mit Einsen besetzt ist.

In diesem Abschnitt betrachten wir den Fall, dass die Spalten von X einen funktionalen Zusammenhang erfüllen. Konkret gehen wir davon aus, dass die i -te Zeile von X von der Form $\tilde{x}_i = (1, x, x^2, \dots, x^p) \in \mathbb{R}^{p+1}$, $i = 1, \dots, n$ ist.

Unsere Designmatrix ist also zeilenweise von Polynomgestalt. Dies ist genau der Fall, den wir im Beispiel und den Simulationen, die am Ende der Abschnitte stehen, betrachtet haben.

Das Ziel ist wieder Konfidenzbänder der Form

$$\tilde{x}'\beta \in \tilde{x}'\hat{\beta} \pm c\hat{\sigma}\sqrt{\tilde{x}'(X'X)^{-1}\tilde{x}} \quad \text{für alle } x \in A = (a, b) \quad (9)$$

zu bestimmen. Dann können wir sie direkt mit den Konfidenzbändern aus den Abschnitten 1.3 und 1.5 vergleichen. Dazu reicht es nur die kritischen Konstanten c miteinander zu vergleichen.

Dabei ist die kritische Konstante c wieder so zu bestimmen, dass

$$\mathbb{P}\left(\tilde{x}'\beta \in \tilde{x}'\hat{\beta} \pm c\hat{\sigma}\sqrt{\tilde{x}'(X'X)^{-1}\tilde{x}} \quad \text{für alle } x \in A = (a, b)\right) = 1 - \alpha$$

gilt.

Es können allerdings nicht direkt die Ergebnisse aus dem letzten Abschnitt benutzt werden. Der Grund dafür sind die folgenden zwei Eigenschaften, die [Liu64, S. 180] entnommen sind.

1. Angenommen es seien zwei polynomiale Modelle $\tilde{x}'\beta_1$ und $\tilde{x}'\beta_2$ gegeben und das erste ist näher an dem wahren Modell als das zweite. Näher meint in diesem Fall, dass

$$\sup_{x \in \mathbb{R}} \frac{|\tilde{x}'(\hat{\beta}_1 - \hat{\beta})|}{\hat{\sigma}\sqrt{\tilde{x}'(X'X)^{-1}\tilde{x}}} < \sup_{x \in \mathbb{R}} \frac{|\tilde{x}'(\hat{\beta}_2 - \hat{\beta})|}{\hat{\sigma}\sqrt{\tilde{x}'(X'X)^{-1}\tilde{x}}}$$

Zu erwarten ist, dass das erste Modell sinnvoll ist, wenn auch das zweite sinnvoll ist. Immerhin ist das erste Modell näher an dem wahren Modell, als das zweite.

Dies ist allerdings für die bisher betrachteten Konfidenzbänder nicht der Fall, da die bisher betrachteten Verfahren die spezielle polynomiale Struktur nicht berücksichtigen.

2. Die kritische Konstante c kann kleiner als im letzten Abschnitt gewählt werden. Der Grund dafür ist, dass das Konfidenzband nur auf einer Teilmenge

$$\{\tilde{x} : \tilde{x} = (1, x, \dots, x^p) \text{ für ein } x \in \mathbb{R}\} \subset \mathbb{R}^{p+1}$$

die Wahrscheinlichkeit $1 - \alpha$ haben muss.

Als Lösung wird in diesem Abschnitt eine Methode, die auf Simulation basiert, vorgestellt. Das Vorgehen ist also ähnlich zu dem Vorgehen in Abschnitt 1.5. Diese Methode orientiert sich an [Liu64, S. 183,184].

Im Satz 1.3.2 haben wir das Konfidenzniveau vom Modell $Y = X\beta + e$, mit Y, X, β, e wie üblich, als $\mathbb{P}(S \leq c)$ bestimmt. Dabei ist

$$\begin{aligned} S &= \sup_{x_0 \in A} \frac{|\tilde{x}'(\hat{\beta} - \beta)/\hat{\sigma}|}{\sqrt{\tilde{x}'(X'X)^{-1}\tilde{x}}} \\ &= \sup_{x_0 \in A} \frac{|\tilde{x}'N/(\hat{\sigma}/\sigma)|}{\sqrt{\tilde{x}'(X'X)^{-1}\tilde{x}}} \\ &= \sup_{x_0 \in A} \frac{|\tilde{x}'T|}{\sqrt{\tilde{x}'(X'X)^{-1}\tilde{x}}} \\ &= K_{2h}(T, (X'X)^{-1}, (a, b)) \end{aligned}$$

Man beachte, dass in diesem Fall $A = [a, b]$ mit $a, b \in \mathbb{R}$ gilt. Der Grund dafür ist, dass $\tilde{x} = (1, x, x^2, \dots, x^p)$ ist.

In den obigen Umformungen wurden die Definitionen $N := (\hat{\beta} - \beta)/\sigma \sim \mathcal{N}_{p+1}(0, (X'X)^{-1})$, $T := N/(\hat{\sigma}/\sigma) \sim \tau_{p+1,v}(0, (X'X)^{-1})$ und für $K_{2h}(T; (X'X)^{-1}; A)$

$$K_{2h}(T; (X'X)^{-1}; (a, b)) := \sup_{x_0 \in A} \frac{|\tilde{x}'T|}{\sqrt{\tilde{x}'(X'X)^{-1}\tilde{x}}}$$

benutzt. Es ist also klar, dass c das $1 - \alpha$ Quantil von $S = K_{2h}(T; (X'X)^{-1}; (a, b))$ ist. Allerdings ist die Verteilung von $K_{2h}(T; (X'X)^{-1}; (a, b))$ nur schwer explizit zu bestimmen. Um c trotzdem zu ermitteln, kann man r unabhängige Realisationen von S berechnen und von diesen S_1, \dots, S_r das $1 - \alpha$ Quantil benutzen.

Um S für vorgegebene N und $\hat{\sigma}/\sigma$ zu bestimmen, muss man $K_{2h}(T; (X'X)^{-1}; (a, b))$ mit $T = N/(\hat{\sigma}/\sigma)$ berechnen.

Um $K_{2h}(T; (X'X)^{-1}; (a, b))$ zu berechnen, kann man folgendes Verfahren aus [Liu64, Appendix E] benutzen.

Betrachten wir die Definition von $K_{2h}(T; (X'X)^{-1}; (a, b))$, ist klar, dass wir ein Supremum suchen. Gesucht ist das Supremum der Funktion

$$g_h(x) = \frac{\tilde{x}'t}{\sqrt{\tilde{x}'(X'X)^{-1}\tilde{x}}}$$

Dazu gibt es zwei Möglichkeiten. Zum einen kann man das Supremum mittels einer Gittersuche näherungsweise bestimmen. Das heißt man gibt sich ein Gitter vor und berechnet dann auf allen Gitterpunkten den Funktionswert. Als Supremum der Funktion wählt man dann den größten Wert aus.

Diese Möglichkeit hat den Nachteil, dass eventuell nicht das Maximum, sondern ein zu kleiner Wert benutzt wird. Dafür hat es den Vorteil, dass diese Methode leicht zu implementieren ist und eine gute Laufzeit aufweist.

Zum anderen gibt es die Möglichkeit das Supremum durch ableiten der Funktion zu bestimmen. Von der Ableitung muss man dann die Nullstellen bestimmen.

Das Bestimmen der Nullstellen einer Funktion hat den Nachteil, dass es nicht einfach zu implementieren ist. Außerdem kann die Laufzeit, gerade bei rechenintensiven Simulationen zur Bestimmung der Überdeckungswahrscheinlichkeit, sehr lang werden. Dafür müsste allerdings immer ein bessere Wert für das Maximum der Funktion berechnet werden.

In [Liu64] wird der zweite Weg vorgeschlagen, der im Folgenden beschrieben wird. Bei den Berechnungen in den späteren Kapiteln dieser Bachelorarbeit wird allerdings ausschließlich der zweite Weg gewählt.

Zuerst beachte, dass das Supremum entweder an a oder an b oder den stationären Punkten der Funktion g_h angenommen werden kann, da g_h eine glatte Funktion auf A ist. Leitet man g_h nach x ab, erhält man

$$\frac{dg_h}{dx}(x) = \left(\left(\frac{d\tilde{x}'}{dx} \right) t \left(\tilde{x}'(X'X)^{-1}\tilde{x} \right) - (\tilde{x}'t) \left(\frac{d\tilde{x}'}{dx} \right) (X'X)^{-1}\tilde{x} \right) \left(\tilde{x}'(X'X)^{-1}\tilde{x} \right)^{-3/2}$$

Also sind die stationären Punkte von $g_h(x)$ durch die reellen Wurzeln des $(3p-2)$ -gradigen Polynoms

$$h(x) = \left(\frac{d\tilde{x}'}{dx} \right) t \left(\tilde{x}'(X'X)^{-1}\tilde{x} \right) - (\tilde{x}'t) \left(\frac{d\tilde{x}'}{dx} \right) (X'X)^{-1}\tilde{x}$$

gegeben.

Um diese Wurzeln zu bestimmen, wird auf das Intervall $[0, 1]$ ein Gitter aus 100 Punkten gelegt. Dann wird der Wert von $g.prime$ auf den Punkten dieses Gitters bestimmt und gespeichert.

Dabei entspricht $g.prime$ der Funktion h , die für die Zwecke der Berechnung die Rolle von $\frac{dg_h}{dx}(x)$ übernimmt.

Anschließend werden aufeinanderfolgende Funktionswerte von h miteinander multipliziert. Tritt ein Vorzeichenwechsel auf, so ist in dem Intervall zwischen den Funktionswerten mindestens eine Nullstelle. Diese Nullstelle werden dann mit der R-internen Funktion *uniroot* bestimmt.

Dieses Vorgehen orientiert sich an dem Vorgehen des R-Paketes *rootsolve* und ist notwendig, da die Funktion *uniroot* nur eine Wurzel und nicht alle Wurzeln in dem gegebenen Intervall findet.

Zusammenfassend geht man bei der Bestimmung von c also wie folgt vor:

1. Simuliere $N \sim \mathcal{N}_{p+1}(0, (X'X)^{-1})$ und $\hat{\sigma}/\sigma \sim \sqrt{\chi_v^2/v}$ mit $v = n - p - 1$ und p der Anzahl an unabhängigen Parametern.
2. Berechne $T = N/(\hat{\sigma}/\sigma)$.
3. Bestimme $K_{2h}(T; (X'X)^{-1}; (a, b))$. Dazu bestimmt man die Maxima von g_h . Um dies zu erreichen bestimmt man die kritischen Stellen der Ableitung von g_h . Dazu reicht es die Nullstellen von h zu bestimmen. Hierzu
 - (a) Berechne die Funktionswerte von h auf dem Gitter von a bis b mit Feinheit 100.
 - (b) Multipliziere je zwei aufeinander folgende Funktionswerte.
 - (c) Liegt ein Vorzeichenwechsel vor befindet sich in dem Intervall eine Nullstelle.
 - (d) Benutze die R-Funktion *uniroot* um die Nullstelle in diesem Intervall zu bestimmen
4. Hat man die Nullstellen r_1, \dots, r_n von h gefunden, berechnet man die Funktionswerte der Funktion g_h an diesen Stellen.
5. Damit ist $S = \max\{g(a), g(b), g(r_1), \dots, g(r_n)\}$

Als nächstes wiederholt man diese Schritte q mal, um $K = \{S_1, \dots, S_q\}$ zu erhalten. Dann ist c das $1-\alpha$ -Quantil von K .

Beispiel 1.6.1. Jetzt wird das Beispiel aus dem Abschnitt 1 fortgeführt, indem zu der Regression mit Grad Drei, dem Konfidenzband auf ganz \mathbb{R}^2 und dem auf $[0,1]$ noch ein Konfidenzband auf $[0,1]$ für Polynome in die Abbildung 5 eingezeichnet wird.

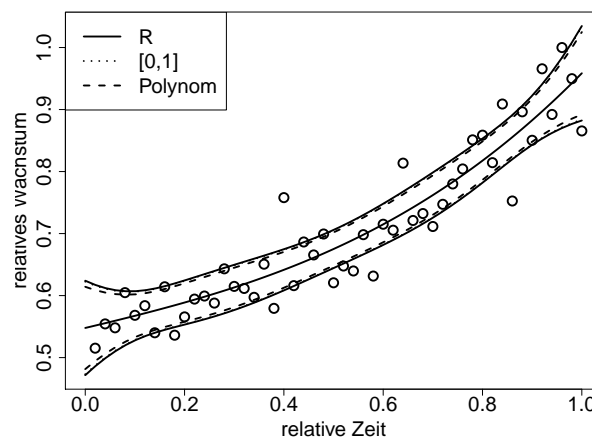


Abbildung 5: Daten aus Kapitel 1 mit den drei Konfidenzbandarten

Die kritischen Werte für das Konfidenzband auf ganz \mathbb{R}^3 und für das Konfidenzband auf $[0,1]$ wurden in den vorherigen Abschnitten bereits berechnet. Es ergaben sich die Werte 3.20876 und 3.143099. Berechnet man den kritischen Wert für das Konfidenzband auf $[0,1]$, welches die Polynomgestalt berücksichtigt, erhält man den Wert 2.79563. Dieser Wert ist noch kleiner als der Wert von 3.143099, was auch zu erwarten war.

Als Seed wurde weiterhin 4 benutzt.

Simulation 1.6.2. Die Simulationstabelle aus den letzten beiden Abschnitten wird durch Überdeckungswahrscheinlichkeiten für die Konfidenzbänder auf einem Intervall A beendet. In diesem Fall ist wieder $A = \{(x_1, \dots, x_p)' : 0 \leq x_i \leq 1, i = 1, \dots, p\}$ aufgrund der Struktur des Modells. Außerdem wird davon ausgegangen, dass die Designmatrix Polynomgestalt hat.

	unabhängig	AR bekannt	AR
Konfidenzband auf ganz \mathbb{R}^p	0.953	0.94	1.00
Konfidenzband auf einem Polyeder	.956	0.93	1.00
Konfidenzband auf einem Polyeder für Polynome	0.951	0.90	1.00

2 Vergleich von zwei Regressionsmodellen

Da das Ziel dieser Arbeit der Vergleich von Regressionsmodellen ist, wird in diesem Kapitel gezeigt, wie man Regressionsmodelle vergleicht.

