

Was ist Regression?

Statistik II

Wiederholung
Regression
Parameterschätzung für die lineare Regression
Zusammenfassung

Übersicht

Wiederholung

Literatur

Regression

Was ist Regression?

Wiederholung: Standardmodell der
linearen Regression

Nomenklatur

Wiederholung:

Wahrscheinlichkeitsverteilungen

Beschreibung und Inferenz

Parameterschätzung für die lineare
Regression

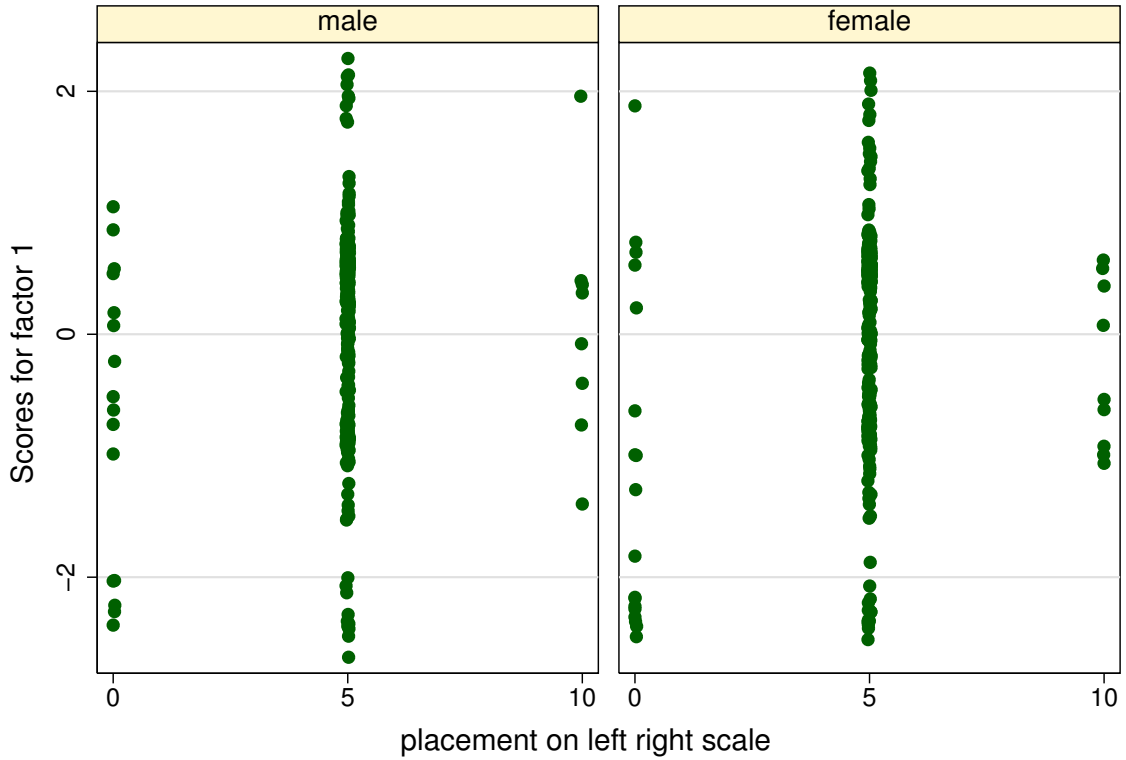
Zusammenfassung

Literatur für heute

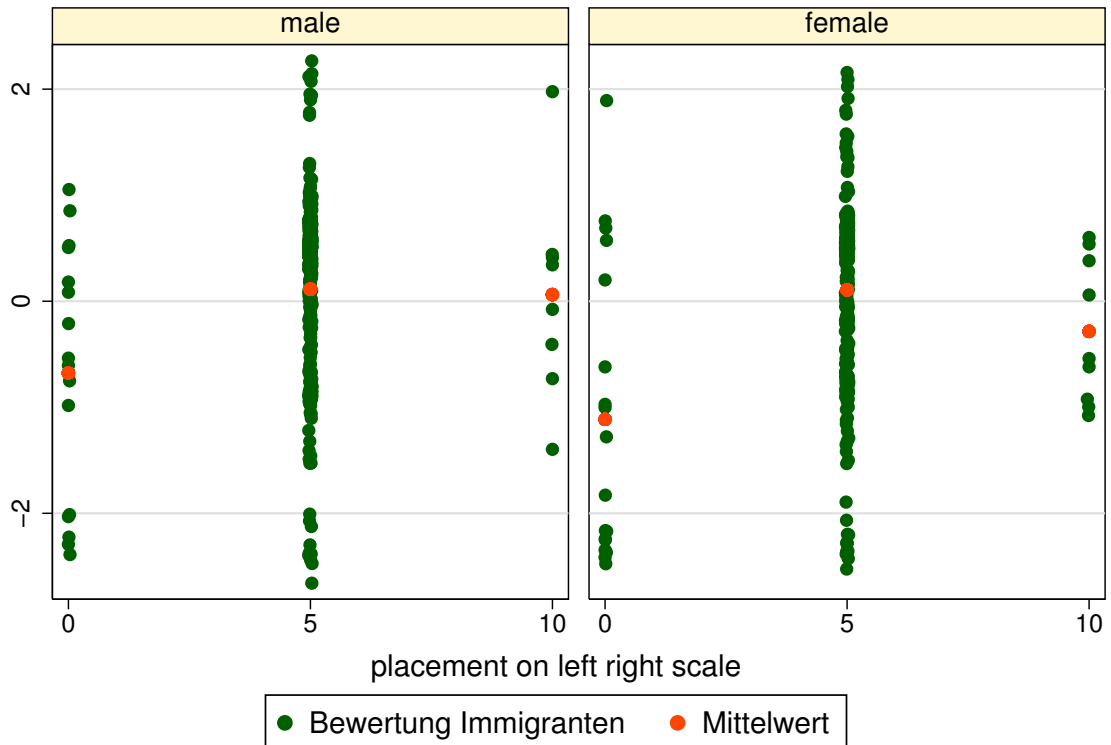
- ▶ Berk (2004, S. 13-17, 39-56) und
- ▶ Fox (1997, S. 86-88, 101, 204-205, 212-213)
- ▶ (beides im ReaderPlus)

