Statistik für Digital Humanities

Lineare Regression

Dr. Jochen Tiepmar

Institut für Informatik
Computational Humanities
Universität Leipzig

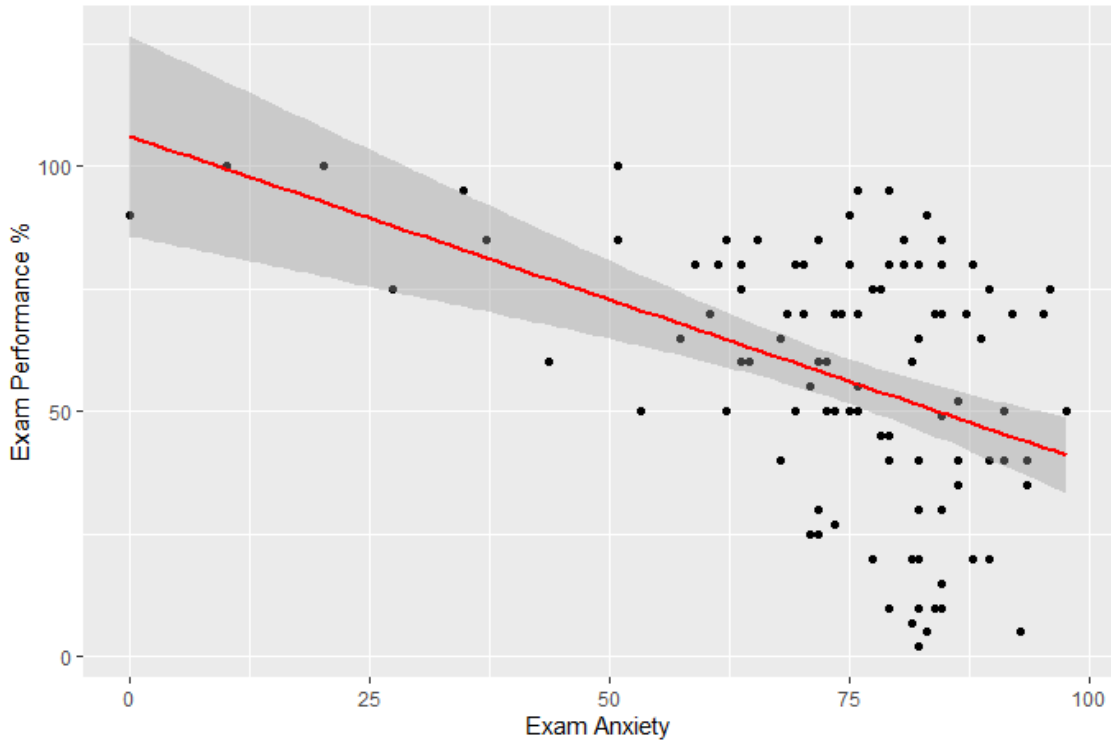
18. Mai 2020

[Letzte Aktualisierung: 17/05/2020, 22:15]

Regression

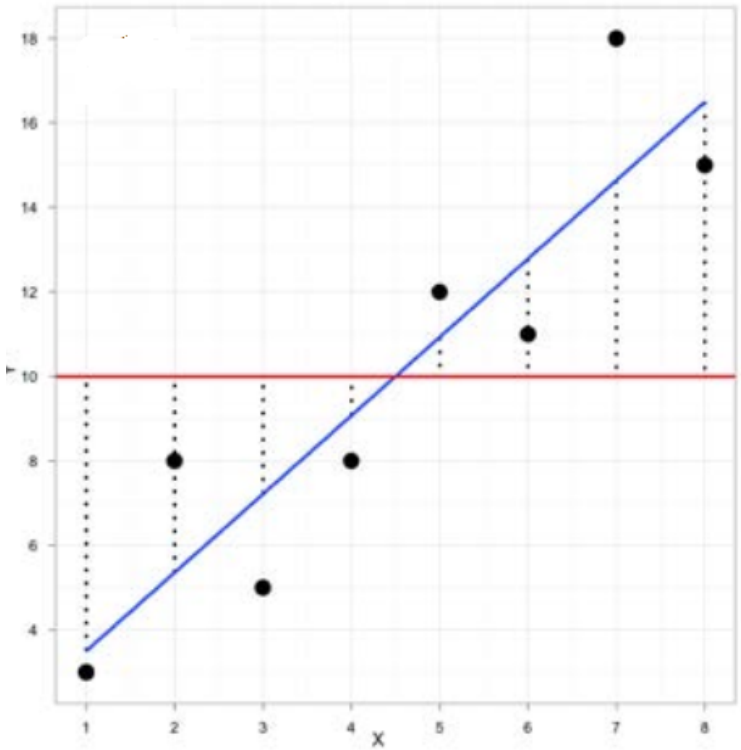
- Statistisches Modell zur Vorhersage einer abhängigen Variable auf Basis von unabhängigen Variablen
  - → Step 1: Modellfitting auf Daten
  - → Step 2: REGRESSION
  - → Step 3: Outcome für neuen Prädiktor errechnet
- Wie viel Angst haben Studierende 10, 5 oder 2 Minuten vor der Prüfung?
- Wie viele Personen werden zu einer öffentlichen jährlich wiederholten Veranstaltung erwartet?
- Wie viele Alben verkaufen wir, wenn wir x Euro für Werbung ausgeben?
- Einfache Regression
  - → 1 Prädiktor
- Multiple Regression
  - → Mehr als 1 Prädiktor

