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

Matrikelnummer:
21340619

Studiengang:
B. Sc. Mathematik

Abgabetermin:
11.07.2017

Erstgutachter:
Prof. Dr. Tatyana Krivobokova

Zweitgutachter:
Jun.-Prof. Dr. Andrea Krajina

Inhaltsverzeichnis

Einleitung

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

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

gegeben. Dabei seien $Y_i = (Y_{i,1}, \ldots, Y_{i,n_i})'$ zwei Vektoren mit Beobachtungen.
Weiterhin sei X_i eine $n_i \times (p+1)$ Designmatrix mit festem Design, dass heißt, für $i \in \{1, \ldots, \min(n_1, n_2)\}$ ist die l-te Zeile von X_1 gleich der l-ten Zeile von X_2.
Weiterhin ist $\beta_i = (\beta_{i,0}, \ldots, \beta_{i,p})'$ ein Koeffizientenvektor und $e_i = (e_{i,1}, \ldots, e_{i,n_i})'$ ein Vektor mit Zufallsfehlern. Also sind Y_1 und Y_2 zwei Gruppen von Beobachtungen, die von den selben Kovariaten 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_i, 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 $e_i \sim \mathcal{N}_{n_i}(0, \sigma^2 I_{n_i})$. Es stellt sich die Frage zu testen, ob

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

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, dass heißt die l-te ($1 \leq l \leq n_i$) Zeile der Designmatrix X ist von der Form $(1, x_{i,l}, x_{i,l}^2, \ldots, x_{i,l}^p)$.
- Den Fehlern liegt ein AR(1)-Prozess zugrunde, dass 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 Null ist.

Motivation für diese Betrachtungen sind der Vergleich von verschiedenen Regressionsmodellen für Stammzellen.
Das Kapitel 1 stellt die Konzepte der Regression und der Konfidenzbänder vor. Außerdem werden Methoden angeben, 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-Test angegeben und dann mithilfe der Methoden aus dem Kapitel 1 bessere Möglichkeiten einen Teil des Konfidenzbandes zu überprüfen, angegeben.
Das Kapitel 4 beschäftigt sich mit der Fragestellung, wie Regression und Konfidenzbänder bei abhängigen Daten durchzuführen ist. 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 den ersten Kapiteln folgt auf die dort behandelte Theorie oder Methode ein Beispiel. Dieses Beispiel zieht sich durch die gesamten ersten Kapitel und wird jeweils ergänzt oder modifiziert.
Das beiliegenden R-Paket ist in der beiliegenden README Datei beschrieben.
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 und dürfte gut zu lesen sein.
Bei den Simulationen erhält man folgende Ergebnisse:

$\times$	unabhängig	AR bekannt	AR
Konfidenzband auf ganz $\mathbb{R}^p$	1.00	0.94	1.00
Konfidenzband auf einem Polyeder	.99	0.93	1.00
Konfidenzband auf einem Polyeder für Polynome	0.93	0.86	0.99

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 $A \subset \mathbb{R}^p$, Konfidenzbänder auf einem Polyeder A, wenn das Modell Polynomgestalt hat.

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 Teilabschnitt $\mathrm{A} = \{(x_1, \ldots, x_p)' : a_i \leq x_i \leq b_i, i = 1, \ldots, p\} \subset \mathbb{R}^p$ für $-\infty \leq a_i < b_i \leq \infty, i = 1, \ldots, p$. Dabei bezeichnet x' die Transponierte von x während die Ableitung von x mit $\frac{d}{dx}x$ bezeichnet wird.
Dabei ist $x_i = (1, x_0) \subset \mathbb{R}^{p+1}$ die i-te Zeile von 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 betrachten wir den allgemeinen 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 von [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 ein 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 $\mathrm{e} \in \mathbb{R}^n$ 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 Regression benutzt. Das heißt, es liegt die Annahme zugrunde, dass $f(X) = X \cdot \beta$ ist. Dabei sind die $\beta = (\beta_0, \ldots, \beta_p)$ unbekannte Konstanten.
Ein besonderes Modell, dass im Abschnitt 1.6 und dem letzten Kapitel über Datenbeschreibung und Resultate wichtig wird, ist das Polynom-Modell. Bei diesem Modell ist die l-te Zeile von X gegeben durch $\tilde{x_l} = (1, x_l^1, x_l^2, \ldots, x_l^p)$.
Damit wird das Regressionsmodell zu

$$Y = X\beta + e \tag{1}$$

Zusätzlich ist in dieser Ausarbeitung für dieses Kapitel und die Kapitel über den Vergleich und das Prüfen von Regressionsmodellen 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 Werte bekannt sind und mit großen griechischen Buchstaben, wenn ihre Werte 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 $\widehat{\sigma^2}$ bezeichnet.
Der folgende Satz gibt ein Möglichkeit Schätzer $\hat{\beta}$ und $\hat{e}$ beziehungsweise $\widehat{\sigma^2}$ konkret zu berechnen.

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

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

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

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

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

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

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

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

Also muss gelten

$$X'X\hat{\beta} = X'Y \tag{6}$$

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] □

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 äuqidistantes 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 $\tilde{x_l} \in G, l \in \{1, \ldots, 50\}$. Als nächstes wird $\beta = (10, 5, -4, 7)$ und $\sigma = 1$ gewählt.
Dann wird $e \sim \mathcal{N}_n(0, \sigma * I_n)$ generiert und Y nach dem Modell (1) berechnet.
In folgender Graphik 1 sind die Datentupel (Y, G) eingezeichnet:

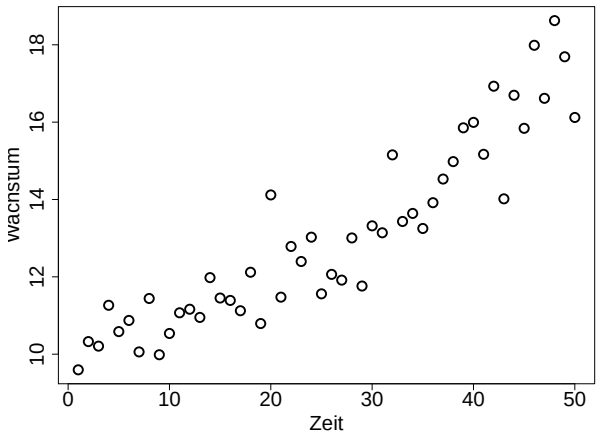

Abbildung 1: Plot der Beispieldaten

In den folgenden Kapiteln wird mit einem normierten Y gearbeitet, indem durch $\max(Y)$ geteilt wird. Dadurch erhält man Werte zwischen Null und Eins, wodurch sich die Berechnungen vereinfachen.
Man kann dies machen, 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 multiple 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 $\widehat{\sigma^2}$=0.04515636. In folgender Graphik 2 sind die Daten mit der Regressionsgerade eingezeichnet.

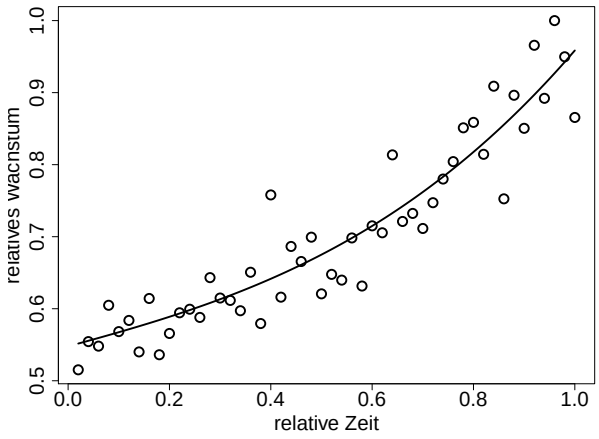

Abbildung 2: Plot der Beispieldaten mit Regressionsgerade

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 punktweisen und gleichmäßigen Konfidenzbändern wird erklärt. Die folgende Definition orientiert sich an [Geo09, S. 229].

Definition 1.2.1. Sei $(\chi, \mathscr{F}, \mathbb{P}_\vartheta : \vartheta \in \Theta)$ ein stochastisches Modell, Σ eine beliebige Menge, $\tau : \Theta \to \Sigma$ eine zu ermittelnde Kenngröße für den Parameter, und $0 < \alpha < 1$. Dabei bezeichnet $\mathbb{P}(...)$ ein Wahrscheinlichkeitsmaß.
Eine Abbildung $C : \chi \to \mathscr{P}(\Sigma)$, 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 - α), 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 $C(x) = \{y \in \mathbb{R}^n : x'\hat{\beta} - c\ \hat{\sigma}\sqrt{x'(X'X)^{-1}x} \leq y \leq x'\hat{\beta} + c\ \hat{\sigma}\sqrt{x'(X'X)^{-1}x}\}$ ist. Dabei ist c der sogenannte kritische Wert.
Für den Rest der Ausarbeitung wird dies durch

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

abgekürzt. Eine andere mögliche Schreibweise ist

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

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 betrachten.

