

Regressionsmodelle für Politikwissenschaftler: Seminarüberblick und Einführung

Regressionsmodelle für Politikwissenschaftler

Formalia
Seminarablauf
Grundsätzliches zur Regression
Wiederholung: Parameterschätzung für die lineare Regression

Überblick

Formalia

Seminarablauf

Grundsätzliches zur Regression

Was ist Regression?

Wiederholung: Das Standardmodell der linearen Regression

Nomenklatur

Wiederholung: Wahrscheinlichkeitsverteilungen

Beschreibung und Inferenz

Wiederholung: Parameterschätzung für die lineare Regression

Formales zum Scheinerwerb

- ▶ Sie beteiligen sich am Seminargespräch. Voraussetzung dafür ist die Lektüre der Pflichttexte, die zu jeder Sitzung angegeben sind. „Pflichttexte“ bedeutet: Die Lektüre ist verbindlich, deshalb überprüfe ich gelegentlich Ihren Kenntnisstand.
- ▶ Auch mit einer „Entschuldigung“ dürfen Sie maximal zwei Sitzungen versäumen (siehe Studienordnung).
- ▶ Sie fertigen eine Hausarbeit an, die in thematischem Zusammenhang mit dem Seminar steht. *Die Arbeit muß eine explizite Fragestellung verfolgen.*

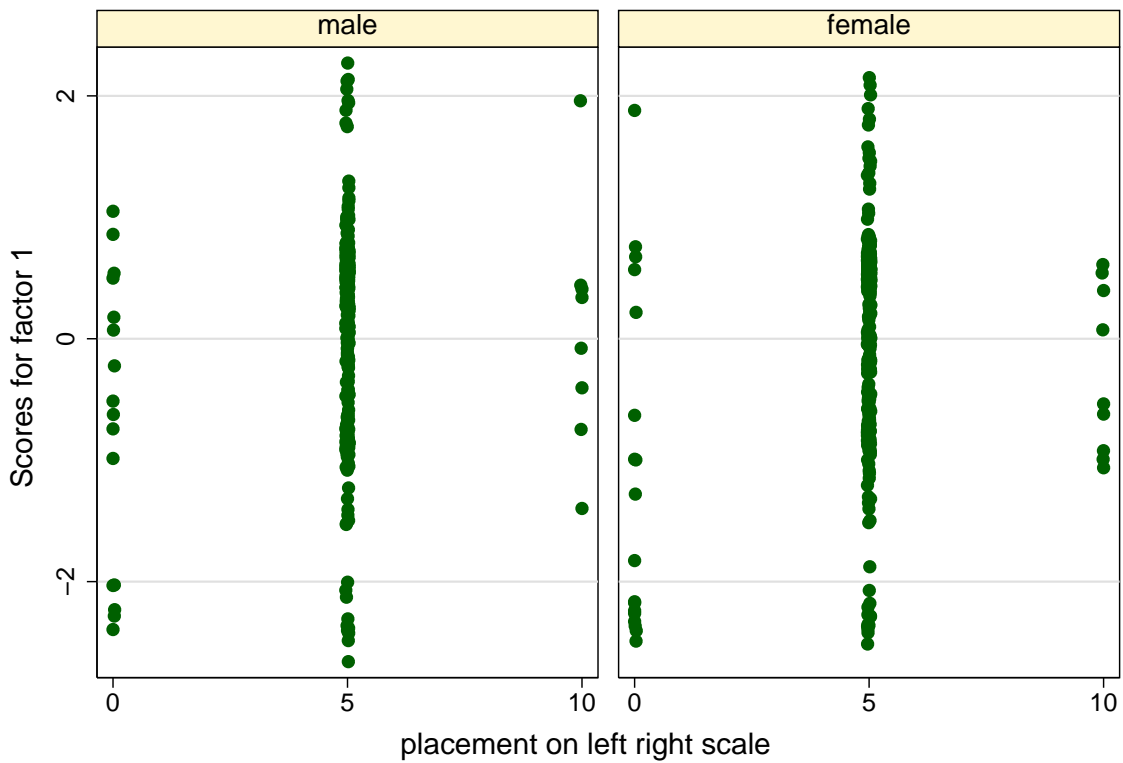
Formales zur Hausarbeit I

- ▶ Gliederung, Zitierweise, Literaturverzeichnis etc. bitte entsprechend den üblichen Standards, z. B.
<http://www.kai-arzheimer.com/Lehre-BRD/ha.html>
[/Lehre-BRD/ha.html](http://www.kai-arzheimer.com/Lehre-BRD/ha.html)
- ▶ Schriftart: Times, Schriftgröße: 11 Punkt (für Fußnoten 10 Punkt), Zeilenabstand: anderthalbfach, für Fußnoten und Literaturverzeichnis einfach, Satz: Blocksatz mit automatischer Silbentrennung, Umfang: ca. 7 000-9 000 Worte, was bei diesen Einstellungen etwa 20-25 reinen Textseiten entspricht.

