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

Was ist ein Regressionsmodell? I

Mittels eines Regressionsmodells werden *Zusammenhänge zwischen mehreren Merkmalen* durch ein *mathematisches Modell* abgebildet. Dabei wird grundsätzlich zwischen abhängigen und unabhängigen Variablen unterschieden.

Üblicherweise werden Modelle mit einer abhängigen und einer oder mehreren unabhängigen Variablen geschätzt. Es lassen sich auch Modelle mit mehreren abhängigen Variablen formulieren, diese sind jedoch in den Sozialwissenschaften eher selten und übersteigen auf jeden Fall den Horizont einer Einführung.

Das lineare Regressionsmodell ist für *metrische abhängige Variable* geeignet. Grundsätzlich sind die unabhängigen Variablen ebenfalls metrisch, allerdings lassen sich nicht-metrische Variablen auf eine bestimmte Art und Weise im Modell „unterbringen“.

Es wird unterstellt, dass der Einfluss der unabhängigen Variablen auf die abhängige Variable *linear* ist (wie der Name des Modells schon sagt). Nicht-lineare Einflüsse lassen sich allerdings ebenfalls durch geeignetes Vorgehen modellieren.

Was ist ein Regressionsmodell? II

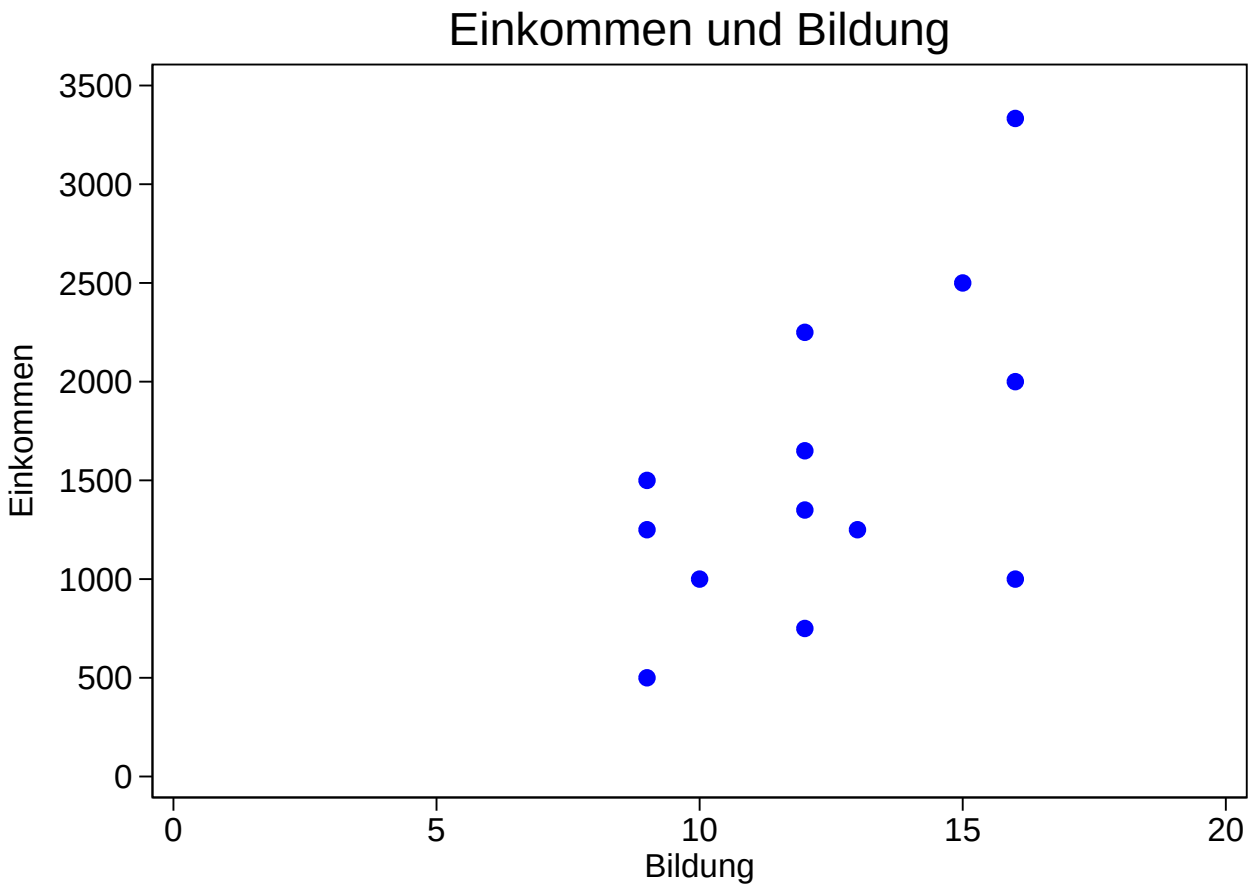
Man unterscheidet zwischen der

- Einfachregression (Erklärung einer abhängigen durch *eine* unabhängige Variable) und der
- Mehrfachregression oder multiplen Regression (Erklärung einer abhängigen durch *mehrere* unabhängige Variable).

