

# Willkommen zur Vorlesung Statistik

Thema dieser Vorlesung:  
Das lineare Regressionsmodell

Prof. Dr. Wolfgang Ludwig-Mayerhofer

Universität Siegen – Philosophische Fakultät, Seminar für Sozialwissenschaften

# Lineare Regression

- Einführung
- Die lineare Einfachregression
- Voraussetzungen
- Multiple Regression









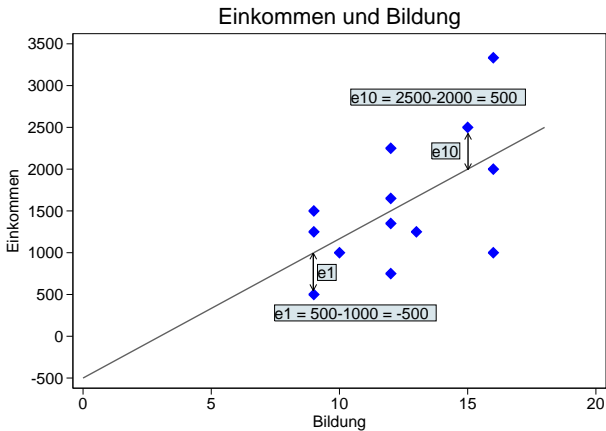








## Die Residuen



# Varianzzerlegung

In Analogie zur Varianzanalyse gilt:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (3)$$

$$QS_{\text{total}} = QS_{\text{Schätzwerte}} + QS_{\text{Residuen}}$$

Die gesamte Streuung in den Daten setzt sich also zusammen aus der Streuung, die wir auf die Modellvorhersage zurückführen können („erklärte Varianz“) und die nicht erklärte Varianz oder Residualvarianz.

# Der Determinationskoeffizient R-Quadrat

In Analogie zu  $\eta^2$  aus der Varianzanalyse wird ein Maß  $R^2$  berechnet, welches das Ausmaß angibt, in dem die Streuung in den Daten durch das Modell erklärt wird:

- Kennen wir die Messwerte der Individuen in  $X$  nicht, ist bester Schätzwert der Gesamtmittelwert (Schwerpunkteigenschaft des arithmetischen Mittels). Der Fehler ( $E_1$ ) ist dann  $QS_{\text{total}}$ .
- Kennen wir den Wert eines Individuums in der unabhängigen Variablen  $X$ , ist der beste Schätzwert für das Individuum  $\hat{y}$ . Der Fehler ( $E_2$ ) ist dann  $QS_{\text{Residuen}}$ .

$$R^2 = \frac{E_1 - E_2}{E_1} = \frac{QS_{\text{total}} - QS_{\text{Residuen}}}{QS_{\text{total}}} = \frac{QS_{\text{Schätzwerte}}}{QS_{\text{total}}} \quad (4)$$

bzw.

$$R^2 = 1 - \frac{E_2}{E_1} = 1 - \frac{QS_{\text{Residuen}}}{QS_{\text{total}}} \quad (5)$$

## Residuen im Beispiel

Einkommen	Bildung	$\hat{y}$	$y - \hat{y}$	$y - \bar{y}$
500,00	9	1000,00	-500,00	-1064,1
1000,00	10	1166,66	-166,66	-564,1
1250,00	13	1666,66	-416,66	-314,1
750,00	12	1500,00	-750,00	-814,1
2000,00	16	2166,66	-166,66	435,9
1500,00	9	1000,00	500,00	-64,1
1250,00	9	1000,00	250,00	-314,1
1650,00	12	1500,00	150,00	85,9
1350,00	12	1500,00	-150,00	-214,1
2500,00	15	2000,00	500,00	935,9
2250,00	12	1500,00	750,00	685,9
3333,33	16	2166,66	1166,66	1769,2
1000,00	16	2166,66	-1166,66	-564,1

$$QS_{\text{total}} = 7\,352\,692; \quad QS_{\text{Schätzwerte}} = 2\,418\,803; \quad QS_{\text{Residuen}} = 4\,933\,889$$

$$R^2 = \frac{7\,352\,692 - 4\,933\,889}{7\,352\,692} \approx 0,3290 \text{ (bzw. } 0,33)$$

