SE Geographie und Ökonomie

Einheit 5:
Lineares Regressionsmodell

Dieter Pennerstorfer

dieter.pennerstorfer@jku.at

Institut für Volkswirtschaftslehre

WS 2021/22

Motivation

- Bisher haben wir den **Zusammenhang zweier Variablen** anhand der Korrelation gemessen.
- Der **Korrelationskoeffizient** bildet die Stärke dieses Zusammenhanges ab.
- Korrelationen sagen aber nichts über Wirkungszusammenhänge aus.
- Um solch einen **Wirkungszusammenhang** darzustellen, benötigen wir sogenannte **Regressionsmodelle**.

- Literaturgrundlage: von Auer (2003 oder 2016): Ökonometrie – Eine Einführung

Lernziele der Einheit 5

- Sie können ein einfaches **Regressionsmodell mit zwei Variablen** aufstellen.
- Sie verstehen, unter welchen Umständen unser einfaches Regressionsmodell einen **kausalen Wirkungszusammenhang** abbildet.
- Sie können ein lineares Regressionsmodell **in Excel berechnen und interpretieren**.

Das lineare Regressionsmodell

- Das **einfache lineare Regressionsmodell** hat die Form:

$$y_i = \alpha + \beta x_i + u_i$$

  für $i = 1, 2, ..., n$ Beobachtungen
- Wir erklären y_i **linear** durch x_i. Wir gehen davon aus, dass x_i auf y_i wirkt (und nicht umgekehrt).
- y_i wird als **Regressand**, oder als **endogene**, **abhängige** oder **erklärte Variable** bezeichnet.
- x_i wird als **Regressor**, oder als **exogene**, **unabhängig** oder **erklärende Variable** bezeichnet.
- α und β werden als **(Regressions)-Parameter** oder als **Koeffizienten** bezeichnet.
- u_i ist der **Fehler**, die **Störgröße** oder der **Störterm**.
- y_i und x_i werden beobachtet, α, β und u_i hingegen nicht.

Der Störterm

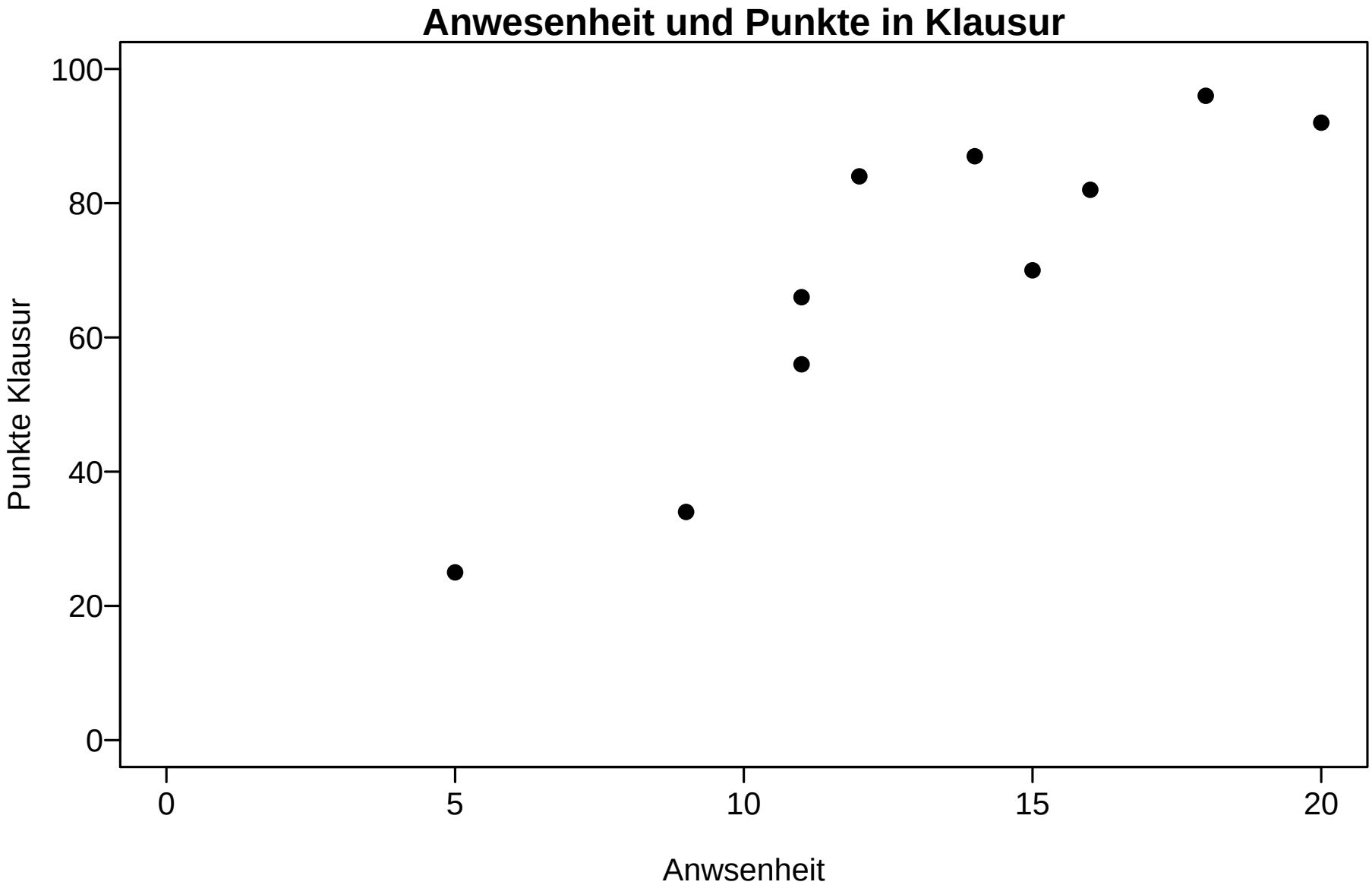
- Wie wir später sehen werden spielt der **Störterm** u_i eine besondere Rolle.
- Es gibt mehrere Gründe, warum wir in unserem Modell explizit einen **Störterm** haben:
  - Für die eigentlich relevanten Variablen liegen keine Daten vor und wir müssen uns mit Proxyvariablen behelfen. Das kann zu **unsystematischen Erhebungs- und Messfehlern** führen.
  - Bestimmte exogene **Variablen sind nicht** im ökonomischen Modell **berücksichtigt**, da sie entweder nicht beobachtbar sind oder nur zu prohibitiv hohen Kosten erhoben werden können.
  - Das **Verhalten** einzelner Individuen, das dem Modell zu Grunde liegt, ist **unberechenbar** und enthält selbst Zufallselemente.
- Zentrale Annahmen in unserem Modell ist immer, dass u_i **keine Systematik** aufweist.

Beispiel: Anwesenheit und Punkte bei Klausur

Für $n = 10$ Studierende (i) liegen folgende Beobachtungen für die Teilnahme x_i und erreichte Punkte bei der Abschlussklausur y_i vor:

i	x_i	y_i
1	15	70
2	12	84
3	20	92
4	9	34
5	11	56
6	16	82
7	18	96
8	5	25
9	11	66
10	14	87

Beispiel: Anwesenheit und Punkte bei Klausur



Das lineare Regressionsmodell

- **Ziel** des linearen Regressionsmodell ist es, eine lineare Kurve (Gerade) durch die Punktwolke zu legen, sodass **der Abstand zwischen den Punkten und der Linie am kleinsten ist**.
- Dadurch erhalten wir die **geschätzten Parameter** $\widehat{\alpha}$ und $\widehat{\beta}$ für die **wahren Parameter** α und β. $\widehat{\alpha}$ und $\widehat{\beta}$ werden oft verkürzt als **Schätzer** bezeichnet.
- Die Gleichung

$$\widehat{y}_i = \widehat{\alpha} + \widehat{\beta} x_i$$

  wird als **geschätztes Modell** bezeichnet. $\widehat{y}_i$ gibt den Wert an, den y_i annehmen müsste, wenn es keine Störeinflüsse gäbe.
- Wie erhalten wir die geschätzten Parameter $\widehat{\alpha}$ und $\widehat{\beta}$?