Wie der Name schon sagt, handelt es sich bei der linearen Regression um ein *Modell*. Dessen Angemessenheit für die Analyse gegebener Daten ist jeweils zu prüfen.

Lineare Einfachregression: Der Ausgangspunkt

Es sind zwei metrische Variablen gegeben, von denen die eine (Bildung) als Ursache der anderen (Einkommen) angenommen werden kann.



Lineare Einfachregression: Der Ausgangspunkt

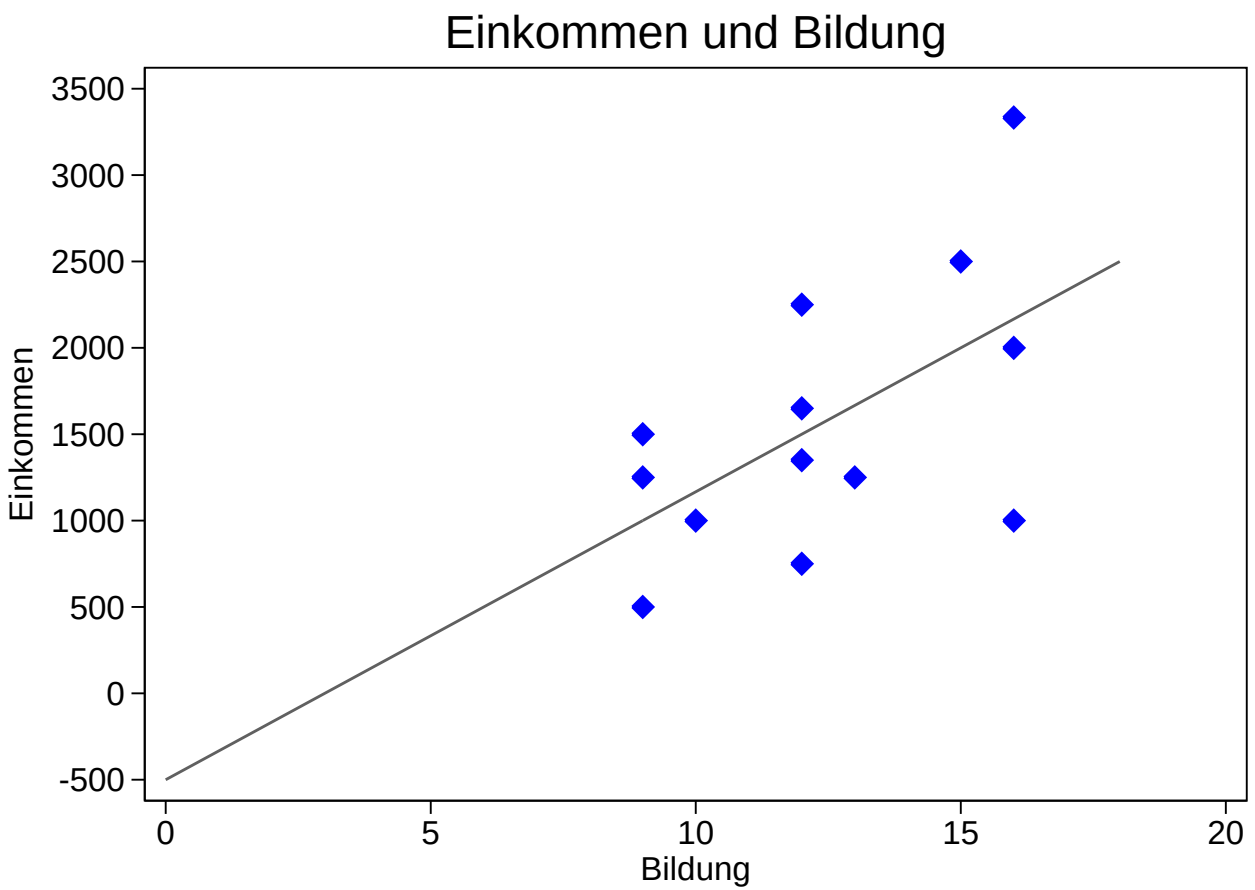
Die Daten

Einkommen (y)	Bildung (x)
500,00	9
1000,00	10
1250,00	13
750,00	12
2000,00	16
1500,00	9
1250,00	9
1650,00	12
1350,00	12
2500,00	15
2250,00	12
$3333,\overline{33}$	16
1000,00	16

$\bar{y} \approx 1564{,}1$ $\bar{x} \approx 12{,}4$

Lineare Einfachregression: Das Modell

Die Grundidee der linearen Regression: Wir suchen eine Gerade, die den Zusammenhang zwischen den beiden Merkmalen möglichst gut beschreibt.