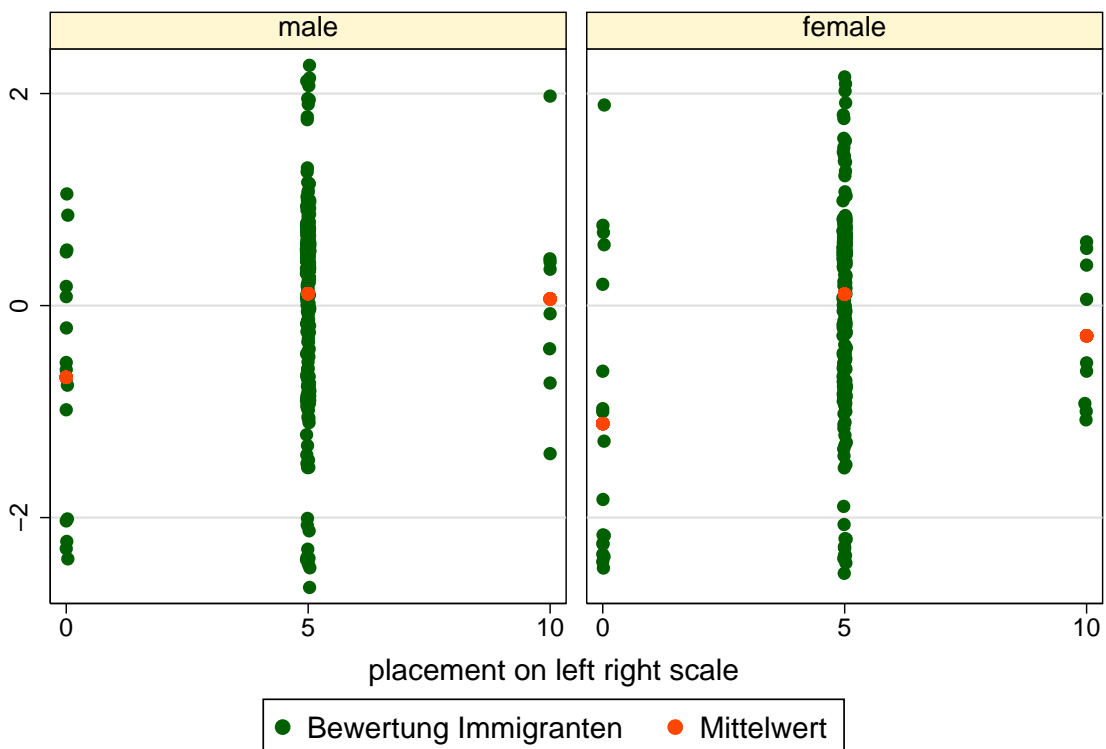
Formales zur Hausarbeit II

- ▶ Eine Vorlage für WinWord mit entsprechenden Einstellungen finden Sie hier: <http://www.kai-arzheimer.com/Vorlage-HS.doc>
[/Vorlage-HS.doc](#)
- ▶ Die Arbeit kann und soll nach Absprache mit mir bereits während des Semesters begonnen werden.
- ▶ **Letzter Abgabetermin: Freitag, 21. August 2009** (Zentralsekretariat / Poststempel). Eine Verschiebung des Termins ist nur mit einem ärztlichen Attest möglich.

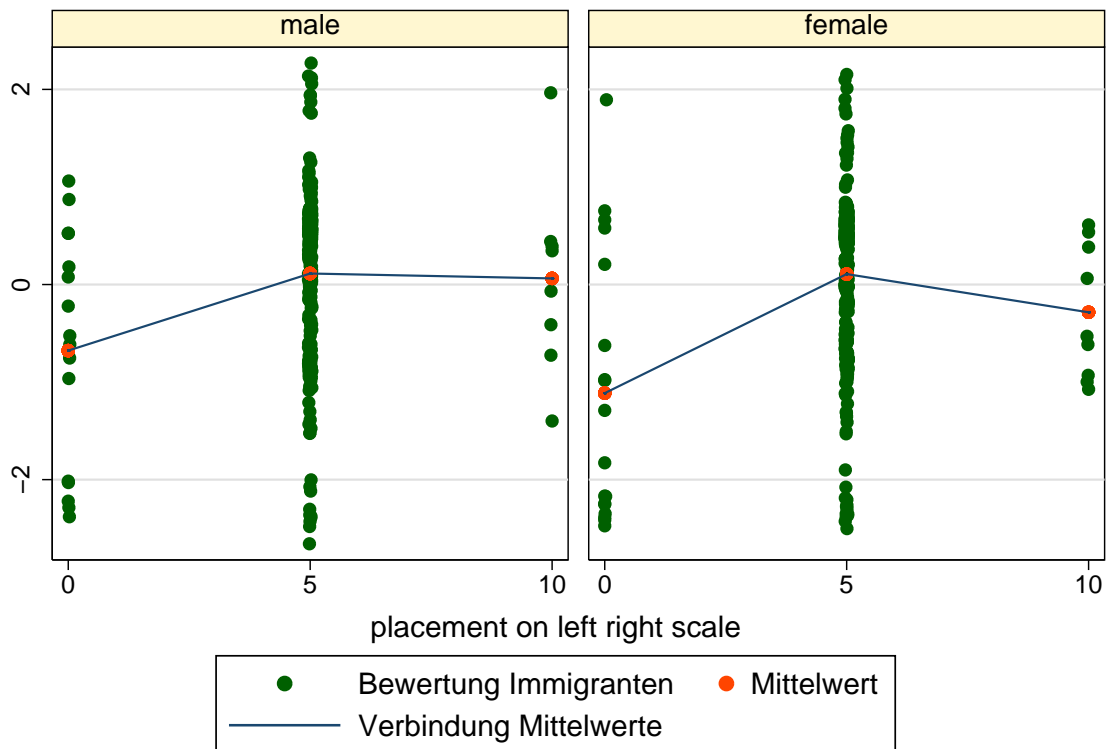
1. Grundlagen: Wiederholung zum linearen Modell, Annahmenverletzungen und ihre Konsequenzen, verallgemeinertes Modell und ML-Schätzung
2. Besondere Modelle für kategoriale Daten und Missing Data
3. Ereignisdaten (Grundlagen)
4. TSCS/Panel (Grundlagen)
5. Strukturgleichungsmodelle
6. Mehr-Ebenen-Analyse



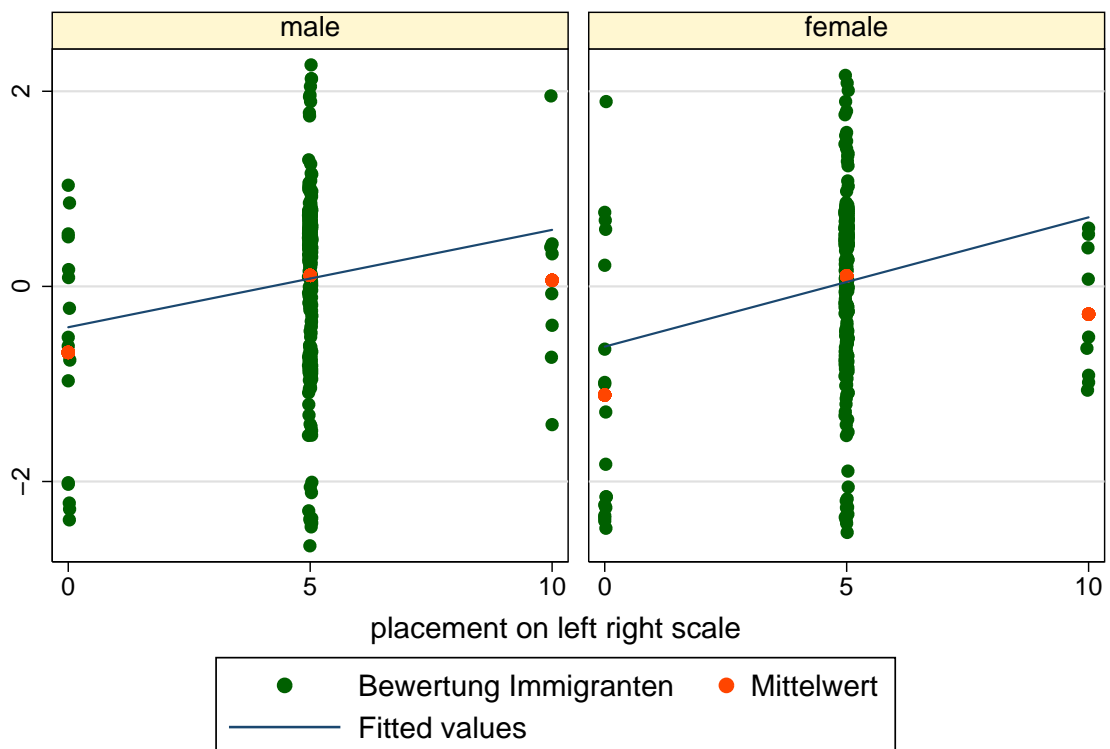
Graphs by gender



Graphs by gender



Graphs by gender



Graphs by gender

Was ist Regression?