Es seien also, ähnlich wie in [Liu64, S. 113] und [Liu64, S. 178], zwei Regressionsmodelle

$$Y_i = X_i\beta_i + e_i \text{ mit } i = 1, 2 \quad (10)$$

gegeben.

Dabei seien $Y_i = (Y_{i,1}, \dots, Y_{i,n_i})$ für $i = 1, 2$ zwei Vektoren mit zufälligen Beobachtungen, die die abhängigen Daten darstellt.

Weiterhin seien X_i für $i = 1, 2$ zwei $n_i \times (p + 1)$ Designmatrix mit vollem Zeilenrang und festem Design. Dabei sei die j -te Zeile von X_1 mit der j -ten Zeile von X_2 für $j = 1, \dots, \min(n_1, n_2)$ identisch. Diese Bedingung bedeutet, dass beide Designmatrizen für $j = 1, \dots, \min(n_1, n_2)$ dasselbe, feste Design haben.

Außerdem seien β_i zwei $(p + 1) \times 1$ Koeffizientenvektoren, die den funktionalen Zusammenhang zwischen Y_i und X_i darstellen.

In diesem Kapitel sei $e_i \sim \mathcal{N}_{n_i}(0, \sigma^2 I_{n_i})$. Diese Annahme ist wichtig, da sich die Modelle somit nur in β_i unterscheiden.

Die zugrunde liegenden Modelle sind also gleich, wenn $\beta_1 = \beta_2$

Da die Werte von β_i unbekannt sind und nur geschätzt werden können, testet man

$$H_0 : \beta_1 = \beta_2 \text{ vs. } H_1 : \beta_1 \neq \beta_2 \quad (11)$$

Dabei bezeichnet H_0 die Nullhypothese und H_1 die Alternativhypothese.

Dazu ist es entweder möglich einen Test zu konstruieren oder Konfidenzbänder zu benutzen. In Abschnitt 2.1 wird ein Test konstruiert und im Abschnitt 2.2 werden Konfidenzbänder benutzt.

Es ist mit der in diesem Kapitel eingeführten Methode auch möglich zwei Modelle mit verschiedenen abhängigen Daten zu vergleichen.

Ein Beispiel für solch ein Vorgehen ist der Vergleich des Blutdrucks von Männern und Frauen. Konkret kann man mit dieser Methode beispielsweise die Frage beantworten, ob sowohl für Männer als auch für Frauen der Zusammenhang zwischen Alter und Blutdruck gleich ist.

Beispiel 2.0.3. Es wird das Beispiel aus Kapitel 1 fortgeführt, indem eine zweite Designmatrix X_2 einführt wird. Die Designmatrix X_2 stellt ein polynomiales Modell vom Grad Vier dar.

Das heißt, im Beispiel zu diesem Kapitel wird es darum gehen, ob es für die Daten, die bereits in Kapitel 1 vorgestellt wurden, einen Unterschied macht, ein polynomiales Modell vom Grad Drei oder vom Grad Vier zugrunde zu legen.

In der Abbildung 6 sind die beiden Modelle mit den Daten eingezeichnet.

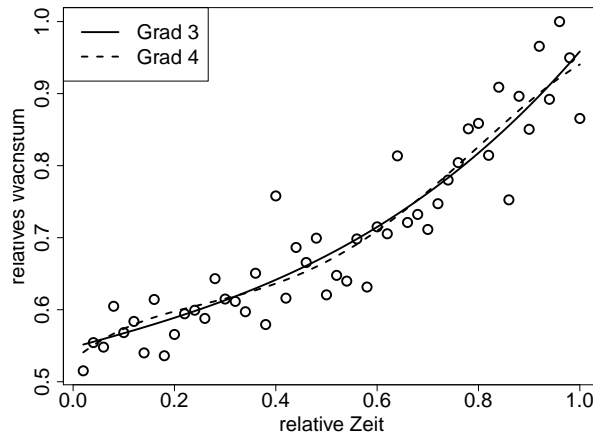


Abbildung 6: polynomielle Regression von Grad Drei und Grad Vier im Vergleich

2.1 F-Test zum Vergleich von zwei Regressionsmodellen

Dieser Abschnitt beruht auf [Liu64, S. 9-15], [Liu64, S. 114-115] und [DS98].

Zuerst betrachten wir den Test, der prüft, ob der Regressionskoeffizient β eine lineare Einschränkung $Z\beta = d$ erfüllt. Dabei ist Z eine gegebene $r \times (p+1)$ Matrix mit vollem Zeilenrang r mit $1 \leq r \leq p+1$ und d ist ein vorgegebener Vektor in \mathbb{R}^r . Wir wollen also testen, ob

$$H_0 : Z\beta = d \text{ vs. } H_a : Z\beta \neq d \quad (12)$$

Von Interesse ist also, ob der Parameter in einem gegebenen Regressionsmodell eine lineare Einschränkung erfüllt.

Dazu wird zuerst der OLS-Schätzer $\hat{\beta}_Z$ unter der Bedingung, dass $Z\beta = d$ gilt, benötigt. Das heißt, $\hat{\beta}_Z$ minimiert

$$L(\beta) = (Y - X\beta)'(Y - X\beta) = \|Y - X\beta\|^2$$

über alle $\beta \in \mathbb{R}^{p+1}$, die die Bedingung $Z\beta = d$ erfüllen. Man kann $\hat{\beta}_Z$ mittels Lagrange-Multiplikatoren finden.

Betrachtet wird dazu der nachfolgende Satz.

Satz 2.1.1. Unter den oben genannten Bedingungen ist der OLS-Schätzer $\hat{\beta}_Z$ gegeben durch

$$\hat{\beta}_Z = (X'X)^{-1}(X'Y + Z'f) = \hat{\beta} + (X'X)^{-1}Z'f$$

mit $f = (Z(X'X)^{-1}Z')^{-1}(d - Z\hat{\beta})$

Beweis. Siehe [Liu64, S. 13] □

Im Folgenden werden zwei Sätze aus [Liu64] angegeben, die wir zur Konstruktion des Test im weiteren Verlauf dieses Abschnittes benötigen.

Satz 2.1.2. Unter den oben genannten Bedingungen gilt

1. $\|X\hat{\beta} - X\hat{\beta}_Z\|^2 \sim \sigma^2 \chi_r^2(\delta)$ mit nicht zentralem Parameter
 $\delta = \|X\beta - X\mathbb{E}(\hat{\beta}_Z)\|^2 / \sigma^2 = (\beta - \mathbb{E}(\hat{\beta}_Z))' X' X (\beta - \mathbb{E}(\hat{\beta}_Z)) / \sigma^2$
2. $\|Y - X\hat{\beta}_Z\|^2 \sim \sigma^2 \chi_{n-(p+1)+r}^2(\delta)$ mit δ wie in 1.
3. $\|X\hat{\beta} - X\hat{\beta}_Z\|^2$ und $\|Y - X\hat{\beta}_Z\|^2$ sind unabhängig.
4. Außerdem ist

$$\frac{\|X\hat{\beta}_Z - X\hat{\beta}\|^2 / r}{\|Y - X\hat{\beta}\|^2 / (n - p - 1)} \sim f_{p+1, n-p-1}$$

Beweis. [Liu64, S. 11] □

Damit kann ein Test zum Niveau $1 - \alpha$ konstruiert werden:

$$\text{Es ist } H_0 \text{ genau dann abzulehnen, wenn } \frac{\|X\hat{\beta}_Z - X\hat{\beta}\|^2 / r}{\|Y - X\hat{\beta}\|^2 / (n - p - 1)} > f_{p+1, n-p-1}^\alpha \quad (13)$$

Dabei ist $f_{p+1, n-p-1}^\alpha$ das α -Quantil der F-Verteilung mit den Freiheitsgraden $p + 1$ und $n - p - 1$.

Gleich wird noch folgender Satz aus [Liu64, S. 13] benötigt:

Satz 2.1.3. Es gilt

$$\|X\hat{\beta} - X\hat{\beta}_Z\|^2 = (Z\hat{\beta} - d)' (Z(X'X)^{-1}Z')^{-1} (Z\hat{\beta} - d)$$

Beweis. [Liu64, S. 13] □

Jetzt wird dieses Resultat auf den Vergleich von Regressionsmodellen angewendet. Dabei wird sich an [Liu64, S. 114] orientiert.

Betrachten wir wieder Modell (10) vom Anfang dieses Kapitels. Außerdem wird wieder das Testproblem (11) betrachtet:

$$H_0 : \beta_1 = \beta_2 \text{ vs. } H_1 : \beta_1 \neq \beta_2$$

Um einen solchen Test durchzuführen, benötigt man eine Dummyvariable z :

$$z = \begin{cases} 1 & \text{falls Y aus Modell 1 ist.} \\ 0 & \text{falls Y aus Modell 2 ist.} \end{cases}$$

Mit z können die beiden Modelle zu einem vereint werden, indem

$$Y = x'c_1 + zx'c_2 + e \quad (14)$$

mit $x = (1, x_1, \dots, x_p)'$, $c_1 = \beta_2$ und $c_2 = \beta_1 - \beta_2$ gesetzt wird. Dass dieses Modell mit den beiden Modellen aus (10) übereinstimmt, sieht man daran, dass

$$Y = x'(c_1 + c_2) + e = x'\beta_1 + e$$

wenn Y aus dem ersten Modell und

$$Y = x'c_1 + e = x'\beta_2 + e$$

wenn Y aus dem zweiten Modell stammt. Damit kann man den Hypothesentest (11) umformulieren zu

$$H_0 : c_2 = \beta_1 - \beta_2 = 0 \text{ vs. } H_1 : c_2 \neq 0$$

Unter H_0 reduziert sich also das Gesamtmodell (14) zu

$$Y = x'c_1 + e$$

Jetzt kann (13) benutzt werden und man erhält als Teststatistik :

$$\text{Es ist } H_0 \text{ genau dann zu verwerfen, wenn } \frac{(\hat{\beta}_1 - \hat{\beta}_2)' D (\hat{\beta}_1 - \hat{\beta}_2) / (p+1)}{\widehat{\sigma^2}} > f_{p+1, n-p-1}^\alpha \quad (15)$$

Dabei ist $\widehat{\sigma^2}$ die mittlere Varianz des Modells (14), $\hat{\beta}_i = (X_i' X_i)^{-1} X_i' Y$ und außerdem bezeichnet $D = (X_1' X_1)^{-1} + (X_2' X_2)^{-1}$.

Um zu sehen, dass (15) tatsächlich aus (13) hergeleitet werden kann, ist folgende Rechnung nach [Liu64, S. 115] zu betrachten. Dabei erfolgt in der Betrachtung allerdings ausschließlich die Berechnung für den Nenner.

Den Nenner der linken Seite von (15) kann man schreiben als

$$\begin{aligned} \widehat{\sigma^2} &= \frac{\|Y - X\hat{c}\|}{n_1 + n_2 - 2(p+1)} \\ &= \frac{\|Y_1 - X_1\hat{\beta}_1\|^2 + \|Y_2 - X_2\hat{\beta}_2\|^2}{n_1 + n_2 - 2(p+1)} \\ &= \frac{n_1 - p - 1}{n_1 + n_2 - 2(p+1)} \frac{\|Y_1 - X_1\hat{\beta}_1\|^2}{n_1 - p - 1} + \frac{n_2 - p - 1}{n_1 + n_2 - 2(p+1)} \frac{\|Y_2 - X_2\hat{\beta}_2\|^2}{n_2 - p - 1} \\ &= \frac{n_1 - p - 1}{n_1 + n_2 - 2(p+1)} \widehat{\sigma_1^2} + \frac{n_2 - p - 1}{n_1 + n_2 - 2(p+1)} \widehat{\sigma_2^2} \end{aligned}$$

Es wird nun das Beispiel aus der Einleitung zu diesem Kapitel fortgeführt.

Beispiel 2.1.4. Berechnet man die Teststatistik (15) erhält man als Teststatistik den Wert 14.0476 und als kritischen Wert 2.574035. Also verwirft man die Hypothese. Die Modelle sind also verschieden.

Allerdings macht solch ein Test keinerlei Aussage über die Größe des Unterschiedes zwischen den Modellen. Deshalb werden im nächsten Abschnitt die Ergebnisse aus Kapitel 1 benutzt.

2.2 Vergleich von Regressionsmodellen mit Konfidenzbändern

Dieser Abschnitt basiert auf [Liu64, S. 119-121]. Die grundlegende Idee ist, die Modelle voneinander abzuziehen und dann zu sehen, ob die Nullfunktion in einem Konfidenzband um die Differenz der beiden Modelle liegt.

Es wird dies nach [Liu64, S. 122] formalisiert:

Ein zweiseitiges, hyperbolisches, gleichmäßiges Konfidenzband für $x'\beta_2 - x'\beta_1$ über der Region A hat die Form

$$x'\beta_2 - x'\beta_1 \in x'\hat{\beta}_1 - x'\hat{\beta}_2 \pm c \hat{\sigma} \sqrt{x'Dx} \quad \text{für alle } x \in A = (a, b) \quad (16)$$

wobei c eine kritische Konstante ist, sodass das Konfidenzniveau des Konfidenzbandes $1 - \alpha$ beträgt. Dabei ist $D = (X_1'X_1)^{-1} + (X_2'X_2)^{-1}$

Sei P die eindeutig bestimmte Wurzel aus $D = (X_1'X_1)^{-1} + (X_2'X_2)^{-1}$ und definiere weiterhin $T = P^{-1}(\hat{\beta}_2 - \beta_2 - \hat{\beta}_1 + \beta_1)/\hat{\sigma}$, welches wieder die $\tau_{p+1,v}$ Verteilung besitzt. Es folgt wie in Lemma 1.3.2, dass das simultane Konfidenzband (16) gegeben ist durch $\mathbb{P}(S < c)$ mit

$$\begin{aligned} S &= \sup_{x \in A} \frac{x'(\hat{\beta}_2 - \beta_2 - \hat{\beta}_1 + \beta_1)}{\hat{\sigma} \sqrt{x'Dx}} \\ &= \sup_{x \in A} \frac{(Px)'(P^{-1}(\hat{\beta}_2 - \beta_2 - \hat{\beta}_1 + \beta_1)/\hat{\sigma})}{\sqrt{(Px)'(Px)}} \\ &= \sup_{x \in A} \frac{|(Px)'T|}{\|Px\|} \end{aligned}$$

Also kann die kritische Konstante c genauso wie in Kapitel 1.5 beziehungsweise wie in Kapitel 1.6 gefunden werden.