Regressionsgerade




```
data<-read.delim("Exam Anxiety.dat", header=TRUE)
graph<-ggplot(data, aes(Anxiety, Exam))
graph + geom_point(method="lm") + geom_smooth() +
  labs(x = "Exam Anxiety", y = "Exam Performance %")
```

Regressionsgerade vs Mittelwert



Oversimplified:
Im Grunde versuchen wir die Mittelwertgerade zu kippen um dann y in Abhängigkeit von x zu berechnen statt immer dieselben y für alle x

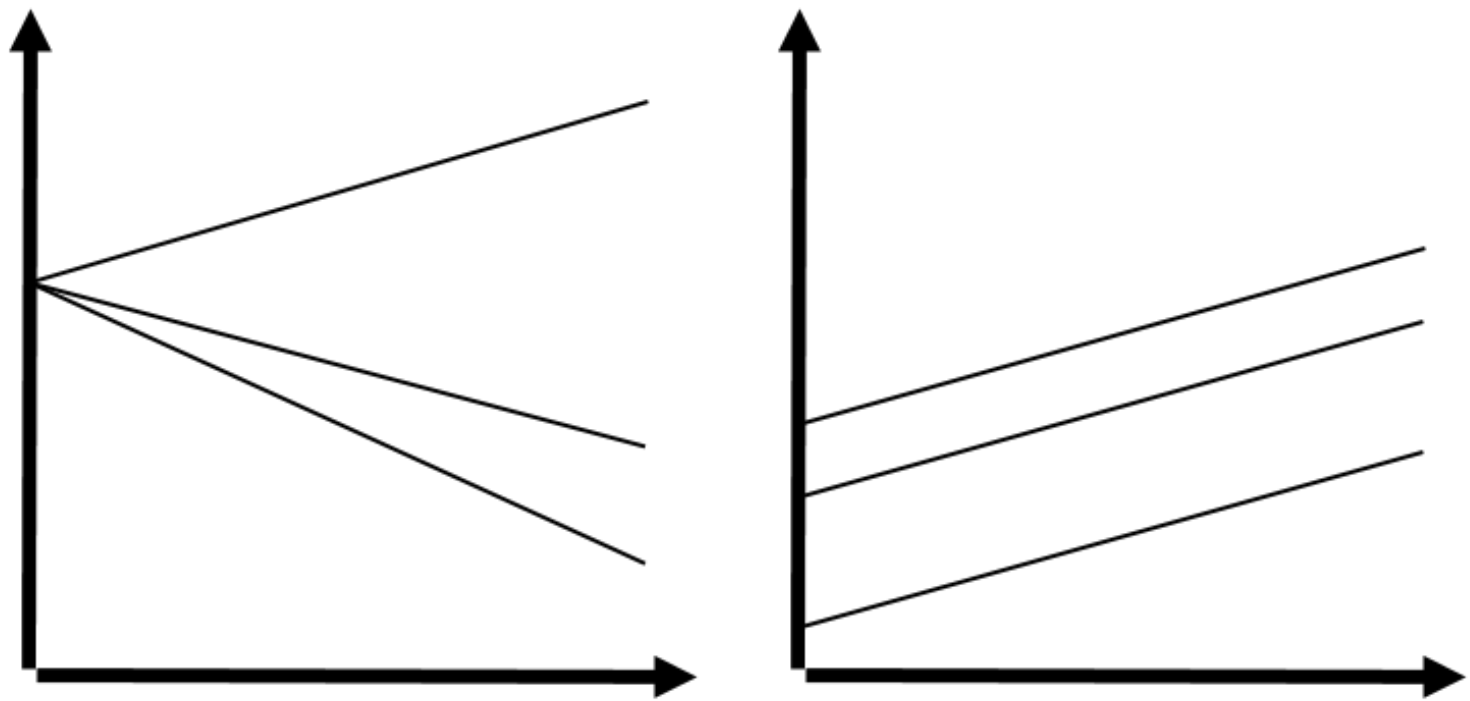
Ausflug Gerade Linien

Gerade Linien durch 2 Parameter bestimmt

- a: Schnittpunkt mit Y-Achse (Intercept)
- b: Winkel (Slope, Gradient)

$Y = a + b * X$

Gleicher Intercept vs. Gleicher Gradient



Regression als statistisches Modell

Kombiniere:

- $Ergebnis = Modell + Fehler$
- $Y = a + b * X$

Regressionsformel

- $\hat{Y} = (b_0 + b_1 * X) + \varepsilon_i$
- $\hat{Y}$ = vorhergesagtes Outcome
- X = Prädiktoren
- **Regressionskoeffizienten**
  - $\rightarrow$ b_0 = Schnittpunkt mit Y-Achse
  - $\rightarrow$ b_1 = Winkel der Geraden
- ε = **Residual Term**
  - $\rightarrow$ oft nicht explizit angegeben

Regressionskoeffizienten

Regressionskoeffizienten

- b_0 = Schnittpunkt mit Y-Achse
  - → Position des Modells im geometrischen Raum
- b_1 = Winkel der Geraden
  - → Richtung der Beziehung zwischen Prädiktor und Outcome
  - → positiv: Je höher x, desto höher y
    Übungszeit → Sprachverständnis
  - → negativ: Je höher x desto niedriger y
    Übungszeit → Freizeit
  - → Je extremer b_1, desto mehr ändert sich y bei einer Verschiebung von x
- b meint meistens b_1