Einfaches Regressionsmodell

Eine Gerade in einem cartesischen Koordinatensystem können wir durch die Gleichung

$$y_i = b_0 + b_1 x_i \tag{1}$$

beschreiben. Im linearen Regressionsmodell schreiben wir

$$\widehat{y}_i = b_0 + b_1 x_i \ , \tag{2}$$

um zu verdeutlichen, dass die $\widehat{y}_i$-Werte (sprich: y-Dach) nicht die empirisch beobachteten, sondern die aus dem Modell vorhergesagten Werte sind.

Die gesuchte Gerade soll so gewählt werden, dass die quadrierten Abstände zwischen den durch das Modell vorhergesagten und den empirisch beobachteten Werten minimiert werden; es soll also gelten:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \text{Minimum}$$

. Die Werte $e_i = y_i - \widehat{y}_i$ werden als Residuen bezeichnet.

Das Modell

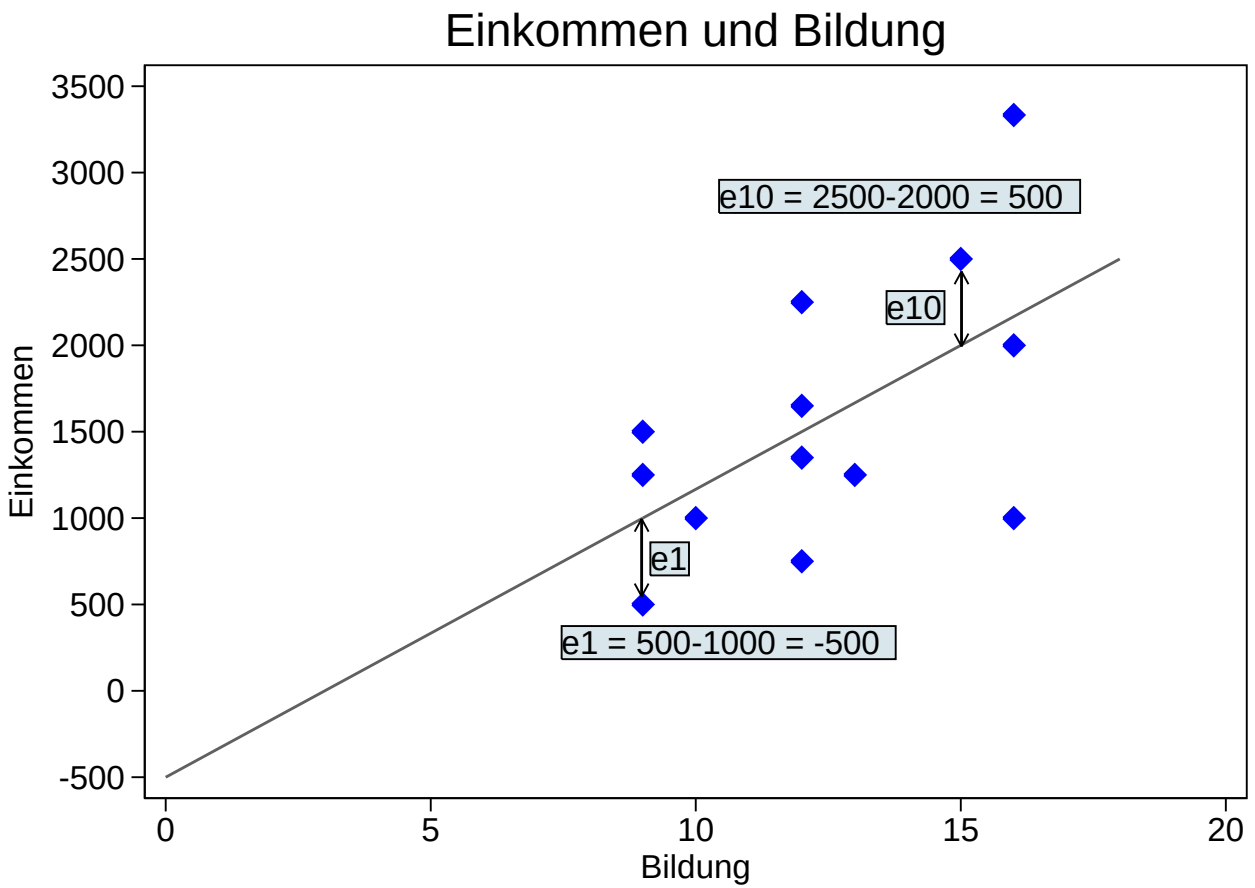
Bei den Größen b_0 und b_1 handelt es sich um Schätzungen für die unbekannten Parameter β_0 und β_1. (Dementsprechend werden sie oft auch als $\widehat{\beta}_0$ und $\widehat{\beta}_1$ bezeichnet.) In unserem Fall ergibt sich