# Inferenzstatistik I: Der F-Test

Für das Gesamtmodell lässt sich ein F-Test ganz in Analogie zu dem der Varianzanalyse durchführen. Der Test prüft die Nullhypothese, dass das gesamte Modell keine Verbesserung gegenüber einer einfachen Schätzung des arithmetischen Mittels bringt. Anders gesagt: Es wird die Nullhypothese geprüft, dass sämtliche Steigungskoeffizienten gleich 0 sind.

Die Formel für die Teststatistik lautet

$$F = \frac{MQS_{\text{Schätzwerte}}}{MQS_{\text{Residuen}}} \quad (6)$$

Dabei entspricht die Zahl der Freiheitsgrade für  $MQS_{\text{Schätzwerte}}$  der Anzahl  $k$  der unabhängigen Variablen; die Zahl der Freiheitsgrade für  $MQS_{\text{Residuen}}$  beträgt  $n - 1 - k$ .

## Inferenzstatistik I: Der F-Test im Beispiel

	QS	df	MQS
Total	7 352 692		
Schätzwerte	2 418 803	1	2 418 803
Residuen	4 933 889	11	448 535

$$F = \frac{2\,418\,803}{448\,535} \approx 5,3927$$

Der kritische Wert der F-Verteilung mit (1,11) Freiheitsgraden beträgt (bei  $\alpha = 0,05$ ) 4,844. Die errechnete Teststatistik übersteigt den kritischen Wert; die  $H_0$  wird also abgelehnt.

# Inferenzstatistik I: Erläuterung zum F-Test

Da es sehr viele F-Verteilungen (je nach Freiheitsgraden) gibt, wäre die Ermittlung des kritischen Wertes jeweils ziemlich aufwändig.

Bei Berechnung mit Statistik-Software wird aber ohnehin anders vorgegangen: Die Statistik-Software gibt in der Regel nicht direkt aus, ob eine Teststatistik im Ablehnungsbereich liegt oder nicht (dazu müsste die Software „wissen“, welches Signifikanzniveau gewählt wurde). Sie gibt vielmehr aus, welchem exakten Quantil  $(1 - \alpha)$  der Verteilung die Teststatistik entspricht (sog. p-Wert oder empirisches Signifikanzniveau).

Haben wir beispielsweise ein Signifikanzniveau von  $\alpha = 0,05$  gewählt, so prüfen wir:

Gilt  $p < 0,05$ , wird die  $H_0$  abgelehnt.

Gilt  $p \geq 0,05$ , wird die  $H_0$  beibehalten.

Im Beispiel: Für den F-Wert von 5,3927 ermittelt der Computer einen p-Wert von  $0,0404 < 0,05$ .



## Inferenzstatistik II: Test der einzelnen Koeffizienten

Für die einzelnen Regressionkoeffizienten können Standardfehler  $SE_b$  berechnet werden. Die Teststatistik

$$\frac{b - \beta_{(H_0)}}{SE_b} \quad (7)$$

folgt einer t-Verteilung (mit  $n - k$  Freiheitsgraden) bzw. bei großen Fallzahlen einer Standardnormalverteilung. Dabei ist  $\beta_{(H_0)}$  der Regressionskoeffizient, der in der  $H_0$  postuliert wird (meist 0) und  $k$  die Zahl der unabhängigen Variablen im Modell. (Zur groben Orientierung hinsichtlich der kritischen Werte kann man sich hier nach der Vorlesung Inferenzstatistik, Abschnitt Signifikanztests, Schritt 3 richten.)

Ebenso lassen sich anhand der Standardfehler Konfidenzintervalle für die Regressionkoeffizienten berechnen.