Literatur für nächste Woche

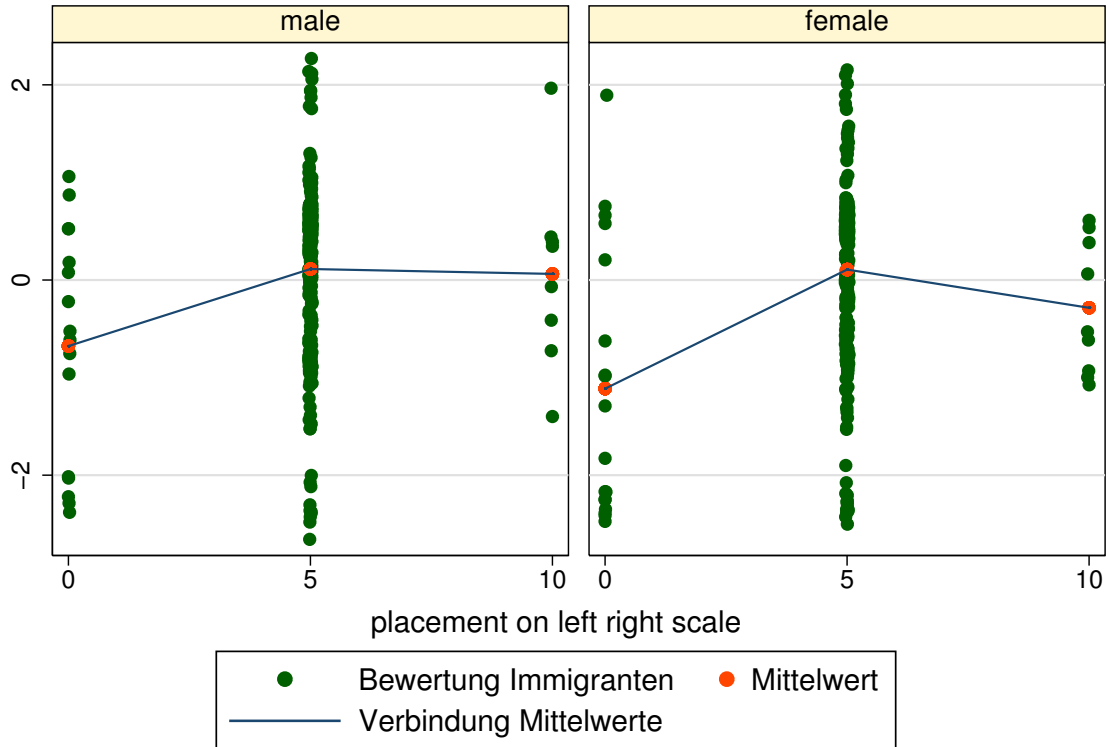
- ▶ Agresti ch. 10



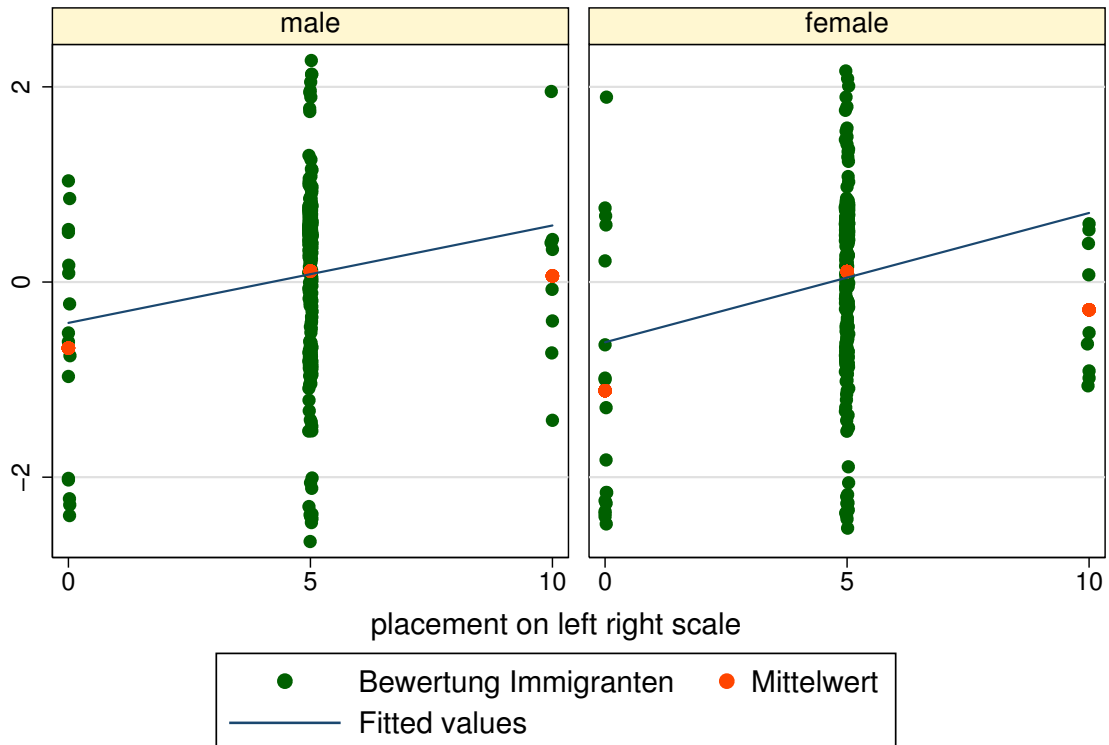
Graphs by gender



Graphs by gender



Graphs by gender



Graphs by gender

Was ist Regression?