Berechnung

Methode Andy Field:

- Suche einen kleinen bärtigen Zauberer namens Nephwick the Line Finder (Frage ein Statistikprogramm)

Mathematischereres Vorgehen:

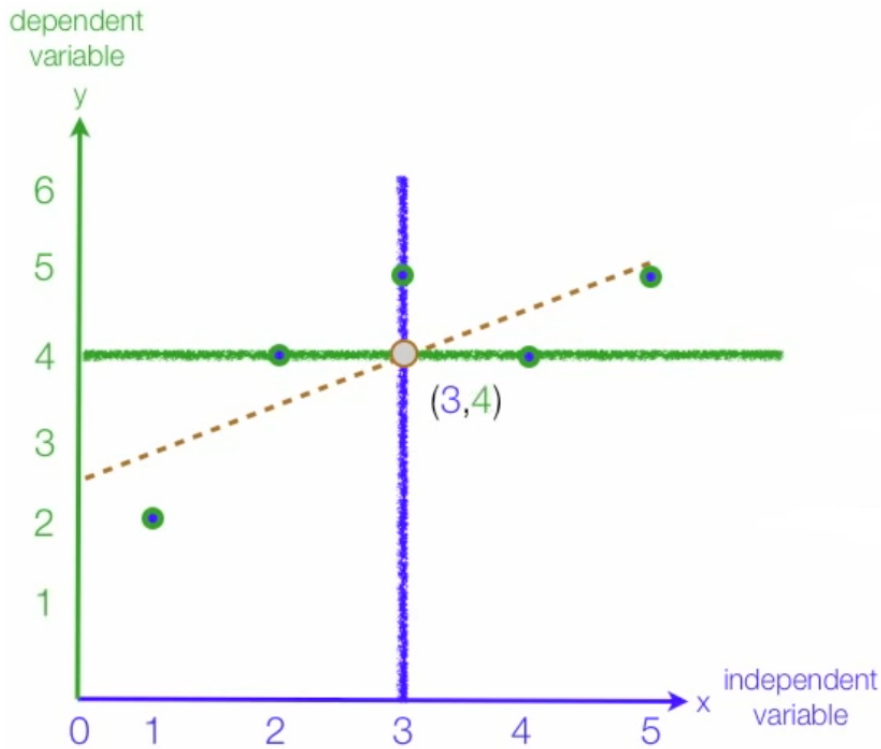
- Youtube StatisticsFun "How to calculate linear regression using least square method"
- `https://www.youtube.com/watch?v=JvS2triCgOY`
- $b_1 = \frac{\sum (x_i - \overline{x}) * (y_i - \overline{y})}{\sum (x_i - \overline{x})^2}$
- $b_0 = \overline{y} - b_1 * \overline{x}$

Methode der kleinsten Quadrate

Gegeben:

X	Y
1	2
2	4
3	5
4	4
5	5

Gesucht:



- Blaue vertikale Line $= \overline{x}$
- Grüne horizontale Line $= \overline{y}$
- Braune diagonale Line = Regressionslinie

Methode der kleinsten Quadrate

	X	Y	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x})^2$	$(x_i - \overline{x}) * (y_i - \overline{y})$
	1	2	-2	-2	4	4
	2	4	-1	0	1	0
	3	5	0	1	0	0
	4	4	1	0	1	0
	5	5	2	1	4	2
Mean:	3	4		Sum:	10	6

$b_1 = \frac{\sum (x_i - \overline{x}) * (y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = \frac{6}{10} = 0.6$

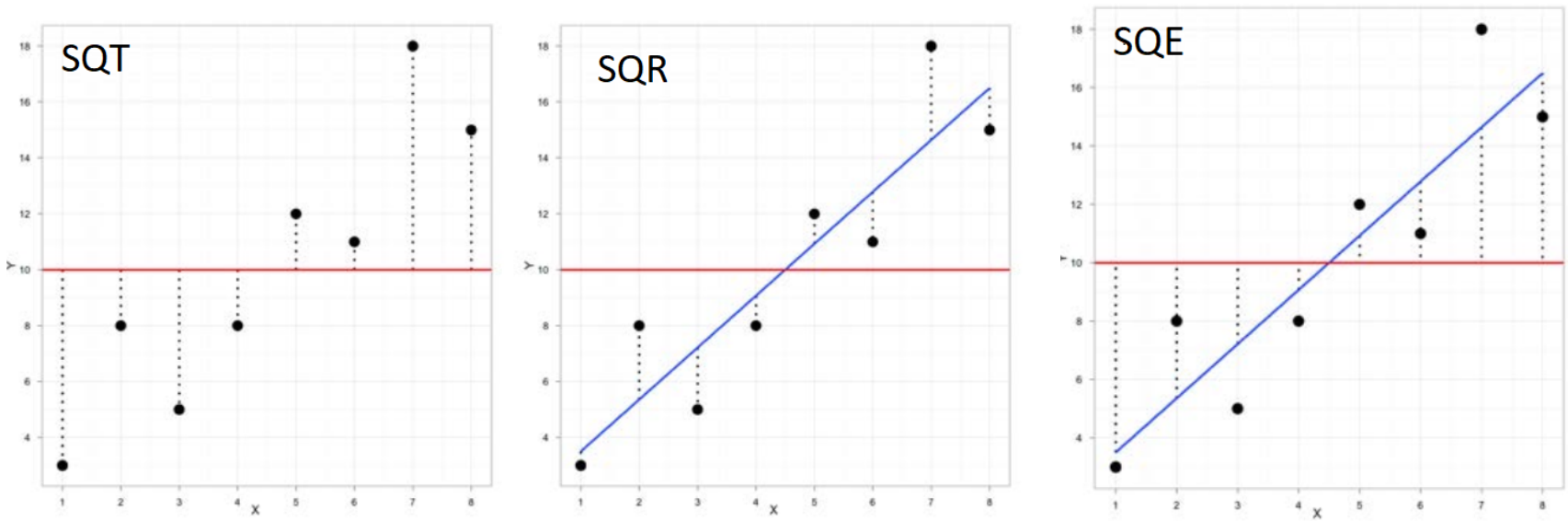
$b_0 = \overline{y} - b_1 * \overline{x} = 4 - 0.6 * 3 = 2.2$