# Voraussetzungen des linearen Regressionsmodells I

Zwei wichtige inhaltliche Voraussetzungen des linearen Regressionsmodells:

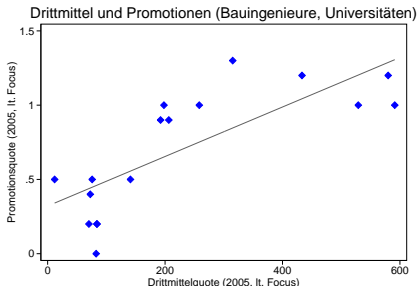
- 1 Alle relevanten Einflussgrößen müssen im Modell enthalten sein.  
Streng genommen muss „nur“ die Bedingung erfüllt sein, dass nicht im Modell enthaltene Einflüsse nicht mit den Residuen korrelieren. Dies dürfte aber nur selten zutreffen.
- 2 Die untersuchten Zusammenhänge müssen (annähernd) linear sein oder linear gemacht werden.

Im Fall der Einfachregression dürfte die erste Bedingung meist verletzt sein (außer bei einem experimentellen Design).

Die zweite Bedingung lässt sich bei der Einfachregression anhand des Streudiagramms von unabhängiger und abhängiger Variablen prüfen.

# Voraussetzungen des linearen Regressionsmodells I

Nicht-Linearität: Das lineare Regressionsmodell schätzt immer eine Gerade, solange wir nicht explizit einen nicht-linearen Zusammenhang modellieren (siehe die folgenden aus der Vorlesung zu Korrelation bekannten Daten):



Über das Umgehen mit nicht-linearen Zusammenhängen mehr in der nächsten Woche (Sie kennen das aber aus der Schule!)

# Voraussetzungen des linearen Regressionsmodells II

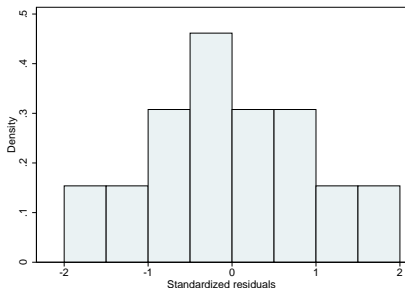
Außerdem sind drei Voraussetzungen für die Gültigkeit der inferenzstatistischen Tests zu prüfen:

- 1 Die Residuen sollten untereinander nicht korrelieren.
- 2 Die Residuen sollten (annähernd) normalverteilt sein.
- 3 Die Streuung der Residuen sollte über den ganzen Wertebereich der abhängigen Variablen in etwa konstant sein (sog. Homoskedastizität).

Ist die **erste Voraussetzung** verletzt, spricht man von Autokorrelation. Für diese gibt es eine Test-Statistik (Durbin-Watson-Statistik), wichtiger ist jedoch Wissen um das Zustandekommen der Daten. Hat man es mit Individualdaten aus einer echten Zufallsstichprobe zu tun, kann man davon ausgehen, dass keine Autokorrelation vorliegt.

# Voraussetzungen des linearen Regressionsmodells II

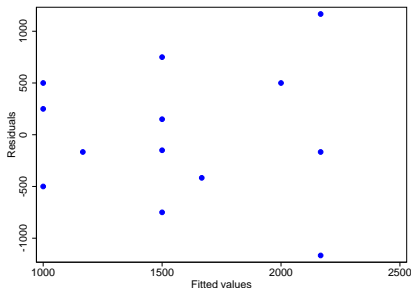
**Inferenzstatistische Voraussetzung 2:** Die Residuen sollten (annähernd) normalverteilt sein. Prüfung anhand eines Diagramms der Residuen (Stamm-Blatt-Diagramm, Kern-Dichte-Schätzer oder – wie hier – Histogramm):



Bei  $n = 13$  dürfen wir keine perfekte Übereinstimmung mit einer Normalverteilung erwarten.