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

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

gilt. Dabei ist A ein Bereich, dem aus bestimmten Gründen besonders interessiert 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] benutzten wir diese Form von Konfidenzbändern.

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

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

Dabei ist $f^{\alpha}_{p+1;n-p-1}$ das α-Quantil der F-Verteilung mit Parametern $p+1$ und $n-p-1$.
Dieser Satz eröffnet uns eine einfache Möglichkeit, Konfidenzbänder auf $\mathbb{R}^p$ zu bestimmen. Da das Konfidenzband in Satz 1.3.1 von hyperbolischer Form ist, betrachteten 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

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

mit

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

Beweis.

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

□

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 das 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. $\hat{\sigma}^2 \sim \frac{\sigma^2}{n-p-1}\chi^2_{n-p-1} = \frac{\sigma^2}{v}\chi^2_{n-p-1}$. Dabei ist $n - p - 1 :=$ v, was auch in Zukunft so beibehalten wird. Weiterhin ist χ^2_v 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 $e \sim \mathcal{N}_n(0, \sigma^2)$ so ist $Y = X\beta + e \sim \mathcal{N}_n(X\beta, \sigma^2 I)$. 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 reicht den Erwartungswert und die Kovarianz zu bestimmen.

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

und

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

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:

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

und

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

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

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

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 $e = Y - X\beta \sim \mathcal{N}_n(0, \sigma^2 I)$. Sei $z = Te$ mit $z \sim \mathcal{N}_n(0, \sigma^2 I)$. Solch ein T existiert, da T orthogonal ist.
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

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

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

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

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.

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

□

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

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

gilt.
Sei P die eindeutige Wurzel aus $(X'X)^{-1}$ und somit $(X'X)^{-1} = P^2$. Weiterhin definiere N=$P^{-1}(\hat{\beta}-\beta)/\sigma$. Dann hat N eine multivariate Normalverteilung $\mathcal{N}_{p+1}(0, I)$.

Damit gilt

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

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 folgenden 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. Als kritischen Parameter für das Konfidenzband hat sich 2.526154 ergeben.

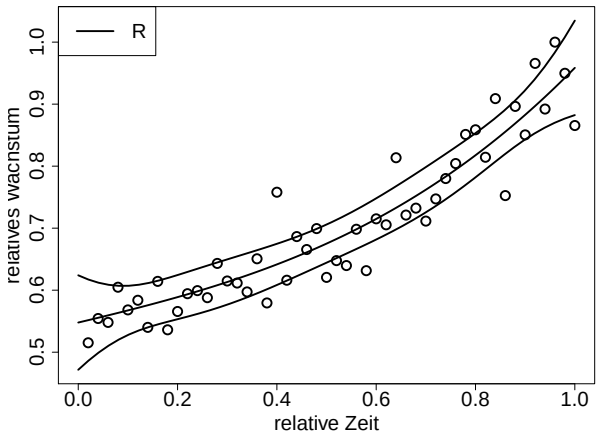

Abbildung 3: Weiterführung des Beispiels durch Konfidenzband auf ganz $\mathbb{R}^p$

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 wurde als Grundmodell ein multiples lineares, homoskedastisches Regressionsmodell zugrunde gelegt. 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 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 beziehungsweise mit einem AR(1) transformierten OLS-Schätzer berechnet. Dann wurde gezählt, wie oft das wahre Modell in dem Konfidenzband liegt.
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, Designmatrizten und weiter 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 Konfidenzbänder bei denen Simulationen durchgeführt werden müssen beträgt ungefähr 15 Minuten bei den empfohlenen 5.000 Wiederholungen.
Wie genau bei einem AR(1) Modell OLS-Schätzer bestimmt werden können wird in dem Kapitel 4 erläutert.
Die Ergebnisse sind in der folgenden Tabelle zusammengestellt:

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

Die erste Spalte gibt die Überdeckungswahrscheinlichkeit bei dem homoskedastischen Modell an, 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. Dies liegt daran, 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 weiter Zeilen mit den anderen Methoden Konfidenzbänder zu erzeugen erweitert, nachdem die neuen Methoden eingeführt sind. Dabei wird die bereits eingeführte Methode Konfidenzbänder auf ganz $\mathbb{R}^p$ zu erzeugen, mit R bezeichnet.

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 einen unabhängige Variable x_1 haben vereinfacht sich das Konfidenzband 1.3 zu

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

Dabei und für diesen Abschnitt gelten die folgenden Definitionen

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

Aus [Liu64, S. 19-20] hat man folgende Ergebnisse

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

dabei ist

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

ein 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

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

Man muss also

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

nach c lösen und kann dann das Konfidenzband (7) bestimmen. Diese Berechnung wird hier nicht durchgeführt.

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

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

Außerdem definieren wir $x_0 = (x_1,\ldots,x_p)'$.
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 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

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

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

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

gilt.
Es ist aus Satz 1.3.2 aus dem vorherigen Abschnitt bekannt, dass

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

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.
Dies folgt aus Satz 1.3.3. Dort wurde gezeigt, dass $\hat{\beta} \sim \mathcal{N}_{p+1}(\beta, \sigma^2(X'X)^{-1})$ und $\hat{\sigma} \sim \sigma^2/(n-p-1) \cdot \chi^2_{n-p-1}$. 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 $i = 1, \ldots, p$ 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

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

wobei $P = (p_0, \ldots, p_p)$.
Allerdings ist es für ein allgemeines $p \geq 1$ sehr schwierig, eine genaue Formel für die Verteilung von S explizit zu bestimmen, 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 $p_0 + c_1 p_1 + \ldots + c_p p_p$ aufgespannt wird, wobei c_i entweder a_i oder b_i für $i = 1, \ldots, p$ 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 $N \sim \mathcal{N}_{p+1}(0, I)$ und $\hat{\sigma}/\sigma \sim \sqrt{\chi^2_v / v}$ 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 R mal, so erhält man $S_1, \ldots, S_R$. Dann ist c das $1 - \alpha$ Quantil von $S_1, \ldots, S_R$.
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 $P^{-1}v = \lambda x = (\lambda, \lambda x_1, \ldots, \lambda x_p)'$ 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 $j = 1, \ldots p$ oder gleichwertig

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

Diese Beschränkungen kann man zusammenfassen zu

$$Av \leq 0$$

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

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

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 dem 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 $A = \{x_0 \in \mathbb{R}^3 : x_i \in [0,1] \text{ für alle } i = 1,2,3\}$.

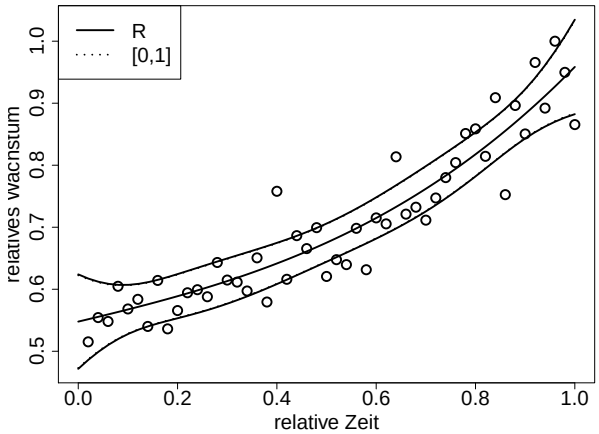

Abbildung 4: Weiterführung des Beispiels durch Konfidenzband auf [0,1]

Der kritischen Wert für das Konfidenzband auf ganz $\mathbb{R}^3$, der bereits in Abschnitt 1.3 berechnet wurde war 2.526154. Für das Konfidenzband auf A ergibt sich 2.074654. Dieser Wert ist kleiner als der bisherige, was auch zu erwarten war.

Dies sieht man auch daran, dass das Konfidenzband auf A schmaler ist, als das Konfidenzband auf ganz $\mathbb{R}^3$.
Da bei der Berechnung des kritischen Wertes für das Konfidenzband auf A ein Simulation verwendet wird und Reproduktivität gewährleisten sein soll, wird an dieser Stelle die Seed 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 wird mit Minmax bezeichnet. Auch diese Überdeckungswahrscheinlichkeiten liegen über den geforderten 0.95.