$\hat{Y} = (b_0 + b_1 * X) = \underline{2.2 + 0.6 * X}$

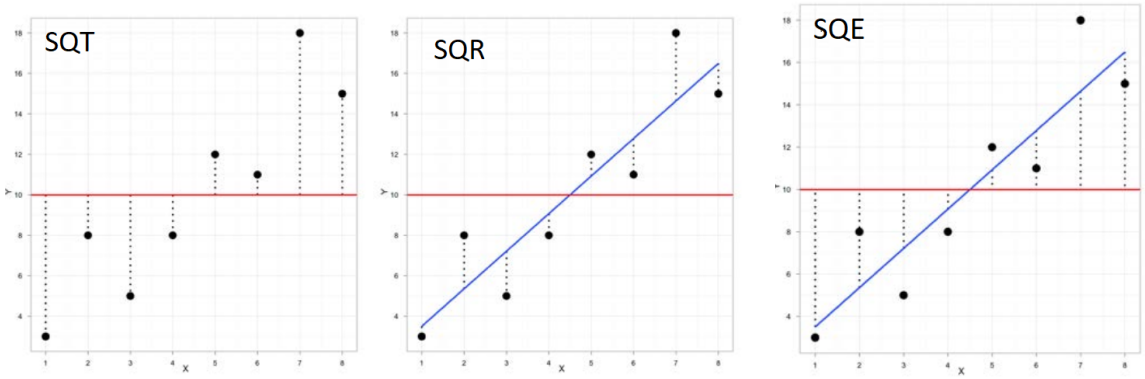
Fitness einer Regression

- Regressionsline gilt als *Best Fit* für ein Regressionsmodell...
- ...aber muss kein guter Fit sein

Fitness einer Regressionslinie



Fitness einer Regressionslinie



Abstände von Regression zu Beobachtung sind Residuen (Residuum)

- **Quadratsumme der Abweichungen** $SQT = \sum(y_i - \overline{y})^2$
- **Residuenquadratsumme** $SQR = \sum(y_i - \hat{y}_i)^2$
- **Erklärte Quadratsumme** $SQE = \sum(\overline{y} - \hat{y}_i)^2$
- $R^2 = \frac{SQE}{SQT}$

Interpretation

- hohes SQE bedeutet hohe Verbesserung des Regressionsmodells gegenüber dem Mittelwert
- R^2 ist der Anteil der Variation im Outcome, der durch das Modell erklärt wird

  $\rightarrow$ Fun Fact: Bei einfacher Regression gilt $\sqrt{R^2}$ = Pearsons r

Fitness einer Regressionslinie

Alternativ F-Test

- MQx = Mittelwert der Quadrate von x
- $MQE = \frac{SQE}{Variablenanzahl}$
- $MQR = \frac{SQR}{Beobachtungen - Regressionskoeffizienten}$
- F-Ratio $F = \frac{MQE}{MQR}$
- H_0 = Alle Regressionskoeffizienten sind 0 Die Regressionsline hat keine Vorhersagekraft
- Je höher F, desto besser das Modell
- Dazu später mehr...

Fitness von Prädiktoren

t-Test:

- Allgemein: $t = \frac{b_{observed} - b_{expected}}{SE_b}$ SE = Standardfehler = $\frac{s}{\sqrt{n}}$
- $H_0 : b == 0 \; // b_{expected}$ ist bei uns also 0
- $\rightarrow t_{b==0} = \frac{b_{observed}}{SE_b}$
- t_{kr} aus Tabelle ablesen ($df = n - anz_{predictors} - 1 \rightarrow n - 2$ für einfache Regression)
- $abs(t) < t_{kr} \rightarrow H_0$ angenommen $\rightarrow$ wahrscheinlich kein Effekt, der Unterschied zwischen $b_{observed}$ und 0 ist nicht signifikant
- Dazu später mehr...