$$\widehat{y_i} = -500 + 166,\overline{66}x_i \text{ bzw.}$$
$$y_i = -500 + 166,\overline{66}x_i + e_i$$

b_0 und b_1 werden als Regressionskoeffizienten bezeichnet; dabei ist b_0 die (Regressions-)Konstante und b_1 der Koeffizient (auch: Regressionsgewicht, engl. *regression weight*) für die Variable X_1 (also die erste – und in unserem Falle einzige – unabhängige Variable).

Für die Berechung von b_0 und b_1 in der linearen Einfachregression finden Sie Formeln in verschiedenen Lehrbüchern. Da jedoch diese Formeln im Falle der (weitaus wichtigeren) Mehrfachregression nicht funktionieren (in diesem Fall ist die Berechnung mit der üblichen mathematischen Hausapotheke nicht mehr zu bewerkstelligen), verzichte ich gleich auf die Angabe der Formeln.

Die Residuen



Varianzzerlegung

In Analogie zur Varianzanalyse gilt:

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

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

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 η^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_{total}.
- Kennen wir den Wert eines Individuums in der unabhängigen Variablen X, ist der beste Schätzwert für das Individuum $\widehat{y}$. Der Fehler (E_2) ist dann QS_{Residuen}.

$$R^2 = \frac{E_1 - E_2}{E_1} = \frac{\text{QS}_{\text{total}} - \text{QS}_{\text{Residuen}}}{\text{QS}_{\text{total}}} = \frac{\text{QS}_{\text{Schätzwerte}}}{\text{QS}_{\text{total}}} \tag{4}$$

bzw.

$$R^2 = 1 - \frac{E_2}{E_1} = 1 - \frac{\text{QS}_{\text{Residuen}}}{\text{QS}_{\text{total}}} \tag{5}$$

Residuen im Beispiel

Einkommen	Bildung	$\widehat{y}$	$y - \widehat{y}$	$y - \bar{y}$
500,00	9	1000,00	−500,00	−1064,1
1000,00	10	1166,$\overline{66}$	−166,$\overline{66}$	−564,1
1250,00	13	1666,$\overline{66}$	−416,$\overline{66}$	−314,1
750,00	12	1500,00	−750,00	−814,1
2000,00	16	2166,$\overline{66}$	−166,$\overline{66}$	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,$\overline{33}$	16	2166,$\overline{66}$	1166,$\overline{66}$	1769,2
1000,00	16	2166,$\overline{66}$	−1166,$\overline{66}$	−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

$$F = \frac{\text{MQS}_{\text{Schätzwerte}}}{\text{MQS}_{\text{Residuen}}} \tag{6}$$

Dabei entspricht die Zahl der Freiheitsgrade für $\text{MQS}_{\text{Schätzwerte}}$ der Anzahl k der unabhängigen Variablen; die Zahl der Freiheitsgrade für $\text{MQS}_{\text{Residuen}}$ 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

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

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 Voraussetzunges 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.

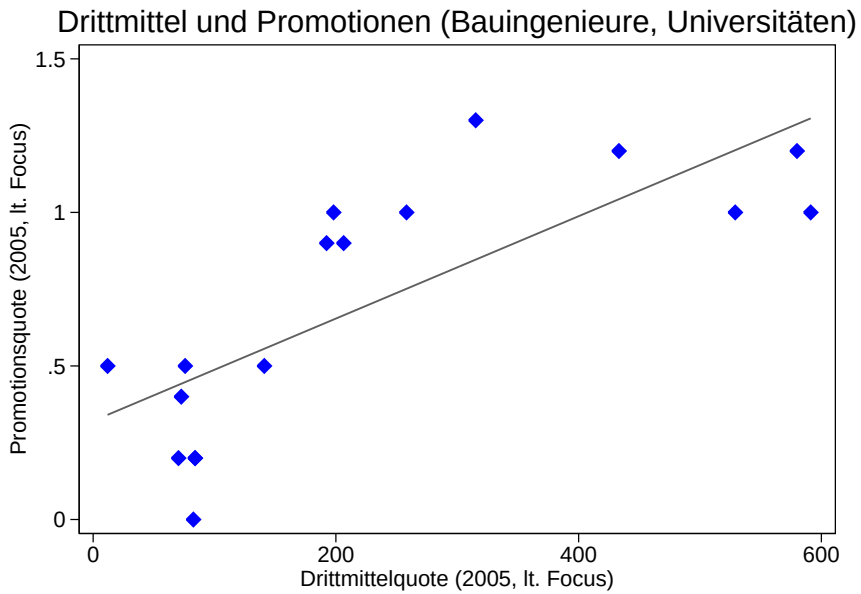
2. Die untersuchten Zusammenhänge müssen (annähernd) linear sein oder linear gemacht werden.