×	unabhängig	AR bekannt	AR
Konfidenzband auf ganz $\mathbb{R}^p$	1.00	0.94	1.00
Konfidenzband auf einem Polyeder	.99	0.93	1.00

1.6 Konfidenzbänder auf 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 $e \sim \mathcal{N}_n(0, I_n\sigma^2)$ 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. In dem 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 $\tilde{x}_i = (1, x, x^2, \ldots, x^p) \in \mathbb{R}^{p+1}, i = 1, \ldots, n$ ist.
Unsere Designmatrix ist also zeilenweise von Polynomgestalt.
Dies ist genau der Fall, den wir in dem Beispiel und den Simulationen, das am Ende der Abschnitte steht, betrachtet haben.
Das Ziel ist wieder Konfidenzbänder der Form

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

zu bestimmen. Dann können wir sie direkt mit den Konfidenzbändern aus den Abschnitten 1.3 und 1.5 vergleichen, indem wir die kritischen Konstanten c vergleichen.
Dabei ist die kritische Konstante c wieder so zu bestimmen, dass

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

gilt.
Es können allerdings nicht direkt die Ergebnisse aus dem letzten Abschnitt benutzen werden. Dies liegt daran, dass es folgende ungünstige Eigenschaften, die [Liu64, S. 180] entnommen sind, gibt:

1. Angenommen es seien zwei 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. Dies liegt daran, dass das Konfidenzband nur auf einer Teilmenge

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

   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, hatten wir das Konfidenzniveau von dem 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. Dies liegt daran, dass $\tilde{x}= (1, x, x^2, \ldots, x^p)$ 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 $S_1, \ldots, S_r$ das $1 - \alpha$ Quantil benutzen.
Um S für vorgegebenes 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.
Betrachtet man die Definition von $K_{2h}(T, (X'X)^{-1}, (a, b))$, ist klar, dass wir ein Supremum suchen. Das Supremum kann entweder an a oder an b oder den stationären Punkten der Funktion

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

angenommen werden, da g_h eine glatte Funktion auf A ist.
Leitet man g_h ab, erhält man

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

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

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

Um diese Wurzeln zu bestimmen, wurde auf das Intervall $[0, 1]$ ein Grid aus 100Punkten gelegt. Dann wurde der Wert von *g.prime* auf den Punkten dieses Grids bestimmt und gespeichert.
Dabei entspricht *g.prime* der Funktion h, die für die Zwecke der Berechnung die Rolle von $\frac{dg_h(x)}{dx}$ 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 wurde 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 ein Wurzel findet und nicht alle Wurzeln in dem gegebenen Intervall.
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 $\hat{\sigma}/\sigma \sim \sqrt{\chi^2_v/v}$ 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 Funtionswerte.
   (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 von h, bezeichnet mit $r_1, \ldots, r_n$, gefunden, berechnet man den Funktionswert an diesen Stellen.
5. Damit ist $S = \max\{g(a), g(b), g(r_1), \ldots g(r_n)\}$

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

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

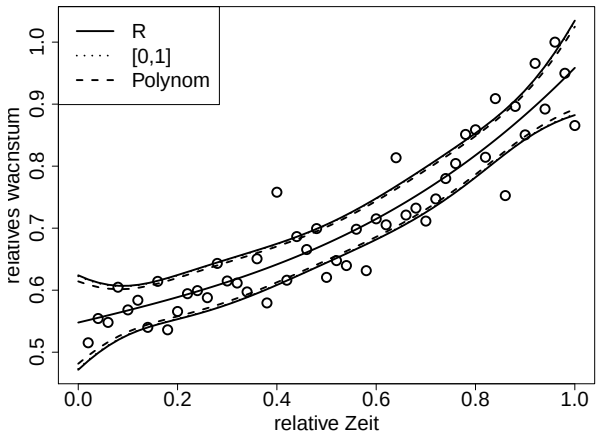

Abbildung 5: Weiterführung des Beispiels durch Konfidenzband auf [0,1] für Polynome

Die kritischen Werte für das Konfidenzband auf ganz $\mathbb{R}^3$ und für das Konfidenzband auf [0,1] wurde in den vorherigen Abschnitten bereits berechnet. Es ergaben sich die Werte 2.526154 und 2.074654. Berechnet man den kritischen Wert für das Konfidenzband auf [0,1], welches die Polynomgestalt berücksichtigt, erhält man den Wert 2.706617. Dieser Wert ist noch kleiner als der Wert von 2.074654, was auch zu erwarten war.
Als seed wurde weiterhin 4 benutzt.

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

$\times$	unabhängig	AR bekannt	AR
Konfidenzband auf ganz $\mathbb{R}^p$	1.00	0.94	1.00
Konfidenzband auf einem Polyeder	.99	0.93	1.00
Konfidenzband auf einem Polyeder für Polynome	0.93	0.86	0.99

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

$$Y_i = X_i\beta_i + e_i \tag{10}$$

mit $i = 1, 2$ gegeben.
Dabei ist $Y_i = (Y_{i,1}, \ldots, Y_{i,n_i})$ ein für $i = 1, 2$ ein Vektor aus zufälligen Beobachtungen der die abhängigen Daten darstellt.
Weiterhin sei X_i für $i = 1, 2$ eine $n_i \times (p+1)$ Designmatrix mit vollem Zeilenrang und festem Design. Dabei ist die j-te Zeile von X_1 mit der j-ten Zeile von X_2 für $j = 1, \ldots, \min(n_1, n_2)$ identisch. Diese Bedingung bedeutet, dass beide Designmatrizen für $j = 1, \ldots, \min(n_1, n_2)$ dasselbe, feste Design haben.
Außerdem sind β_i zwei $(p+1) \times 1$ Koeffizientenvektoren, die den funktionalen Zusammenhang zwischen Y_i und X_i darstellen.
In diesem Kapitel ist $e_i \sim \mathcal{N}_{n_i}(0, \sigma^2 I_{n_i})$. 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, ob

$$H_0 : \beta_1 = \beta_2 \textbf{ vs. } H_1 : \beta_1 \neq \beta_2 \tag{11}$$

Dabei bezeichnet H_0 die Nullhypothese und H_1 die Alternativhypothese.
Dazu ist es entweder möglich einen Test zu konstruieren oder Konfidenzbänder zu benutzen. Im nächsten Abschnitt wird einen Test konstruieren und im übernächsten Abschnitt 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 die Frage beantworten, ob sowohl für Männer, als auch für Frauen, der Zusammenhang zwischen zum Beispiel 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 im Kapitel 1 vorgestellt wurden, einen Unterschied macht, ein polynomiales Modell vom Grad Drei oder vom Grad Vier zugrunde zu legen.
In der folgenden Graphik 6 sind die beiden Modelle mit den Daten eingezeichnet.

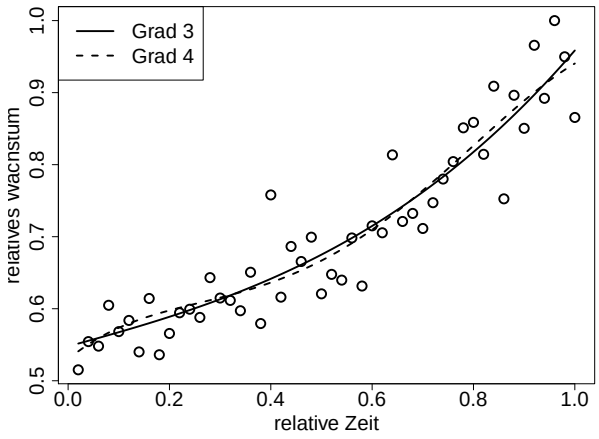

Abbildung 6: Regression von Grad Drei und Grad Vier im Vergleich

2.1 F-Test

Dieser Abschnitt beruht auf [Liu64, S. 9-15] ,[Liu64, S. 114-115] und [DS98].
Zuerst betrachten wir den Test, zu prüfen, ob der Regressionskoeffizient β eine lineare Einschränkung $Z\beta = d$ erfüllt. Dabei ist Z eine gegebene $r \times (p+1)$ Matrix mit vollem Zeilenrank r mit $1 \leq r \leq p+1$ und d ist ein vorgegebener Vektor in $\mathbb{R}^r$. Wir wollen testen, ob

$$H_0 : Z\beta = d \textbf{ vs. } H_a : Z\beta \neq d \tag{12}$$

Von Interesse ist, ob der Parameter in einem gegebenen Regressionsmodell eine lineare Einschränkung erfüllt.
Dazu wird zuerst der OLS-Schätzer $\hat{\beta}_Z$ unter der Bedingung, dass $Z\beta = d$ 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 dazu nachfolgender 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. [Liu64, S. 13] □

Bedeutung hat im Weiteren der folgende Satz:

Satz 2.1.2. Unter den oben genannten Bedingungen gilt

1. $\|X\hat{\beta} - X\hat{\beta}_Z\|^2 \sim \sigma^2\chi^2_r(\delta)$ mit nicht zentralem Parameter
   $\delta = \|X\beta - XE(\hat{\beta}_Z)\|^2/\sigma^2 = (\beta - E(\hat{\beta}_Z))'X'X(\beta - E(\hat{\beta}_Z))/\sigma^2$
2. $\|Y - X\hat{\beta}_Z\|^2 \sim \sigma^2\chi^2_{n-(p+1)+r}(\delta)$ mit δ wie in 1.
3. $\|X\hat{\beta} - X\hat{\beta}_Z\|^2$ und $\|Y - X\hat{\beta}_Z\|^2$ sind unabhängig.
4. $$\frac{\|X\hat{\beta}_Z - X\hat{\beta}\|^2/r}{\|Y - X\hat{\beta}\|^2/(n-p-1)} \sim f_{p+1,n-p-1}$$

Beweis. [Liu64, S. 11] □

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

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

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 anwendet. Dabei wird sich an [Liu64, S. 114] orientiert.
Betrachten wir wieder das Modell 10 aus der Einleitung. Außerdem wird wieder das Testproblem (11) betrachtet:

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

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

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

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

$$Y = x'c_1 + zx'c_2 + e \tag{14}$$

mit $x = (1, x_1, \ldots, x_p)'$, $c_1 = \beta_2$ 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

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

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 :

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

Dabei ist $\widehat{\sigma^2}$ die mittlere Varianz des Modells (14), $\hat{\beta}_i = (X_i' X_i)^{-1} X_i Y$ und 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.
Für den Nenner betrachtet man

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

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 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

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

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 $D = (X_1'X_1)^{-1} + (X_2'X_2)^{-1}$ wie im ersten Kapitel und definiere $T = P^{-1}(\hat{\beta}_2 - \beta_2 - \hat{\beta}_1 + \beta_1)/\hat{\sigma}$, 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.
Zum Schluss dieses Abschnittes zeigen wir noch ein Resultat, dass zeigt, dass es keinen Unterschied macht, ob ein Test durchgeführt oder ein Konfidenzband auf ganz $\mathbb{R}^p$ benutzt wird.
Dabei handelt es sich um den Test

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

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]

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

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

Satz 2.2.1. Der Test (17) und Konfidenzbandmethode (1.3.1), das ist das Konfidenzband auf ganz $\mathbb{R}^p$, 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 das man β mit β_0 ersetzt, erhält man

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

Dies entspricht der Teststatistik (17), also sind die beiden Test äquivalent. □

Dies bedeutet, dass es von einem Interferenzstandpunkt her immer besser ist, ein Konfidenzband auf einem Intervall zum Vergleich von zwei Modellen zu verwenden.
Weiterhin wurde bereits im ersten Kapitel gezeigt, dass Konfidenzbänder auf einem Intervall und erst recht Konfidenzbänder für Polynome auf einem Intervall eine bessere Inferenz zulassen als Konfidenzbänder auf ganz $\mathbb{R}^p$. Also ist es auch in dieser Hinsicht von Vorteil, Konfidenzbänder auf Intervallen zu verwenden.

Beispiel 2.2.2. Es wird jetzt das Beispiel aus den vorherigen Abschnitten fortgeführt, indem die Differenz der Regressionsmodelle geplottet wird und ein Konfidenzbänder auf A für Polynome einfügt wird:

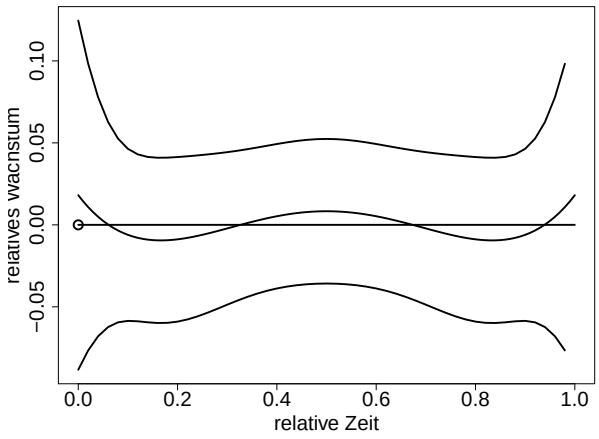

Abbildung 7: Differenz des Beispiels mit Konfidenzband auf A für Polynome

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 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

$$Y = X\beta + e \tag{18}$$

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 $\beta_1 = (\beta_0, \ldots, \beta_{p-k})$ und $\beta_2 = (\beta_{p-1+1}, \ldots, \beta_p)$ für $1 \leq k \leq p$ eine gegebene natürliche Zahl.
Auf die gleiche Art und Weise kann man die Spalten x von X in $x_1 = (1, x_1, \ldots, x_{p-k})$ und $x_2 = (1, x_{p-k+1}, \ldots, x_p)$ unterteilen.
Ist $\beta_2 = 0$ so haben die unabhängigen Variablen $x_{p-k+1}, \ldots, x_p$ keinen Einfluss auf Y und man kann das Modell 18 auf

$$Y = X_1\beta_1 + e \tag{19}$$

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

3.1 F-Fest

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

$$H_0 : \beta_2 = 0 \textbf{ vs. } H_1 : \beta_2 \neq 0 \tag{20}$$

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

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

mit $v = n - p - 1$, $\tilde{I}$ der $k \times (p + 1)$ Matrix die durch $\tilde{I} = (0, I_k)$ gegeben ist und $f^{\alpha}_{k,v}$ dem α Quantil der f Verteilung mit k und v Freiheitsgraden.
Dabei ist der MS residual von Modell (18) gegeben durch

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

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. Dies 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 gibt genau solch ein Theorem an

Satz 3.1.1. Es gilt

$$\mathbb{P}(x_2'\beta_2 \in x_2'\hat{\beta}_2 \pm \sqrt{kf_{k,n-p-1}^{\alpha}}\hat{\sigma}\sqrt{x_2'Vx_2}) = 1-\alpha \tag{21}$$

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 Information. 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 Teil eines Regressionsmodells auf einem Intervall überprüfen

Dieser Abschnitt orientiert sich an [Liu64, S. 102-105].
Wir beschränken uns auf die rechteckige Region

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

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

$$x_2'\beta_2 \in x_2'\hat{\beta}_2 \pm c\hat{\sigma}\sqrt{x_2'Vx_2} \text{ für alle } x_2 \in A_2 \tag{22}$$

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

$$\begin{aligned} S &= \sup_{x_2 \in A_2} \frac{|x_2'(\hat{\beta}_2 - \beta_2)|}{\hat{\sigma}\sqrt{x_2'Vx_2}} \\ &= \sup_{x_2 \in A_2} \frac{|(Wx_2)'W^{-1}(\hat{\beta}_2 - \beta_2)/\hat{\sigma}|}{\sqrt{(Wx_2)'(Wx_2)}} \\ &= \sup_{v \in C(W,A_2)} \frac{|v'T_2|}{\|v\|} \end{aligned}$$

wobei $C(W, A_2) = \{\lambda W x_2 = \lambda(x_{p-k+1}w_1 + \ldots + x_p w_k) : \lambda > 0 \text{ and } x_2 \in A_2\}$ 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 $\hat{\sigma}/\sigma \sim \sqrt{\chi^2_v/v}$
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 $S_1, \ldots, S_R$.
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 letzten Kapitel Datenbeschreibung und Resultate 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 multilinearen Regression und einer Regression mit AR(1)-Kovarianzmatrix durchgeführt:

×	unabhängig	AR bekannt	AR
Konfidenzband auf ganz $\mathbb{R}^p$	1.00	0.94	1.00
Konfidenzband auf einem Polyeder	.99	0.93	1.00
Konfidenzband auf einem Polyeder für Polynome	0.93	0.86	0.99
Teil des Modells auf $\mathbb{R}^p$ überprüfen	0.99	-	1.00

3.3 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

$$\text{reject } H_0 \leftrightarrow \frac{\hat{\beta}_2{}' V^{-1} \hat{\beta}_2 / k}{\widehat{\sigma^2}} > f^{\alpha}_{k,v} \tag{23}$$

mit $\hat{\beta}_2$ ist der Schätzer für $\beta_2 = (b_{p-k+1}, \ldots b_p)'$ welcher die letzten k Komponenten ist und 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, dass heißt $V = W^2$.
Analog zum letzten Abschnitt kann wieder gezeigt werden, dass dieser Test dem Konfidenzband

$$x_2'\beta_2 \in x_2'\hat{\beta}_2 \pm \sqrt{k f^{\alpha}_{k,v}}\hat{\sigma}\sqrt{x_2' V x_2} \text{ für alle } x_2 \in \mathbb{R}^k \tag{24}$$

entspricht. Das heißt liegt die Nullfunktion nicht vollständig in dem 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, dass die polynomstruktur berücksichtigt.
Dazu betrachtet man, dass das Konfidenzniveau von (24) für c anstatt $\sqrt{kf_{k,v}^{\alpha}}$ durch $\mathbb{P}(S \leq c)$ gegeben ist. Dabei ist

$$\begin{aligned} S &= \sup_{x \in A} \frac{|\tilde{x_2}'(\hat{\beta}_2 - \beta_2)}{\hat{\sigma}\sqrt{\tilde{x_2'} V \tilde{x_2}}} \\ &= \sup_{x \in A} \frac{|(1, \ldots, x^{k-1})(\hat{\beta}_2 - \beta_2}{\hat{\sigma}\sqrt{(1, \ldots, x^{k-1}) V (1, \ldots x^{k-1})'}} \end{aligned}$$

mit $\tilde{x_2} = x^{p-k+1}(1, x, \ldots, x^{k-1})$.
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 $\mathbb{R}^p$	1.00	0.94	1.00
Konfidenzband auf einem Polyeder	.99	0.93	1.00
Konfidenzband auf einem Polyeder für Polynome	0.93	0.86	0.99
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 $Y = (y_1, \ldots, y_n) \in \mathbb{R}^n$, $\beta \in \mathbb{R}^{p+1}$ 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 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 betrachten wir die konkrete Funktionsweise dieser Funktionen anhand des Buches [PB00],
Im letzten Abschnitt dieses Kapitels betrachten wir, 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 das Modell (25) auf den Fall von abhängigen Daten verallgemeinert. Daten sind abhängig, wenn $\mathrm{Cov}(y_i, y_j) \neq 0$ 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 ist eine Menge von Messungen z_t, an einem bestimmten Zeitpunkten t.
Bei einer diskreten Zeitreihe ist $t = 1, \ldots, T$ mit $T \in \mathbb{N}$ fest.