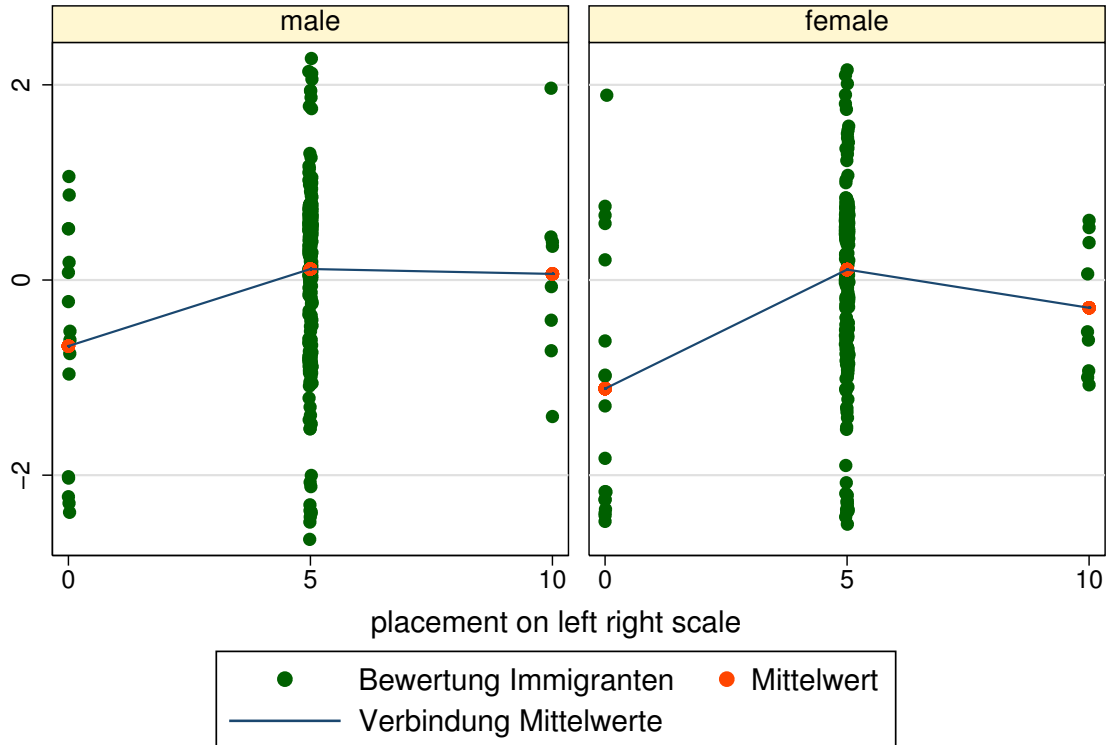
- ▶ Regression ist der Oberbegriff für Verfahren, ...
- ▶ die die *konditionale* Verteilung einer Variablen y ...
- ▶ in Abhängigkeit von einer oder mehreren anderen Variablen $x_1, x_2 \dots x_k$ beschreiben