Jetzt zeigen wir noch ein Resultat, das zeigt, dass es keinen Unterschied macht, ob ein Test durchgeführt oder ein Konfidenzband auf ganz \mathbb{R}^p benutzt wird.

Betrachten wir den Test

$$H_0 : \beta = \beta_0 \text{ vs. } H_1 : \beta \neq \beta_0$$

Das heißt, es wird getestet, ob das Regressionsmodell $x'\beta$ dem wahren Modell $x'\beta_0$ entspricht. Dieser Test entspricht einem F-Test mit $Z = I_{p+1}$ und $\beta_0 = d$. Die Teststatistik ist nach [Liu64, S. 17]

Es ist H_0 genau dann zu verwerfen, wenn
$$\frac{(\beta_0 - \hat{\beta})'(X'X)^{-1}(\beta_0 - \hat{\beta})}{(p+1)\|Y - X\hat{\beta}\|^2/(n-p-1)} > f_{p+1, n-p-1}^\alpha \quad (17)$$

Satz und Beweis orientieren sich an [Liu64, S. 67].

Satz 2.2.1. Der Test (17) und Konfidenzbandmethode 1.3.1 akzeptieren und widerlegen H_0 immer dann, wenn auch die andere Methode akzeptiert beziehungsweise widerlegt.

Beweis. Benutzt man die Definition von N wie bisher, außer dass man β mit β_0 ersetzt, erhält man

$$\begin{aligned} \sqrt{(p+1)f_{p+1, n-p-1}^\alpha} &< \sup_{x_{(0)} \in \mathbb{R}^p} \frac{|x'(\hat{\beta} - \beta_0)|}{\hat{\sigma} \sqrt{x'(X'X)^{-1}x}} \\ &= \sup_{x_{(0)} \in \mathbb{R}^p} \frac{|(Px)'N|}{(\hat{\sigma}/\sigma) \sqrt{(Px)'(Px)}} \\ &= \frac{\|N\|}{(\hat{\sigma}/\sigma)} \left(\sup_{x_{(0)} \in \mathbb{R}^p} \frac{|(Px)'N|}{\|Px\| \|N\|} \right) \\ &= \frac{\|N\|}{(\hat{\sigma}/\sigma)} \\ &= \sqrt{\frac{(\hat{\beta} - \beta_0)' P^{-1} P^{-1} (\hat{\beta} - \beta_0)}{\widehat{\sigma^2}}} \\ &= \sqrt{\frac{(\hat{\beta} - \beta_0)' (X'X)^{-1} (\hat{\beta} - \beta_0)}{\widehat{\sigma^2}}} \end{aligned}$$

Dies entspricht der Teststatistik (17), also sind die beiden Tests äquivalent. \square

Man beachte, dass in diesem Satz das Konfidenzband auf ganz \mathbb{R}^p benutzt wurde.

Das bedeutet, dass man mehr Informationen über ein gegebenes Entscheidungsproblem erhält, wenn man ein Konfidenzband berechnet, als wenn man einen Test durchführt. Dies liegt daran, dass das Konfidenzband auch eine Aussage über die Größe des Unterschiedes zwischen den Modellen macht.

Weiterhin wurde bereits im ersten Kapitel gezeigt, dass Konfidenzbänder auf einem Intervall und erst recht Konfidenzbänder für Polynome auf einem Intervall schmäler als Konfidenzbänder auf ganz \mathbb{R}^p sind. Also ist es auch in dieser Hinsicht von Vorteil, Konfidenzbänder auf Intervallen zu verwenden.

Beispiel 2.2.2. Es wird jetzt das Beispiel aus dem Abschnitt 2.1 fortgeführt, indem die Differenz der Regressionsmodelle geplottet wird und ein Konfidenzbänder auf A für Polynome eingefügt wird. Siehe dazu Abbildung 7.

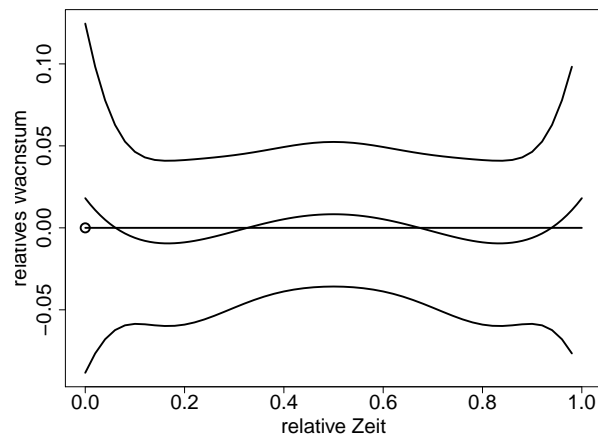


Abbildung 7: Differenz des Beispiels mit Konfidenzband auf A für Polynome

Man sieht, dass die konstante Nullfunktion ganz im Konfidenzband enthalten ist. Also wird in diesem Fall die Nullhypothese nicht verworfen. Dies kann daran liegen, dass in diesem Test nur auf dem Intervall $[0,1]$ getestet wird. Außerdem respektiert dieser Test die besondere Polynomstruktur des zugrunde liegenden Modells.

3 Einen Teil eines Regressionsmodells überprüfen

In diesem Abschnitt geht es um den Spezialfall, zu prüfen, ob ein Teil des Regressionsmodells signifikant von Null verschieden ist oder nicht. Das heißt, wir gehen von dem Modell

$$Y = X\beta + e \quad (18)$$

aus und wollen prüfen, ob einige der Koeffizienten β_i gleich Null sind oder nicht. Ist β_i gleich Null haben die zugehörigen unabhängigen Variablen x_i keinen Einfluss auf die abhängige Variable Y .

Zu diesem Zweck unterteilen wir den Vektor $\beta = (\beta_1, \beta_2)$ mit $\beta_1 = (\beta_0, \dots, \beta_{p-k})$ und $\beta_2 = (\beta_{p-k+1}, \dots, \beta_p)$ für $1 \leq k \leq p$ eine gegebene natürliche Zahl.

Auf die gleiche Art und Weise kann man die Spalten x von X in $x_1 = (1, x_1, \dots, x_{p-k})$ und $x_2 = (1, x_{p-k+1}, \dots, x_p)$ unterteilen. Dabei wird das selbe k für β und für X verwendet.

Ist $\beta_2 = 0$ so haben die unabhängigen Variablen x_{p-k+1}, \dots, x_p keinen Einfluss auf Y und man kann Modell (18) zu

$$Y = X_1\beta_1 + e \quad (19)$$

vereinfachen. Dabei ist X_1 die Matrix, die von den ersten $p - k + 1$ Spalten von X geformt wird.

3.1 F-Test zum Überprüfen eines Teils eines Regressionsmodells

In diesem Abschnitt wird der F-Test als Methode, um zu überprüfen, ob $\beta_2 = 0$ ist, vorgestellt. Dieser Abschnitt orientiert sich an [Liu64, S. 100-102].

Ein Ansatz, um zu überprüfen, ob $\beta_2 = 0$ ist, ist der Test

$$H_0 : \beta_2 = 0 \text{ vs. } H_1 : \beta_2 \neq 0 \quad (20)$$

der auch in [DS98] vorgeschlagen wird. Analog zu Abschnitt 2.1 erhält man die Teststatistik

$$\frac{(\tilde{I}\hat{\beta})'(\tilde{I}(X'X)^{-1}\tilde{I}')^{-1}(\tilde{I}\hat{\beta})/k}{\text{mean square residual von Modell (18)}} > f_{k,v}^\alpha$$

mit $v = n - p - 1$, \tilde{I} der $k \times (p + 1)$ Matrix die durch $\tilde{I} = (0, I_k)$ gegeben ist und $f_{k,v}^\alpha$ dem α Quantil der f Verteilung mit k und v Freiheitsgraden.

Dabei ist das mean squared residual von Modell (18) gegeben durch

$$MS = \|Y - X\hat{\beta}\|^2 / (n - p - 1) = \hat{\sigma}$$

Lehnt man bei diesem Test die Nullhypothese ab, heißt dies, dass es nicht genug statistisch gesicherte Hinweise gibt, um anzunehmen, dass $\beta_2 = 0$ ist. Das heißt allerdings nicht, dass wir davon ausgehen können, dass $\beta_2 \neq 0$ ist.

Benutzt man stattdessen ein Konfidenzband hat man ein Maß für den Unterschied der Modelle und kann eher entscheiden, ob $\beta_2 \neq 0$. Der nächste Satz macht genau solch eine Aussage.

Satz 3.1.1. Es gilt

$$\mathbb{P} \left(x'_2 \beta_2 \in x'_2 \hat{\beta}_2 \pm \sqrt{k f_{k,n-p-1}^\alpha \hat{\sigma} \sqrt{x'_2 V x_2}} \right) = 1 - \alpha \quad (21)$$

Dabei ist $V = \tilde{I}(X'X)^{-1}\tilde{I}'$, $\hat{\beta}_2 = \tilde{I}\hat{\beta}$ mit \tilde{I} wie oben.

Dieser Satz kann ähnlich wie Satz 1.3.1 bewiesen werden. Außerdem gilt folgender Satz.

Satz 3.1.2. Test (20) und das Konfidenzband (21) verwerfen und akzeptieren H_0 zur selben Zeit.

Das Konfidenzband (21) gibt einem also immer mehr Informationen. Allerdings ist das Konfidenzband auf ganz R^k definiert. Eine noch besser Aussage über H_0 kann man also treffen, wenn man stattdessen ein Konfidenzband auf A betrachtet.

3.2 Einen Teil eines Regressionsmodells auf einem Intervall überprüfen

Dieser Abschnitt orientiert sich an [Liu64, S. 102-105].

Wir beschränken uns auf die rechteckige Region

$$A_2 = \{x_2 \in \mathbb{R}^k = (x_{2,p-k+1}, \dots, x_{2,p}) : x_{2,i} \in [a_i, b_i], i = p - k + 1, \dots, p\}$$

Dabei sind $-\infty \leq a_i \leq b_i \leq \infty, i = p-k+1, \dots, p$ gegebene Konstanten. Ein hyperbolisches Konfidenzband auf A_2 ist gegeben durch

$$x'_2 \beta_2 \in x'_2 \hat{\beta}_2 \pm c \hat{\sigma} \sqrt{x'_2 V x_2} \text{ für alle } x_2 \in A_2 \quad (22)$$

dabei muss man die kritische Konstante c so bestimmen, dass das gleichmäßige Konfidenzniveau $1 - \alpha$ ist.

Sei W die Wurzel aus V , also sei $V = W^2$. Bezeichne $N_2 = W^{-1}(\hat{\beta}_2 - \beta_2)/\sigma \sim \mathcal{N}_k(0, I)$ und $T_2 = N_2/(\hat{\sigma}/\sigma \sim \tau_{k,v})$. Dann ist das Konfidenzlevel von (22) gegeben durch $\mathbb{P}(S < c)$ wobei

$$\begin{aligned} S &= \sup_{x_2 \in A_2} \frac{|x'_2(\hat{\beta}_2 - \beta_2)|}{\hat{\sigma} \sqrt{x'_2 V x_2}} \\ &= \sup_{x_2 \in A_2} \frac{|(W x_2)' W^{-1}(\hat{\beta}_2 - \beta_2)/\hat{\sigma}|}{\sqrt{(W x_2)'(W x_2)}} \\ &= \sup_{v \in C(W, A_2)} \frac{|v' T_2|}{\|v\|} \end{aligned}$$

Dabei gilt $C(W, A_2) = \{\lambda W x_2 = \lambda(x_{p-k+1}w_1 + \dots + x_p w_k) : \lambda > 0 \text{ and } x_2 \in A_2\}$ mit $W = (w_1, \dots, w_k)$.

Die Verteilung von S hängt nicht von β und σ ab. Allerdings hängt sie in komplizierter Weise von der Region A_2 und W durch $C(W, A_2)$ ab.

Für die beiden Spezialfälle $k = 1$ und $0 \in C(W, A_2)$ ergeben sich Vereinfachungen. Im allgemeinen Fall kann man c ähnlich wie in Abschnitt 1.5 finden. Dazu berechnet man R mal S_i auf die folgende Art:

1. Simuliere $N_2 \sim \mathcal{N}_k(0, I)$ und $\hat{\sigma}/\sigma \sim \sqrt{\chi_v^2/v}$
2. Bestimme $\|\pi^*(T_2, W, A_2)\|$ und $\|\pi^*(-T_2, W, A_2)\|$
3. Dann ist $S = \max(\|\pi^*(T_2, W, A_2)\|, \|\pi^*(-T_2, W, A_2)\|)$

Dann ist die kritische Konstante c das $1 - \alpha$ Quantil von S_1, \dots, S_R .

Die Berechnung von $\|\pi^*(T_2, W, A_2)\|$ findet man in [Liu64, Appendix B].

Diese Methode wurde nicht weiter verfolgt, da das Ziel dieser Arbeit ist, Konfidenzbänder zu vergleichen. Außerdem werden im Kapitel 5 vor allem Polynommodelle betrachtet. Wie man in diesem Fall Konfidenzbänder konstruiert ist Thema des nächsten Abschnitts.