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

Definition 4.1.2. 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

$$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q} \tag{26}$$

mit $(Z_t) \sim \mathrm{WN}(0, \sigma^2)$. Dabei sagt man, dass der Prozess $(Z_t) \sim \mathrm{WN}(0, \sigma^2)$ 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

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

schreiben. Dabei sind ζ und θ die Polynome

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

und

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

und B ist der Rückwärtsshift-Operator definiert durch

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

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 4.1.3. 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

$$y_t = \phi y_{t-1} + e_t \tag{27}$$

Dabei ist ϕ ein unbekannter Parameter und $\mathbb{E}(e_t) = 0$, $\mathbb{E}(e_t^2) = \sigma^2 < \infty$. Für den Fall das $\phi \in (-1, 1)$ 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 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:

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

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

$$\Upsilon = \begin{bmatrix}
1 & \phi & \phi^2 & \cdots & \phi^{n-2} & \phi^{n-1} \\
\phi & 1 & \phi & \cdots & \phi^{n-3} & \phi^{n-2} \\
\phi^2 & \phi & 1 & \ddots & \vdots & \vdots \\
\vdots & \vdots & \ddots & \ddots & \phi & \phi^2 \\
\phi^{n-2} & \phi^{n-3} & \cdots & \phi & 1 & \phi \\
\phi^{n-1} & \phi^{n-2} & \cdots & \phi^2 & \phi & 1
\end{bmatrix} \tag{28}$$

Das heißt, das Modell ist gegeben durch

$$Y = X\beta + e \tag{29}$$

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

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

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 ML-Schätzung bestimmt. Dazu fixiert man zuerst ϕ und berechne 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-triangulären Zerlegungen. Sieh hierzu [PB00, S. 68-75].
Um den Wert von $\hat{\phi}$ für vorgegebene Daten zu bestimmen, wurde die Funktion *gls* aus dem Paket *nlme* benutzt. Bei der Benutzung von *gls* um einen Schätzer für ϕ zu bestimmen, muss der Wert *correlation=corAR1()* an *gls* übergeben werden.
Um dann bei gegebenem ϕ beziehungsweise $\hat{\phi}$ einen Schätzer für β und σ zu bestimmen, wird folgender Algorithmus verwendet:

1. Benutze gls 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 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

$$\hat{\beta} = (X^tX)^{-1}X^tY$$

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 ist genau das 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 dem Abschnitt 1.1 fortgeführt, indem eine Regression mit AR(1) für das Polynommodell mit Grad Drei durch geführt 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 Korrelatiosmatrix Υ von der Form

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

mit $\phi = 0.75$ ein fester Wert.
Man erhält als Schätzer für den AR(1) Parameter $\hat{\phi}$ = -0.938988, während der wahre Wert 0.75ist.
Als Schätzung für die Parameter erhält man (0.5142065 ; 0.4247473) und 0.2106673, während die wahren Werte (10,5,-4,7)und 1sind.
Zeichnet man dann dieses Regressionspolynom mit den Daten in eine Graphik erhält man

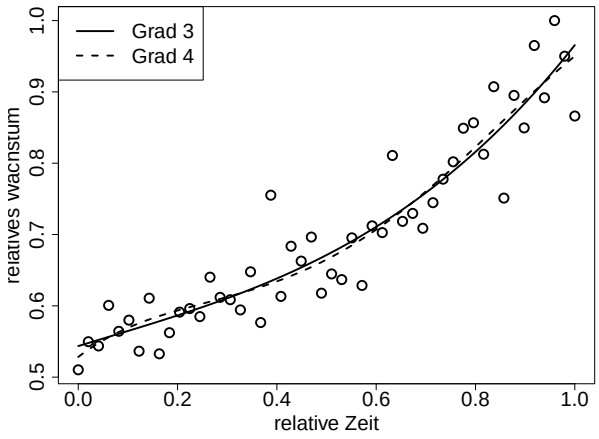

Abbildung 8: Regression für AR(1)

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 angewandt werden und ein Konfidenzband K^* bestimmt werden, sodass

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

Da allerdings nicht $x^{*'}\beta$ interessiert, sondern $x'\beta$, müssen noch beide Seiten der logischen Gleichheit 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.
Konkret wird die folgende Simulation verwendet, um konkret den Wert der kritischen Konstante c bei abhängigen Daten zu bestimmen:

1. Simuliere $N \sim \mathcal{N}_n(0, (\Upsilon^{-1/2}X'\Upsilon^{-1/2}X)^{-1}) = (0, (X^{*'}X^*)$ und $\hat{\sigma}/\sigma \sim \sqrt{\chi^2_v/v}$ mit $v = n - p - 1$.

2. Berechnung von $K_{2h}(T, (X^{*'}X^*)^{-1}, (a, b))$.

Die Berechnung von $K_{2h}(T, \cdot, (a, b))$ läuft analog zum 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 Eins, 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.

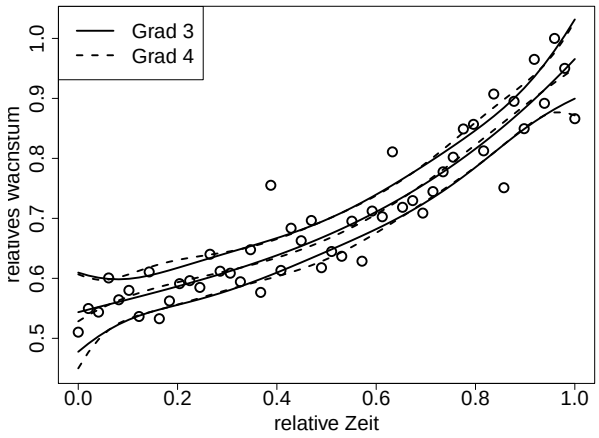

Abbildung 9: Weiterführung des Beispiels durch Konfidenzband auf A unter Berücksichtigung von Polynomstruktur und AR(1)-Struktur

Man erhält als kritischen Wert für das Konfidenzband auf ganz $\mathbb{R}^p$ 2.526154, für das Konfidenzband auf A 2.074654, für das Konfidenzband auf A für Polynome 2.706617und für das Konfidenzband auf A für Polynome das die AR(1)-Struktur berücksichtigt 1.957394.
Da es sich um eine Simulation handelt und Reproduktivität gewährleistet werden soll, wird die seed 4benutzt.
Vergleicht man wieder die Regression mit Grad Eins und mit Grad drei erhält man die folgende Graphik:

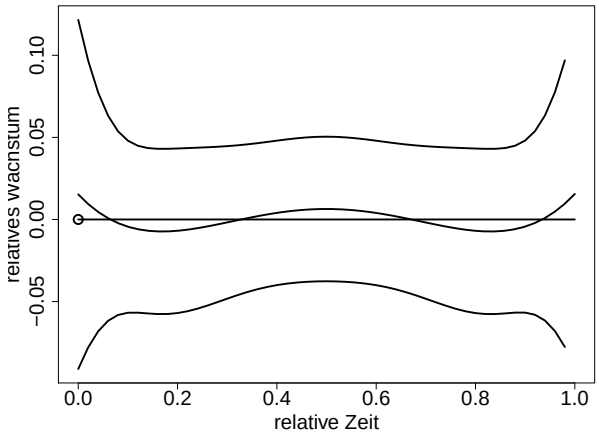

Abbildung 10: Vergleich Regression Grad Eins und Grad drei unter Berücksichtigung der Polynomstruktur und AR(1)-Struktur

Man sieht, dass das Konfidenzband weit von den Daten entfernt ist. Vermutlich liegt dies daran, dass $X'X$ in diesem Fall Eigenwerte nahe Null hat. Dadurch führt die Berechnung von $(X'X)^{-1}$ zu seltsamen Ergebnissen.

5 Datenbeschreibung und Resultate

In diesem Kapitel werden zuerst die Daten beschrieben, die als Motivation für die bisher erklärten Methoden dienten. 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. Im letzten Abschnitt werden 10 Kilopascal (kPa) und 30 kPa Daten verglichen. Die Daten entstammen der Arbeit von [ZR10].
Die zur Berechnung der Parameter, namentlich β, σ und der kritischen Werte $c_{\mathbb{R}}, c_A, c_{AP}$, und zum erzeugen der Graphiken verwendeten R Skripte können unter

https://github.com/fake1884/KBminmaxpoly

eingesehen werden. Eine Übersicht über die verwendeten Skripte befindet sich in der *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 $A \subset \mathbb{R}^p$ für ein polynomielles 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 Nullfuktion 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, 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 verschieden Datensätze. Einmal von Stammzellen die auf ein Oberfläche mit einer Härte von 10 kPa und einmal 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 die Größe gemessen. Da die Fläche der Stammzellen für jeden einzelne Stammzelle stark schwankt, wird der Mittelwert der Flächen gebildet.
Plottet man nun das Mittel der relative Wachstüme Y gegen die relative Zeit von Null bis Eins mit Schrittweite 145, also der Anzahl an Beobachtungen, entstehen folgende Graphiken (11) und (12) .