# Voraussetzungen des linearen Regressionsmodells II

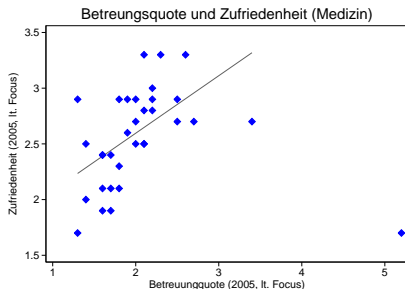
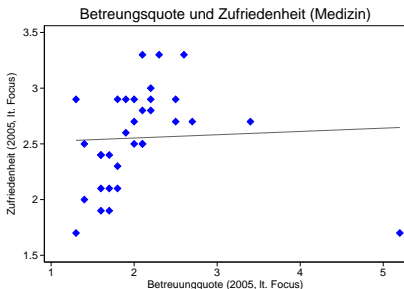
**Inferenzstatistische Voraussetzung 3:** Es sollte Homoskedastizität (oder, anders formuliert: keine Heteroskedastizität) vorliegen. Prüfung durch Streudiagramm der vorhergesagten Werte vs. Residuen:



Dass Heteroskedastizität vorliegt, konnten wir (weil Einfachregression) bereits im Diagramm der abhängigen gegen die unabhängige Variable sehen.

# Voraussetzungen des linearen Regressionsmodells III

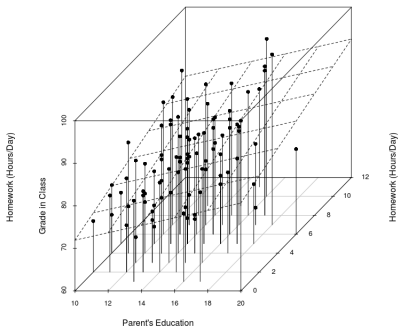
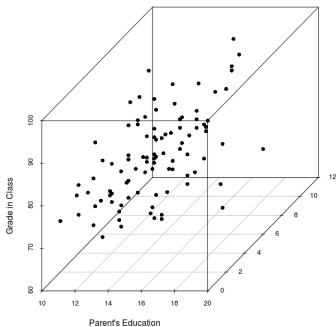
Die Ergebnisse der Modellschätzung sollten nicht zu stark von einzelnen Fällen beeinflusst werden.



Links: Regressionsgerade unter Berücksichtigung aller Fälle; rechts: Regressionsgerade ohne Ausreißer rechts unten.

# Multiple Regression

In der multiplen Regression haben wir es nicht mehr mit einer Regressionsgeraden, sondern einer (u. U. mehr als 3-dimensionalen) Regressions(hyper)ebene zu tun (Graphiken by courtesy of Nathaniel Raley, University of Texas at Austin).





# Das multiple Regressionsmodell

Die rein algebraische Interpretation einer multiplen Regressionsgleichung ist hingegen nicht schwerer als die einer Einfachregression:

$$\hat{y}_i = b_0 + b_1x_{1i} + b_2x_{2i} + \cdots + b_kx_{ki} \quad (8)$$

$k$  = Index für die unabhängigen Variablen.

Ziel der Schätzung multipler Regressionsmodelle (und vieler verwandter Modelle) ist es, die „Netto-Einflüsse“ mehrerer unabhängiger Variablen ermitteln zu können, also die „reinen“ Einflüsse einzelner Variabler, die übrig bleiben, wenn

# Multiple Regression: Die Daten

Wir fügen Daten zur Betriebszugehörigkeitsdauer (in Jahren) und dem Geschlecht (0 = Mann, 1 = Frau) hinzu.

Einkommen	Bildung	Dauer	Geschlecht
500,00	9	21	1
1000,00	10	2	0
1250,00	13	8	1
750,00	12	33	0
2000,00	16	17	1
1500,00	9	15	0
1250,00	9	3	1
1650,00	12	23	0
1350,00	12	25	1
2500,00	15	12	1
2250,00	12	14	0
3333,33	16	20	1
1000,00	16	8	0

## Multiple Regression: Modell mit zwei uV

Nehmen wir zusätzlich zu Bildung ( $x_1$ ) die Betriebszugehörigkeitsdauer ( $x_2$ ) als uV in das Modell auf, erhalten wir folgende Gleichung:

$$\hat{y}_i = -617,8 + 124,5x_{1i} + 41,4x_{2i}$$
$$R^2 = 0,534$$

Die Steigungskoeffizienten (oder Regressionsgewichte)  $b_1$  und  $b_2$  geben den jeweils um die andere Variable „bereinigten“ Einfluss wieder; man sagt, der Einfluss der anderen Variablen wurde „auspartialisiert“.