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?

- ▶ Nomenklatur oft wenig einheitlich
- ▶ Grundregeln:
 1. y für „abhängige“ Variable, x für „unabhängige“ Variable
 2. Variablen, Parameter und Untersuchungseinheiten kann man mit einem Index durchnummerieren: $x_1, x_2 \dots x_k$
 3. Lateinische Buchstaben für Variablen und Parameter in der Stichprobe,
 4. Griechische Buchstaben für die unbekannt Parameter der Grundgesamtheit
 5. Variablen erkennt man am *Kursivdruck*
 6. Für Vektoren verwendet man (griechische oder lateinische) Kleinbuchstaben in **Fettdruck**
 7. Für Matrizen verwendet man (griechische oder lateinische) Großbuchstaben in **Fettdruck**
 8. 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?

- ▶ Zufallsvariablen Ergebnis von Zufallsexperimenten
- ▶ Zufallsvariablen im Regressionsmodell
 - ▶ Zufällige Einflüsse auf einen Fall
 - ▶ Zufällige Variation der Schätzungen bei wiederholter Stichprobenziehung
- ▶ 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
- ▶ *Einfaches lineares Regressionsmodell: stetige Zufallsvariablen wichtig*

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)

Regressionsmodell

- ▶ Hochgradig vereinfachte
- ▶ Nicht unbedingt realistische
- ▶ Mathematisch formalisierte
- ▶ Beschreibung der sozialen Wirklichkeit als Funktion von
 - ▶ systematischen und

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 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 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)

Regressionsmodell

- ▶ Hochgradig vereinfachte
- ▶ Nicht unbedingt realistische
- ▶ Mathematisch formalisierte
- ▶ Beschreibung der sozialen Wirklichkeit als Funktion von
 - ▶ systematischen und

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
 - ▶ **Erfordert Annahmen über Zustandekommen der Daten**
→ **klassische Inferenz**

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

Konfidenzintervall

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

- ▶ Habe ich eine der 95 „glücklichen“ Stichproben gezogen?

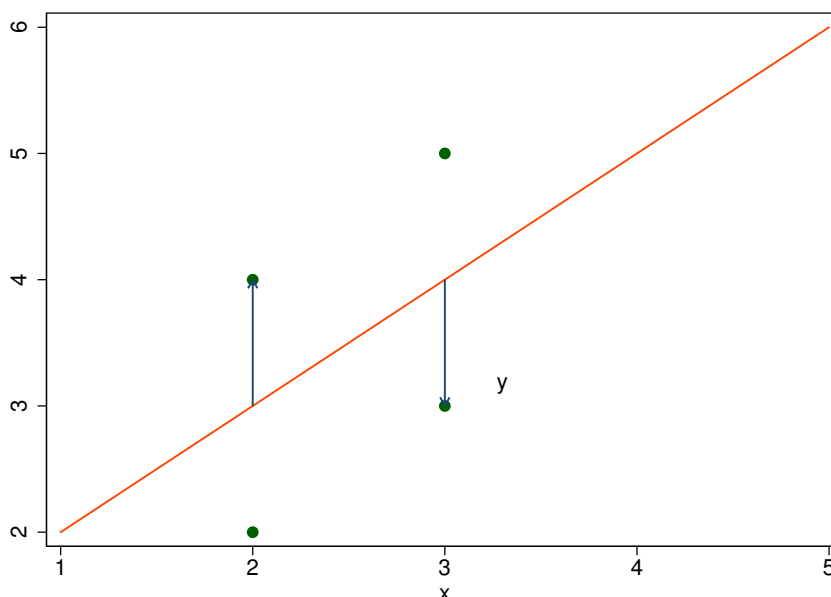
Und wenn ich keine Zufallsstichprobe habe?