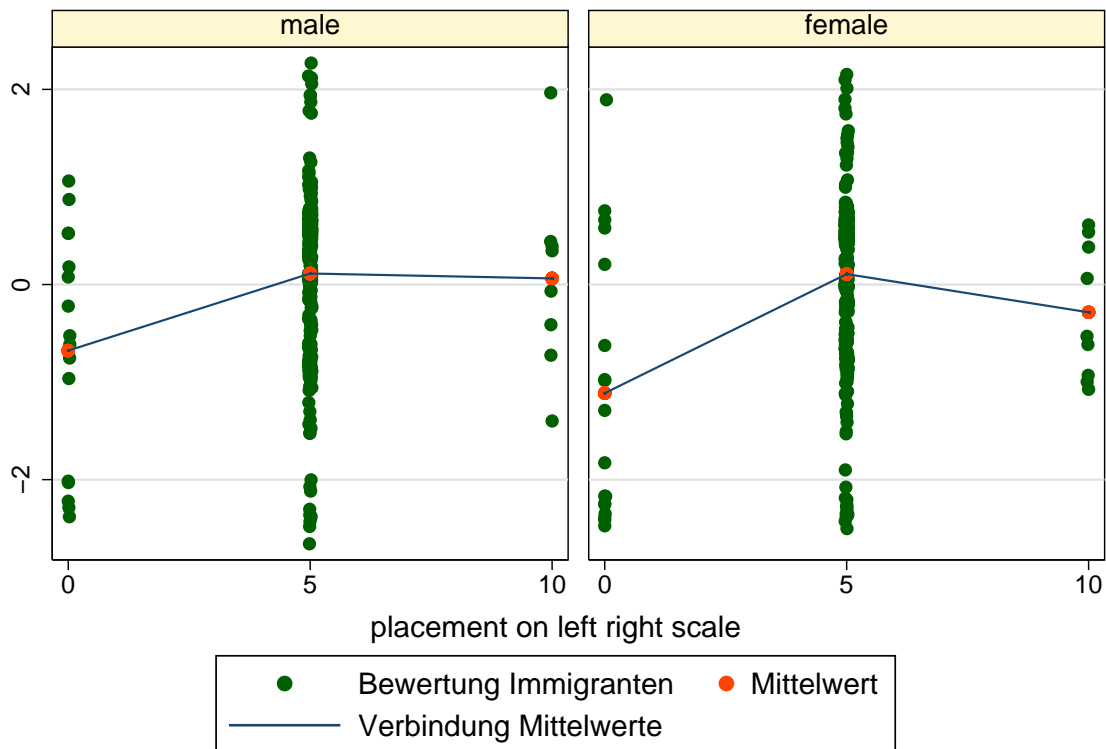
- ▶ Regression ist der Oberbegriff für Verfahren, ...
- ▶ die die *konditionale* Verteilung einer Variablen y ...
- ▶ in Abhängigkeit von einer oder mehreren anderen Variablen $x_1, x_2 \dots x_k$ beschreiben

Was ist eine „konditionale Verteilung“?

- ▶ Verteilung von y (Mittelwert, Streuung etc.) ...
- ▶ innerhalb von Subgruppen, die durch $x_1, x_2 \dots x_k$ definiert sind

Was ist Regression?

- ▶ Die konditionalen Mittelwerte können durch eine glatte Linie beschrieben werden
- ▶ Übergang zum Modell: Annahmen über die Eigenschaften der Linie kommen von außen
- ▶ „Abhängige“ / „unabhängige“ Variable kommen ebenfalls von außen
- ▶ Das Beispiel zeigt u. a.
 - ▶ Mehrere unabhängige Variablen
 - ▶ Kategoriale unabhängige Variablen
 - ▶ Interaktion
 - ▶ Probleme mit der Linearitätsannahme



Graphs by gender

Wie sieht das Standardmodell aus?

$$\begin{aligned}
 y &= \alpha + \beta_1 x_1 + \beta_2 x_2 \cdots + \epsilon \\
 &= \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 \cdots + \epsilon \\
 &\text{mit } x_0 = 1 \text{ für alle Einheiten}
 \end{aligned}$$

Welche Symbole werden verwendet?

- ▶ Leider ist die Nomenklatur oft wenig einheitlich
- ▶ Grundregeln:
 - ▶ y für „abhängige“ Variable, x für „unabhängige“ Variable
 - ▶ Variablen, Parameter und Untersuchungseinheiten kann man mit einem Index durchnummerieren: $x_1, x_2 \dots x_k$
 - ▶ Lateinische Buchstaben für Variablen und Parameter in der Stichprobe,
 - ▶ Griechische Buchstaben für die unbekannt Parameter der Grundgesamtheit
 - ▶ Übersicht über das griechische Alphabet: <http://www.kai-arzheimer.com/Lehre-Regression/greek.pdf>

Welche Symbole werden verwendet?

- ▶ Leider ist die Nomenklatur oft wenig einheitlich
- ▶ Grundregeln:
 - ▶ Variablen erkennt man am *Kursivdruck*
 - ▶ Für Vektoren verwendet man (griechische oder lateinische) Kleinbuchstaben in **Fettdruck**
 - ▶ Für Matrizen verwendet man (griechische oder lateinische) Großbuchstaben in **Fettdruck**
 - ▶ Ein „Dach“ über einem Parameter (z. B. $\hat{\beta}$) zeigt an, daß es sich um eine Schätzung handelt (wird oft weggelassen)

Was ist eine Zufallsvariable?