Im vorliegenden Fall zeigt sich, dass der Einfluss von Bildung deutlich geringer ist als in der Einfachregression. Bildung und Betriebszugehörigkeitsdauer hängen offenbar zusammen; ist letztere nicht im Modell enthalten, so wird derjenige Teil ihres Einflusses, der mit Bildung zusammenhängt, dem Koeffizienten für Bildung „zugerechnet“.

## Multiple Regression: Standardisierte Regressionskoeffizienten

Da die Variablen „Bildung“ und „Betriebszugehörigkeitsdauer“ unterschiedliche Dimensionen haben (Bildung variiert von 9 bis 16, Betriebszugehörigkeitsdauer von 2 bis 33), sind die Regressionskoeffizienten nicht gut vergleichbar. Abhilfe schafft die Schätzung eines Modells mit standardisierten Variablen; man erhält dadurch standardisierte Regressionskoeffizienten.

Die standardisierten Regressionskoeffizienten lassen sich auch direkt aus den unstandardisierten Koeffizienten und den Standardabweichungen der Variablen berechnen, im Beispiel:

$$b_1^* = b \frac{\hat{\sigma}_{x1}}{\hat{\sigma}_y} \quad b_2^* = b \frac{\hat{\sigma}_{x2}}{\hat{\sigma}_y}$$

mit  $b_1^*$  und  $b_2^*$  als den standardisierten Koeffizienten und  $\hat{\sigma}_y$ ,  $\hat{\sigma}_{x1}$  und  $\hat{\sigma}_{x2}$  den Standardabweichungen der Variablen im Modell.

# Multiple Regression: Standardisierte Regressionskoeffizienten

In unserem Beispiel erhalten wir:

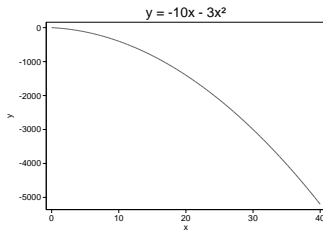
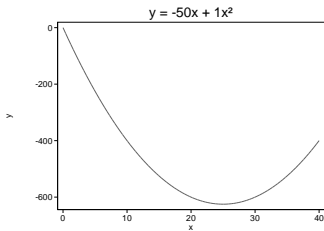
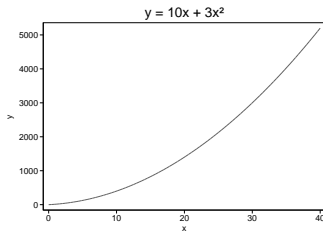
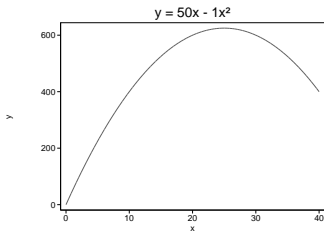
$$b_1^* = 124,5206 \cdot \frac{2,693772}{782,7671} = 0,42852 \approx 0,43$$

$$b_2^* = 41,37846 \cdot \frac{8,996438}{782,7671} = 0,47557 \approx 0,48$$

Der Einfluss der Betriebszugehörigkeitsdauer ist (in unseren fiktiven Daten) sogar etwas größer als der der Bildung.

# Nicht-lineare Zusammenhänge

Nicht-lineare Zusammenhänge können durch Polynome modelliert werden. Oft genügt ein Polynom 2. Grades.