Simulation 3.2.1. Es werden für das Konfidenzband auf ganz \mathbb{R}^p wieder eine Simulation für die Überdeckungswahrscheinlichkeit für Daten aus einer gewöhnlichen multilinear Regression und einer Regression mit AR(1)-Kovarianzmatrix durchgeführt:

	unabhängig	AR bekannt	AR
Konfidenzband auf ganz \mathbb{R}^p	0.953	0.94	1.00
Konfidenzband auf einem Polyeder	.956	0.93	1.00
Konfidenzband auf einem Polyeder für Polynome	0.951	0.90	1.00
Teil des Modells auf \mathbb{R}^p überprüfen	0.99	-	1.00

3.3 Einen Teil eines Regressionsmodells überprüfen, wenn das Modell Polynomgestalt hat

Dieser Abschnitt orientiert sich an [Liu64, S. 190-192] und [DS98].

Für den Fall, dass man Modelle mit Polynomgestalt betrachtet erhält man als F-Test

$$\text{lehne } H_0 \text{ ab} \Leftrightarrow \frac{\hat{\beta}_2' V^{-1} \hat{\beta}_2 / k}{\widehat{\sigma^2}} > f_{k,v}^\alpha \quad (23)$$

Dabei ist $\hat{\beta}_2$ der Schätzer für $\beta_2 = (b_{p-k+1}, \dots, b_p)'$ welcher die letzten k Komponenten von β ist. Dieser Schätzer hat Verteilung $\mathcal{N}(\beta_2, \sigma^2, V)$. Außerdem ist V die $k \times k$ Matrix, die von den letzten k Zeilen und den letzten k Spalten von $(X'X)^{-1}$ erzeugt wird. Da V nicht singulär ist, sei W die eindeutig bestimmte Wurzel von V , das heißt $V = W^2$.

Analog zum letzten Abschnitt kann wieder gezeigt werden, dass dieser Test dem Konfidenzband

$$x_2' \beta_2 \in x_2' \hat{\beta}_2 \pm \sqrt{k f_{k,v}^\alpha} \hat{\sigma} \sqrt{x_2' V x_2} \text{ für alle } x_2 \in \mathbb{R}^k \quad (24)$$

entspricht. Das heißt, liegt die Nullfunktion nicht vollständig im Konfidenzband (24), kann man die Nullhypothese aus Test (23) ablehnen.

Analog zu Abschnitt 1.6 kann man wieder ein Konfidenzband auf einem Intervall A konstruieren, das die Polynomstruktur berücksichtigt.

Dazu betrachtet man, dass das Konfidenzniveau von (24) für c anstatt $\sqrt{k f_{k,v}^\alpha}$ durch den Ausdruck $\mathbb{P}(S \leq c)$ gegeben ist. Dabei ist

$$\begin{aligned} S &= \sup_{x \in A} \frac{|\tilde{x}_2'(\hat{\beta}_2 - \beta_2)|}{\hat{\sigma} \sqrt{\tilde{x}_2' V \tilde{x}_2}} \\ &= \sup_{x \in A} \frac{|(1, \dots, x^{k-1})(\hat{\beta}_2 - \beta_2)|}{\hat{\sigma} \sqrt{(1, \dots, x^{k-1}) V (1, \dots, x^{k-1})'}} \end{aligned}$$

mit $\tilde{x}_2 = x^{p-k+1}(1, x, \dots, x^{k-1})$.

Also ist die kritische Konstante c genau so wie in Abschnitt 1.6 berechenbar, außer dass man $p = k - 1$ und $(X'X)^{-1}$ mit V ersetzen muss. Die Simulation ändert sich allerdings nicht.

Simulation 3.3.1. Auch für die Methode aus diesem Abschnitt wurden Simulationen durchgeführt:

	unabhängig	AR bekannt	AR
Konfidenzband auf ganz \mathbb{R}^p	0.953	0.94	1.00
Konfidenzband auf einem Polyeder	.956	0.93	1.00
Konfidenzband auf einem Polyeder für Polynome	0.951	0.90	1.00
Teil des Modells auf \mathbb{R}^p überprüfen	0.99	-	1.00
Teil des Modells auf A für Polynome überprüfen	0.97	-	1.00

4 Regression und Konfidenzbänder für abhängige Daten

Alle bisherigen Ergebnisse beruhen auf dem homoskedastischen, multiplen linearen Regressionsmodell. Das heißt, mit $Y = (y_1, \dots, y_n) \in \mathbb{R}^n$, $\beta \in \mathbb{R}^{p+1}$ und $X \in \mathbb{R}^{n \times (p+1)}$ wurde von dem funktionalem Zusammenhang

$$Y = X\beta + e \tag{25}$$

ausgegangen, wobei $e \sim \mathcal{N}_n(0, \sigma^2 I_n)$ war. Es wurden für dieses Modell Konfidenzbänder mit verschiedenen Designmatritzen X sowohl auf \mathbb{R}^p als auch auf einem Intervall A bestimmt. In diesem Kapitel wird der Fall betrachtet, dass die Fehler untereinander korreliert sind. Zuerst werden wir autoregressive Modelle und im besonderen Zeitreihen, die einem AR(1) Prozess folgen, eingeführt. Dieser erste Abschnitt orientiert sich an [Han15] und [BD91]. Danach betrachten wir die Berechnung von Schätzern für die Parameter von Zeitreihen. Dazu sollen Funktionen aus dem R Paket *nlme*, dies steht für *nonlinear mixed-effects models*, verwendet werden. Dazu wird die Quelle [PB00] verwendet.

In Abschnitt 4.3 dieses Kapitels werden Methoden angegeben, wie wir für abhängige Daten Konfidenzbänder konstruieren können. Das heißt, wir betrachten wie die beiden vorhergehenden Abschnitte die Überlegungen aus den Abschnitten 1.6 und 3.3 ändern.

4.1 Autoregressive Modelle und AR(1)

In diesem Abschnitt wird Modell (25) auf den Fall von abhängigen Daten verallgemeinert. Daten sind abhängig, wenn $\text{Cov}(y_i, y_j) \neq 0$ für $i \neq j$.

Beginnen wir mit ein paar Definitionen. Die erste orientiert sich an [BD91, S. 1].

Definition 4.1.1. Zeitreihe

Eine Zeitreihe X ist eine Menge von Messungen x_t an bestimmten Zeitpunkten t .

Bei einer diskreten Zeitreihe ist $t = 1, \dots, T$ mit $T \in \mathbb{N}$ fest.

Weiterhin brauchen wir die folgende Definition, die [Han15] entnommen ist.

Definition 4.1.2. stationäre Zeitreihe

Eine Zeitreihe $X = (X_t)$ nennt man stationär, wenn $\mathbb{E}(X_t) = \mu$ unabhängig von t ist und $\text{Cov}(X_t, X_{t-k}) = \gamma(k)$ unabhängig von t für alle k ist.

Eine besondere Zeitreihe ist ein so genannter *autoregressive-moving average* (ARMA) Prozess. Die folgende Definition orientiert sich an [BD91, S. 78].

Definition 4.1.3. ARMA(p,q) Prozess

Den Prozess (X_t) nennt man einen ARMA(p,q) Prozess, wenn (X_t) stationär ist und für jedes t gilt

$$X_t - \phi_1 X_{t-1} - \dots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \dots + \theta_q Z_{t-q} \quad (26)$$

mit $(Z_t) \sim \text{WN}(0, \sigma^2)$. Dabei sagt man, dass der Prozess $(Z_t) \sim \text{WN}(0, \sigma^2)$ ist, wenn (Z_t) Erwartungswert Null und eine Kovarianzfunktion der Form

$$\gamma(h) = \begin{cases} \sigma^2 & , \text{ falls } h = 0 \\ 0 & , \text{ falls } h \neq 0 \end{cases}$$

Die Gleichung (26) kann man kompakter als

$$\phi(B)X_t = \theta(B)Z_t, t \in 0, \pm 1, \pm 2, \dots$$

schreiben. Dabei sind ϕ und θ die Polynome

$$\phi(z) = 1 - \phi_1 z - \dots - \phi_p z^p$$

und

$$\theta(z) = 1 + \theta_1 z + \dots + \theta_q z^q$$

und B ist der Rückwärtsshift-Operator definiert durch

$$B^j X_t = X_{t-j}, \quad j = 0, \pm 1, \pm 2, \dots$$

Es gibt verschiedene Arten von ARMA Prozessen. Uns werden im folgenden nur so genannte autoregressive Prozesse erster Ordnung (AR(1)), ein Spezialfall von allgemeinen ARMA Prozessen, interessieren.

Definition 4.1.4. AR(p)

Ist in der obigen Definition $\theta(x) \equiv 1$, so ist

$$\phi(B)X_t = Z_t$$

und man nennt den Prozess einen autoregressiven Prozess von Ordnung p (oder AR(p)).

Ein Spezialfall von AR(p) Prozessen ist der AR(1) Prozess, bei dem gilt

$$y_t = \phi y_{t-1} + e_t \quad (27)$$

Dabei ist ϕ ein unbekannter Parameter und $\mathbb{E}(e_t) = 0$, $\mathbb{E}(e_t^2) = \sigma^2 < \infty$. Für den Fall, dass $\phi \in (-1, 1)$ ist, nennt man die Zeitreihe streng stationär und ergodisch.

Da die Messungen nacheinander getätigt werden, steht zu vermuten, dass y_t und y_{t+1} auf irgendeine Art und Weise nahe beieinander sind. Ist dies der Fall, sind y_t und y_{t+1} nicht unabhängig. Unabhängigkeit war allerdings eine der zentrale Annahmen im ersten Kapitel. Deshalb muss man die Schätzer für die Parameter des Modells auf eine andere Art bestimmen.

Bei AR(1)-Prozessen hängt der Wert zur Zeit t offenbar vom Wert der Zeitreihe zur Zeit $t - 1$ ab. Somit ist hier $\text{Cov}(y_i, y_j) \neq 0$ für $i \neq j$ und die Ergebnisse aus Kapitel 1 können nicht angewendet werden.

Man kann $\text{Cov}(y_i, y_j)$ durch den funktionalen Zusammenhang $\mathbb{E}(Y_i) = \phi^{i-j}\mathbb{E}(Y_j)$ für $i > j$ finden. Für $j < i$ ist $\mathbb{E}(Y_j) = \phi^{i-j}\mathbb{E}(Y_i)$.

Beispielsweise erhält man für $i = 1$ den Zusammenhang $y_1 = \phi^j y_j$ und somit

$$\begin{aligned} \text{Cov}(y_1, y_j) &= \mathbb{E}((\mathbb{E}(y_1) - y_1)(\mathbb{E}(y_j) - y_j)) \\ &= \mathbb{E}(\mathbb{E}(y_1)\mathbb{E}(y_j) - \mathbb{E}(y_1)y_j - \mathbb{E}(y_j)y_1 + y_j y_1) \\ &= \phi^j \mathbb{E}(\mathbb{E}(y_j^2) - 2\mathbb{E}(y_j)y_j + y_j^2) \\ &= \phi^j \mathbb{E}(\mathbb{E}(y_j^2) - \mathbb{E}(y_j)^2) \\ &= \phi^j \sigma^2 \end{aligned}$$

Damit hat die Korrelationsmatrix Υ von e in diesem Fall die Form

$$\Upsilon = \begin{bmatrix} 1 & \phi & \phi^2 & \dots & \phi^{n-2} & \phi^{n-1} \\ \phi & 1 & \phi & \dots & \phi^{n-3} & \phi^{n-2} \\ \phi^2 & \phi & 1 & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \phi & \phi^2 \\ \phi^{n-2} & \phi^{n-3} & \dots & \phi & 1 & \phi \\ \phi^{n-1} & \phi^{n-2} & \dots & \phi^2 & \phi & 1 \end{bmatrix} \quad (28)$$

Das heißt, das Modell ist gegeben durch

$$Y = X\beta + e \quad (29)$$

mit $e \sim \mathcal{N}_n(0, \Upsilon)$.

4.2 AR(1) und das *nlme* Paket

In diesem Abschnitt wird vorgestellt, wie die Schätzer für β , σ und ϕ konkret berechnet werden können. Dieser Abschnitt orientiert sich an [PB00, S. 203-205].

Um einen ML-Schätzer für Modell (29) zu erhalten, kann man folgenden Trick anwenden: Man multipliziert beide Seiten der Gleichung mit $\Upsilon^{-1/2}$. Dann wird aus Modell (29) wieder Modell (25), wenn auch mit anderen Werten. Konkret erhält man

$$(\Upsilon^{-1/2})Y = Y^* = X^*\beta + e^* = (\Upsilon^{-1/2})X + (\Upsilon^{-1/2})e \quad (30)$$

mit $e^* = (\Upsilon^{-1/2})e \sim \Upsilon^{-1/2} \cdot \mathcal{N}_n(0, \Upsilon) = \mathcal{N}_n(0, \Upsilon^{-1}\Upsilon) = \mathcal{N}_n(0, I)$. Die Schwierigkeit besteht also darin, die Konstante ϕ zu bestimmen.

Die Konstante ϕ wird mittels *profiled-maximum-likelihood*-Schätzung bestimmt. Dazu fixiert man zuerst ϕ und berechnet die ML-Schätzer für $\hat{\beta}$ und $\widehat{\sigma^2}$ mittels:

$$\begin{aligned}\hat{\beta}(\phi) &= \left((X^*)' X^* \right)^{-1} (X^*)' Y^* \\ \widehat{\sigma^2}(\phi) &= \frac{\|Y^* - X^* \hat{\beta}(\phi)\|^2}{n - p - 1}\end{aligned}\tag{31}$$

Danach setzt man (31) in die normale Likelihoodfunktion von Modell (30) ein und erhält:

$$l(\phi|Y) = \text{const} - (n - p - 1) \log \|Y^* - X^* \hat{\beta}(\phi)\| - \frac{1}{2} \sum_{i=1}^n \log(\det(\Upsilon))\tag{32}$$

Man findet dann $\hat{\phi}$ indem man das Argmax von (32) für ϕ bestimmt. Dies geschieht in dem Paket *nlme* numerisch mittels orthogonal-triangularen Zerlegungen. Sieh hierzu [PB00, S. 68-75].