Fitness einer Regression

$b_1 = 0.6,\ b_0 = 2.2,\ \hat{Y} = (b_0 + b_1 * X) = 2.2 + 0.6 * X$

	X	Y	$\hat{Y}$	$(y_i - \overline{y})^2$	$(y_i - \hat{y})^2$	$(\overline{y} - \hat{y})^2$
	1	2	2.8	4	0.64	1.44
	2	4	3.4	0	0.77	0.36
	3	5	4.0	1	1	0
	4	4	4.6	0	0.77	0.36
	5	5	5.2	1	0.04	1.44
Mean:	3	4	Sum:	8	3.22	3.6

- $SQT = 8,\ SQR = 3.22,\ SQE = 3.6$
- $R^2 = \frac{SQE}{SQT} = \frac{3.6}{8} = 0.45 \rightarrow 45\%$ der Variation von Y durch X erklärbar
- $t = \frac{b_1}{\sigma_b} = \frac{0.6}{0.2828} = 2.12 < t_{kr} = 3.18 \rightarrow$ Effektstärke des Prädiktors nicht signifikant
  //Standardfehler σ_b kommt vom R-Skript, kann aber auch analog zu vorher errechnet werden

Lineare Regression in *R*

```
geubt<-c(1,2,3,4,5)
punkte<-c(2,4,5,4,5)
data<-data.frame(geubt, punkte)
regression<-lm(data$punkte ~ data$geubt)
summary(regression)
```

```
Residuals:
   1    2    3    4    5
-0.8  0.6  1.0 -0.6 -0.2
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.2000     0.9381   2.345    0.101
data$geubt    0.6000     0.2828   2.121    0.124
Residual standard error: 0.8944 on 3 degrees of freedom
Multiple R-squared:    0.6,        Adjusted R-squared:  0.4667
F-statistic:   4.5 on 1 and 3 DF,  p-value: 0.124

//12% Zufallswahrscheinlich (F Test)
//Mit 0 Ubung sagt das Modell 2.2 Punkte vorraus (Intercept)
```

Vorhersage

Wir erinnern uns:

- Regressionsformel $\hat{Y} = (b_0 + b_1 * X) = 2.2 + 0.6 * X$

Wie kann man jetzt Vorhersagen treffen?
Einfach X einsetzen.

- $2.2 + 0.6 * 5$ *Übungszeit* $= 5.2$ *Punkte*
- $2.2 + 0.6 * 0$ *Übungszeit* $= 2.2$ *Punkte*
- $2.2 + 0.6 * 13$ *Buechergelesen* $= 9$ *Bibelzitate*

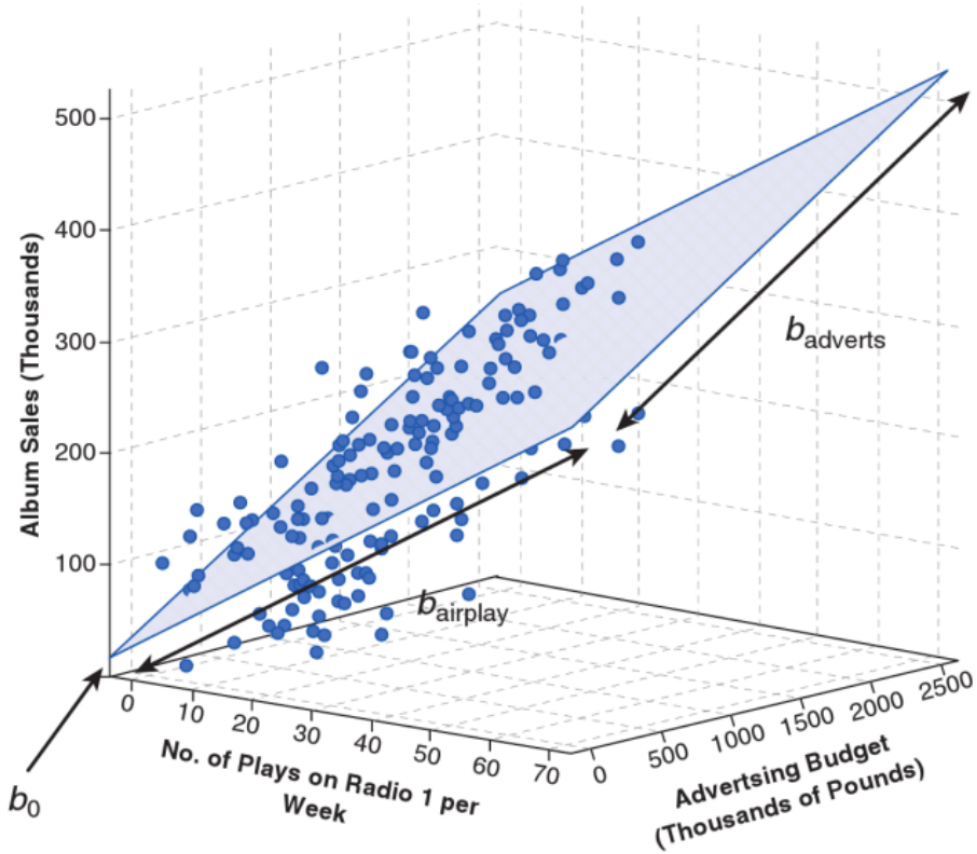
Multiple Regression

- Statistisches Modell zur Vorhersage einer abhängigen Variable auf Basis von mehreren unabhängigen Variablen
- Outcome = (model) + Fehler
- $\hat{Y} = (b_0 + b_1 * X_1 + b_2 * X_2 + \cdots + b_n * X_n) + \varepsilon_i$
- $\hat{Y}$ = vorhergesagtes Outcome
- X_i = Prädiktoren
- **Regressionskoeffizienten**

  → b_0 = Schnittpunkt mit Y-Achse
  → b_i = Koeffizient des Prädiktors X_i

  → Wie viel Angst haben Studierende 10, 5 oder 2 Minuten vor der Prüfung in wie großen Gruppen?
  → Wie viele Alben verkaufen wir, wenn wir x Euro für Werbung ausgeben und einen Song y mal im Radio spielen lassen?