- ▶ Zufallsexperimente
 - ▶ Können theoretisch beliebig oft wiederholt werden
 - ▶ Einzelergebnisse hängen vom Zufall ab, Verteilung der Ergebnisse ist aber bekannt
 - ▶ Bei häufiger Wiederholung nähert sich die empirische Verteilung der theoretischen Verteilung an
- ▶ Ziehung einer Zufallsstichprobe ist ein Zufallsexperiment
- ▶ Deshalb sind Stichprobenkennwerte und Modellparameter ebenfalls Zufallsvariablen

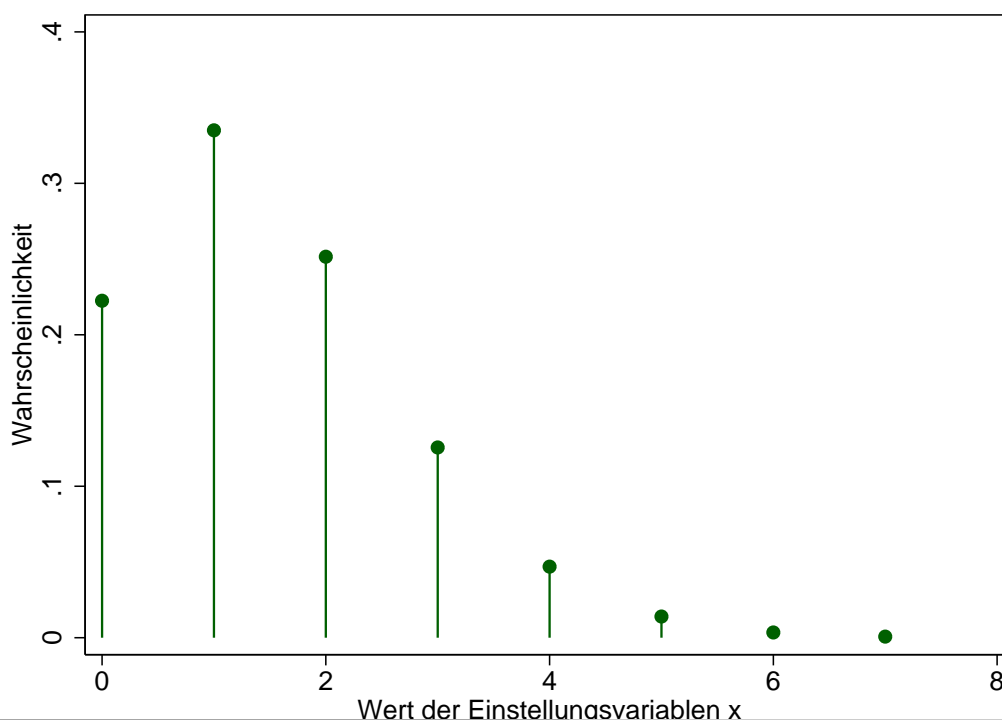
Was ist eine Zufallsvariable?

- ▶ Im Einzelfall weiß man nicht, welchen Wert die Variable annimmt
- ▶ Aber: Ausprägungen von Zufallsvariablen sind nicht willkürlich, sondern höchst regelmäßig verteilt
- ▶ Die Form der Verteilung der Werte einer Zufallsvariablen ist in der Regel bekannt / wird angenommen
- ▶ Zufallsvariablen (und ihre Verteilungen) können diskret oder stetig sein

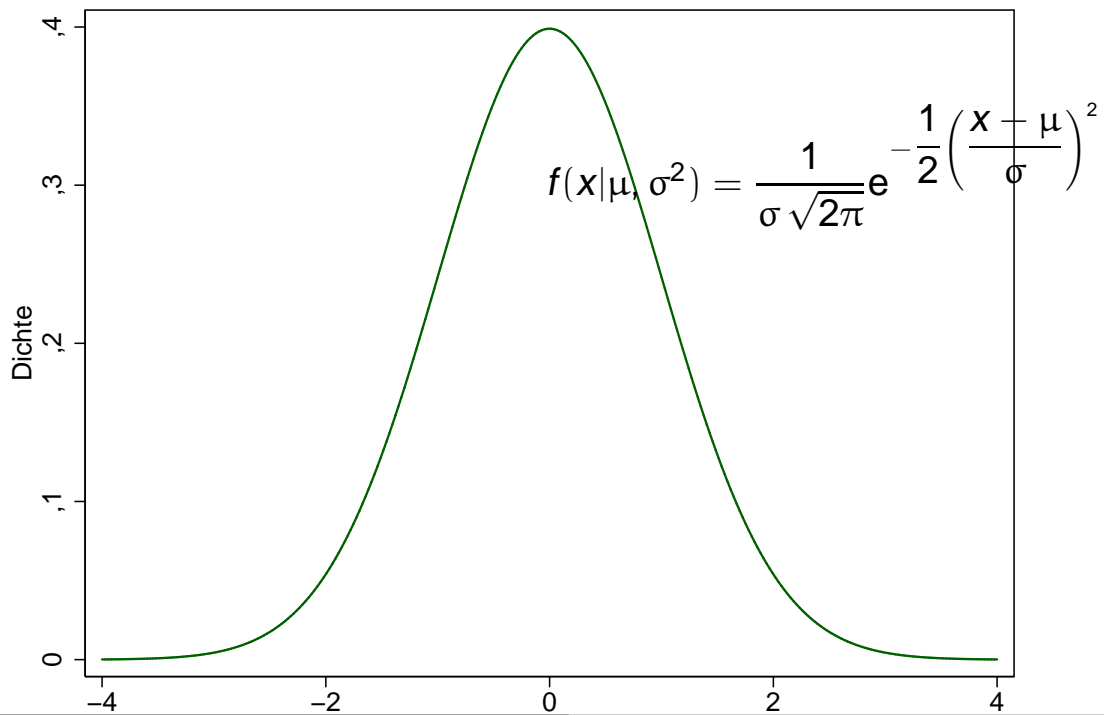
Was ist eine Zufallsvariable?

- ▶ Zufallsvariablen können durch eine Verteilungsfunktion beschrieben werden
 - ▶ Bei diskreten Variablen besteht eine solche Verteilung aus „Stäbchen“ oder Säulen, die die Wahrscheinlichkeit eines Ereignisses repräsentieren
 - ▶ Die Summe der Säulen ergibt 1
 - ▶ Stetige Zufallsvariablen werden durch eine ebenfalls stetige Verteilungsfunktion („Dichte“) beschrieben
 - ▶ Die Wahrscheinlichkeit eines bestimmten Wertes ist in diesem Fall gleich null
 - ▶ Berechnen lassen sich aber Wahrscheinlichkeiten für Intervalle, indem das Integral (Fläche unter der Dichtekurve) bestimmt wird
 - ▶ Die Gesamtfläche ist 1