Methode der Kleinsten Quadrate: Illustration

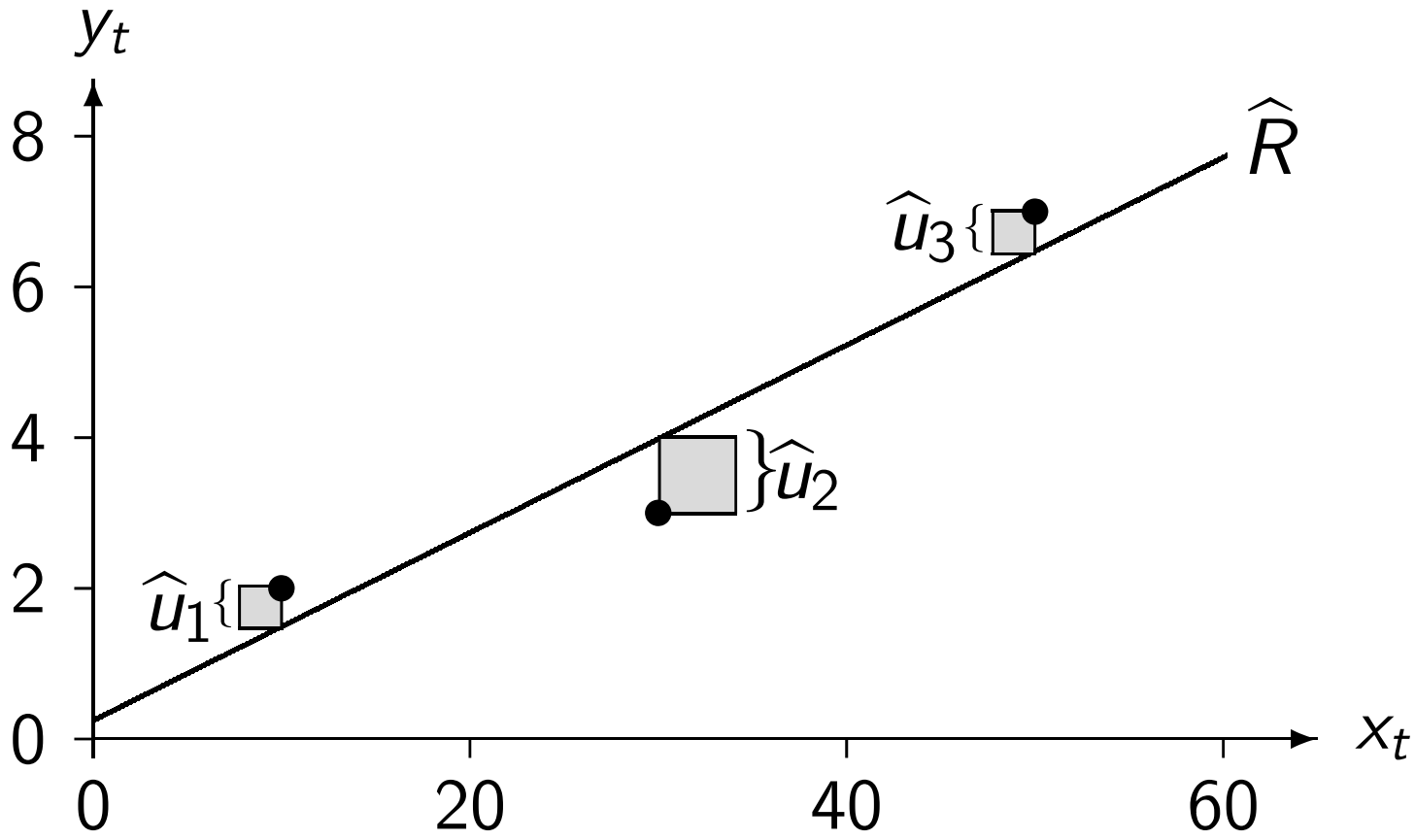


Abbildung aus von Auer (2016)

$y_i = \alpha + \beta x_i + u_i \Leftrightarrow u_i = y_i - \alpha - \beta x_i$
$\widehat{u}_i = y_i - \widehat{\alpha} - \widehat{\beta} x_i$

Anmerkungen: von Auer (2016) verwendet für einzelne Beobachtungen das Subskript t statt i.

Methode der Kleinsten Quadrate: Vorgangsweise

- Um die **beste gerade Linie** durch unsere Punktewolke zu bekommen, gehen wir wie folgt vor:
  - Wir berechnen die **vertikalen (lotrechten) Abstände** zwischen unserer Geraden und den Punkten. Dies ergibt die Residuen $\widehat{u}_i$. Die Residuen $\widehat{u}_i$ sind die Schätzwerte der (unbeobachtbaren) Störgrößen u_i.
  - Wir **quadrieren die Residuen** $\widehat{u}_i^2$ und summieren sie über alle Beobachtungen auf $\sum_{i=1}^{n} \widehat{u}_i^2$.
  - In einem letzten Schritt **minimieren** wir die **Summe der Quadrate**, wodurch wir die Schätzer $\widehat{\alpha}$ und $\widehat{\beta}$ erhalten.
  - Ökonometrische Software erledigt diese Schritte für uns in der Praxis.

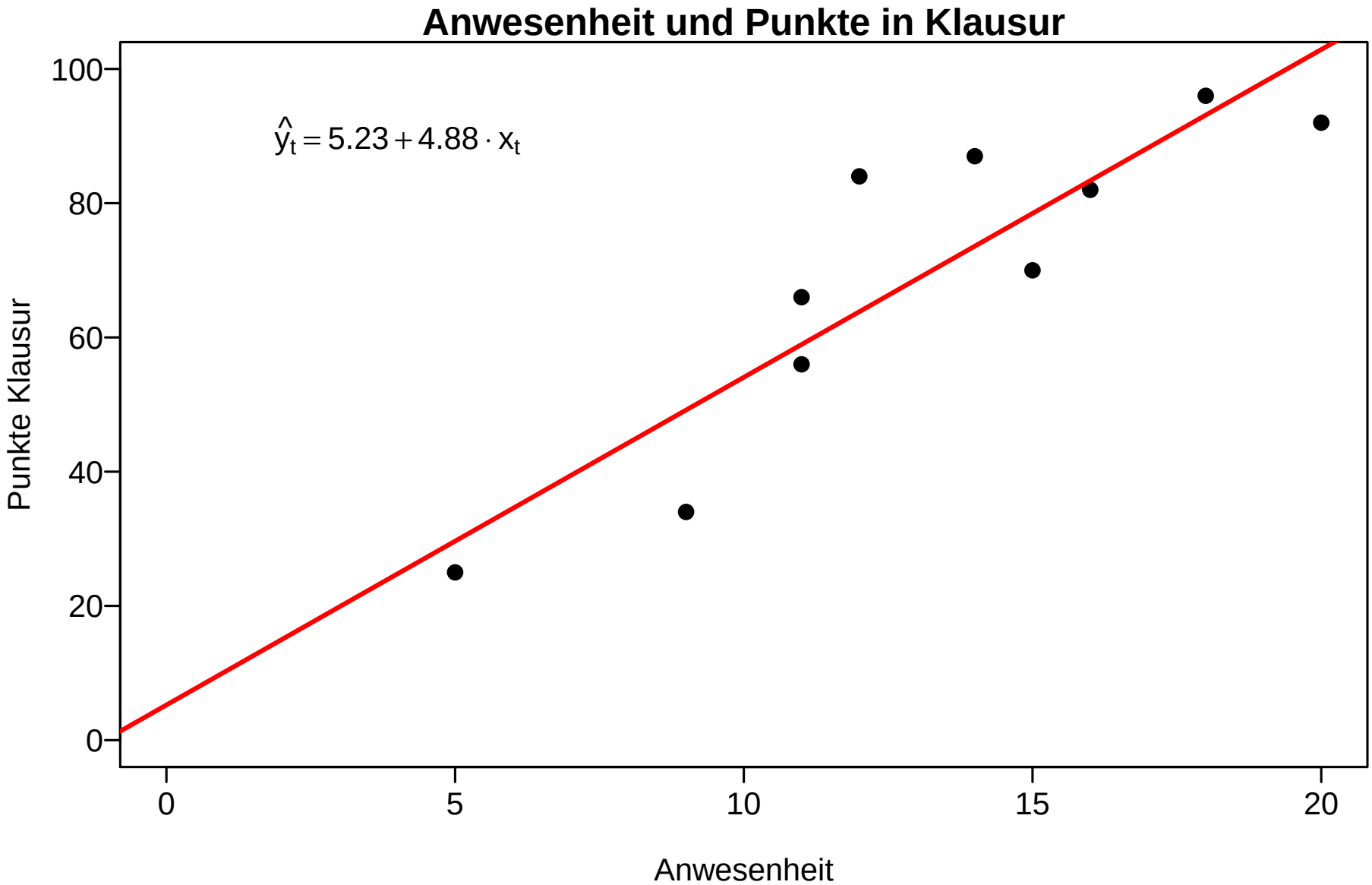
Methode der Kleinsten Quadrate: Begründung

- Wir nehmen hier die **vertikalen Abstände**, da wir annehmen, dass die **Ursache** x einen **Einfluss** auf y hat.
- Wir verwenden **Quadrate**, um besonders **starke Abweichungen** besonders **stark zu gewichten**.
- Die Summe aller Quadrate mathematisch einfach zu minimieren.
- Oft wird der englische Begriff **Ordinary Least Squares** (OLS) verwendet $\Rightarrow$ Wir erhalten den OLS-Schätzer.