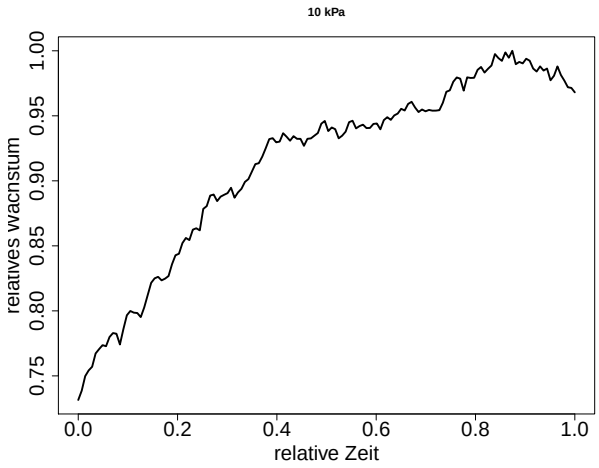

Abbildung 11: 10 kPa Daten

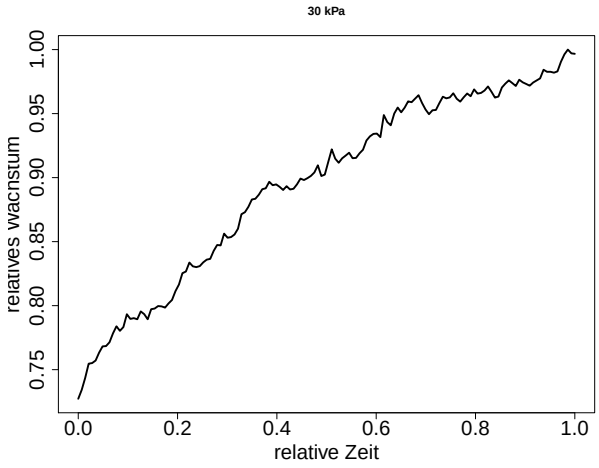

Abbildung 12: 30 kPa Daten

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

5.2 Vergleich verschiedener Konfidenzbänder

In diesem Abschnitt werden die drei Typen von Konfidenzbänder sowohl für 30kPa als auch für 10kPa 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 $\mathbb{E}(e_i) = \sigma_e^2 < \infty$
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

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

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

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

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

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

Als kritische Parameter erhält man $c_{\mathbb{R}} = 3.604075, c_A = 3.505922, c_{AP} = 2.902647$.
Setzt man dasselbe Modell wie oben für die 30kPa Daten für die 10kPa 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 schmaler.
Zeichnet man die Regressionsmodelle und die Konfidenzbänder jeweils in eine Graphik erhält man die beiden folgenden Graphiken (13) und (14).

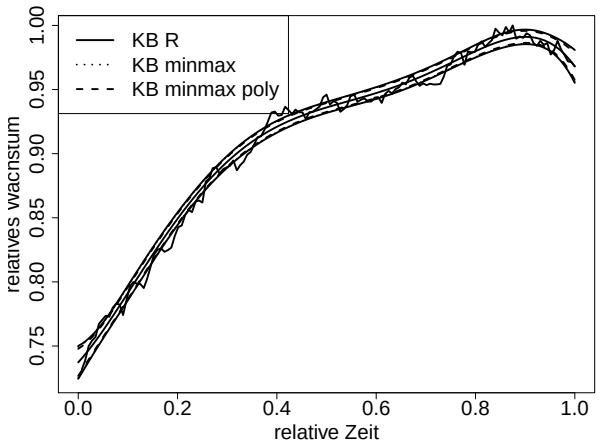

Abbildung 13: Vergleich von Konfidenzbändern 10 kPa

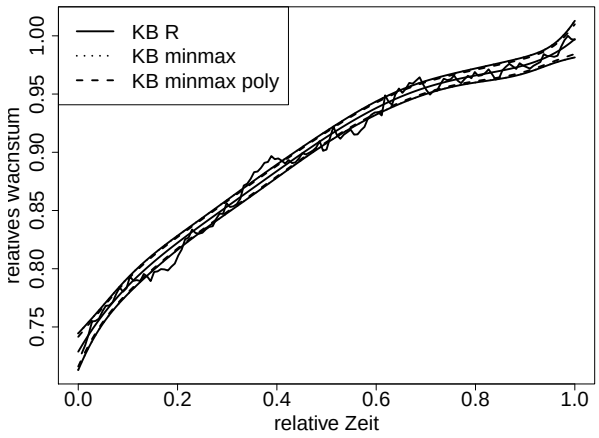

Abbildung 14: Vergleich von Konfidenzbändern 30 kPa

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 faktisch kein Unterschied erkennbar.

5.3 Vergleich von Polynomen mit verschiedenem Grad

Ziel dieses Abschnittes ist es, zu testen, ob Polynomregressionen von verschiedenem Grad für die beiden Stammzelldatensätze sich statistisch signifikant unterscheiden.
Dazu wird zuerst die Regression durchgeführt, bevor die Testhypothese formalisiert wird.
Als Regressionsmodell wird das aus Kapitel 4 Regressionsmodell

$$Y = X\beta + e$$

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 10kPa Daten die Werte:

$$\begin{aligned}
\hat{\beta}_4 &= (0.7314; 0.8401; -1.5883; 1.8877; -0.9029) \\
\hat{\sigma}_4 &= 0.4943004 \\
\hat{\phi}_4 &= 0.9999575
\end{aligned}$$

$$\begin{aligned}
\hat{\beta}_5 &= (0.7372; 0.4219; 1.8730; -7.6992; 9.7893; -4.1542) \\
\hat{\sigma}_5 &= 0.007545373 \\
\hat{\phi}_5 &= 0.8225374
\end{aligned}$$

$$\begin{aligned}
\hat{\beta}_6 &= (0.7352; 0.6265; -0.6151; 3.0831; -11.1962; 14.6223; -6.2906) \\
\hat{\sigma}_6 &= 0.007205863 \\
\hat{\phi}_6 &= 0.8065796
\end{aligned}$$

und für die 30kPa Daten die Werte:

$$\begin{aligned}
\hat{\beta}_4 &= 0.73110.5245 - 0.35460.03900.0546 \\
\hat{\sigma}_4 &= 0.009576459 \\
\hat{\phi}_4 &= 0.9090669
\end{aligned}$$

$$\begin{aligned}
\hat{\beta}_5 &= 0.72880.7345 - 2.15945.2097 - 5.90292.3864 \\
\hat{\sigma}_5 &= 0.009266163 \\
\hat{\phi}_5 &= 0.9039573
\end{aligned}$$

$$\begin{aligned}
\hat{\beta}_6 &= 0.72730.9495 - 4.766416.2343 - 26.756620.5211 - 5.9124 \\
\hat{\sigma}_6 &= 0.01673707 \\
\hat{\phi}_6 &= 0.9701881
\end{aligned}$$

Plottet man die drei Regressionsmodelle in je eine Graphik für die beiden Datensätze entstehen die beiden Graphiken (15) und (16):

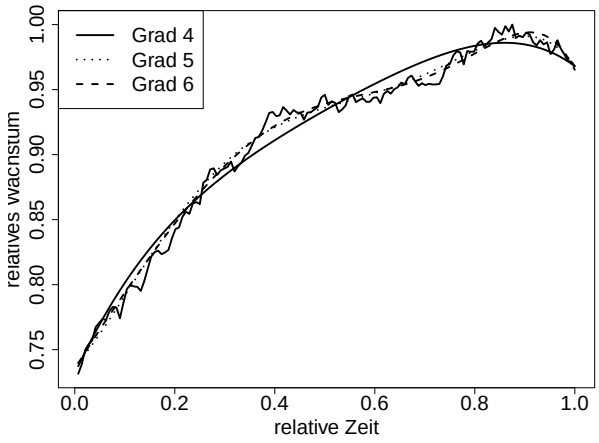

Abbildung 15: 10kPa Verschiedene Polynomgrade

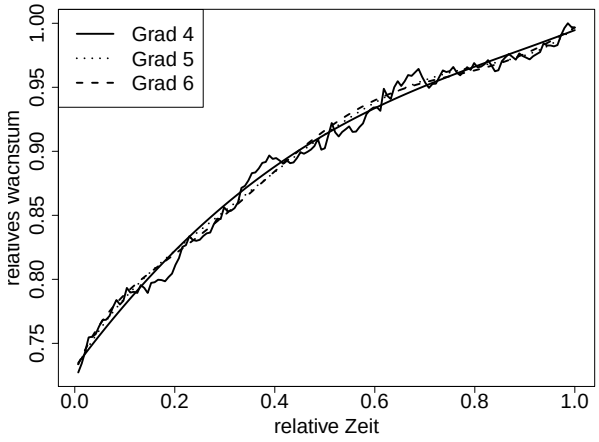

Abbildung 16: 30kPa Verschiedene Polynomgrade

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

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

Dazu benutzen wir die Konfidenzbandmethode aus Kapitel 2 beziehungsweise ihre Abwandlung in Abschnitt 4.3.
Zeichnen wir die Differenz der von $\hat{\beta}_4$ und $\hat{\beta}_5$ mit dem zugehörigen Konfidenzband in je eine Graphik für die beiden Datensätze erhalten wir die beiden folgenden Graphiken (17) und (18).