Visualisierung



- Visualisierung schwierig
- 3 Prädiktoren (also ein Würfel) bereits schwer eindeutig darstellbar.

Fitness

- SQT, SQR,SQE analog zu linearer Regression berechenbar
- R = Korrelation zwischen beobachteten Y und berechneten Y
- Multiples R^2 = Maßzahl für Fitness ($1 \rightarrow$ Perfekter Fit)

Aber: R^2 steigt mit Anzahl der Prädiktoren, bevorteilt also Modelle mit mehr Prädiktoren, deshalb Sparsamkeitsbedachte Werte (Parsimony)
Akaike Information Criterion

- $AIC = n * \ln(\frac{SQR}{n}) + 2k$
- n = Anzahl der Fälle
- k = Anzahl der Prädiktoren

Interpretation nur im direkten Vergleich bei Modelle mit gleichen Daten, absolute Werte bedeutungslos
Interpretation: Je höher desto schlechter der Fit
Bayesian Information Criterion (Berechnung via R)

Auswahl der Prädiktoren

Prädiktoren korrelieren meist und haben Wechselwirkungen im Modell, deshalb Auswahl der Prädiktoren entscheidend

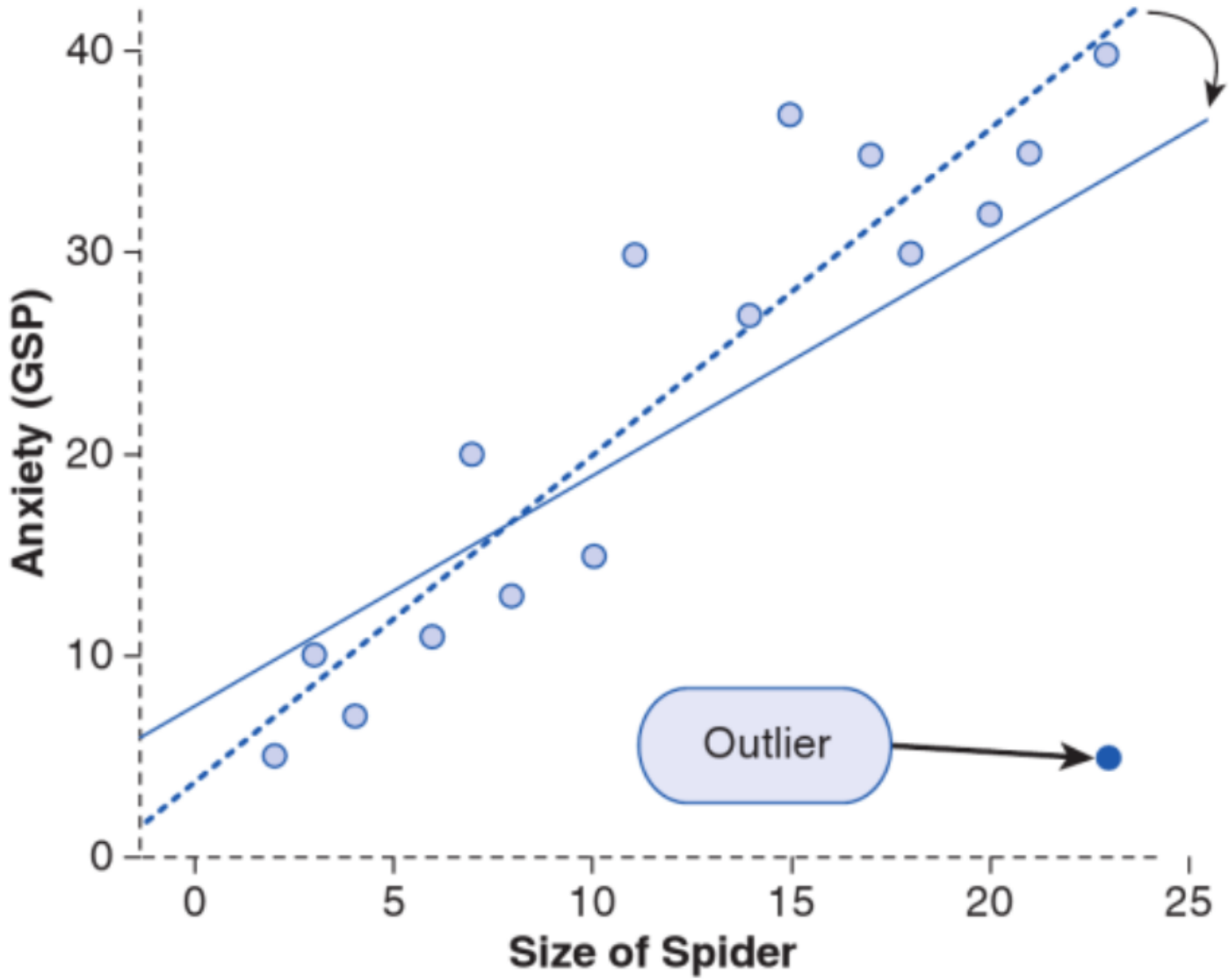
- Hierarchisch
  - → Nach Einfluss auf Modell
  - → Bekannte Prädiktoren zuerst (Bspw. Vorarbeiten)
  - → Weitere gleichzeitig oder schrittweise oder wieder hierarchisch
- Erzwungen (Alle auf einmal)
- Schrittweise (Greedy)
  - → *vorwärts*: Wähle Prädiktor, der am meisten erklärt solange *AIC* besser wird
  - → *rückwärts*: Füge alle Prädiktoren ein und losche die, deren Löschung *AIC* verbessert
  - → *beidseits*: *Greedy vorärts* mit *Greedy rückwärts* in jedem Schritt
  - → Nachteil am Beispiel Anziehsachen: Wähle die wärmsten Kleidungsstücke → Unterwäsche vergessen
- Alle Teilmengen
  - → Bewertung aller möglichen Permutationen
  - → 2 Prädiktoren: 4 Permutationen, 3 Prädiktoren: 8 Permutationen
  - → Fitnessbewertung mittels *Mallows* C_p