Methode der Kleinsten Quadrate: Formale Darstellung

- lineares Regressionsmodell: $y_i = \alpha + \beta x_i + u_i$
- lineares Schätzmodell: $\widehat{y}_i = \widehat{\alpha} + \widehat{\beta} x_i$
- Wir schätzen das Modell, indem wir die Summe der quadrierten Residuen $\sum_{i=1}^{n} \widehat{u}_i^2$ minimieren. Die Residuen $\widehat{u}_i$ sind die Schätzwerte der unbeobachteten Störgrößen u_i:
  $\widehat{u}_i = y_i - \widehat{y}_i = y_i - \widehat{\alpha} - \widehat{\beta} x_i$
- Die Summe der quadrierten Residuen ist daher:
  $S_{\widehat{u}\widehat{u}} \equiv \sum_{i=1}^{n} \widehat{u}_i^2 = \sum_{i=1}^{n} (y_i - \widehat{\alpha} - \widehat{\beta} x_i)^2$
- Um die aggregierten Residuenquadrate zu minimieren, wählt man daher jene Schätzer $\widehat{\alpha}$ und $\widehat{\beta}$, sodass die 1. Ableitung (die Veränderung) von $S_{\widehat{u}\widehat{u}}$ gleich 0 ist:
  $\frac{\partial S_{\widehat{u}\widehat{u}}}{\partial \widehat{\alpha}} = \sum 2(y_i - \widehat{\alpha} - \widehat{\beta} x_i)(-1) = 0$
  $\frac{\partial S_{\widehat{u}\widehat{u}}}{\partial \widehat{\beta}} = \sum 2(y_i - \widehat{\alpha} - \widehat{\beta} x_i)(-x_i) = 0$
- Durch Umformen dieser beiden Gleichungen erhalten wir die geschätzten Parameter:
  $\widehat{\beta} = \frac{\widehat{s}_{xy}}{\widehat{s}_x^2}$; $\qquad \widehat{\alpha} = \bar{y} - \widehat{\beta} \bar{x}$

Ergebnis der OLS Schätzung: Illustration



Interpretation der geschätzten Parameter

- Wenn wir unsere Methode der Kleinsten Quadrate anwenden, erhalten wir:

$$\widehat{y}_i = \underbrace{5.23}_{=\widehat{\alpha}} + \underbrace{4.88}_{=\widehat{\beta}} x_i$$

- Wie können wir $\widehat{\beta}$ interpretieren?
  - $\widehat{\beta}$ ist der Steigungsparameter unseres geschätzten Modells.
  - $\widehat{\beta}$ gibt an, **um wie viele Einheiten wir erwarten, dass sich y ändert, wenn sich x um eine Einheit erhöht**.
  - In unserem Beispiel heißt dies also, dass die Anwesenheit in einer weiteren Einheit im Durchschnitt zu 4.88 mehr Punkten bei der Klausur führt.
- Wie können wir $\widehat{\alpha}$ interpretieren?
  - $\widehat{\alpha}$ stellt den **y-Achsenabschnitt** dar.
  - $\widehat{\alpha}$ gibt den geschätzten Wert für y_i (also $\widehat{y}_i$) an, wenn x_i den Wert 0 annimmt. Oft ist dies nicht plausibel und eine direkte Interpretation von $\widehat{\alpha}$ daher nicht sinnvoll.
  - Meistens ist man an $\widehat{\alpha}$ nicht sonderlich interessiert, und $\widehat{\alpha}$ kann als technische Größe verstanden werden, die im Modell berücksichtigt werden muss.

Umsetzung in EXCEL

- **Daten** $\rightarrow$ **Analyse** $\rightarrow$ **Datenanalyse** $\rightarrow$ **Regression**
- Beachten Sie, dass Beobachtungen mit fehlenden Werten gelöscht werden müssen (oder zumindest nicht markiert sein dürfen).

	Koeffizienten	*Standardfehler*	*t-Statistik*	*P-Wert*
Schnittpunkt	5.23289994	12.1172435	0.43185564	0.6772495
Anwesenheit	4.88298474	0.88070108	5.54442917	0.00054452

Annahmen in unserem Modell

- Die Ergebnisse zeigen jedenfalls einen Zusammenhang (**Korrelation**) zwischen zwei Merkmalen (Variablen).
- In unserem ökonomischen Modellen möchten wir eine klare **Ursachen-Wirkung Beziehung** darstellen. In anderen Worten: Wir möchten einen **kausalen Zusammenhang** zwischen x und y schätzen.
- Wir müssen daher verschiedene Annahmen über unser Schätzmodell treffen:
  - Annahmen, die unser **ökonometrisches Modell** betreffen – sogenannte A-Annahmen nach von Auer.
  - Annahmen **über den Störterm** u_i – sogenannte B-Annahmen nach von Auer.
  - Um einen kausalen Zusammenhang in unserem ökonometrisches Modell festzustellen, müssen die A-Annahmen erfüllt sein, sowie die Annahme B1.
  - Die B-Annahmen sind notwendig, um **Rückschluss** von der Stichprobe **auf die Grundgesamtheit** zu ziehen (d.h. zu beurteilen, wie wahrscheinlich es ist, dass es den gefundenen Zusammenhang auch in der Grundgesamtheit gibt – mehr dazu später).
  - Hinweis: Manchen Annahmen könnten durch schwächere Formen ersetzt werden. Eine Diskussion hierüber würde aber den Rahmen des Kurses übersteigen.

A-Annahme 1: Vollständigkeit und Relevanz

A1: Vollständigkeit und Relevanz

In unserem ökonometrischen Modell fehlen keine relevanten exogenen Variablen, es ist also vollständig. Darüber hinaus sind alle benutzen Variablen x_i relevant.

- Der erste Teil von Annahme A1 (**Vollständigkeit**) sagt aus, dass wir alle ökonomisch relevanten Variablen beobachten und auch in unserem ökonometrischen Modell verwenden.
- Wäre nicht nur die Anwesenheit, sondern auch der Notendurchschnitt bei der Matura relevant für die erreichte Punktezahl bei der Klausur, so müssten wir diese Variable in unser Modell mit aufnehmen (wir besprechen Regressionen mit mehreren Variablen später).
- Der zweite Teil von Annahme A1 (**Relevanz**) sagt aus, dass zwischen der erklärenden Variable x und der erklärten Variable y auch tatsächlich eine Ursachen-Wirkung-Beziehung existiert.
- In der Praxis basiert die Argumentation für/gegen Annahme A1 oft auf ökonomischer Theorie und institutionellem Wissen. Insbesondere die Annahme der Vollständigkeit ist sehr wichtig und sollte gut begründet werden.