Um den Wert von $\hat{\phi}$ für vorgegebene Daten zu bestimmen, wurde die Funktion *gl*s aus dem Paket *nlme* benutzt. Bei der Benutzung von *gl*s, um einen Schätzer für ϕ zu bestimmen, muss der Wert *correlation=corAR1()* an *gl*s übergeben werden.

Um dann bei gegebenem ϕ beziehungsweise $\hat{\phi}$ einen Schätzer für β und σ zu bestimmen, wird folgender Algorithmus verwendet:

1. Benutze *gl*s um ϕ zu bestimmen.
2. Bestimme Υ und mittels der R-Funktion *solve* Υ^{-1}
3. Führe eine Eigenwertzerlegung von Υ^{-1} durch, um die Matrix Υ^{-1} zu diagonalisieren und nehme dann von der Hauptdiagonalenmatrix elementenweise die Wurzel, um $\Upsilon^{-1/2}$ zu bestimmen.
4. Transformiere Y nach Y^* und X nach X^* mittels Linksmultiplikation mit $\Upsilon^{-1/2}$
5. Benutze eine normale OLS um $D = (X'^* X^*)^{-1}$, β und σ^2 zu berechnen.

Führt man diese Schritte durch, das heißt man transformiert die Daten und berechnet dann den OLS Schätzer

$$\hat{\beta} = (X^t X)^{-1} X^t Y$$

für die transformierten Daten

$$\Upsilon^{-1/2} Y = \Upsilon^{-1/2} X \beta + \Upsilon^{-1/2} e$$

erhält man

$$\begin{aligned}
\hat{\beta} &= ((\Upsilon^{-1/2}X)^t \Upsilon^{-1/2}X)^{-1} (\Upsilon^{-1/2}X)^t \Upsilon^{-1/2}Y \\
&= (X^t \Upsilon^{-1/2} \Upsilon^{-1/2})^{-1} X^t \Upsilon^{-1/2} \Upsilon^{-1/2}Y \\
&= (X^t \Upsilon^{-1} X)^{-1} X^t \Upsilon^{-1} Y
\end{aligned}$$

Was genau der *generalised least squares* (GLS) Schätzer ist.

Dies entspricht dem Vorgehen der Funktion *gls* aus dem *nlme* Paket. Deswegen wird in Zukunft immer direkt der *gls* Schätzer benutzt.

Beispiel 4.2.1. Jetzt wird das Beispiel aus Abschnitt 1.1 fortgeführt, indem eine Regression mit AR(1) für das Polynommodell mit Grad Drei durchgeführt wird.

Die Daten werden wie im ersten Abschnitt dieser Arbeit erzeugt. Das heißt, es werden β, σ und X initialisiert. Dann wird e simuliert. In diesem Beispiel gehe wir allerdings davon aus, dass $e \sim \mathcal{N}_n(0, \sigma^2 \Upsilon)$. Dabei ist die Korrelationsmatrix Υ von der Form

$$\Upsilon = \begin{bmatrix} 1 & \phi & \phi^2 & \dots & \phi^{n-2} & \phi^{n-1} \\ \phi & 1 & \phi & \dots & \phi^{n-3} & \phi^{n-2} \\ \phi^2 & \phi & 1 & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \phi & \phi^2 \\ \phi^{n-2} & \phi^{n-3} & \dots & \phi & 1 & \phi \\ \phi^{n-1} & \phi^{n-2} & \dots & \phi^2 & \phi & 1 \end{bmatrix}$$

mit dem fest gewählten Wert $\phi = 0.75$.

Man erhält als Schätzer für den AR(1) Parameter $\hat{\phi} = -0.938988$, während der wahre Wert 0.75 ist.

Als Schätzung für die Parameter erhält man (0.54357174 ; 0.21282372 ; -0.03915277 ; 0.24840114) und 0.04455856, während die wahren Werte (10,5,-4,7) und 1 sind.

Zeichnet man dann dieses Regressionspolynom mit den Daten in eine Abbildung erhält man

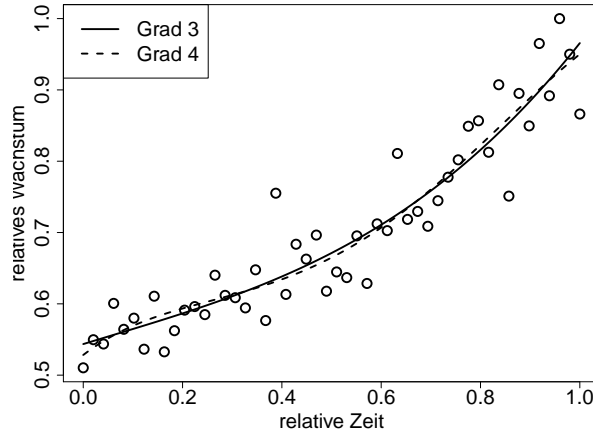


Abbildung 8: Regression für AR(1)

Man sieht, dass auch in diesem Fall ein deutlicher Unterschied zwischen den beiden Regressionspolynomen besteht.

4.3 Konfidenzbänder für AR(1)

Im letzten Abschnitt wurde gezeigt, dass man einen Schätzer für β und σ^2 bei korrelierten Daten erhalten kann, indem man beide Seiten mit der Kovarianzmatrix Υ multipliziert. Wie schon gezeigt, ist $e^* = (\Upsilon^{-1/2})e \sim \mathcal{N}_n(0, I)$. Die Fehler dieses abgeänderten Modells sind also wieder unabhängig verteilt. Das heißt, es können die Ergebnisse aus Kapitel 1 angewendet und ein Konfidenzband K^* bestimmt werden, sodass

$$\mathbb{P}(x^{*\prime}\beta \in K^*) = 1 - \alpha$$

Da uns allerdings nicht $x^{*\prime}\beta$ interessiert, sondern $x'\beta$, muss Bedingung an die Menge in $\mathbb{P}()$ mit $\Upsilon^{1/2}$ multipliziert werden. Auf diese Art erhält man dann ein Konfidenzband für das ursprüngliches Modell (29).

Nun betrachten wir, wie dies die Simulation aus Abschnitt 1.6 beeinflusst. Es ändert sich nur die Simulation von N .

Im Speziellen wird die folgende Simulation verwendet, um konkret den Wert der kritischen Konstante c bei abhängigen Daten zu bestimmen:

1. Simuliere $N \sim \mathcal{N}_n(0, (\Upsilon^{-1/2}X'\Upsilon^{-1/2}X)^{-1}) = (0, (X^{*\prime}X^*)^{-1})$ und $\hat{\sigma}/\sigma \sim \sqrt{\chi_v^2/v}$ mit $v = n - p - 1$.
2. Berechnung von $K_{2h}(T; (X^{*\prime}X^*)^{-1}; (a, b))$.

Die Berechnung von $K_{2h}(T, \cdot, (a, b))$ läuft analog zu Kapitel 1.6, da sich die Funktion $K_{2h}(T; \cdot; (a, b))$ nicht ändert.

Beispiel 4.3.1. Jetzt wird das Beispiel aus dem ersten Kapitel fortgeführt, indem zu der Regression mit Grad Drei, dem Konfidenzband auf ganz \mathbb{R}^p , dem auf A und dem auf A für Polynome noch ein Konfidenzband auf A für Polynome unter Berücksichtigung der AR(1)-Struktur bestimmt wird.

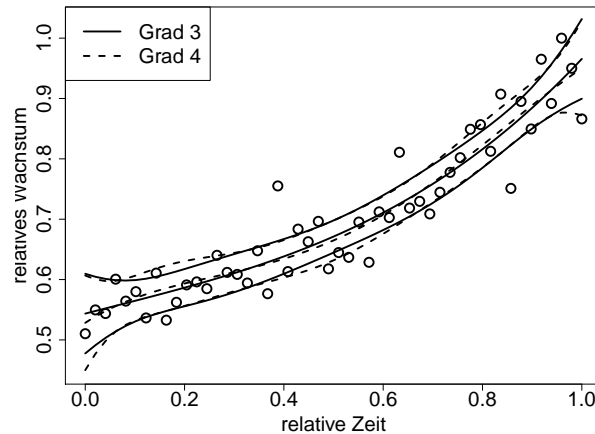


Abbildung 9: Weiterführung des Beispiels durch Konfidenzband auf A unter Berücksichtigung von Polynomstruktur und AR(1)-Struktur

Man erhält als kritischen Wert für das Konfidenzband auf ganz \mathbb{R}^p 3.20876, für das Konfidenzband auf A 3.143099, für das Konfidenzband auf A für Polynome 2.79563 und für das Konfidenzband auf A für Polynome das die AR(1)-Struktur berücksichtigt 2.817034.

Da es sich um eine Simulation handelt und Reproduktivität gewährleistet werden soll, wird die Seed 4 benutzt.

Vergleicht man die Regression mit Grad Drei und mit Grad Vier erhält man die Abbildung 10.

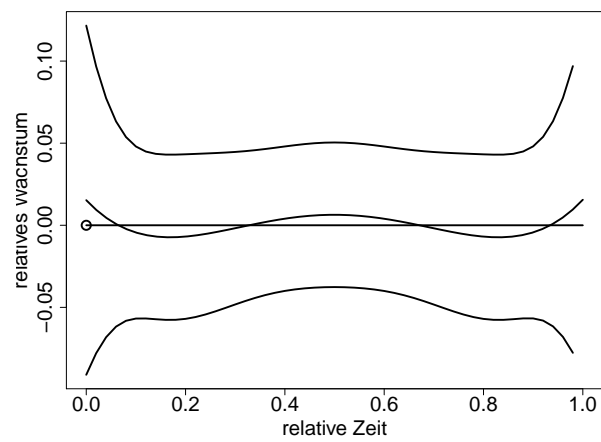


Abbildung 10: Vergleich Regression Grad Drei und Grad Vier unter Berücksichtigung der Polynomstruktur und AR(1)-Struktur

5 Datenbeschreibung und Resultate

In diesem Kapitel werden zuerst die Daten beschrieben, die als Motivation für die bisher erklärten Methoden dienen. Danach werden die Methoden auf die Daten angewendet. Zuerst werden die verschiedenen Methoden Konfidenzbänder zu bestimmen verglichen. Danach werden Polynommodelle von verschiedenem Grad miteinander verglichen. Die Daten entstammen der Arbeit von [ZR10].

Die zur Berechnung der Parameter β und σ sowie der kritischen Werte $c_{\mathbb{R}}, c_A, c_{AP}$, und zum Erzeugen der Abbildungen verwendeten R Skripte können unter

<https://github.com/fake1884/KBminmaxpoly>

eingesehen werden. Eine Übersicht über die verwendeten Skripte befindet sich in der dem R Paket beiliegenden *README.md* Datei in oben genanntem Verzeichnis.

Dabei bezeichnet c immer einen kritischen Wert für die Konstruktion von Konfidenzbändern und der Index bestimmt, um welchen kritischen Wert es sich handelt. Dabei steht \mathbb{R} für ein Konfidenzband auf ganz \mathbb{R}^p , A steht für ein Konfidenzband auf einer Region $A \subset \mathbb{R}^p$ und AP steht für ein Konfidenzband auf einer Region $A \subset \mathbb{R}^p$ mit einem polynomiellen Regressionsmodell.

Es stellt sich in diesem Kapitel häufig die Frage, ob die Nullfunktion in einem bestimmten Konfidenzband enthalten ist. Diese Frage wurde immer durch hinschauen entschieden, da die Unterschiede recht groß sind. Sollte eine genauere Auswertung nötig sein, kann die Funktion *Test.function* verwendet werden, die bei den Simulationen geprüft hat, wie oft das wahre Modell im Konfidenzband liegt.

5.1 Datenbeschreibung

Die Motivation dieser Arbeit ist, wie eine Stammzelle entscheidet, zu welcher Art Gewebe sie wird. Es wird vermutet, dass Stammzellen diese Entscheidung treffen, wenn sie gerade nicht wachsen. Deshalb wurden Stammzellen auf verschiedene Untergründe gesetzt und zu jeweils bestimmten äquidistanten Zeitpunkten unter einem Mikroskop fotografiert. Danach wurde ihre Fläche bestimmt. Mit Hilfe dieser Daten wird eine Regression mit der Zeit als unabhängige Variable durchgeführt. Es wird eine Polynomregression verwendet, da die Ableitungen von Polynomen gut zu berechnen sind und man so die kritischen Punkte des Regressionsgraphen einfach bestimmen kann.

Es geht darum, das beste Polynom-Modell für die Daten zu wählen. Ein Polynom-Modell ist durch den Grad des Polynoms eindeutig definiert. Das Problem ist also den besten Polynomgrad zu finden.

Um dieses Problem zu lösen, betrachtet man, ob sich die Polynome statistisch signifikant unterscheiden. Dazu schaut man die Differenz der Polynome an und berechnet für diese gleichmäßige Konfidenzbänder. Ist die Nullfunktion vollständig in diesem Konfidenzband enthalten, ist es nicht möglich eine Aussage über den Unterschied der Modelle zu machen. Liegt die Nullfunktion allerdings nicht vollständig im Konfidenzband so, ist der Unterschied signifikant.

Die konkrete Fragestellung meiner Bachelorarbeit ist, ob die Polynome, die man durch die Daten legt, sich signifikant unterscheiden.

Es gibt zwei verschiedenen Datensätze. Zum einen gibt es einen Datensatz mit Stammzellen, die auf eine Oberfläche mit einer Härte von 10 kPa gesetzt wurden. Zum anderen gibt es einen Datensatz aus Stammzellen, die auf eine Oberfläche von 30 kPa gesetzt wurden.

Es handelt sich um Daten von 53 Stammzellen bei den 30 kPa Daten und um 51 Stammzellen bei den 10 kPa Daten. An jeder Stammzelle wurde 145 Mal die Größe gemessen. Da die Fläche der Stammzellen für jede einzelne Stammzelle stark schwankt, wird der Mittelwert der Flächen gebildet.