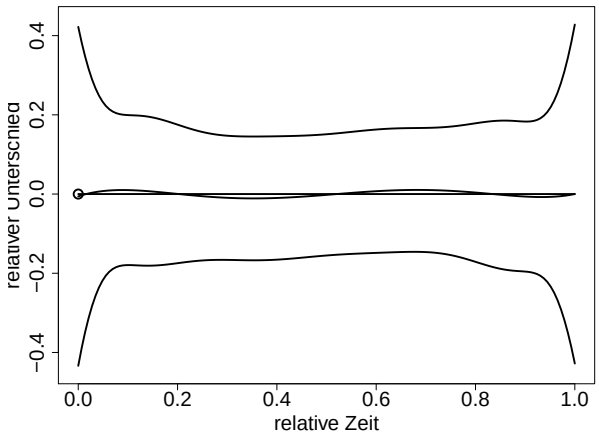

Abbildung 17: 10kPa Vergleich von Regressionsmodellen 4-5

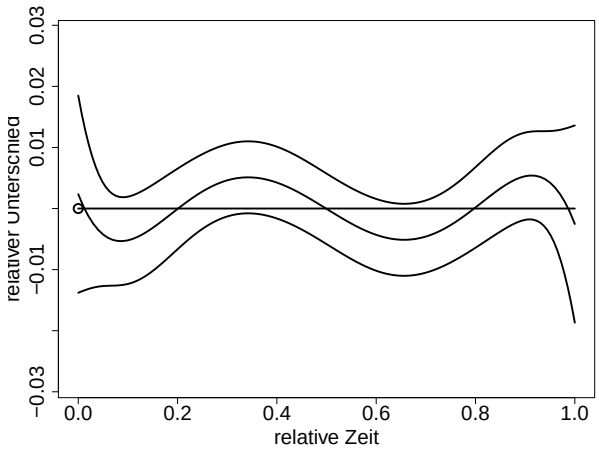

Abbildung 18: 30kPa Vergleich von Regressionsmodellen 4-5

Wie man sieht, ist in im Fall der 10kPa Daten die Nullfunktion eindeutig vollständig im Konfidenzband enthalten und wir können die Nullhypothese nicht ablehnen.
Bei den 30kPa Daten ist die Nullhypothese auch knapp ganz im Konfidenzband enthalten und wir können die Nullhypothese wieder nicht ablehnen.
In den folgenden Vier Graphiken testen wir die analogen Hypothesen

$$H_0 : \beta_4 = \beta_6 \quad \textbf{vs.}\ H_1 : \beta_4 \neq \beta_6$$
$$H_0 : \beta_5 = \beta_6 \quad \textbf{vs.}\ H_1 \beta_5 \neq \beta_6$$

für die 10kPa beziehungsweise 30kPa Daten. Für die erste Hypothese erhält man die Graphiken (19) und (20)

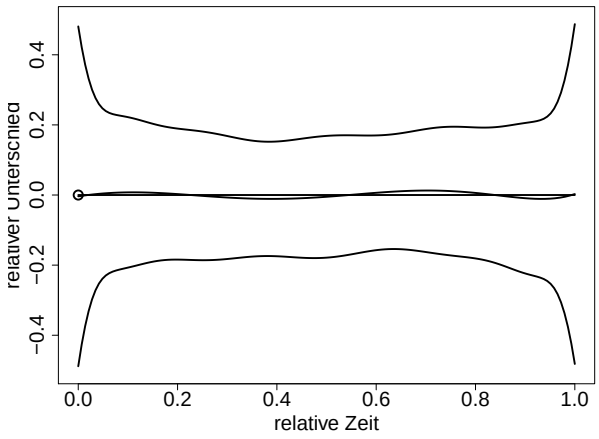

Abbildung 19: 10kPa Vergleich von Regressionsmodellen 4-6

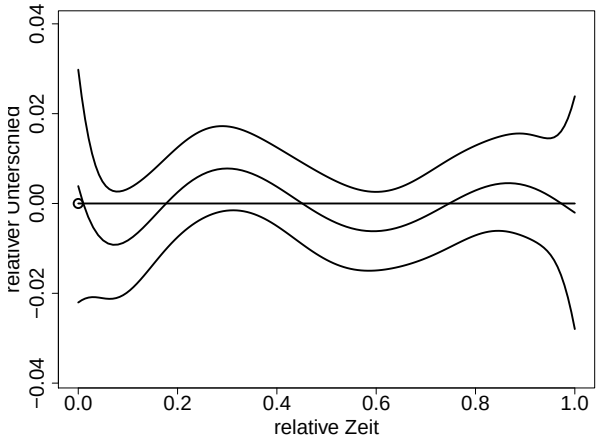

Abbildung 20: 30kPa Vergleich von Regressionsmodellen 4-6

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

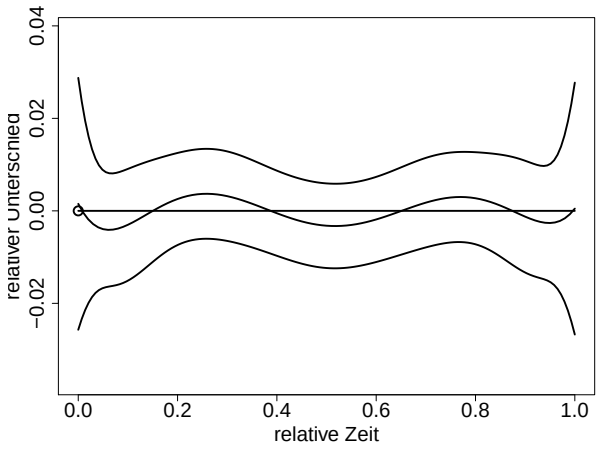

Abbildung 21: 10kPa Vergleich von Regressionsmodellen 5-6

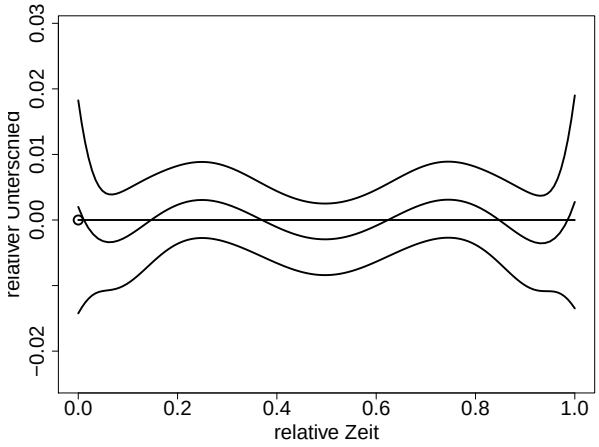

Abbildung 22: 30kPa Vergleich von Regressionsmodellen 5-6

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:

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

Dabei ist β_{10kPa} der Koeffizientenvektor einer Polynomregression vom Grad Fünf und β_{30kPa} der Koffizientenvektor 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, dass heißt wir gehen davon aus, dass beide Prozesse das gleiche ϕ und σ besitzen.
Ersteinmal plotten wir die beiden Datensätze mit den Regressionspolynomen in die selbe Graphik, um einen Überblick über ihre Ähnlichkeit zu bekommen. Dabei entsteht die folgende Graphik (23).