A-Annahme 2: Linearität

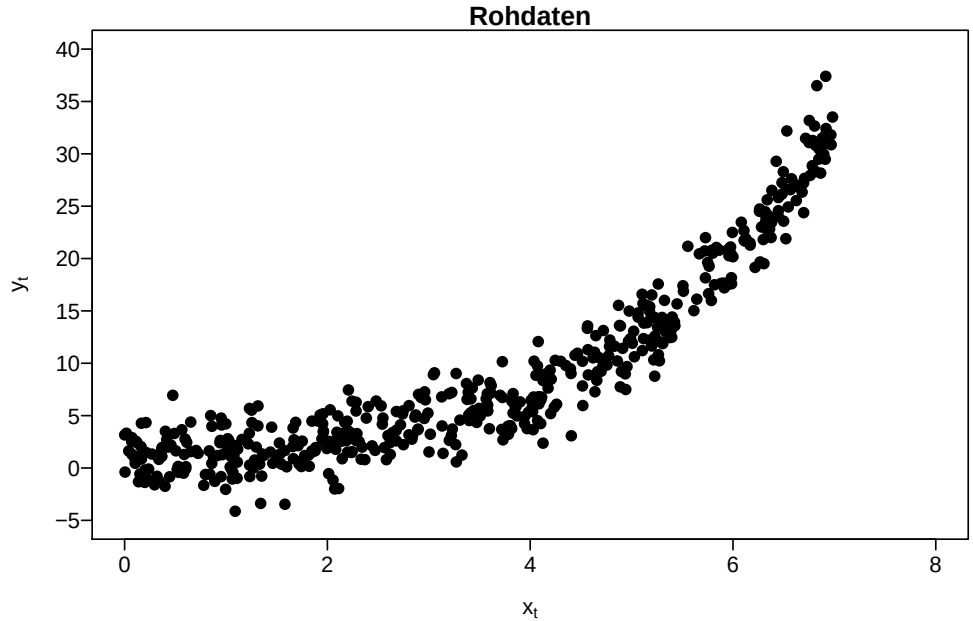
A2: Linearität

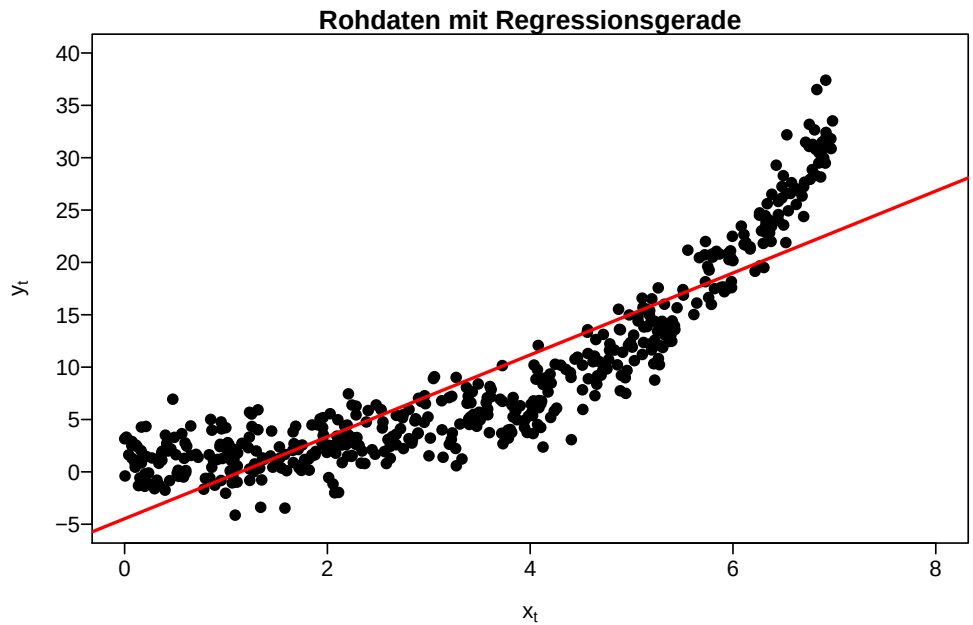
In unserem Modell ist der Zusammenhang zwischen x_i und y_i linear.

$$y_i = \alpha + \beta x_i + u_i$$

- Annahme A2 lässt sich im einfachsten Fall durch **grafische Darstellung einer Punktewolke** überprüfen.
- Ob die Annahme eines linearen Zusammenhangs plausibel ist, wird oft mit Argumenten der **ökonomischen Theorie** untermauert.
- In der Praxis argumentiert man oft (nicht immer!), dass ein lineares Modell die Wirklichkeit hinreichend gut approximiert.

Verletzung von A2





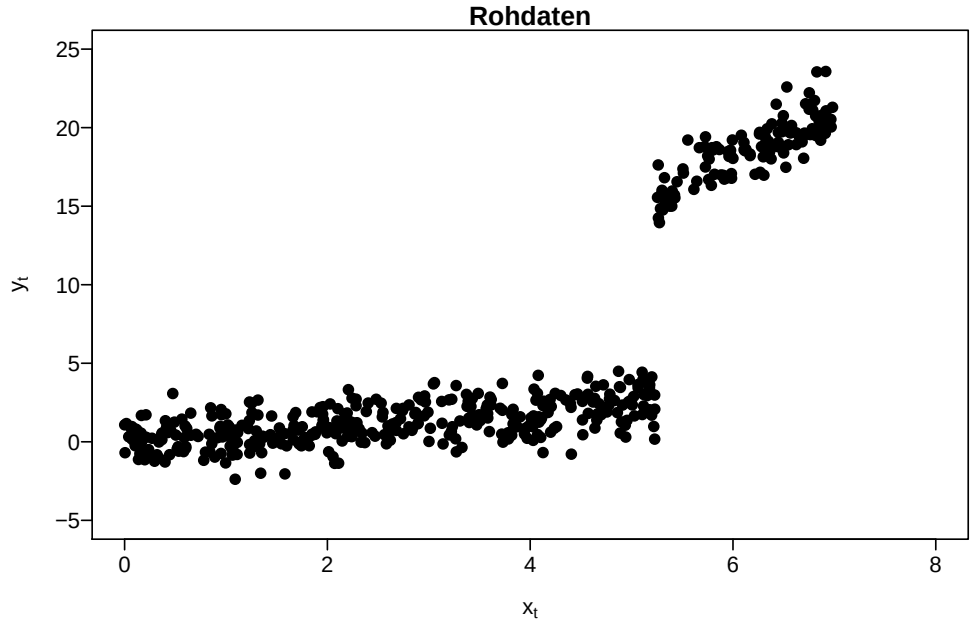
A-Annahme 3: Konstante Parameter α & β

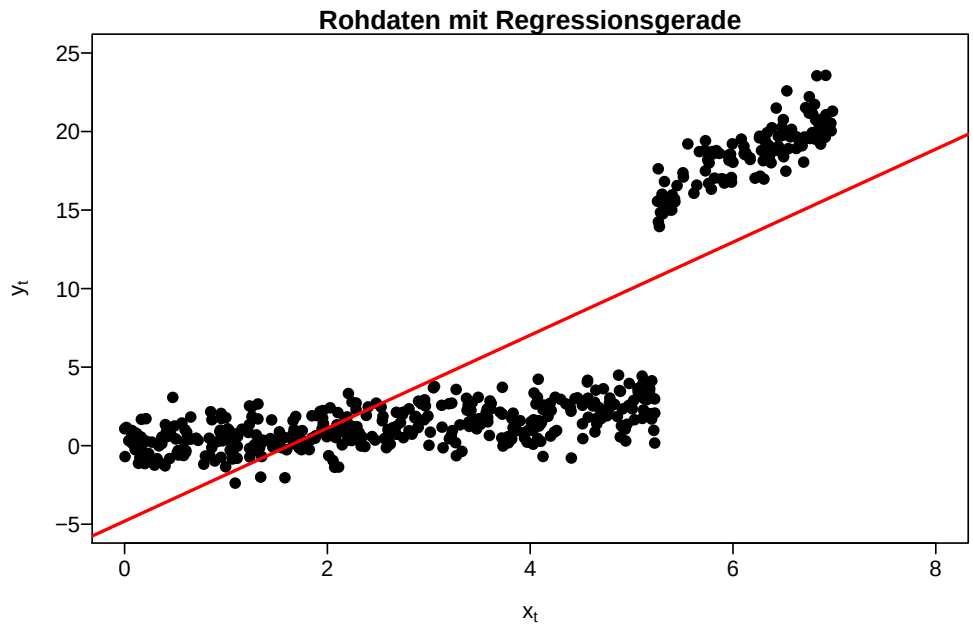
A3: Konstante Parameter

Die Parameter α und β sind für alle n Beobachtungen von x_i und y_i konstant.

- Annahme A3 schließt **Strukturbrüche** in unseren Daten aus.
- Wie bei A2 lässt sich dies im einfachsten Fall durch Darstellung einer Punktewolke grafisch überprüfen.
- Wenn wir wissen, wo der Strukturbruch entsteht, so können komplexere Modelle diesen Strukturbruch berücksichtigen. Dies ist oft in der sogenannten Zeitreihenanalyse der Fall.

Verletzung von A3





A-Annahmen für unser Modell

- Die A-Annahmen zielen darauf ab, dass unser ökonometrisches Modell den Wirkungszusammenhang **funktional korrekt abbildet**.
- Ob unsere A-Annahmen zutreffen, kann in unserem einfachen Rahmen oft durch Visualisierungen erörtert werden.
- Eine besondere und wichtige Ausnahme ist A1 (Vollständigkeit), auf die wir später nochmals genauer eingehen werden.
- Im Gegensatz hierzu betreffen die B-Annahmen den Störterm u_i.