Evaluation von Regressionen

2 Schritte zur Bewertung der Korrektheit

- Schritt 1: Fitness bezogen auf eigene Daten (Extremwerte und Einflusstarke Werte)
- Schritt 2: Generalisierbarkeit, Lässt sich das Modell auf andere Daten übertragen?

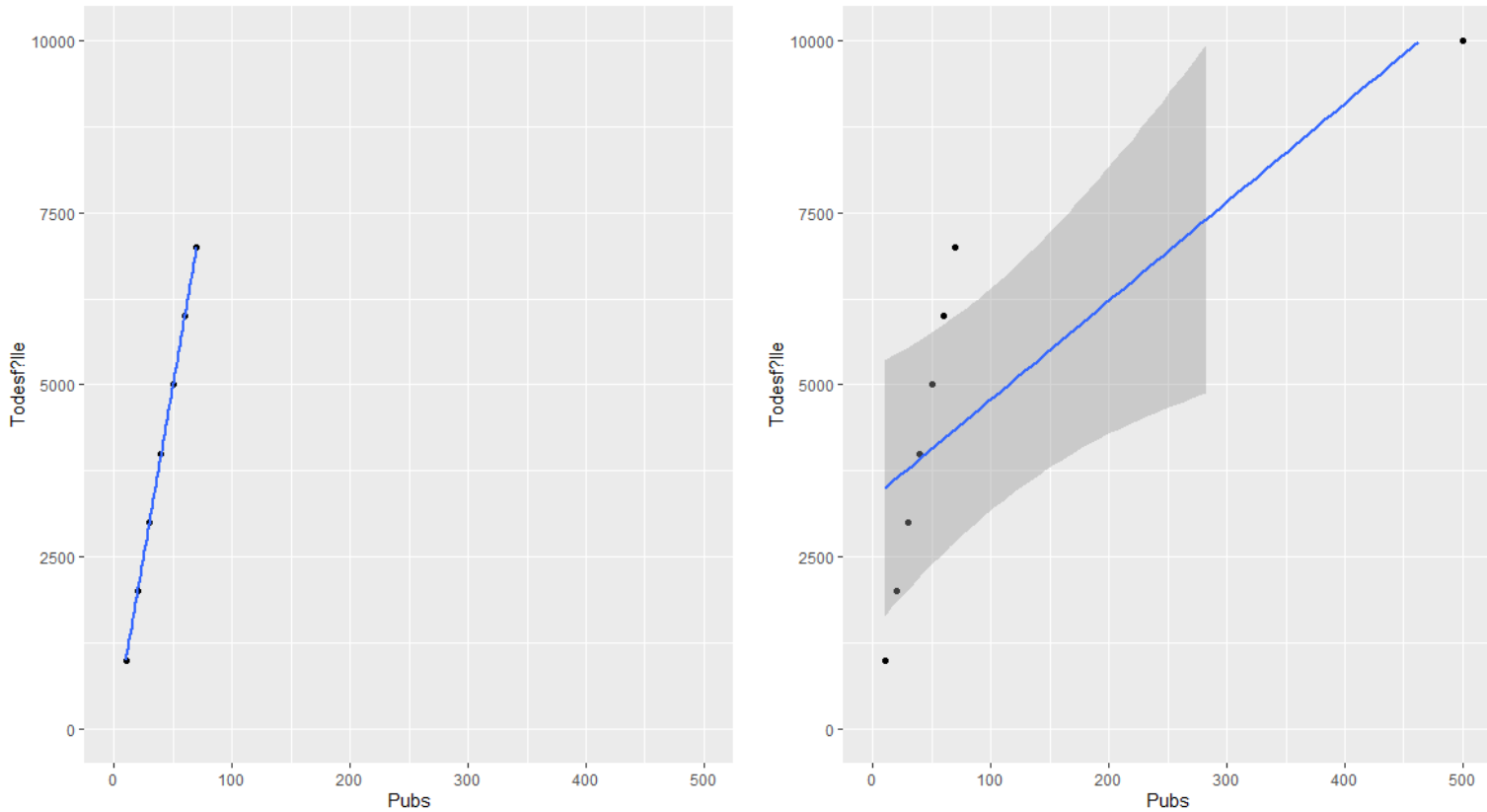
Extremwerte



Extremwerte

- Extremwerte kippen Regressionsgerade und erzeugen (wenn unpassend) Bias im Modell
- **Residuum R** = Abstand zwischen Regression und Beobachtung
- Extremwerte sind auffällig große Residuen
- Aber:Toleranz des absoluten Residuenabstand vom Modell abhängig
- $\rightarrow$ **Standardisierte Residuen SR** $= \frac{R}{s_R}$
- Merkregeln, die aus Umwandlung in z-Scores folgen:
  - $\rightarrow$ $SR > 3.29$ sind auffällig und unüblich
  - $\rightarrow$ Wenn mehr als 1% der SR über 2.58 liegen, passt das Modell schlecht zu den Daten
  - $\rightarrow$ Wenn mehr als 5% der SR über 1.96 liegen, passt das Modell schlecht zu den Daten