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

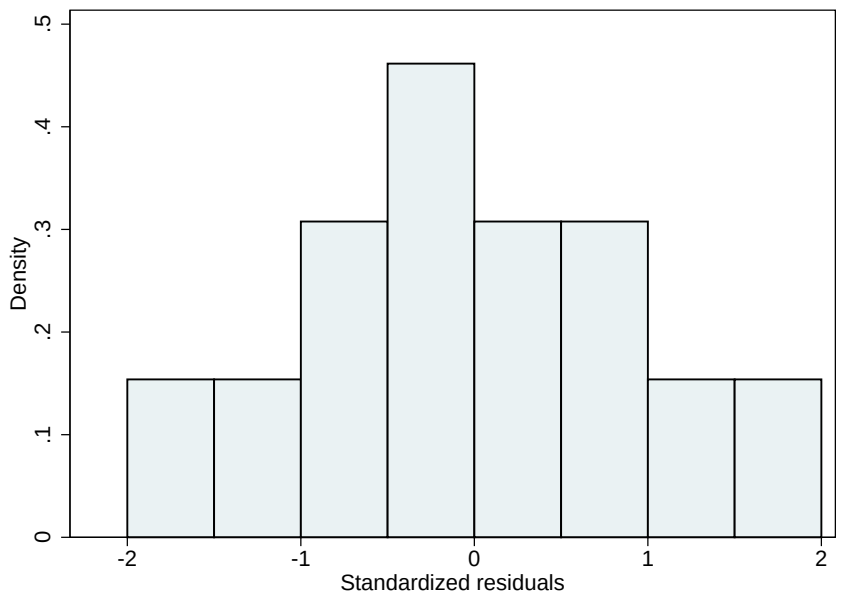
Außerdem sind drei Voraussetzungen für die Gültigkeit der inferenzstatistischen Tests zu prüfen:

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

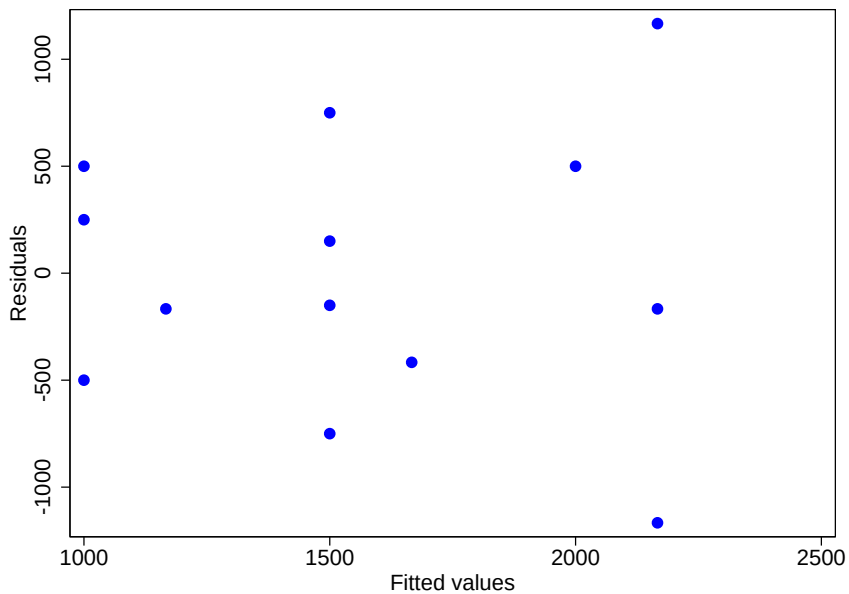
Inferenzstatistische Voraussetzung 2: Die Residuen sollten (annähernd) normalverteilt sein. Prüfung anhand eines Diagramms der Residuen (Stamm-Blatt-Diagramm, Kern-Dichte-Schätzer oder – wie hier – Histogramm):