- ▶ Schulbezirke, OECD-Staaten, Studierende an einer bestimmten Universität
- ▶ Strategie I: Die Daten werden wie eine Grundgesamtheit behandelt
Regression dient nur zur *Beschreibung*
- ▶ Strategie II (mit Varianten): Annahmen über Natur, Superpopulation, ...
 - ▶ Standardfehler werden „als ob“ berechnet
 - ▶ Innerhalb des klassischen Ansatzes nicht ok
 - ▶ Erfordert andere statistische Annahmen
- ▶ Extreme Vorsicht mit Standardfehlern bei Non-Samples

Wie komme ich zu meinen Schätzungen?

- ▶ Wie lege ich die Gerade durch die Punkte (gute Beschreibung/gute Schätzung)?
- ▶ Standardmethode: „Kleinste-Quadrate-Schätzung“ (Ordinary Least Squares, OLS) Abweichungsquadrate?
 - ▶ Welche Koeffizienten minimieren die SAQ?
 - ▶ Gute Beschreibung/Anpassung
- ▶ Und (in diesem Fall) auch gute *Schätzung* für Grundgesamtheit

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 Parameterschätzungen
- ▶ Gesucht sind *Parameterschätzungen*, die SAQ minimieren

Wie minimiere ich die SAQ?

- ▶ Möglichkeit I:
 - ▶ Durch systematisches Variieren der Parameter
 - ▶ Entspricht in etwa den iterativen Verfahren
- ▶ Möglichkeit II:
 - ▶ Es existiert eine analytische Lösung
 - ▶ Funktion hat globales Minimum
 - ▶ Notwendige Bedingung für einen Extremwert: 1. Ableitung gleich 0 (Tangente ist an dieser Stelle flach)
 - ▶ Funktion hat zwei Variablen \rightarrow zwei partielle Ableitungen (nach b_0 und b_1) betrachten
 - ▶ „Normalgleichungen“

Wie sehen die Normalgleichungen aus?

$$b_0 \times n + b_1 \sum x_{1i} + b_2 \sum x_{2i} + \dots + b_k \sum x_{ki} = \sum y_i \quad (3)$$

$$b_0 \sum x_{1i} + b_1 \sum x_{1i}^2 + b_2 \sum x_{1i}x_{2i} + \dots + b_k \sum x_{1i}x_{ki} = \sum x_{1i}y_i \quad (4)$$

⋮

$$b_0 \sum x_{ki} + b_1 \sum x_{ki}x_{1i} + b_2 \sum x_{ki}x_{2i} + \dots + b_k \sum x_{ki}^2 = \sum x_{ki}y_i \quad (5)$$

Nur zur Illustration, muß nicht auswendig gelernt werden

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):

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} \quad (6)$$

- ▶ Matrix mit einer Spalte: Spaltenvektor; Matrix mit einer Zeile: Zeilenvektor [▶ weiter](#)

Wie kann man mit Matrizen rechnen?

- ▶ Der Stoff auf den nächsten Folien dient Ihrem Verständnis, ist aber nicht klausurrelevant
- ▶ 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

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \quad (7)$$

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 $\mathbf{e}'\mathbf{e}$ (**warum?** – siehe Matrix-Multiplikation drei Folien vorher)

$$\text{SAQ} = \mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) \quad (8)$$

$$= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\mathbf{b} - \mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} \quad (9)$$

$$= \mathbf{y}'\mathbf{y} - (2\mathbf{y}'\mathbf{X})\mathbf{b} + \mathbf{b}'(\mathbf{X}'\mathbf{X})\mathbf{b} \quad (10)$$

Muß nicht auswendig gelernt werden, aber Sie sollten es in groben Zügen verstehen

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: $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$

Muß nicht auswendig gelernt werden, aber Sie sollten es in groben Zügen verstehen

Zusammenfassung

- ▶ Regression betrachtet konditionalen Mittelwert einer Variablen
- ▶ Mittelwert folgt in Abhängigkeit von unabhängigen Variablen einem Pfad
- ▶ Im klassischen Modell entspricht dieser Pfad einer Linie/Fläche/Hyperfläche, die die SAQ minimiert
- ▶ Das Gleichungssystem läßt sich analytisch lösen, um die optimalen Parameter zu finden
- ▶ Matrix muß genug unabhängige Informationen enthalten
- ▶ OLS gutes Mittel zur Datenverdichtung – auch ein gutes Schätzverfahren?