Einflusstarke Werte




```
pubs <- c(10,20,30,40,50,60,70,500)
deaths <- c(1000,2000,3000,4000,5000,6000,7000,10000)
pubsdeaths <-data.frame(pubs,deaths)
graph<-ggplot(pubsdeaths, aes(pubs, deaths))
graph + geom_point() + xlim(0,500) +ylim(0,10000)
    + labs(x = "Pubs", y = "Todesf?lle") + geom_smooth(method="lm")
```

Einflusstarke Werte

- Einflusstarke Werte machen das Modell instabil
  - $\rightarrow$ $DFFit_i$ = Differenz zwischen y_i mit und ohne Fall i
  - $\rightarrow$ Studentisiertes Residuum = Differenz zwischen y_i und ohne Fall i geteilt durch Standardfehler
  - $\rightarrow$ Cooks Distance gibt Einflussstärke eines Falles auf Vorhersagen aller anderen Fälle wieder ($> 1 \rightarrow$ Problemwert)
  - $\rightarrow$ Hat-Value (Leverage/Hebelkraft): Durchschnitt berechnen $\frac{k+1}{n}$. Je mehr Abstand (Leverage) des Falls i zum Durchschnitt hat, desto höher ist der Einfluss

Einflusstarke Werte

Achtung:

- Influenzanalyse dient zur Bewertung eines Modells
- ...nicht zur Rechtfertigung einer Löschung eines Falls
- Gegenteil möglich: "Fall i ist Extremwert, aber da Cook Distance < 1 muss er nicht gelöscht werden."

Generalisierbarkeit

Lässt sich das Modell auf andere Daten übertragen? Das Modell hat weniger Bias, desto besser es folgende Annahmen erfüllt

- Prädiktoren haben Varianz > 0
- Keine hohe Korrelation zwischen Prädiktoren (Multikollinearität)
- Prädiktoren korrelieren nicht mit externen Variablen
- Homoskedastizität (gleichmäßige Varianz der Residuen)
- Normalverteilung der Residuen mit Mittelwert 0
- Unabhängigkeit der Outcomes
- Linearität der Outcomes
- Variablentypen
  - $\rightarrow$ Prädiktoren: Intervall oder 2 Kategorien
  - $\rightarrow$ Outcome: Intervall, stetig, uneingeschränkt (Spanne von Y sollte Spanne der Datenpunkte nicht überschreiten)
- Unabhängigkeit der Fehler
  - $\rightarrow$ Autokorrelation
  - $\rightarrow$ Durbin-Watson Test

Kreuzvalidierung

- Je ähnlicher die Vorhersagekraft des Modells für verschiedene Samples, desto generalisierbarer ist es
- R^2 nach Stein: (Achtung, Adjusted R^2 in der Sprache R nach Wherry passt hier nicht)
  - $\rightarrow$ Adjusted $R^2 = 1 - [\frac{n-1}{n-k-1} * \frac{n-2}{n-k-2} * \frac{n+1}{n}] * (1 - R^2)$
  - $\rightarrow$ je höher, desto besser kreuzvalidiert das Modell
- Data Splitting
  - $\rightarrow$ Daten zufällig teilen
  - $\rightarrow$ Modell für Teilsamples berechnen
  - $\rightarrow$ Generalisierbare Modelle sollten jetzt ähnliche Koeffizienten haben

Strichprobengröße

- Je mehr, desto besser
- Oversimplified: Mindestens 10 bis 15 mal die Anzahl der Prädiktoren
- Green, Samuel B (1991): *How Many Subjects Does It Take to Do a Regression Analysis?*

  $\rightarrow$ Bei Modelltests $n_{min} = 50 + 8 * k$
  $\rightarrow$ Bei fallbezogenen Tests $n_{min} = 104 + k$
  $\rightarrow$ Regelfall (Beides) : Maximum beider Werte

Zusammenfassung

- Regression erlaubt Abschätzen von Y für neue Werte aus X
- Zur Beschreibung benötigen wir Winkel und Schnittpunkt der Linie

  $\rightarrow$ Methode der kleinsten Quadrate
- Regressionsformel $\hat{Y} = (b_0 + b_1 * X)$
- b_0 und b_1 sind Regressionskoeffizienten
- Als statistisches Modell hat eine Regressionslinie eine Fitness

  $\rightarrow$ Residuenquadratsumme, Erklärte Quadratsumme, R^2 = Verhältnis beider
  $\rightarrow$ F-Test möglich um Modell zu bewerten
  $\rightarrow$ t-Test möglich um Einflußstärke des Prädiktors zu bewerten
- 1 Prädiktor $\rightarrow$ Einfache Regression, Mehr als Prädiktor $\rightarrow$ Multiple Regression

  $\rightarrow$ Auswahl der Prädiktoren entscheidend
- Fitness der Regression zu Daten, Generalisierbarkeit
- Übersprungen: Multikollinearität, Annahmenbruch (Transformation der Residuen / Bootstrapping)