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

Voraussetzungen des linearen Regressionsmodells II

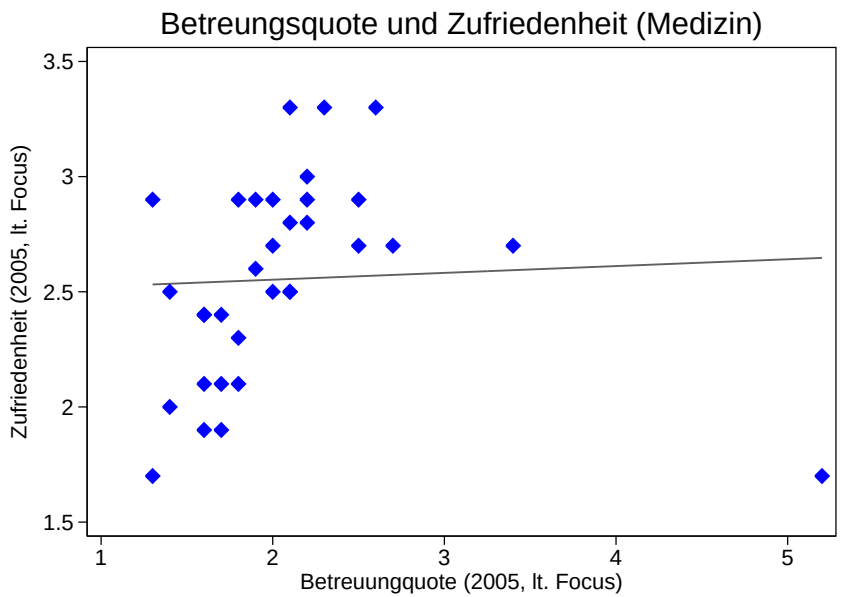
Inferenzstatistische Voraussetzung 3: Es sollte Homoskedastizität (oder, anders formuliert: keine Heteroskedastizität) vorliegen. Prüfung durch Streudiagramm der vorhergesagten Werte vs. Residuen:

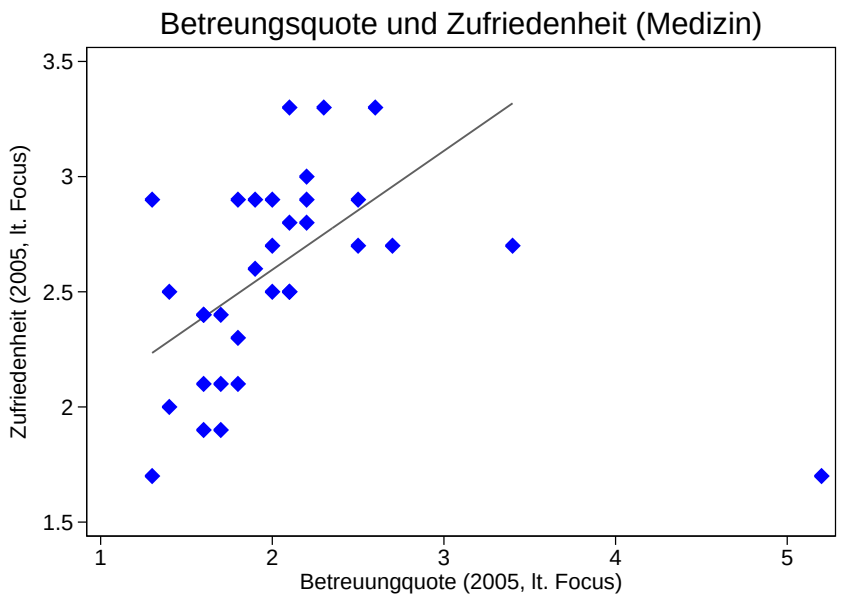


Dass Heteroskedastizität vorliegt, konnten wir (weil Einfachregression) bereits im Diagramm der abhängigen gegen die unabhängige Variable sehen.