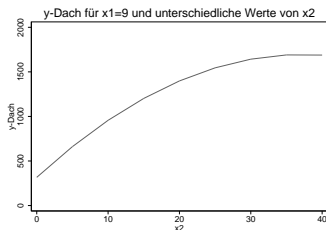
## Nicht-lineare Zusammenhänge im Beispiel

Der Einfluss von Betriebszugehörigkeitsdauer ist meist nicht linear – das Gehalt steigt am Anfang rasch, später langsamer. Das Modell

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{2i}^2 \quad (9)$$

ergibt, angewendet auf unsere Daten, folgende Koeffizienten:

$$\hat{y}_i = -786 + 122,5x_{1i} + 73,9x_{2i} - 0,99x_{2i}^2$$



$$x_1 = 9, x_2 = 20: \hat{y}_i = -786 + 122,5 \cdot 9 + 73,9 \cdot 20 - 0,99 \cdot 400 = 1398,5$$

## Binäre unabhängige Variablen

Eine binäre Variable wie Geschlecht ( $x_3$ ) kann unproblematisch als uV in ein lineares Modell aufgenommen werden, da sich durch zwei Punkte eine Gerade legen lässt. Das Modell

$$\hat{y}_i = b_0 + b_1x_{1i} + b_2x_{2i} + b_3x_{3i} \quad (10)$$

ergibt, angewendet auf unsere Daten, folgende Koeffizienten:

$$\hat{y}_i = -546 + 125,4x_{1i} + 41,0x_{2i} - 143,3x_{3i}$$

Wir können sagen: Wenn die Variable  $x_3$  (Geschlecht) „um eine Einheit zunimmt“, nimmt das Einkommen um 143,3 € ab. Da „um eine Einheit zunehmen“ hier bedeutet: weiblich (im Vergleich zu männlich) zu sein, heißt das, dass Frauen um gut 143 € weniger verdienen als Männer – und dies unter Berücksichtigung möglicher Unterschiede zwischen Frauen und Männern in Bildung und Betriebszugehörigkeitsdauer (die in unseren Daten jedoch praktisch nicht vorhanden sind).



## Korrigiertes oder adjustiertes $R$ -Quadrat

Je mehr unabhängige Variablen ein Regressionsmodell enthält, desto höher ist  $R^2$ . Der Einfluss, den die Variablen in der Grundgesamtheit haben, wird jedoch (gerade in kleineren Datensätzen) tendenziell überschätzt. Korrektur:

$$R_{adj}^2 = 1 - \frac{n-1}{n-k-1} (1 - R^2)$$

mit  $k$  als der Zahl der unabhängigen Variablen.

In unserem Beispiel (Modell 10) ergibt sich ein  $R^2$  von 0,5431.

$$R_{adj}^2 = 1 - \frac{13-1}{13-3-1} (1 - 0,5431) = 0,3908$$

Der große Unterschied zwischen  $R^2$  und  $R_{adj}^2$  kommt durch die sehr kleine Stichprobe zustande.

# Prüfung der Modellvoraussetzungen

Im multiplen Regressionsmodell ist noch mehr als bei der Einfachregression die Prüfung der Modellvoraussetzungen erforderlich. Die meisten Voraussetzungen und die Verfahren ihrer Prüfung haben wir bereits kennengelernt.

Im multiplen Regressionsmodell kommt noch eine wesentliche Bedingung hinzu: Bestehen starke Zusammenhänge zwischen den unabhängigen Variablen, spricht man von Multikollinearität. Multikollinearität kann problematisch sein:

- *Perfekte* Multikollinearität (eine der unabhängigen Variablen kann exakt aus den anderen uV vorhergesagt werden) führt dazu, dass das Modell nicht geschätzt werden kann.
- *Hohe* Multikollinearität (eine oder mehrere der unabhängigen Variablen können zu beträchtlichen Teilen aus den anderen uV vorhergesagt werden) kann dazu führen, dass die Standardfehler der Regressionskoeffizienten stark überhöht sind und somit tatsächlich vorhandene Einflüsse nicht statistisch absicherbar sind.

Mehr zum Umgang mit Multikollinearität im Master-Studiengang!