Verteilung einer diskreten Zufallsvariablen



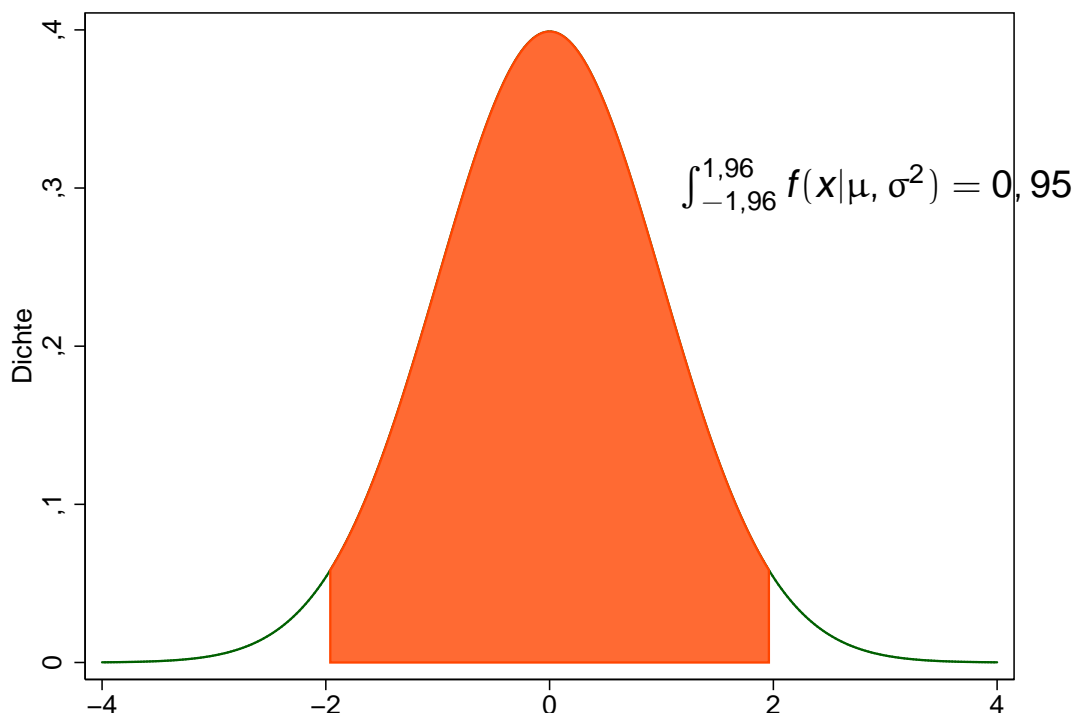
Verteilung einer stetigen Zufallsvariablen



Regressionsmodelle für Politikwissenschaftler

Seminarüberblick und Einführung

Verteilung einer stetigen Zufallsvariablen



Regressionsmodelle für Politikwissenschaftler

Seminarüberblick und Einführung

Was ist eine Wahrscheinlichkeitsverteilung?

- ▶ Können durch wenige (meist ein bis drei) Parameter vollständig beschrieben werden
- ▶ Wichtige Verteilungen:
 - ▶ Bernoulli-Verteilung / Binomial-Verteilung
(Wahrscheinlichkeit von $p = 1$, Zahl der Versuche)
 - ▶ **Normalverteilung** $(f(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \mu, \sigma)$
 - ▶ χ^2 -Verteilung (Summe aus quadrierten standardnormalverteilten Werten, Freiheitsgrade)
 - ▶ t -Verteilung (Quotient $\frac{z}{\sqrt{\chi_n^2/n}}$, Freiheitsgrade)
 - ▶ F-Verteilung $(\frac{\chi_{n1}^2}{\chi_{n2}^2} \times \frac{n_2}{n_1},$ zweimal Freiheitsgrade)

Wozu braucht man Wahrscheinlichkeitsverteilungen?

- ▶ Mathematische Verteilungen können als Modell für reale Zufallsvariablen dienen (lineare Regression: Normalverteilung, t -Verteilung)
- ▶ Z. B. wird angenommen, daß die ϵ im Standardmodell normalverteilt sind
- ▶ **Zentraler Grenzwertsatz:** Unter bestimmten Umständen folgt die *Verteilung der Parameterschätzungen über eine unendliche Zahl von Stichproben hinweg* einer t -beziehungsweise Normalverteilung
- ▶ (Die Verteilung von Abweichungsmaßen folgt idealerweise einer χ^2 -Verteilung)

Was ist der konzeptionelle Status eines Regressionsmodells?

„To err is human, to forgive divine, but to include errors into your design is statistical“ (Leslie Kish)

„All models are wrong. Some are useful“ (George Box)

Was will uns Kish sagen?

- ▶ Abhängige Variable kann niemals vollständig durch $x_1, x_2 \dots x_k$ erklärt werden
- ▶ Zufällige / als zufällig betrachtete Einflüsse sind ebenfalls Bestandteil des Modells (im linearen Modell ϵ)
- ▶ Diese Art von „Fehlern“ ist aus Sicht des Modells völlig unproblematisch

Was will uns Box sagen?

- ▶ Modelle niemals eine vollständige Abbildung der Wirklichkeit, sondern immer extreme Vergrößerung
- ▶ z. B. Auswahl unabhängigen Variablen, Linearitätsannahme
- ▶ Ist das Modell dem Forschungsproblem angemessen?
 - ▶ Instrumentalismus / Idealisierung (Friedman): Gute Prognosen, Problem: Stabilität der Randbedingungen?
 - ▶ Realismus / Abstraktion: Realistische Beschreibung, Problem: Komplexität, „Overfitting“

Was können wir mit den Parametern eines Modells anfangen?

- ▶ *Beschreibung:*
 - ▶ Modell erfasst wesentliche Aspekte einer konkreten Verteilung von Datenpunkten
 - ▶ Keine weitergehenden Schlüsse, Mittel zur Verdichtung der Information
- ▶ *Inferenz:*
 - ▶ Von den konkreten Daten soll auf etwas anderes geschlossen werden, aber auf was?
 - ▶ (Fast völlig) unproblematisch im Fall einer Zufallsstichprobe aus einer großen Grundgesamtheit
 - ▶ Klassische Inferenz, Standardfehler, Konfidenzintervalle, Signifikanztests

Was leistet die klassische Inferenz?

- ▶ Rückschlüsse auf die Verteilung der in der Stichprobe errechneten Schätzungen
- ▶ um die wahren Werte in der Grundgesamtheit
- ▶ wenn Stichprobenziehung unter essentiell identischen Bedingungen
- ▶ unendlich oft wiederholt wird

„Wenn der Wert des Parameters in der Grundgesamtheit tatsächlich null ist, werde ich nur in fünf von 100 Stichproben eine entsprechend große oder größere Schätzung als die hier vorliegende erhalten“

„Ein Intervall, das nach dieser Regel konstruiert wird, wird in 95 von 100 Stichproben den wahren Wert des Parameters mit einschließen“

Und wenn ich keine Zufallsstichprobe habe?

- ▶ Schulbezirke, OECD-Staaten, Studierende an einer bestimmten Universität
- ▶ Strategie I: Die Daten werden wie eine Grundgesamtheit behandelt
 - Regression dient nur zur *Beschreibung*
- ▶ Strategie II: Daten werden behandelt als seien sie eine Zufallsstichprobe
 - ▶ Setzt voraus, daß die (soziale) Natur einem stochastischen Prozeß folgt, der wie die Ziehung einer Zufallsstichprobe funktioniert
 - ▶ Wird in der Regel nicht überprüft, sondern behauptet
 - ▶ Trifft häufig offensichtlich *nicht* zu (Implementationsforschung)
 - ▶ Beispiele im Text: Selektive Ausfälle, Personen handeln nicht unabhängig voneinander

Und wenn ich keine Zufallsstichprobe habe?