B-Annahme 1: Erwartungswert von 0

B1: Erwartungswert von 0

Die Störgröße u_i hat für alle Beobachtungen einen Erwartungswert von 0:

$$E[u_i] = 0$$

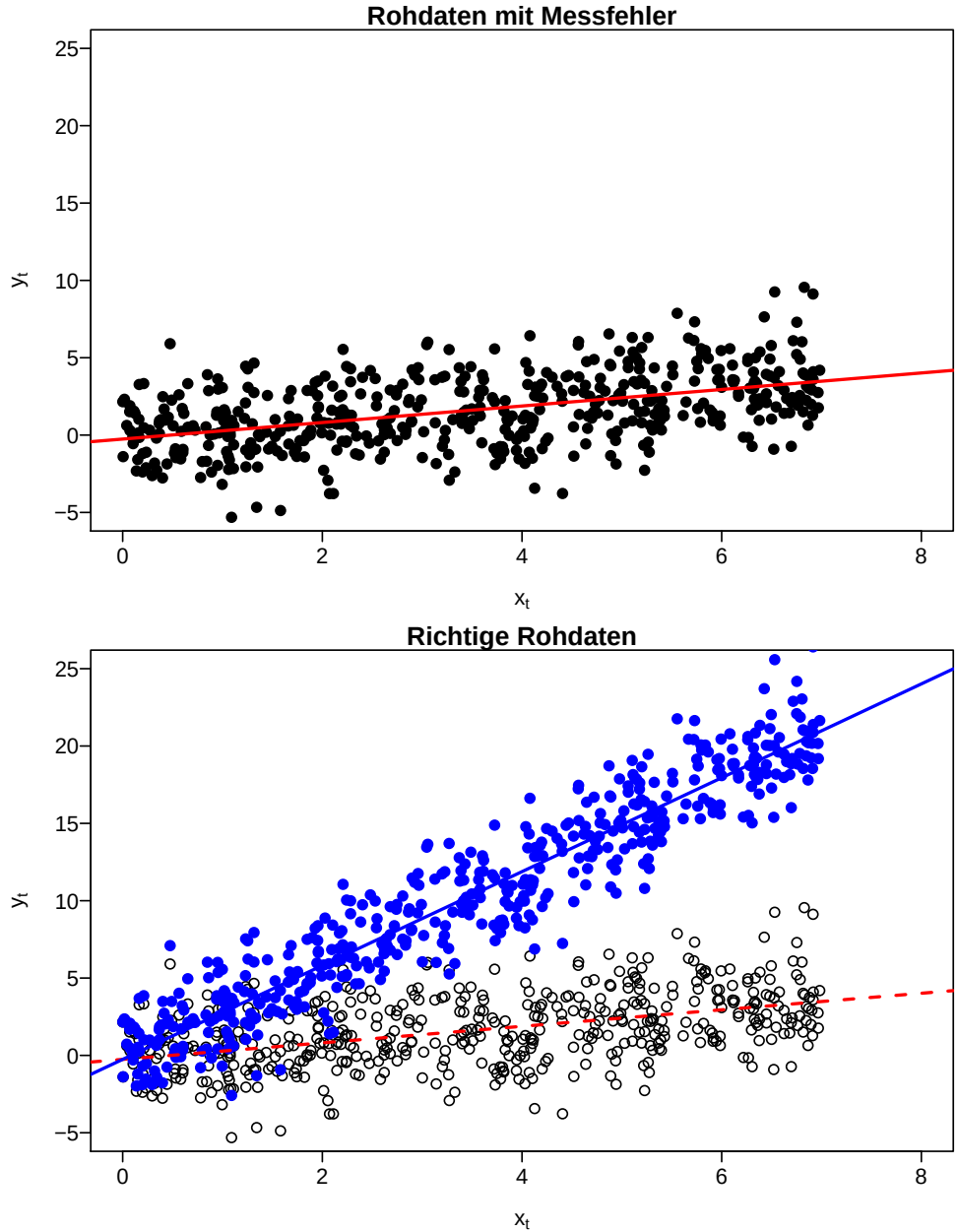
für all Beobachtungen $i = 1, ..., n$

- Man kann y_i als Realisierung eines Prozesses sehen, der zum Teil vom Zufall abhängt. Dieser "Zufallsanteil" wird durch den Störterm u_i abgebildet. Man beobachtet nur eine einzige Realisierung dieses Prozesses. Würde man aber unendlich viele Realisierungen beobachten, dann müsste unter Annahme B1 für jede Beobachtungseinheit der Mittelwert der Residuen über alle Realisierungen gleich 0 sein.
- Damit Annahme B1 erfüllt ist, müssen **alle systematischen Faktoren in unserem Modell berücksichtigt werden**.

B-Annahme 1: Interpretation

- Annahme B1 wäre verletzt, wenn y_i mit einem **systematischen Messfehler** erfasst wird. Zum Beispiel:
  - Punkte auf die Klausur werden systematisch zu hoch angegeben, weil Teilpunkte immer aufgerundet werden.
  - Klausuren mit höherer Punktezahl werden (systematisch) strenger beurteilt als Klausuren mit niedrigerer Punktezahl.
  - Werden hingegen zufällig Klausuren ausgewählt, die zu streng beurteilt werden, während andere zufällig ausgewählte Klausuren zu lasch beurteilt werden, stellt das keine Verletzung der Annahme B1 dar. Es liegt zwar auch in diesem Fall ein Messfehler vor, der Messfehler ist aber **unsystematisch**.
- Annahme **B1 und A1 (Vollständigkeit) sind miteinander verwandt**:
  - Beispiel: Die Klausurergebnisse werden auch durch die kognitiven Fähigkeiten der Studierenden bestimmt. Wenn die kognitiven Fähigkeiten nicht als erklärende Variablen aufgenommen werden, wird die Annahme A1 (Vollständigkeit) verletzt. Wenn eine Studentin besonders hohe kognitive Fähigkeiten aufweist, wird sie wahrscheinlich besser abschneiden, als unser Modell vorhersagt. Das bedeutet, dass u_i in diesem Fall wahrscheinlich größer als 0 ist, d.h. $E[u_i] > 0$. Die Annahme B1 wäre daher ebenfalls verletzt.

Verletzung von B1



B-Annahme 2: Homoskedastische Störgrößen

B2: Homoskedastische Störgrößen

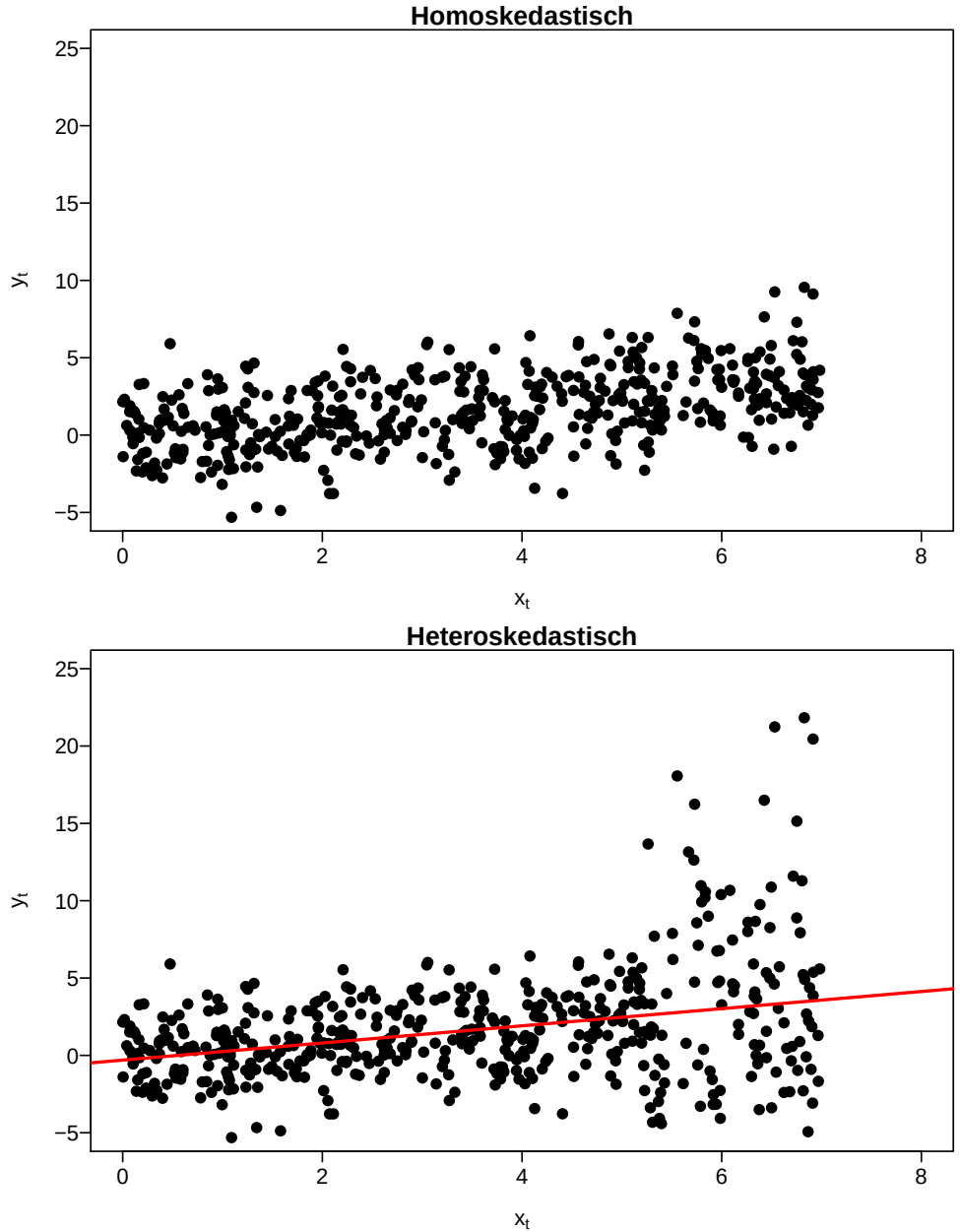
Die Störgröße u_i hat für alle Beobachtungen i eine konstante Varianz:

$$var(u_i) = \sigma^2$$

für all Beobachtungen $i = 1, ..., n$

- Hätten wir Zugang zu unendlich vielen Realisierungen, so müssten unter Annahme B2 die Störgrößen jeder Erhebungseinheit die **gleiche Streuung** besitzen.
- Trifft dies zu, dann spricht man auch von **homoskedastischen Störgrößen**. Ansonsten sind die Störgrößen **heteroskedastisch**.

Verletzung von B2



B-Annahme 3: Keine Korrelation der Störgrößen

B3: Keine Korrelation der Störgrößen

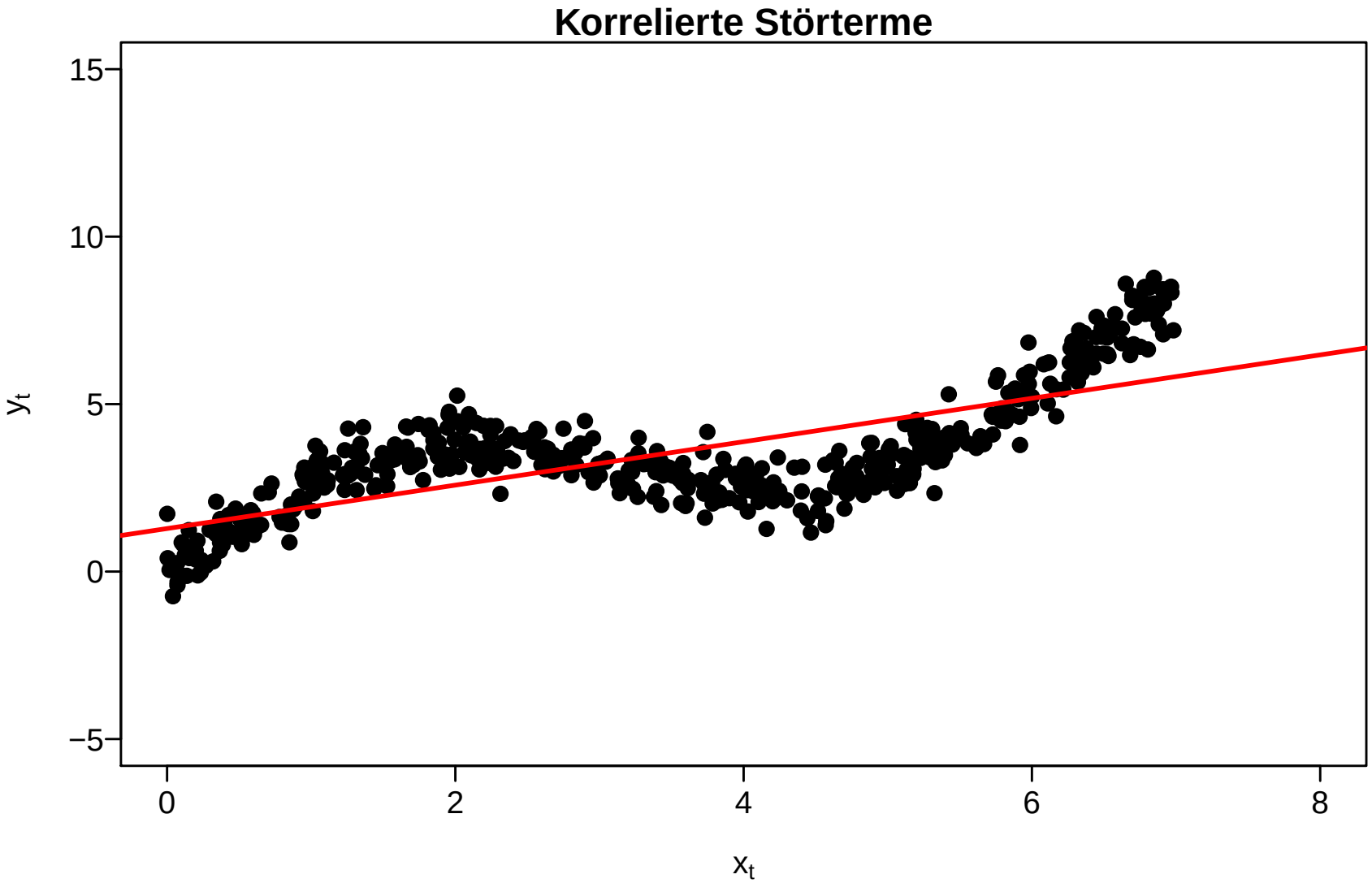
Die Störgröße sind nicht miteinander korreliert:

$$cov(u_i, u_j) = 0$$

für alle $i \neq j$ und $i = 1, ..., n$ sowie $j = 1, ..., n$

- Die Störgröße der Erhebungseinheit i, u_i, und die Störgröße der Erhebungseinheit j, u_j, sind unkorreliert.
- Annahme B3 kann so interpretiert werden, dass zwei unterschiedliche Beobachtungen in unserer Stichprobe unabhängig voneinander sind.
- Annahme B3 wäre beispielsweise verletzt, wenn die Studierenden, die häufig in der Lehrveranstaltung anwesend sind, eine Lerngruppe bilden, und die Studierenden dieser Lerngruppe besser abschneiden, als wir aufgrund ihrer Anwesenheit erwarten würden. Für zwei Mitglieder der Lerngruppe i und j würde dann gelten: $u_i > 0$, $u_j > 0$, und daher $cov(u_i, u_j) > 0$.

Verletzung von B3



Ob Annahme B3 verletzt ist folgt aus der Abbildung nicht eindeutig. Es könnte auch sein, dass unsere Annahme A2 (Linearität) verletzt ist.