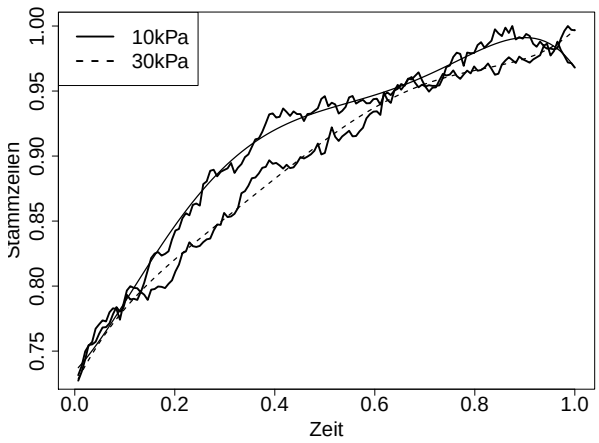

Abbildung 23: Vergleich von Regressionsmodellen

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

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

und für die 30kPa Daten:

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

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

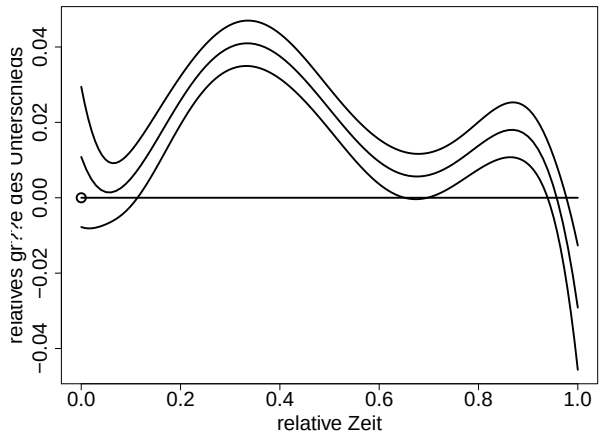

Abbildung 24: Vergleich von Regressionsmodellen

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 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 10kPa als auch für die 30kPa Stammzelldaten je zwei Polynomregressionen durch. Wir legen dabei wieder eine Zeitreihe von der AR(1) Art zugrunde.
Dabei erhalten wir für die 10kPa Daten die Werte:

$$
\begin{aligned}
\hat{\beta}_5 &= (0.7373; 0.4219; 1.8730; -7.6992; 9.7893; -4.1542) \\
\hat{\sigma}_5 &= 0.007545373 \\
\hat{\phi}_5 &= 0.8225374 \\
\hat{\beta}_6 &= (0.7352; 0.6265; -0.6151; 3.0831; -11.1962; 14.6223; -6.2906) \\
\hat{\sigma}_6 &= 0.007205863 \\
\hat{\phi}_6 &= 0.8065796
\end{aligned}
$$

und für die 30 kPa Daten die Werte:

$$
\begin{aligned}
\hat{\beta_5} &= (0.7288; 0.7345; -2.1594; 5.2097; -5.9029; 2.3864) \\
\hat{\sigma_5} &= 0.009266163 \\
\hat{\phi_5} &= 0.9039573 \\
\hat{\beta_6} &= (0.7273; 0.9495; -4.7664; 16.2343; -26.7566; 20.5211; -5.9124) \\
\hat{\sigma_6} &= 0.01673707 \\
\hat{\phi_6} &= 0.9701881
\end{aligned}
$$

Um die Hypothese angeben zu können definiere

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

Dann wollen wir die folgenden Test durchführen:

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

Dazu verwendet man die Methode aus Abschnitt 3.3. Wir verwerfen also die Nullhypothese, wenn die Konfidenzbänder $x_2' * 0$ nicht vollkommen enthalten.
Dabei ist in unserem Fall $x_2' \in \{0, 1/145, 2/145, \ldots, 144/145, 1$
In den folgenden Graphiken (25) und (26) ist $x_2'\beta_{5,5}$ mit zugehörigem Konfidenzband eingezeichnet:

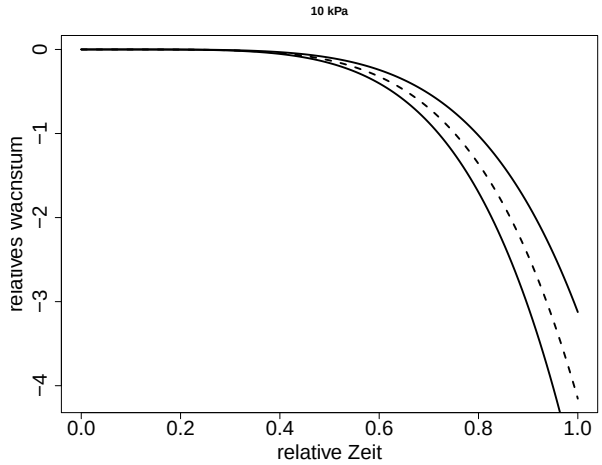

Abbildung 25: 10kPa Grad 5

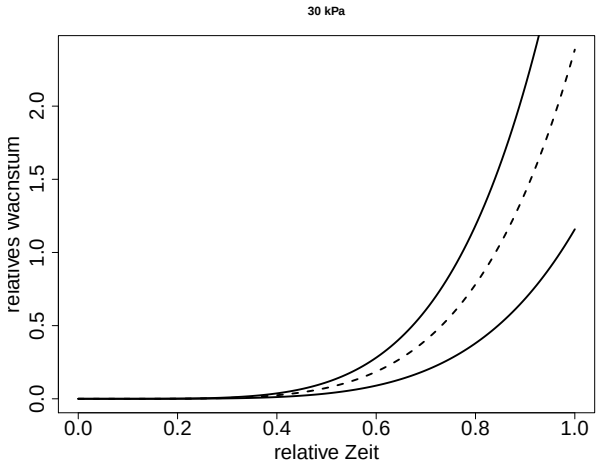

Abbildung 26: 30kPa Grad 5

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

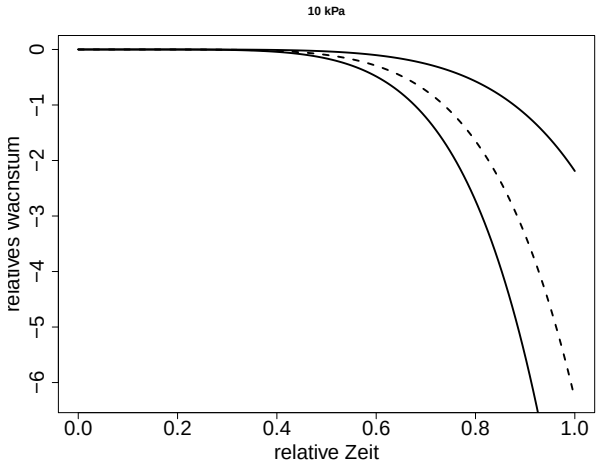

Abbildung 27: 10kPa Grad 6.1

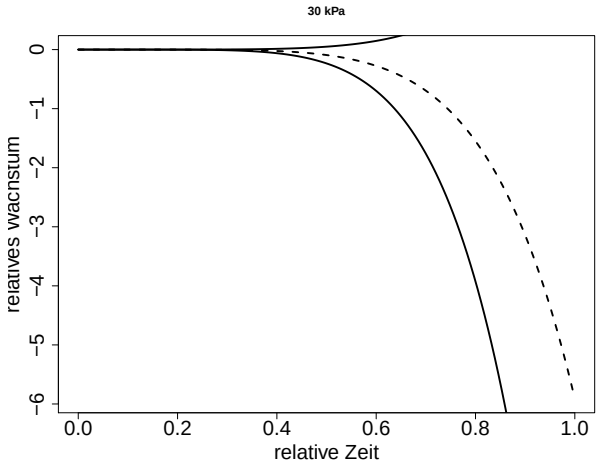

Abbildung 28: 30kPa Grad 6.1

Man sieht, dass für die 10kPa 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 $(\beta_{6,5}, \beta_{6,6}) = (0, 0)$. Man

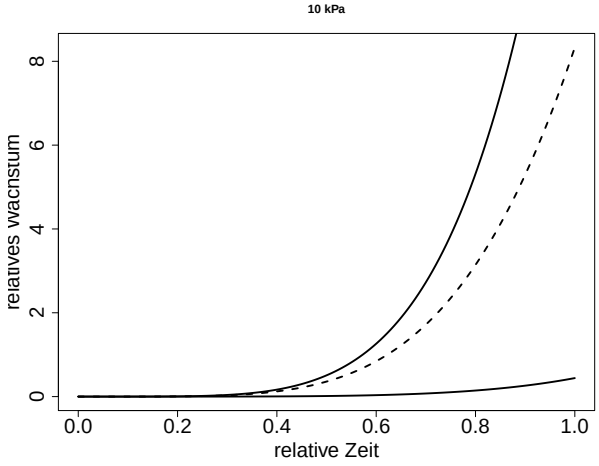

Abbildung 29: 10kPa Grad 6.2

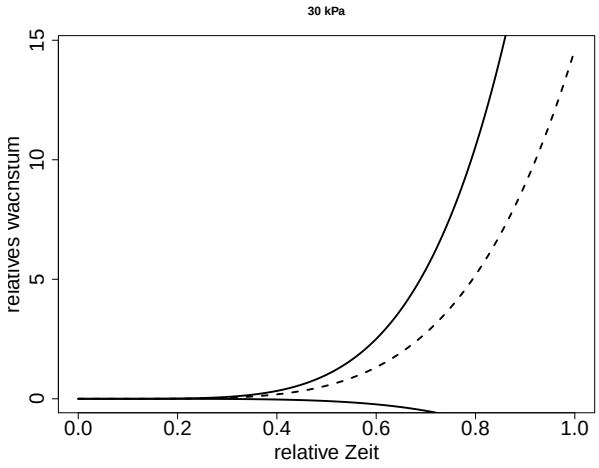

Abbildung 30: 30kPa Grad 6.2

Auch hier sieht man, dass für die 10kPa 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 $\beta_{6,6} \neq 0$.

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 Trancis 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