- ▶ Strategie III: Konstruktion einer hypothetische Grundgesamtheit („superpopulation“)
 - ▶ Zirkuläre Definition: Superpopulation = imaginäre Grundgesamtheit, aus der die Daten stammen würden, wenn sie eine Zufallsstichprobe wären
 - ▶ Impliziter Rückgriff auf einen stochastischen Prozeß in der Natur
 - ▶ In den Sozial- und Wirtschaftswissenschaften extrem weit verbreitet (Standardfehler für ökonomische Zeitreihen)
- ▶ Strategie IV: Model-based sampling
 - ▶ Das Konzept einer Grundgesamtheit wird gänzlich aufgegeben
 - ▶ Statt dessen wird angenommen, daß die Prozesse, die die Wirklichkeit generieren, mit einem (einfachen) linearen Modell identisch sind
 - ▶ Die Inferenz bezieht sich dann direkt auf die Eigenschaften dieses Modells
 - ▶ Interessanter Ansatz, der aber eine Menge komplexer

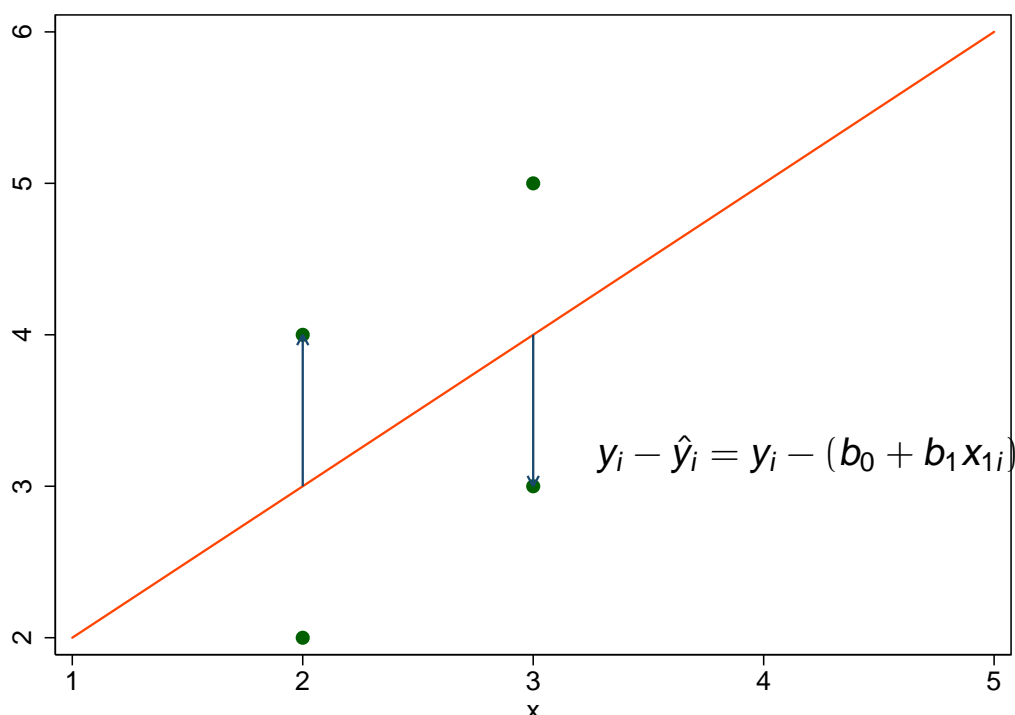
Und wenn ich keine Zufallsstichprobe habe?

Innerhalb der klassischen Analyserahmens (Zufallsstichproben aus sehr großen Grundgesamtheiten) haben Standardfehler eine klare (wenn auch nicht unbedingt befriedigende) Interpretation, und es gibt keine Diskussion über die Gültigkeit des mathematischen Apparats. Ansonsten bewegt man sich auf sehr unsicherem Boden.

Wie komme ich zu meinen Schätzungen?

- ▶ Wie lege ich die Gerade durch die Punkte?
- ▶ Standardmethode: „Kleinste-Quadrate-Schätzung“ (Ordinary Least Squares, OLS)
- ▶ Welche Koeffizienten minimieren die „Summe der Abweichungsquadrate“?
 - ▶ Gerade Linie soll arithmetische Mittel verbinden, arithmetisches Mittel ist der Punkt, für den die Summe der quadrierten Abweichungen minimal ist
 - ▶ OLS findet in diesem speziellen Fall die Parameter, die (wenn es sich um eine Zufallsstichprobe handelt) am ehesten die beobachteten Werte hervorgebracht haben könnten (mehr dazu in zwei Wochen)

Was sind die Abweichungen, die quadriert werden?



Wie komme ich zu meinen Schätzungen?

- ▶ Für alle Datenpunkte $i = 1, 2 \dots n$ Differenz zwischen beobachtetem (y_i) und erwartetem Wert (\hat{y}_i) bestimmen, quadrieren und aufsummieren

$$\text{SAQ} = \sum_{i=1}^n (y_i - (b_0 + b_1 x_{1i}))^2 \quad (1)$$

$$= \sum_{i=1}^n (y_i - b_0 - b_1 x_{1i})^2 \quad (2)$$

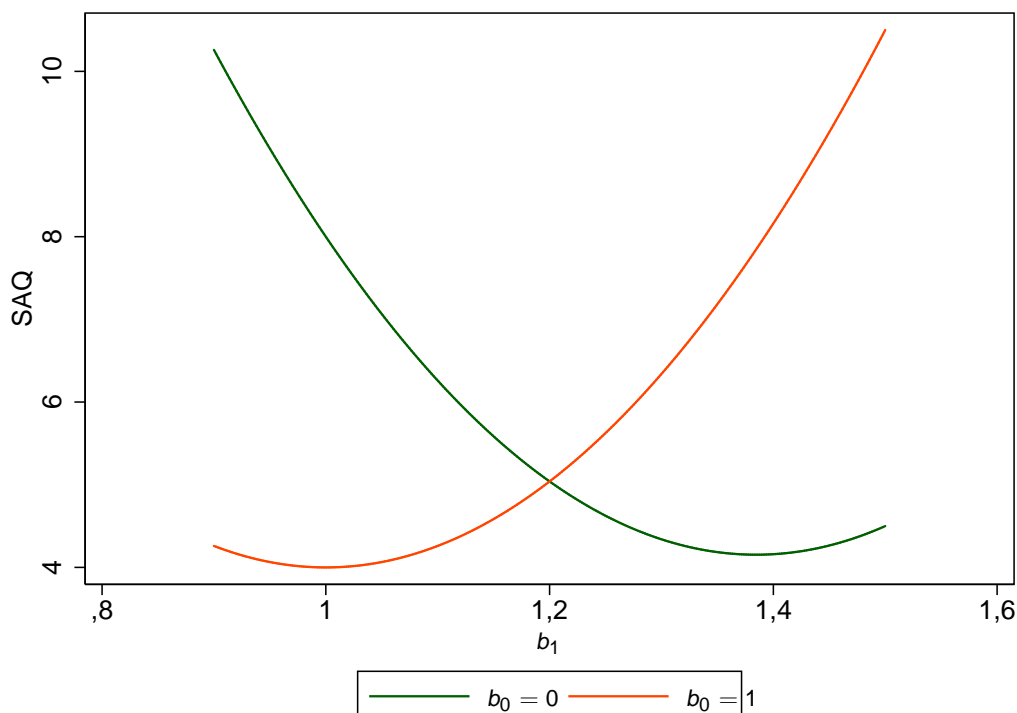
- ▶ Die SAQ in (1) sind eine *Funktion* der Daten und der Parameter
- ▶ Gesucht sind Parameter, die SAQ minimieren