Plottet man nun das Mittel der relative Wachstume Y gegen die relative Zeit von Null bis Eins mit Schrittweite $1/\text{Anzahl an Beobachtungen}$, entstehen die Abbildungen 11 und 12.

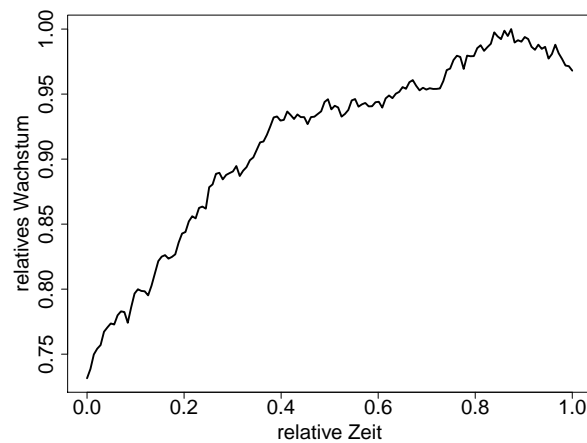


Abbildung 11: 10 kPa Daten

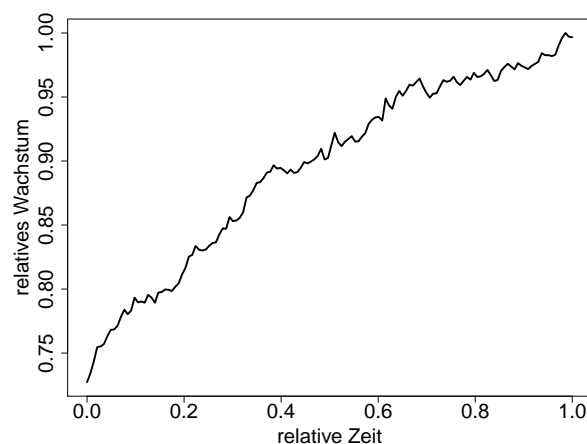


Abbildung 12: 30 kPa Daten

Man sieht, dass die Stammzellen in beiden Fällen im Zeitverlauf wachsen. Dabei scheinen die 10 kPa Stammzellen im Zeitintervall $[0.1, 0.4]$ schneller zu wachsen, als die 30 kPa Stammzellen. Dafür flacht sich das Wachstum im späteren Zeitverlauf bei den 10 kPa Stammzellen im Vergleich zu den 30 kPa Stammzellen wieder etwas ab.

5.2 Vergleich verschiedener Konfidenzbänder

In diesem Abschnitt werden die drei Typen von Konfidenzbändern sowohl für 30 kPa als auch für 10 kPa Daten miteinander verglichen.

Wir gehen an dieser Stelle davon aus, dass es zwischen den Daten eine Autokorrelationsbeziehung der Art AR(1) gibt. Das bedeutet, zwischen der mittleren Fläche der Stammzellen zum Zeitpunkt i und $i - 1$ gibt es folgende Beziehung

$$y_i = \phi y_{i-1} + e_i$$

mit e_i u.i.v., $\mathbb{E}(e_i) = 0$ und $\mathbb{E}(e_i^2) = \sigma_e^2 < \infty$.

Also muss man die Ergebnisse aus Kapitel 4 anwenden. Das Regressionsmodell ist ein Polynom vom Grad Fünf. Das heißt, X ist die 53×5 Matrix mit der l -ten Zeile von der Form

$$\tilde{x}_l = (1, x, x^2, x^3, x^4, x^5) \text{ mit } x \in \{0, 1/145, 2/145, \dots, 144/145, 1\}$$

Klarerweise ist dann $\beta = (1, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$ der gesuchte Koeffizientenvektor und sein Schätzer ist $\hat{\beta} = (1, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5)$. Analoges gilt für σ und $\hat{\sigma}^2$.

Insgesamt gehen wir also von dem folgenden Regressionsmodell aus, wobei die oben eingeführten Bezeichnungen benutzt werden:

$$Y = X\beta + e$$

mit $e \sim \mathcal{N}_{53}(0, \sigma^2 \Upsilon)$. Dabei ist Υ die von dem zugrunde liegendem AR(1)-Prozess erzeugte Korrelationsmatrix (28)

$$\Upsilon = \begin{bmatrix} 1 & \phi & \phi^2 & \dots & \phi^{n-2} & \phi^{n-1} \\ \phi & 1 & \phi & \dots & \phi^{n-3} & \phi^{n-2} \\ \phi^2 & \phi & 1 & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \phi & \phi^2 \\ \phi^{n-2} & \phi^{n-3} & \dots & \phi & 1 & \phi \\ \phi^{n-1} & \phi^{n-2} & \dots & \phi^2 & \phi & 1 \end{bmatrix}$$

die wir bereits aus Kapitel 4.1 kennen.

Die Matrix Υ wird durch den Parameter ϕ bereits vollständig charakterisiert. Der Schätzer für ϕ wird mit $\hat{\phi}$ bezeichnet.

Bei den beiden Simulationen für c_A und c_{AP} werden jeweils 250 Iterationen durchgeführt. Bei den Simulationen auf A für das Modell mit Polynomgestalt wird ein Gridsearch für das Maximum mit Feinheit 145, also der Anzahl an Messungen durchgeführt.

Führt man diese Regression durch, erhält man die Schätzer für die 30 kPa Daten

$$\begin{aligned}\hat{\beta} &= (0.7287814; 0.7345406; -2.1593520; 5.2097479; -5.9029071; 2.3863759) \\ \widehat{\sigma^2} &= 0.009266163 \\ \hat{\phi} &= 0.9039573\end{aligned}$$

Als kritische Parameter erhält man $c_{\mathbb{R}} = 3.604075$, $c_A = 3.505922$ und $c_{AP} = 2.902647$. Setzt man dasselbe Modell wie oben für die 30 kPa Daten für die 10 kPa Daten, nur mit an, erhält man die Werte

$$\begin{aligned}\hat{\beta} &= (0.7372386; 0.4218583; 1.8729639; -7.6991653; 9.7892837; -4.1541509) \\ \widehat{\sigma^2} &= 0.007545373 \\ \hat{\phi} &= 0.8225374\end{aligned}$$

Als kritische Parameter erhält man $c_{\mathbb{R}} = 3.604075$, $c_A = 3.550114$, $c_{AP} = 2.848484$.

Anhand der kritischen Werte sieht man bereits, dass die Interferenz bei dem Konfidenzband auf A für das Polynommodell eindeutig besser ist. Das heißt, der Wert ist kleiner und somit das Konfidenzband schmäler.

Zeichnet man die Regressionsmodelle und die Konfidenzbänder jeweils in eine Abbildung erhält man die beiden Abbildungen 13 und 14.

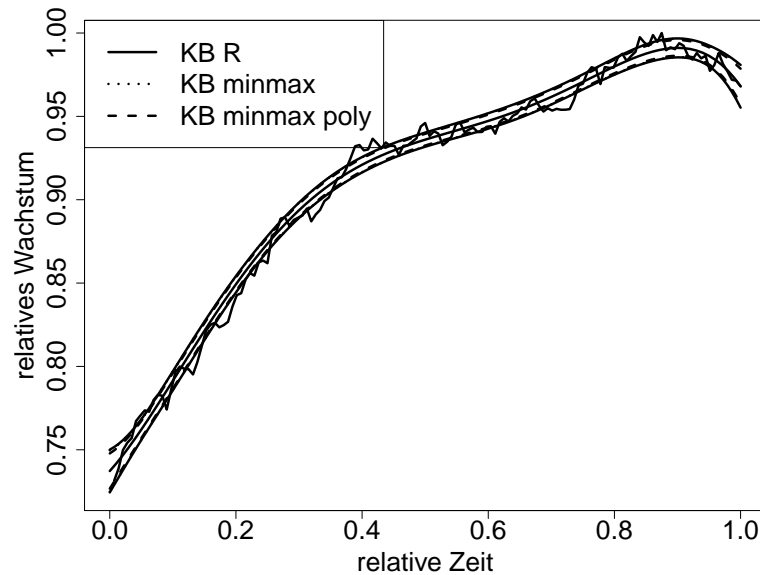


Abbildung 13: Vergleich von Konfidenzbändern 10 kPa

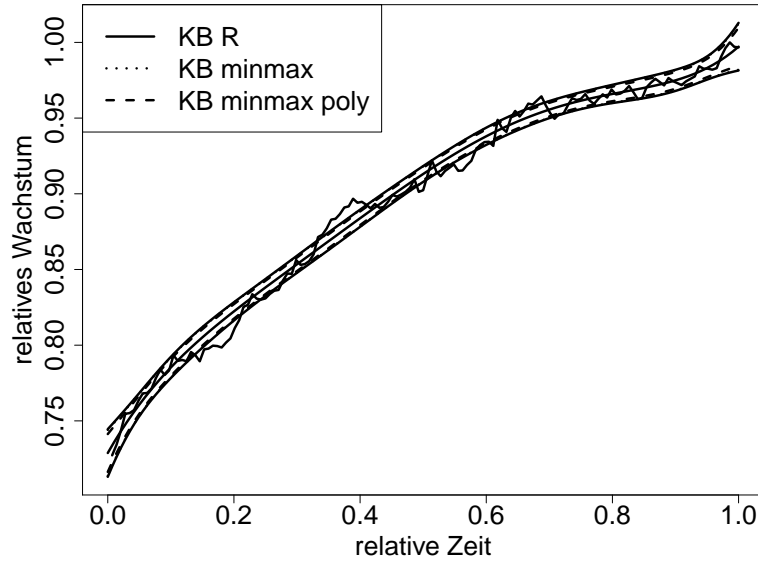


Abbildung 14: Vergleich von Konfidenzbändern 30 kPa

Man sieht, dass das Konfidenzband auf $A = [0, 1] \subset \mathbb{R}^p$, das die Polynomstruktur berücksichtigt, erkennbar schmaler ist. Zwischen den anderen beiden Konfidenzbänder ist kein Unterschied erkennbar.

5.3 Vergleich von polynomen Regressionsmodellen mit Grad Vier, Fünf und Sechs

Ziel dieses Abschnittes ist es zu testen, ob Polynomregressionen von verschiedenem Grad für die beiden Stammzelldatensätze sich statistisch signifikant unterscheiden.

Das heißt, wir führen Regressionen mit einer Designmatrix X bei der die l -te Zeile von X durch $\tilde{x}_l = (1, x, x^2, x^3, x^4)$ gegeben ist durch. Danach führen wir analoge Regressionen für Grad Fünf und Sechs durch.

Dazu wird zuerst die Regression durchgeführt, bevor die Testhypothese formalisiert wird. Als Regressionsmodell wird das Regressionsmodell aus Kapitel 4

$$Y = X\beta + e$$

mit $e \sim \mathcal{N}_n(0, \sigma^2 \Upsilon)$ verwendet. Dabei ist Υ die bekannte von einem AR(1) Prozess erzeugte Korrelationsmatrix, die nur von dem unbekannten Parameter ϕ abhängt.

Führen wir für die Regressionen durch erhalten wir für die 10 kPa Daten die folgenden Werte. Dabei steht der Index für den Grad des zugehörigen Polynommodells

$$\hat{\beta}_4 = (0.7314; 0.8401; -1.5883; 1.8877; -0.9029)$$

$$\hat{\sigma}_4 = 0.4943004$$

$$\hat{\phi}_4 = 0.9999575$$

$$\hat{\beta}_5 = (0.7372; 0.4219; 1.8730; -7.6992; 9.7893; -4.1542)$$

$$\hat{\sigma}_5 = 0.007545373$$

$$\hat{\phi}_5 = 0.8225374$$

$$\hat{\beta}_6 = (0.7352; 0.6265; -0.6151; 3.0831; -11.1962; 14.6223; -6.2906)$$

$$\hat{\sigma}_6 = 0.007205863$$

$$\hat{\phi}_6 = 0.8065796$$

und für die 30 kPa Daten die Werte:

$$\hat{\beta}_4 = 0.73110.5245 - 0.35460.03900.0546$$

$$\hat{\sigma}_4 = 0.009576459$$

$$\hat{\phi}_4 = 0.9090669$$

$$\hat{\beta}_5 = 0.72880.7345 - 2.15945.2097 - 5.90292.3864$$

$$\hat{\sigma}_5 = 0.009266163$$

$$\hat{\phi}_5 = 0.9039573$$

$$\hat{\beta}_6 = 0.72730.9495 - 4.766416.2343 - 26.756620.5211 - 5.9124$$

$$\hat{\sigma}_6 = 0.01673707$$

$$\hat{\phi}_6 = 0.9701881$$

Plottet man die drei Regressionsmodelle in je eine Abbildung für die beiden Datensätze entstehen die beiden Abbildungen 15 und 16:

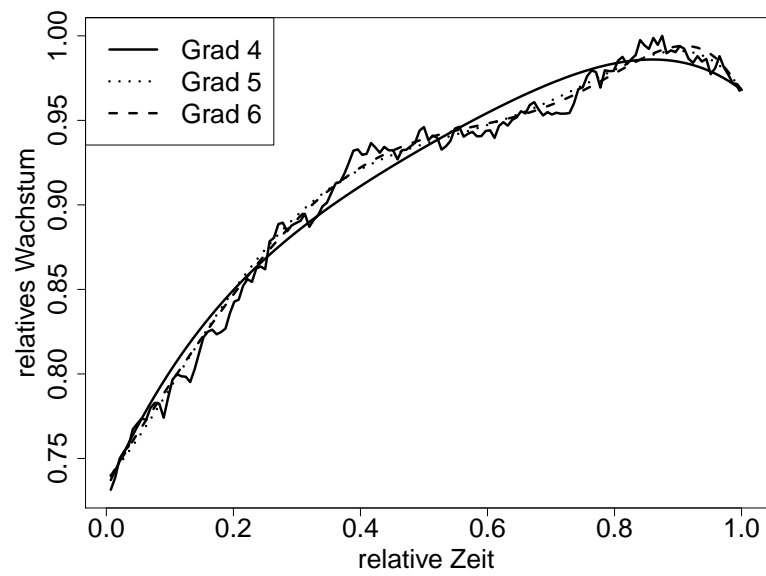


Abbildung 15: 10 kPa Verschiedene Polynomgrade

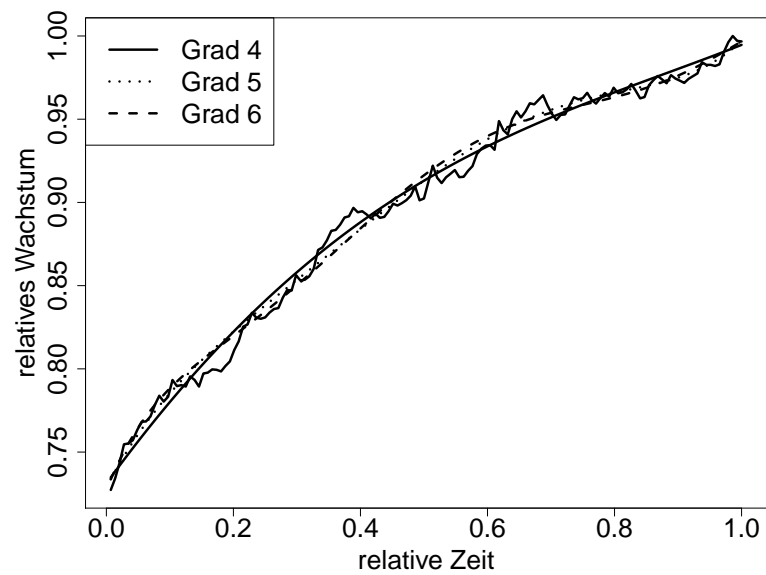


Abbildung 16: 30 kPa Verschiedene Polynomgrade

Man sieht, dass sich die Regressionsmodelle für die 10 kPa Daten relativ stark unterscheiden, während die Regressionsmodelle bei den 30 kPa Daten fast gleich sind.

Als nächstes testen wir, ob sich die Polynomregressionen von Grad Vier und Grad Fünf unter statistischer Unsicherheit gleich sind. Das heißt, wir testen die Hypothese

$$H_0 : \beta_4 = \beta_5 \text{ vs. } \beta_4 \neq \beta_5$$

Dazu benutzen wir die Konfidenzbandmethode aus Kapitel 2 beziehungsweise ihre Abwandlung aus Abschnitt 4.3.

Zeichnen wir die Differenz von $\hat{\beta}_4$ und $\hat{\beta}_5$ mit dem zugehörigen Konfidenzband in je eine Abbildung für die beiden Datensätze erhalten wir die beiden Abbildungen 17 und 18.

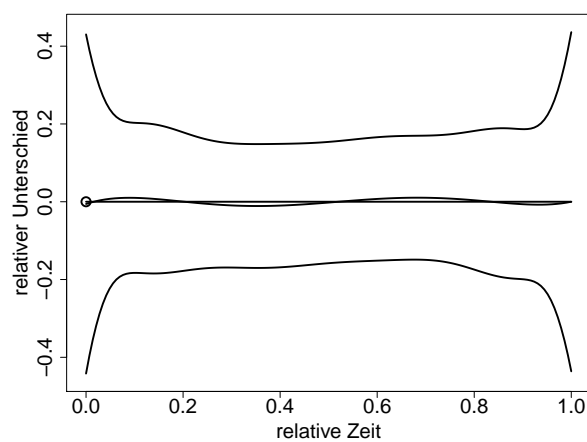


Abbildung 17: 10 kPa Vergleich von Regressionsmodellen mit Grad Vier versus Grad Fünf

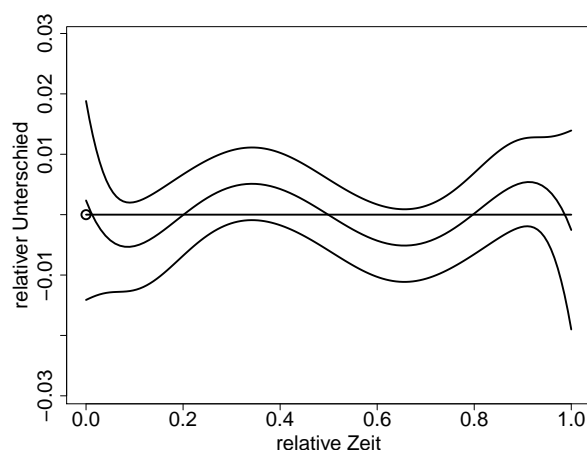


Abbildung 18: 30 kPa Vergleich von Regressionsmodellen mit Grad Vier versus Fünf

Wie man sieht, ist in im Fall der 10 kPa Daten die Nullfunktion eindeutig vollständig im Konfidenzband enthalten und wir können die Nullhypothese nicht ablehnen.

Bei den 30 kPa Daten ist die Nullhypothese auch knapp ganz im Konfidenzband enthalten und wir können die Nullhypothese wieder nicht ablehnen.
In den folgenden Vier Abbildungen testen wir die analogen Hypothesen

$$\begin{aligned} H_0 : \beta_4 = \beta_6 & \quad \text{vs.} \quad H_1 : \beta_4 \neq \beta_6 \\ H_0 : \beta_5 = \beta_6 & \quad \text{vs.} \quad H_1 : \beta_5 \neq \beta_6 \end{aligned}$$

für die 10 kPa beziehungsweise 30 kPa Daten. Für die Hypothese Grad Vier versus Grad Sechs erhält man die Abbildungen 19 und 20

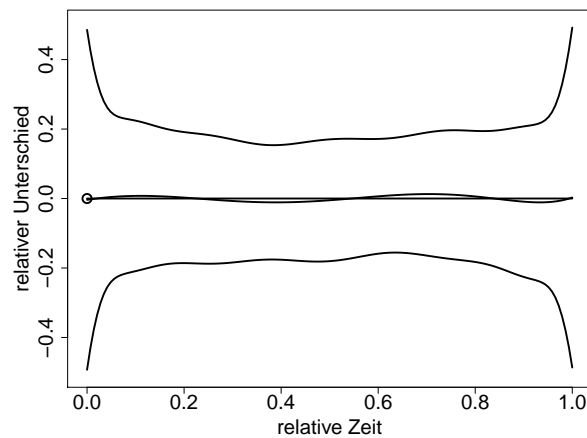


Abbildung 19: 10 kPa Vergleich von Regressionsmodellen mit Grad Vier versus Sechs

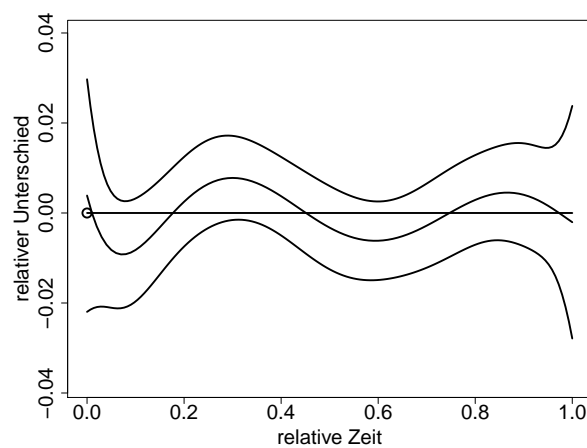


Abbildung 20: 30 kPa Vergleich von Regressionsmodellen mit Grad Vier versus Sechs

Man sieht, dass sich der Trend aus den beiden vorherigen Abbildungen wiederholt. Bei den 10 kPa Daten gibt es keinen Zweifel, dass die Nullhypothese nicht abgelehnt werden kann. Bei den 30 kPa Daten kann die Nullhypothese auch eindeutig nicht abgelehnt werden. Für die Hypothese Grad Fünf versus Grad Sechs erhält man die Abbildungen 21 und 22.

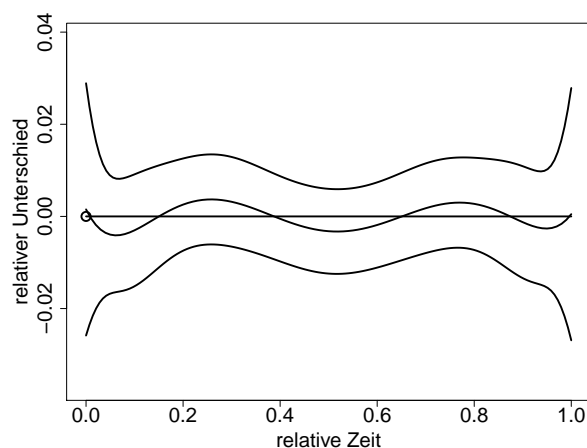


Abbildung 21: 10 kPa Vergleich von Regressionsmodellen Grad Fünf versus Grad Sechs

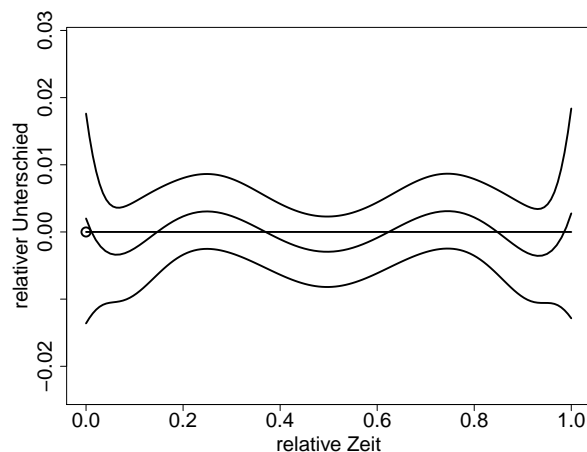


Abbildung 22: 30 kPa Vergleich von Regressionsmodellen Grad Fünf versus Grad Sechs

Man sieht, dass wieder in beiden Fällen die Nullhypothese nicht abgelehnt werden kann.

5.4 Vergleich von 10 kPa und 30 kPa Daten

Ziel dieses Abschnittes ist es die folgende Hypothese zu testen:

$$H_0 : \beta_{10kPa} = \beta_{30kPa} \text{ vs. } H_1 : \beta_{10kPa} \neq \beta_{30kPa}$$

Dabei ist β_{10kPa} der Koeffizientenvektor einer Polynomregression vom Grad Fünf und β_{30kPa} der Koeffizientenvektor einer Polynomregression vom gleichen Grad.

Weiterhin gehen wir davon aus, dass dem Regressionsmodell vom Grad Fünf ein AR(1) Prozess zugrunde liegt. Außerdem verwenden wir die Methode aus Kapitel 2, das heißt wir gehen davon aus, dass beide Prozesse das gleiche ϕ und σ besitzen.

Zuerst plotten wir die beiden Datensätze mit den Regressionspolynomen in dieselbe Abbildung, um einen Überblick über ihre Ähnlichkeit zu bekommen. Dabei entsteht die Abbildung 23.

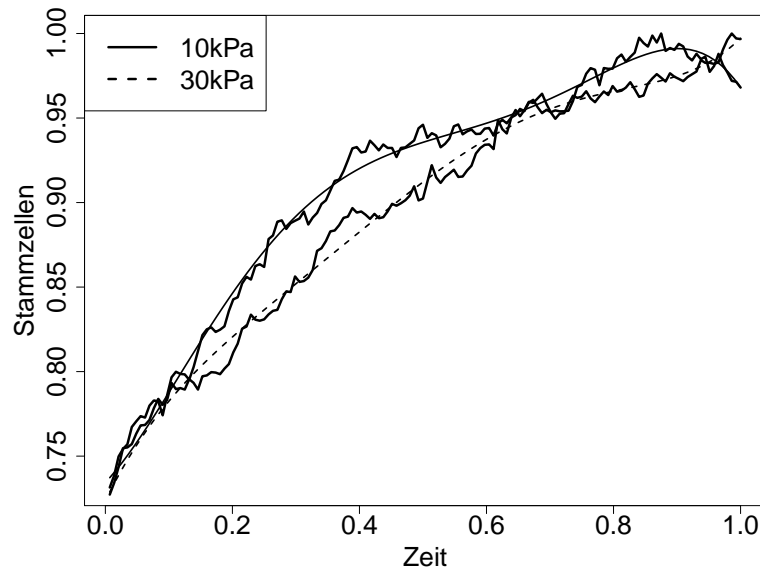


Abbildung 23: Vergleich von Regressionsmodellen

Als Parameter bei der Regression erhält man für die 10 kPa Daten:

$$\begin{aligned}\hat{\beta} &= (0.7343828; 0.3972789; 2.0659657; -8.1435038; 10.2153513; -4.3014467) \\ \widehat{\sigma^2} &= 0.007545373 \\ \hat{\phi} &= 0.8225374\end{aligned}$$

und für die 30 kPa Daten:

$$\begin{aligned}\hat{\beta} &= (0.7235374; 0.7708670; -2.3022522; 5.4896056; -6.1555617; 2.4709908) \\ \widehat{\sigma^2} &= 0.009266164 \\ \hat{\phi} &= 0.9039573\end{aligned}$$

Dies sind genau die selben Regressionsdaten aus Kapitel 5.2, allerdings diesmal mit einer anderen Fragestellung im Hintergrund.

Als nächstes wird die Differenz der Regressionsmodelle mit einem Konfidenzband in eine Abbildung geplottet. Dabei wird zur Erzeugung des kritischen Parameters die Methode für Konfidenzbänder auf einem Intervall, falls das Modell Polynomgestalt hat, angewendet.