Voraussetzungen des linearen Regressionsmodells III

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

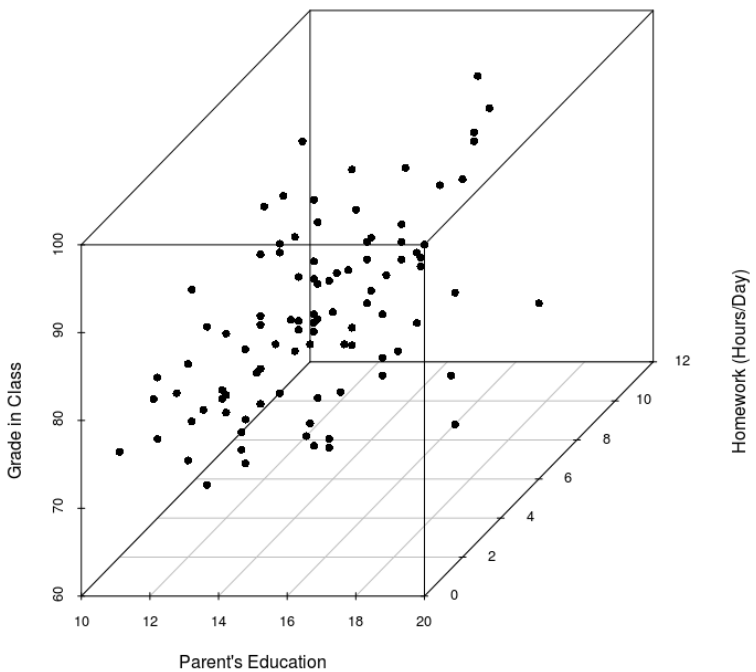


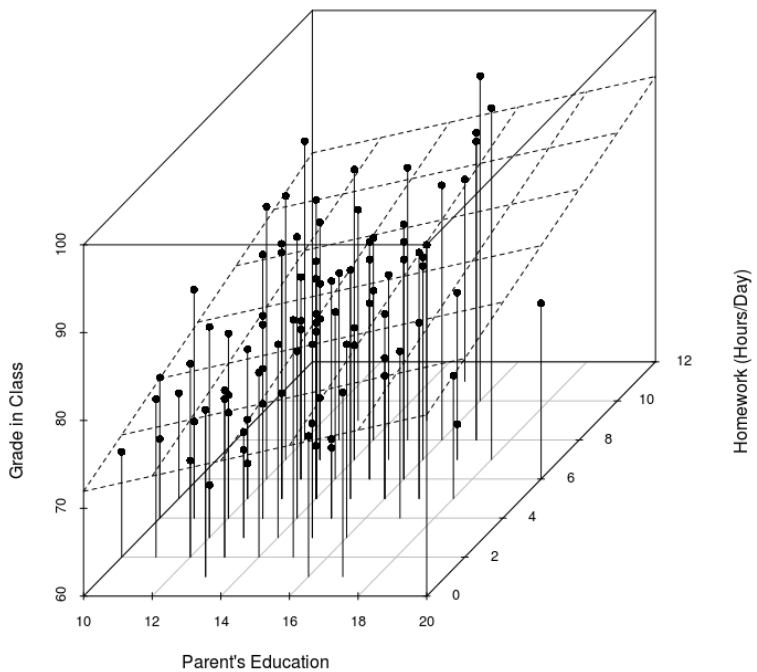


Links: Regressionsgerade unter Berücksichtigung aller Fälle; rechts: Regressionsgerade ohne Ausreißer rechts unten.

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).





Quelle: `http://blog.nathanielraley.com/2014/02/hey-eigenface-diy-face-recognition.html`

Das multiple Regressionsmodell

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

$$\widehat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki} \tag{8}$$

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 die anderen Einflüsse bereits berücksichtigt sind.

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
$3333,\overline{33}$	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:

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

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:

$$b_1^* = b_1 \frac{\widehat{\sigma}_{x1}}{\widehat{\sigma}_y} \qquad b_2^* = b_2 \frac{\widehat{\sigma}_{x2}}{\widehat{\sigma}_y}$$

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

Multiple Regression: Standardisierte Regressionskoeffizienten

In unserem Beispiel erhalten wir:

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

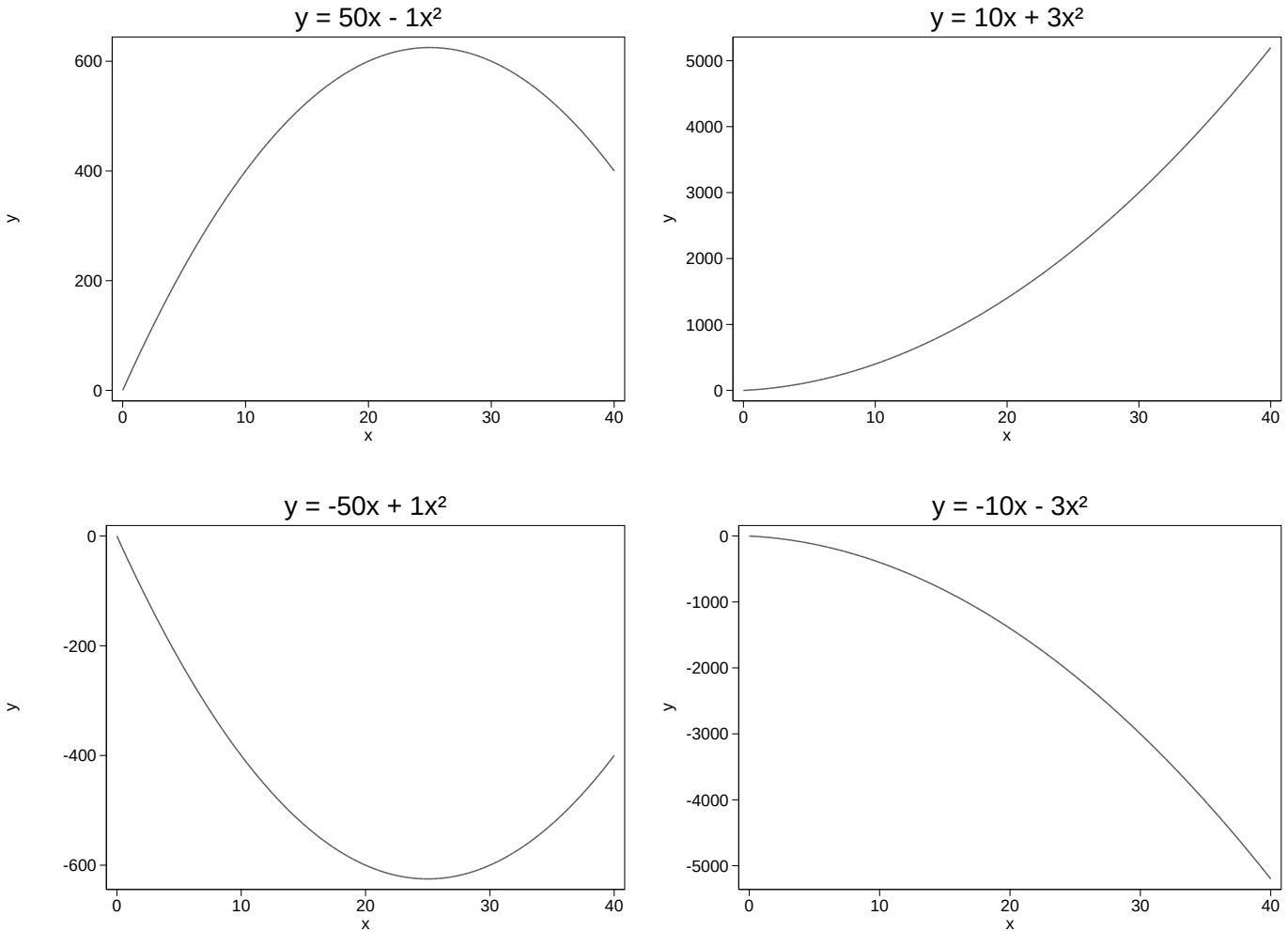
$$b_2^* = 41{,}37846 \cdot \frac{8{,}996438}{782{,}7671} = 0{,}47557 \approx 0{,}48$$

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

Inhaltliche Interpretation: Wenn eine unabhängige Variable um eine Standardabweichung zunimmt, dann ändert sich die abhängige Variable um b_1^* bzw. b_1^* Standardabweichungen.

Nicht-lineare Zusammenhänge

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



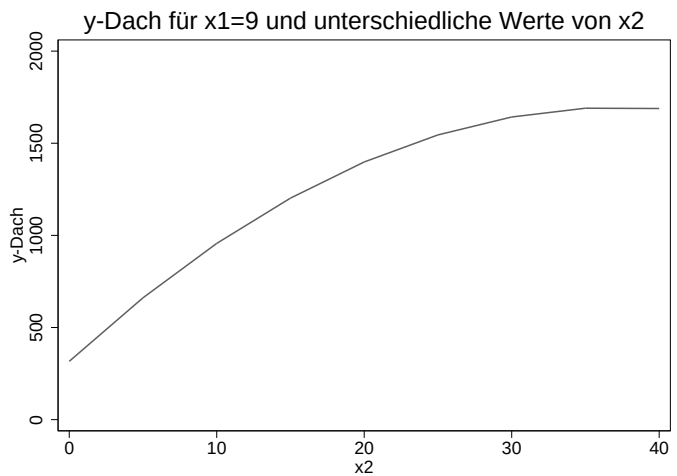
Nicht-lineare Zusammenhänge im Beispiel

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

$$\widehat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{2i}^2 \tag{9}$$

ergibt, angewendet auf unsere Daten, folgende Koeffizienten:

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



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

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

$$\widehat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} \tag{10}$$

ergibt, angewendet auf unsere Daten, folgende Koeffizienten:

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

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:

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

mit k als der Zahl der unabhängigen Variablen.

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

$$R^2_{adj} = 1 - \frac{13-1}{13-3-1}\left(1 - 0{,}5431\right) = 0{,}3908$$

Der große Unterschied zwischen R^2 und R^2_{adj} 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!