Wie minimiere ich die SAQ?

- ▶ Möglichkeit I: Durch systematisches Variieren der Parameter
- ▶ Entspricht in etwa den iterativen Verfahren (übernächste Sitzung)

Index	y	x
1	4	2
2	2	2
3	5	3
4	3	3

Wie minimiere ich die SAQ?



Wie minimiere ich die SAQ?

- ▶ Möglichkeit II: Es existiert eine analytische Lösung
- ▶ Aus der Abbildung ist klar, daß (1) ein globales Minimum hat
- ▶ Notwendige Bedingung für einen Extremwert: 1. Ableitung gleich 0 (Tangente ist an dieser Stelle flach)
- ▶ Funktion hat zwei Variablen → zwei partielle Ableitungen (nach b_0 und b_1) betrachten

Wie finde ich die partiellen Ableitungen?

$$SAQ = \sum_{i=1}^n (y_i - b_0 - b_1 x_{1i})^2 \quad (3)$$

$$\frac{\partial SAQ}{\partial b_0} = \sum_{i=1}^n -1 \times 2 \times (y_i - b_0 - b_1 x_{1i}) = 0 \quad (4)$$

$$\frac{\partial SAQ}{\partial b_1} = \sum_{i=1}^n -x_{1i} \times 2 \times (y_i - b_0 - b_1 x_{1i}) = 0 \quad (5)$$

Wie finde ich die Werte für b_0 und b_1 ?

- (4) und (5) bilden ein Gleichungssystem, das durch Umformen die sogenannten Normalgleichungen ergibt:

$$n \times b_0 + b_1 \sum_{i=1}^n x_{1i} = \sum_{i=1}^n y_i \quad (6)$$

$$b_0 \sum_{i=1}^n x_{1i} + b_1 \sum_{i=1}^n x_{1i}^2 = \sum_{i=1}^n x_{1i} \times y_i \quad (7)$$

Wie finde ich die Werte für b_0 und b_1 ?

- ▶ Durch Auflösen ergibt sich:

$$b_0 = \bar{y} - b_1 \bar{x}_1 \quad (8)$$

$$b_1 = \frac{\sum (x_{1i} - \bar{x}_1)(y_i - \bar{y})}{\sum (x_{1i} - \bar{x}_1)^2} = \frac{SAP_{xy}}{SAQ_x} \quad (9)$$

Was tun, wenn es mehr als ein x gibt?

- ▶ $y = b_0 x_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k$
- ▶ Wieder SAQ berechnen / partielle Ableitungen bilden und auf null setzen, z. B.

$$SAQ = \sum_{i=1}^n (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \dots - b_k x_{ki})^2 \quad (10)$$

- ▶ z. B.:

$$\frac{\partial SAQ}{\partial b_1} = \sum_{i=1}^n -x_{1i} \times 2 \times (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \dots - b_k x_{ki}) = 0 \quad (11)$$

Wie sehen die Normalgleichungen aus?

$$b_0 \times n + b_1 \sum x_{1i} + b_2 \sum x_{2i} + \dots + b_k \sum x_{ki} = \sum y_i \quad (12)$$

$$b_0 \sum x_{1i} + b_1 \sum x_{1i}^2 + b_2 \sum x_{1i}x_{2i} + \dots + b_k \sum x_{1i}x_{ki} = \sum x_{1i}y_i \quad (13)$$

⋮

$$b_0 \sum x_{ki} + b_1 \sum x_{ki}x_{1i} + b_2 \sum x_{ki}x_{2i} + \dots + b_k \sum x_{ki}^2 = \sum x_{ki}y_i \quad (14)$$

Geht das auch etwas übersichtlicher?

- ▶ Schon bei zwei Variablen sehr unübersichtlich
- ▶ Für den multivariaten Fall Darstellung und Berechnung vorzugsweise in Matrix-Schreibweise
- ▶ Matrix: tabellenförmige Darstellung von Zahlen (Elementen der Matrix)
- ▶ **A** ist eine $m \times n$ Matrix (m Zeilen, n Spalten):

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} \quad (15)$$

- ▶ Matrix mit einer Spalte: Spaltenvektor; Matrix mit einer Zeile: Zeilenvektor

Wie kann man mit Matrizen rechnen?

- ▶ Matrizen werden elementweise addiert (Rechenbeispiele aus Wikipedia)
- ▶ Setzt gleiche Zahl von Spalten Zeilen voraus

$$\begin{pmatrix} 1 & 3 & 2 \\ 1 & 2 & 2 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 5 \\ 2 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 1+0 & 3+0 & 2+5 \\ 1+2 & 2+1 & 2+1 \end{pmatrix} = \begin{pmatrix} 1 & 3 & 7 \\ 3 & 3 & 3 \end{pmatrix}$$

Wie kann man mit Matrizen rechnen?

- ▶ Die Multiplikation mit einem Skalar ist einfach:

$$2 \times \begin{pmatrix} 1 & 3 & 2 \\ 1 & 2 & 2 \end{pmatrix} = \begin{pmatrix} 2 \times 1 & 2 \times 3 & 2 \times 2 \\ 2 \times 1 & 2 \times 2 & 2 \times 2 \end{pmatrix} = \begin{pmatrix} 2 & 6 & 4 \\ 2 & 4 & 4 \end{pmatrix}$$

Wie kann man mit Matrizen rechnen?