B-Annahme 4: Normalverteilung der Störgrößen

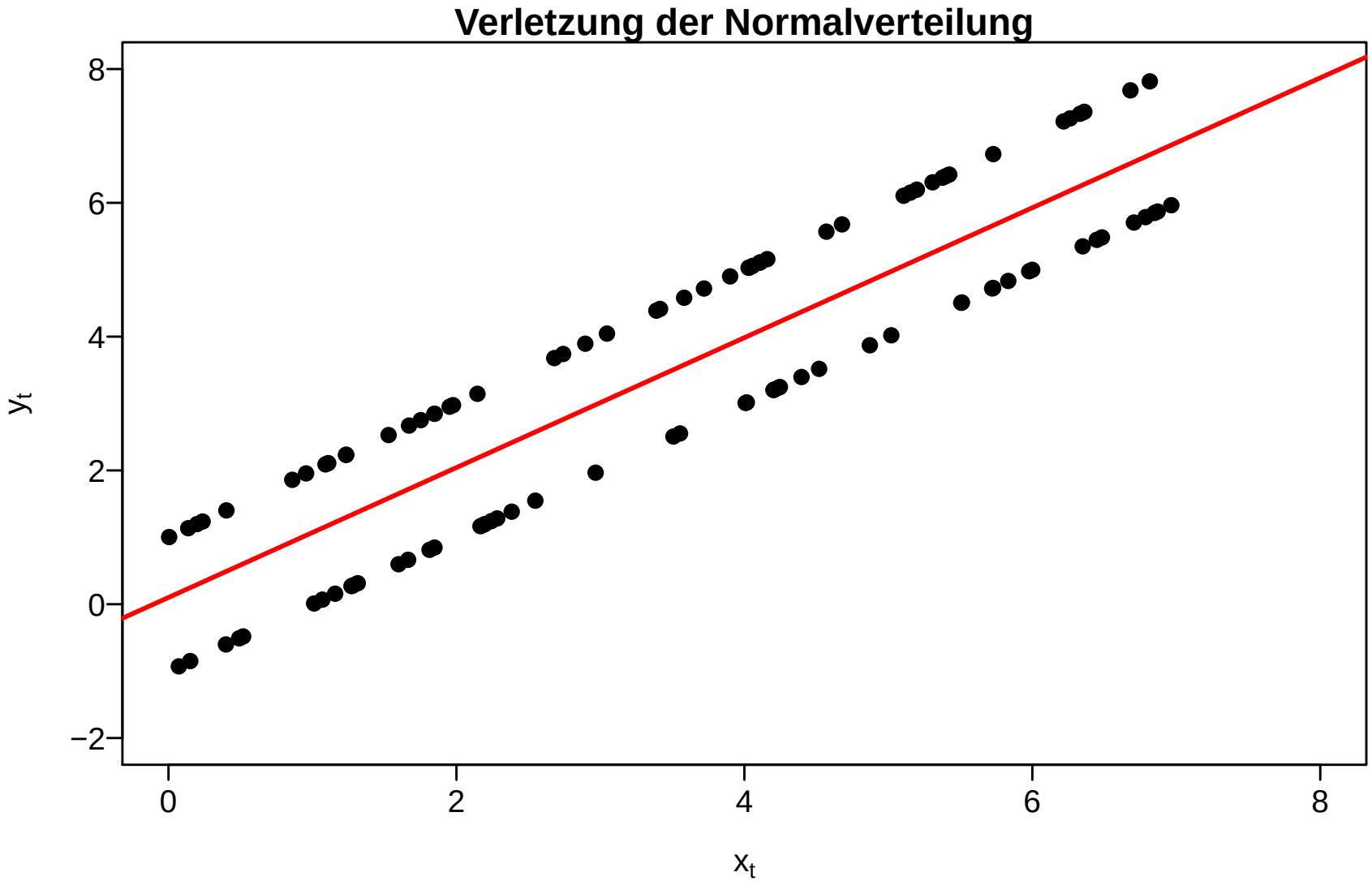
B4: Normalverteilung der Störgrößen

Die Störgrößen sind unabhängig und normalverteilt:

$$u_i \sim N(0, \sigma^2)$$

- In Annahme B4 nützen wir bereits die vorherigen Annahmen B1-B3:
  - Unter B1 ist der Mittelwert 0.
  - Unter B2 ist die Varianz σ^2 konstant.
  - Unter B3 sind alle Beobachtungen unabhängig voneinander.

Verletzung von B4



Beispiel: Anwesenheit und Punkte bei Klausur

- Gehen wir Annahmen A1-A3 und Annahmen B1-B4 anhand unseres Klausur-Beispiels durch.
- Zur Erinnerung: Folgende Daten zu Anwesenheit (x_i) und Punkten (y_i) haben wir erhoben:

i	x_i	y_i
1	15	70
2	12	84
3	20	92
4	9	34
5	11	56
6	16	82
7	18	96
8	5	25
9	11	66
10	14	87

Beispiel: A1 – Vollständigkeit und Relevanz

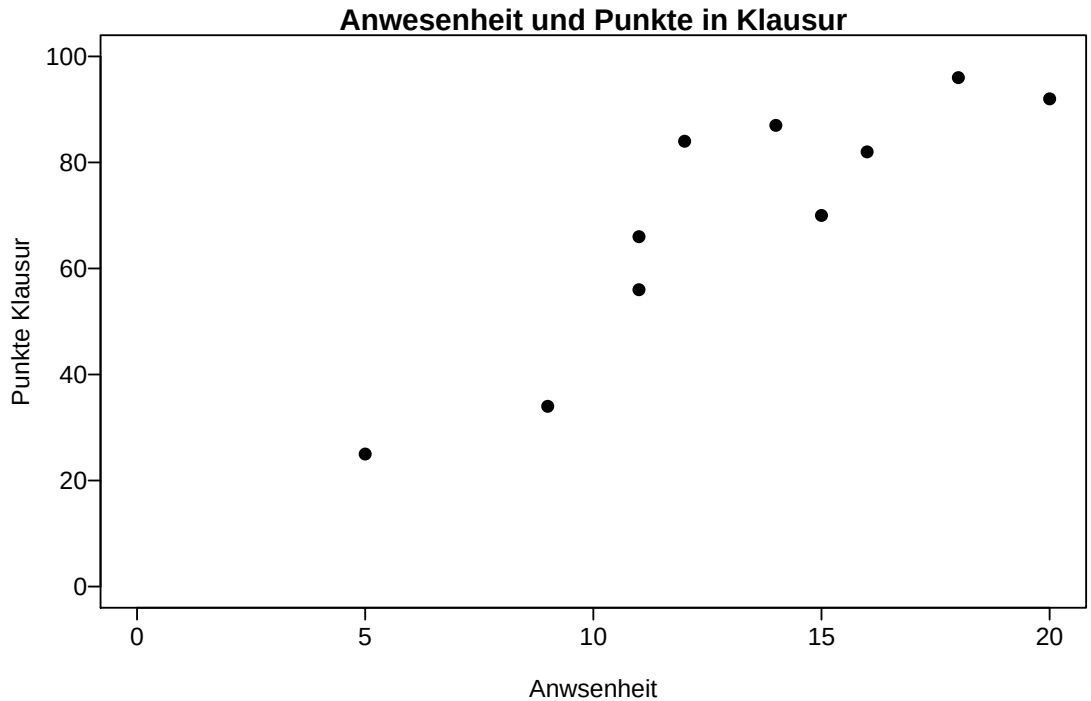
- **Relevanz**: Ist Anwesenheit relevant für Klausurpunkte?
  - Es ist sehr wahrscheinlich, dass Anwesenheit während des Semesters die Klausurpunkte beeinflussen wird.
- **Vollständigkeit**: Sind alle wichtigen exogenen Variablen in unserem Modell vorhanden?
  - Einer der wichtigsten Annahmen in unserem Modell!
  - Es ist wahrscheinlich, dass beispielsweise motiviertere Studenten öfters zur Vorlesung erscheinen.
  - Motiviertere Studenten schneiden wahrscheinlich besser bei der Klausur ab – unabhängig von ihrer Anwesenheit.
  - Wir beobachten die Motivation in unseren Daten nicht $\rightarrow$ das Modell ist daher nicht vollständig.

Beispiel: A1 – Vollständigkeit und Relevanz

- Von unserer kurzen Analyse von A1 wird klar, dass wir in unserem Anwesenheit-Klausur Modell **keinen Wirkungszusammenhang** identifizieren können $\rightarrow$ wir bilden also nur eine **Korrelation** ab.
- Ob ein Kausalzusammenhang oder lediglich eine Korrelation vorliegt, muss bei der Interpretation berücksichtigt werden.
- Das soll nicht heißen, dass Korrelationen nicht interessant sind, wir müssen aber bei den **Schlussfolgerungen** vorsichtig sein:
  - Wenn es einen kausalen Zusammenhang zwischen Anwesenheit und Klausurerfolg gibt, führt eine Erhöhung der Anwesenheit (etwa: Einführung einer Anwesenheitspflicht) zu einer Erhöhung der Punktezahl bei der Klausur. Der geschätzte Parameter $\widehat{\beta}$ gibt Aufschluss über die Größe des (erwarteten) Effekts.
  - Sollte der Erfolg bei der Klausur lediglich von der Motivation abhängen und nicht von der Anwesenheit, motivierte Studierende aber häufiger anwesend sind, dann wird die Einführung einer Anwesenheitspflicht keinen Effekt auf den Klausurerfolg haben. (Weil der Klausurerfolg von der Motivation abhängt, die aber durch die Anwesenheitspflicht nicht gesteigert wird.)

Beispiel: A2 – Linearität

- In unserem einfachen Beispiel mit zwei Variablen können wir Linearität visuell überprüfen.
- Dazu erstellen wir ein Streudiagramm, das x_i auf der (horizontalen) x-Achse und y_i auf der (vertikalen) y-Achse abbildet.