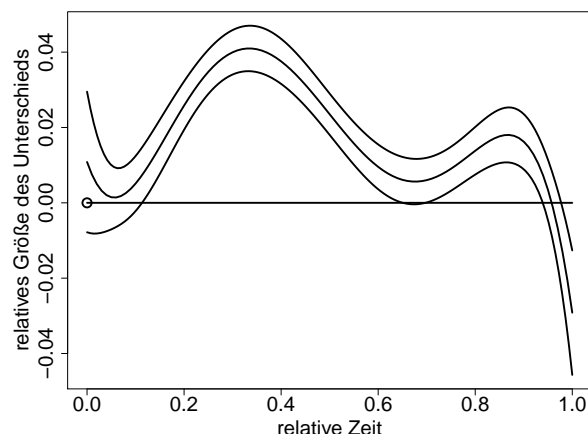


Abbildung 24: Vergleich von 10 kPa und 30 kPa Daten

Man sieht, dass die Nullfunktion nicht komplett im Konfidenzband enthalten ist. Man kann die Nullhypothese also verwerfen. Es gibt einen statistisch signifikanten Unterschied zwischen den beiden Regressionsmodellen.

5.5 Einen Teil eines Regressionsmodells überprüfen

Ziel dieses Abschnittes ist es, zu prüfen, ob bei den Stammzelldaten die ersten Koeffizienten des Koeffizientenvektors signifikant von Null verschieden sind.

Dazu führen wir für sowohl für die 10 kPa als auch für die 30 kPa Stammzelldaten je zwei Polynomregressionen durch. Wir legen dabei wieder eine Zeitreihe von der AR(1) Art zugrunde.

Dabei erhalten wir für die 10 kPa Daten die Werte:

$$\hat{\beta}_5 = (0.7373; 0.4219; 1.8730; -7.6992; 9.7893; -4.1542)$$

$$\hat{\sigma}_5^2 = 0.007545373$$

$$\hat{\phi}_5 = 0.8225374$$

$$\hat{\beta}_6 = (0.7352; 0.6265; -0.6151; 3.0831; -11.1962; 14.6223; -6.2906)$$

$$\hat{\sigma}_6^2 = 0.007205863$$

$$\hat{\phi}_6 = 0.8065796$$

und für die 30 kPa Daten die Werte:

$$\begin{aligned}\hat{\beta}_5 &= (0.7288; 0.7345; -2.1594; 5.2097; -5.9029; 2.3864) \\ \hat{\sigma}_5^2 &= 0.009266163 \\ \hat{\phi}_5 &= 0.9039573 \\ \hat{\beta}_6 &= (0.7273; 0.9495; -4.7664; 16.2343; -26.7566; 20.5211; -5.9124) \\ \hat{\sigma}_6^2 &= 0.01673707 \\ \hat{\phi}_6 &= 0.9701881\end{aligned}$$

Um die Hypothese angeben zu können, definieren wir

$$\beta_i := (\beta_{i,1}, \beta_{i,2}, \dots, \beta_{i,i}) \text{ für } i=5,6$$

Damit können wir die folgenden Test durchführen.

$$\begin{aligned}H_0 : \beta_{5,5} &= 0 & \text{vs. } H_1 : \beta_{5,5} &\neq 0 \\ H_0 : \beta_{6,6} &= 0 & \text{vs. } H_1 : \beta_{6,6} &\neq 0 \\ H_0 : (\beta_{6,5}, \beta_{6,6}) &= (0, 0) & \text{vs. } H_1 : (\beta_{6,5}, \beta_{6,6}) &\neq (0, 0)\end{aligned}$$

Dazu verwendet man die Methode aus Abschnitt 3.3. Wir verwerfen also die Nullhypothese, wenn die Konfidenzbänder $x'_2 \cdot 0$ nicht vollkommen enthalten.

Dabei ist in unserem Fall $x'_2 \in \{0, 1/145, 2/145, \dots, 144/145, 1\}$.

In den Abbildungen 25 und 26 ist $x'_2 \beta_{5,5}$ mit zugehörigem Konfidenzband eingezeichnet.

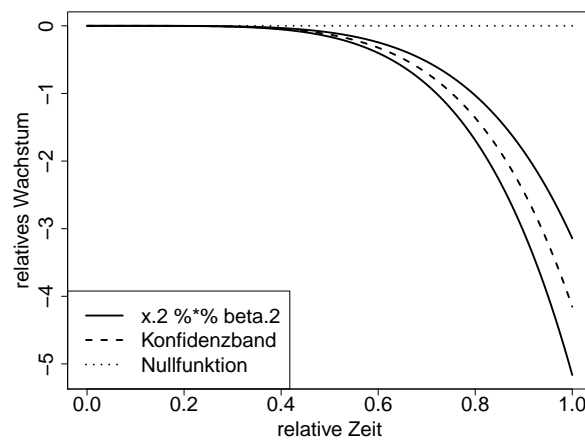


Abbildung 25: 10 kPa, Grad 5

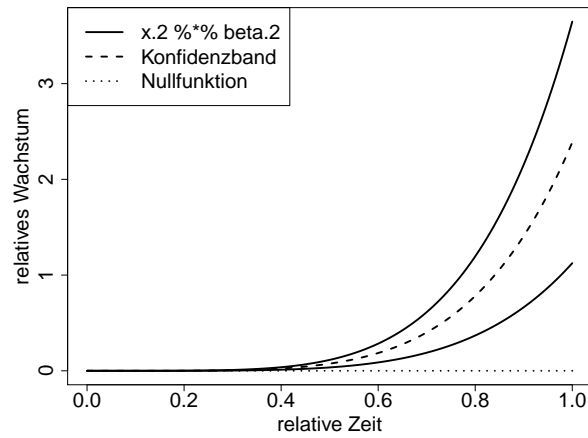


Abbildung 26: 30 kPa, Grad 5

Man sieht deutlich, dass in beiden Fällen die Nullfunktion nicht vollständig im Konfidenzband enthalten ist. Wir können die Nullhypothese H_0 also verwerfen und haben einen Anhaltspunkt, davon auszugehen, dass $\beta_{5,5}$ in einem der beiden Fälle Null ist. Die nächsten Abbildungen 27 und 28 geben die Überprüfung von $\beta_{6,6}$ für die beiden Sätze an Stammzelldaten wieder.

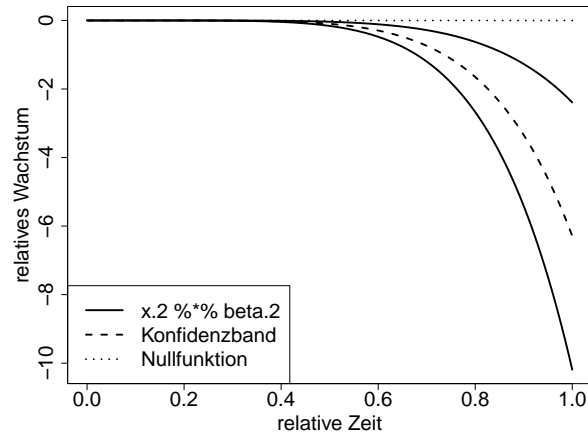


Abbildung 27: 10kPa, Grad 6, Test auf Grad Sechs

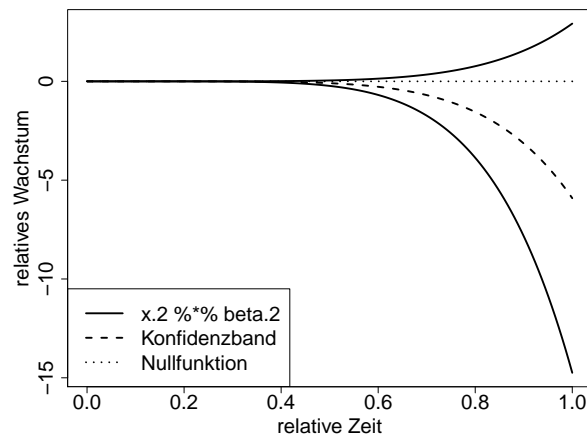


Abbildung 28: 30kPa Grad 6, Test auf Grad Sechs

Man sieht, dass für die 10 kPa Daten die Nullfunktion wieder nicht ganz in dem Konfidenzband enthalten ist.

Bei den 30 kPa Daten ist dies jedoch der Fall. Damit kann die Nullhypothese abgelehnt werden und in diesem Fall gilt $\beta_{6,6} \neq 0$.

Betrachten wir zum Schluss den Test, ob $(\beta_{6,5}, \beta_{6,6}) = (0, 0)$.

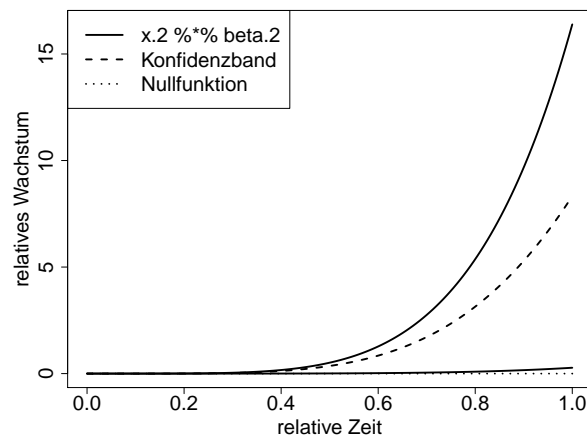


Abbildung 29: 10 kPa, Grad 6, Test auf Grad Fünf

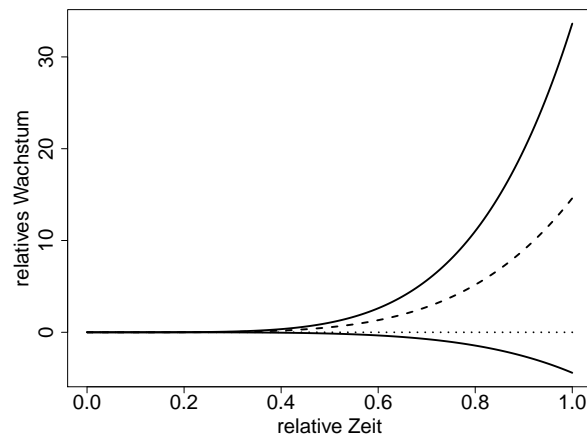


Abbildung 30: 30 kPa, Grad 6, , Test auf Grad Fünf

Auch hier sieht man, dass für die 10 kPa Daten die Nullfunktion wieder nicht ganz in dem Konfidenzband enthalten ist. Deshalb kann auch in diesem Fall H_0 nicht verworfen werden. Bei den 30 kPa Daten ist dies jedoch der Fall. Damit kann die Nullhypothese abgelehnt werden und in diesem Fall gilt $\beta_{6,6} \neq 0$.

Fazit

In dieser Bachelorarbeit haben wir Methoden behandelt, um Regressionsmodelle zu vergleichen. Dazu wurde eine Verbesserung des normalerweise verwendeten F-Tests mittels auf Konfidenzbändern basierender Methoden nach [Liu64] betrachtet.

Dabei wurden die Methoden in [Liu64] auf den Fall von abhängigen Daten mit einer Abhängigkeitsstruktur der Art AR(1) verallgemeinert.

Insbesondere für Regressionsmodelle mit Polynomgestalt auf einem Polyeder waren die durchgeführten Tests aussagekräftig, die diese speziellen Anforderungen an den Test berücksichtigen haben. Dies sieht man daran, dass sich kleinere kritische Konstanten für die Konfidenzbänder ergeben haben.

Allerdings ist die Rechenzeit gerade für diese Methode sehr lang, da eine Monte-Carlo Simulation durchgeführt werden muss. Diese lange Rechenzeit hat sich auch negativ auf die Ergebnisse der simulationsweisen Bestimmung der Überdeckungswahrscheinlichkeiten ausgewirkt.

Die Methode Konfidenzbänder auf ganz \mathbb{R}^p zu bestimmen war schneller und lieferte zum Teil bessere Überdeckungswahrscheinlichkeiten.

Die in dieser Ausarbeitung verwendeten R und C++ Codes sind in dem Paket KBminmaxpoly auf der Seite

<https://github.com/fake1884/KBminmaxpoly>

gefunden werden.

Literatur

- [BD91] P. J. Brockwell und R. A. Davis, *Time Series: Theory and Methods*. Springer-Verlag, 1991.
- [DS98] N. Draper und H. Smith, *Applied Linear Regression, 3rd edition*. Wiley, 1998.
- [Geo09] H.-O. Georgii, *Stochastik*. de Gruyter, 2009.
- [Han15] B. E. Hansen, *Econometrics*. University of Wisconsin, 16.1.2015.
- [Hsu41] P. Hsu, *Canonical reduction of the general regression problem*. Annals of Eugenics, 1941.
- [Kri15] T. Krivobokova, »Angewandte statistik«, Skript, SS 2015.
- [Liu64] W. Liu, *Simultaneous Inference in Regression*. Taylor und Francis Group, Boca Raton, 1964.
- [LLP08] W. Liu, S. Lin und W. Piegorsch, *Construction of exact simultaneous confidence bands for a simple linear regression model*. International Statistical Review, 2008.
- [Mun13] A. Munk, »Introduction to mathematical statistics, From a practical point of view«, Skript, 29.10.2013.
- [Nai87] D. Naiman, *Simultaneous Confidence Bounds in Multiple Regression Using Predictor Variable Constraints*. Journal of the American Statistical Association, 1987.
- [PB00] J. C. Pinheiro und D. M. Bates, *Mixed-Effects Models in S and S-PLUS*. Annals of Eugenics, 2000.
- [WB71] H. Wynn und P. Bloomfield, *Simultaneous confidence bands in regression analysis*. Journal of the Royal Statistical Society, 1971.
- [ZR10] A. Zemel und F. Rehfeldt, »Optimal matrix rigidity for stress-fibre polarization in stem cells«, *Nature Physics*, 2010.

Selbstständigkeitserklärung

Hiermit bestätige ich, dass ich die vorliegende Arbeit selbstständig, nur mit Hilfe meiner Betreuer und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus fremden Quellen (einschließlich elektronischer Quellen) direkt oder indirekt übernommenen Gedanken sind ausnahmslos als solche kenntlich gemacht. Die Arbeit ist in gleicher oder ähnlicher Form oder auszugsweise im Rahmen einer anderen Prüfung noch nicht vorgelegt worden.

Göttingen,

Henning Hause