- ▶ Die Multiplikation von Matrizen ist spannender
- ▶ Nur möglich, wenn die Spaltenzahl der linken mit der Zeilenzahl der rechten Matrix übereinstimmt
- ▶ $\mathbf{A} \times \mathbf{B} \neq \mathbf{B} \times \mathbf{A}$ (normalerweise)

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} \times \begin{pmatrix} 6 & -1 \\ 3 & 2 \\ 0 & -3 \end{pmatrix} =$$

$$\begin{pmatrix} 1 \times 6 + 2 \times 3 + 3 \times 0 & 1 \times (-1) + 2 \times 2 + 3 \times (-3) \\ 4 \times 6 + 5 \times 3 + 6 \times 0 & 4 \times (-1) + 5 \times 2 + 6 \times (-3) \end{pmatrix} =$$

$$\begin{pmatrix} 12 & -6 \\ 39 & -12 \end{pmatrix}$$

Was kann man sonst noch machen?

- ▶ Transponieren, d. h. Zeilen und Spalten vertauschen

$$\begin{pmatrix} 1 & 8 & -3 \\ 4 & -2 & 5 \end{pmatrix}' = \begin{pmatrix} 1 & 4 \\ 8 & -2 \\ -3 & 5 \end{pmatrix}$$

- ▶ Die Inverse suchen (entspricht etwa dem Kehrwert):
 $\mathbf{A} \times \mathbf{A}^{-1} = \mathbf{I}$
- ▶ \mathbf{I} ist die *Einheitsmatrix*
- ▶ Quadratische Matrix mit Einsen auf der Hauptdiagonale, sonst nur Nullen
- ▶ Inverse ermöglicht es, durch Matrix zu teilen; nicht alle Matrizen sind invertierbar

Was hilft uns das?

- ▶ Das lineare Modell kann in Matrix-Schreibweise sehr kompakt formuliert werden

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \text{mit} \quad \begin{cases} \mathbf{y} : & \text{Spaltenvektor mit Werten der abhängigen Variablen} \\ \mathbf{X} : & \text{Matrix mit Werten der unabhängigen Variablen} \\ \boldsymbol{\beta} : & \text{Spaltenvektor mit Koeffizienten} \\ \boldsymbol{\epsilon} : & \text{Spaltenvektor mit zufälligen Einflüssen} \end{cases}$$

dabei ist

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \quad (16)$$

Was hilft uns das?

- ▶ OLS-Schätzung: $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$ (\mathbf{e} ist der Spaltenvektor der Residuen, \mathbf{b} ist der Spaltenvektor der Koeffizienten, \mathbf{X} ist die Datenmatrix)
- ▶ Die Summe der quadrierten Residuen ist $\mathbf{e}'\mathbf{e}$

$$\text{SAQ} = \mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) \quad (17)$$

$$= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\mathbf{b} - \mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} \quad (18)$$

$$= \mathbf{y}'\mathbf{y} - (2\mathbf{y}'\mathbf{X})\mathbf{b} + \mathbf{b}'(\mathbf{X}'\mathbf{X})\mathbf{b} \quad (19)$$

Was hilft uns das?

- ▶ Die partielle Ableitung der SAQ nach \mathbf{b} ist

$$\frac{\partial \text{SAQ}}{\partial \mathbf{b}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b}$$
- ▶ Auf null setzen: $-2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b} = 0$
- ▶ Vektorform der Normalgleichungen: $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$
- ▶ Nach \mathbf{b} auflösen: $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$

Ein Beispiel gefällig?

- ▶ Unsere Datenmatrix ist $\mathbf{X} = \begin{pmatrix} 1 & 2 \\ 1 & 2 \\ 1 & 3 \\ 1 & 3 \end{pmatrix}$
- ▶ die transponierte Matrix ist $\mathbf{X}' = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 2 & 2 & 3 & 3 \end{pmatrix}$
- ▶ $\mathbf{X}'\mathbf{X} = \begin{pmatrix} 4 & 10 \\ 10 & 26 \end{pmatrix}$ (symmetrische Matrix, Dimension entspricht Zahl der Variablen)
- ▶ Und die Inverse zu diesem Produkt ist

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \frac{13}{2} & -\frac{5}{2} \\ -\frac{5}{2} & 1 \end{pmatrix}$$
- ▶ $(\mathbf{X}'\mathbf{X}) \times (\mathbf{X}'\mathbf{X})^{-1} = \mathbf{I}$

Ein Beispiel gefällig?

$$\blacktriangleright \mathbf{y} = \begin{pmatrix} 4 \\ 2 \\ 5 \\ 3 \end{pmatrix}$$

$$\begin{aligned} \mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \begin{pmatrix} \frac{3}{2} & \frac{3}{2} & -1 & -1 \\ -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 4 \\ 2 \\ 5 \\ 3 \end{pmatrix} \\ &= \begin{pmatrix} 1 \\ 1 \end{pmatrix} \end{aligned}$$

Ein Beispiel gefällig?

- ▶ D. h. wie erwartet sind die Werte für Konstante und Steigung jeweils gleich 1
- ▶ Vektor der erwarteten Werte:

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{X}\mathbf{b} \\ &= \begin{pmatrix} 1 & 2 \\ 1 & 2 \\ 1 & 3 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} 3 \\ 3 \\ 4 \\ 4 \end{pmatrix} \end{aligned}$$

Übung für heute

- ▶ Matrix-Algebra vor allem deshalb so effizient, weil man sie heute mit dem Computer betreiben kann (numerisch oder symbolisch)
- ▶ Versuchen Sie, die Berechnung von \mathbf{b} am Computer nachzuvollziehen
- ▶ STATA enthält eine sehr einfache Matrix-Sprache (<http://personal.rhul.ac.uk/uhte/006/ec3327/Matrix%20Commands%20in%20Stata.pdf>)
 - ▶ Melden Sie sich auf dem Terminalserver an und starten Sie STATA 10
 - ▶ Definieren Sie die Datenmatrix X:
 - ▶ Kommata trennen die Elemente innerhalb einer Zeile, der Backslash trennt die Zeilen voneinander
 - ▶ Durch Anhängen von ' (rechts vom „ä“) können Sie eine Matrix transponieren

- ▶ Arbeiten Sie mit Hilfsmatrizen, in denen Sie Zwischenergebnisse speichern

Was ist das Fazit für heute?

- ▶ Regression betrachtet den konditionalen Mittelwert einer Variablen
- ▶ In Abhängigkeit vom Niveau der unabhängigen Variablen folgt dieser Mittelwert einem Pfad
- ▶ Im klassischen linearen Modell entspricht dieser Pfad der Linie / Fläche / Hyperfläche, die die SAQ minimieren → partielle Ableitungen auf null setzen
- ▶ Das Gleichungssystem, mit dessen Hilfe b_0, b_1, \dots gefunden werden, läßt sich mit Hilfe von etwas Matrix-Algebra sehr effizient analytisch lösen
- ▶ Datenmatrix muß genug unabhängige Informationen enthalten → keiner der Spaltenvektoren darf Linearkombination anderer Vektoren darstellen (perfekte Kollinearität)