- Die Abbildung zeigt, dass ein linearer Zusammenhang plausibel erscheint.

Beispiel: A3 – Konstante Parameter α & β

- Ähnlich wie A2 können wir auch A3 grafisch untersuchen.
- Überprüfung der Abbildung auf der vorherigen Folie lässt nicht auf einen Strukturbruch schließen.
- Wir können deshalb mit ziemlicher Sicherheit davon ausgehen, dass in unserem Beispiel A3 erfüllt ist.
- Es sollte nochmals erwähnt werden, dass Visualisierung nur in unserem zwei-Variablen Modell möglich ist $\rightarrow$ bei mehreren Variablen sind die zugrunde liegende ökonomische Theorie, institutionelles Wissen und die statistischen Eigenschaften der Residuen wichtig.

Beispiel: B1 – Erwartungswert von 0

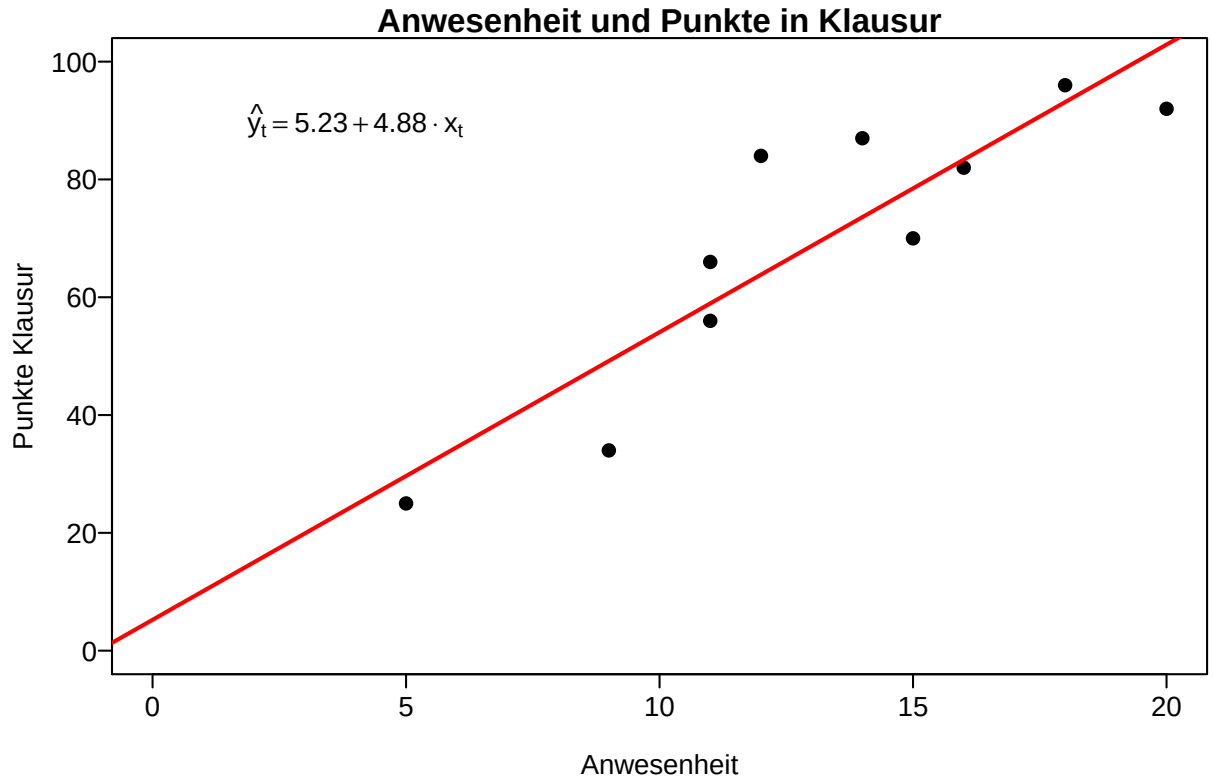
- Können wir in unserem Beispiel davon ausgehen, dass

$$E[u_i] = 0$$

  für alle $i = 1, ..., n$?
- Unglücklicherweise beobachten wir $E[u_i]$ nicht $\rightarrow$ ähnlich wie bei A1 basiert die Diskussion deshalb auf ökonomischer Theorie, institutionellem Wissen und den daraus gezogenen Schlussfolgerungen.
- Können wir davon ausgehen, dass es keinen systematischen Fehler in den Beobachtungen gibt?
  - Dozenten kontrollieren die Anwesenheit und vergeben die Punkte.
  - Dozenten bekommen nicht mehr Geld für höhere Anwesenheitsquoten oder bessere Klausurergebnisse $\rightarrow$ keinen Anreiz, Daten zu manipulieren. Systematische Erhebungsfehler sind daher unwahrscheinlich.
  - Kein Hinweis darauf, dass B1 verletzt ist.

Beispiel: B2 & B3 – Homoskedastische und unkorrelierte Störgrößen

- Grafische Überprüfung von den Annahmen B2 (Homoskedastizität) & B3 (Korrelation):



Beispiel: B2 & B3 – Homoskedastische und unkorrelierte Störgrößen

- Annahme B2 ist wahrscheinlich verletzt, wenn wir ein "auffächern" in unserem Plot sehen würden.
  - Die Daten in unserem Streudiagramm zeigen keine extremen Anzeichen eines "auffächerns".
  - Allerdings ist die Anzahl der Beobachtungen mit 10 sehr klein, so dass sich keine genaue Aussage ableiten lässt.
- Annahme B3 ist wahrscheinlich verletzt, wenn wir ein auffälliges Muster in unserem Plot erkennen können.
  - Die Daten in unserer Abbildung zeigen keine Auffälligkeiten von starker Autokorrelation
  - Die Entscheidung, zur Vorlesung zu gehen, könnte von dem Verhalten der anderen Studierenden abhängen, was zu einer Korrelation der Störgrößen führen kann.
  - Wie bei B2 ist die Anzahl der Beobachtungen zu klein, um klare Aussagen abzuleiten.

Beispiel: B4 – Normalverteilung der Störgrößen

- Um Annahme B4 grafisch zu überprüfen, können wir die Residuen in einem Histogramm darstellen, und das Histogramm mit einer Normalverteilung (mit Mittelwert 0 und Varianz σ^2) vergleichen.
  - Die Zahl der Beobachtungen ist hierfür aber zu gering.
  - Wir besprechen nächste Einheit, wie wir die unbeobachtbare Varianz σ^2 schätzen können.
- In unserem konkreten Beispiel deutet nichts darauf hin, dass u_i nicht normalverteilt ist:
  - Wir beobachten sowohl kürzere wie längere Abstände von Daten zu unserer Geraden.
  - Wir beobachten eine relativ gleiche Anzahl an positiven wie negativen Abweichungen.
- Es sollte erwähnt werden, dass in der Praxis sowohl statistische Testverfahren existieren, um Annahmen B2-B4 zu testen, als auch "robuste" Methoden, bei denen Annahmen B2-B4 nicht erfüllt sein müssen. Das würde aber den Rahmen dieser Einführung übersteigen.

Zusammenfassung

- Die **Ergebnisse** eines einfachen ökonometrischen Modells geben Auskunft darüber, **ob und wie zwei Variable zusammenhängen**.
- Wir sind meist nicht nur an Zusammenhängen, sondern an **Wirkungsbeziehung (kausalen Zusammenhängen)** interessiert: Beeinflusst die Variable x die Variable y kausal? Diese Wirkungsbeziehung erlaubt uns, genauere Aussagen über Zusammenhänge zu tätigen und auch **stärkere Schlussfolgerungen** zu ziehen.
- Damit tatsächlich eine Wirkungsbeziehung vorliegt müssen **mehrere Annahmen erfüllt** sein:
  - A-Annahmen: Wir bilden den Zusammenhang funktional und korrekt ab.
  - B-Annahmen: Der Störterm besitzt gewisse Eigenschaften.
- Ob diese Annahmen zutreffen, ist oft schwierig zu überprüfen. Die ökonomische Theorie, statistische Kennzahlen und institutionelle Kenntnis können uns dafür Argumente liefern.
- Wir haben uns bislang damit beschäftigt, ob wir einen Zusammenhang auch als (kausale) Wirkungsbeziehung interpretieren können. Bisher haben wir aber keine Aussage darüber gemacht, ob wir den Zusammenhang (ob kausal oder nicht), den wir in der Stichprobe finden, auch in der Grundgesamtheit vermuten können (**statistische Signifikanz**) $\rightarrow$ Gegenstand der nächsten Einheit