SE Geographie und Ökonomie

Einheit 6:
Multivariates Lineares Regressionsmodell

Bernhard Schmidpeter

bernhard.schmidpeter@jku.at

Institut für Volkswirtschaftslehre

SoSe 2024

Wiederholung und Motivation

- Bisher haben wir nur den **Zusammenhang** zwischen der erklärenden Variable und **einer unabhängigen Variable** betrachtet (einfaches lineares Regressionsmodell):

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

  für $i = 1, 2, ..., n$ Beobachtungen.
- Dieser Zusammenhang ist aber oft zu restriktiv:
  - Bildungsniveau hängt vom elterlichen Einkommen, Bildung der Eltern usw. ab.
  - Einkommen hängt von Bildung, Arbeitsmarkterfahrung, Region usw. ab.
  - ...
- Wir können unser bisheriges Regressionsmodell und Hypothesentests relativ einfach anpassen.

Lernziele Einheit 6

- Sie können ein **Regressionsmodell mit mehreren Variablen** formulieren, in EXCEL **schätzen**, und die Ergebnisse **interpretieren**.
- Sie können alle **Maßzahlen**, die EXCEL bei einer (multiplen) Regressionsanalyse ausweist, **verstehen und Interpretieren**.
- Sie können **einfache Hypothesen** formulieren und diese **testen**.
- Sie können basierend auf den Schätzergebnissen eine **bedingte Prognose** erstellen.
- Sie verstehen, unter welchen Bedingungen ein geschätzter Parameter als **kausaler Effekt** (**Wirkungszusammenhang**) interpretiert werden kann.

Das Multivariate Lineare Regressionsmodell

Das multivariate lineare Regressionsmodell hat die Form:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + u_i$$

für $i = 1, 2, ..., n$ Beobachtungen

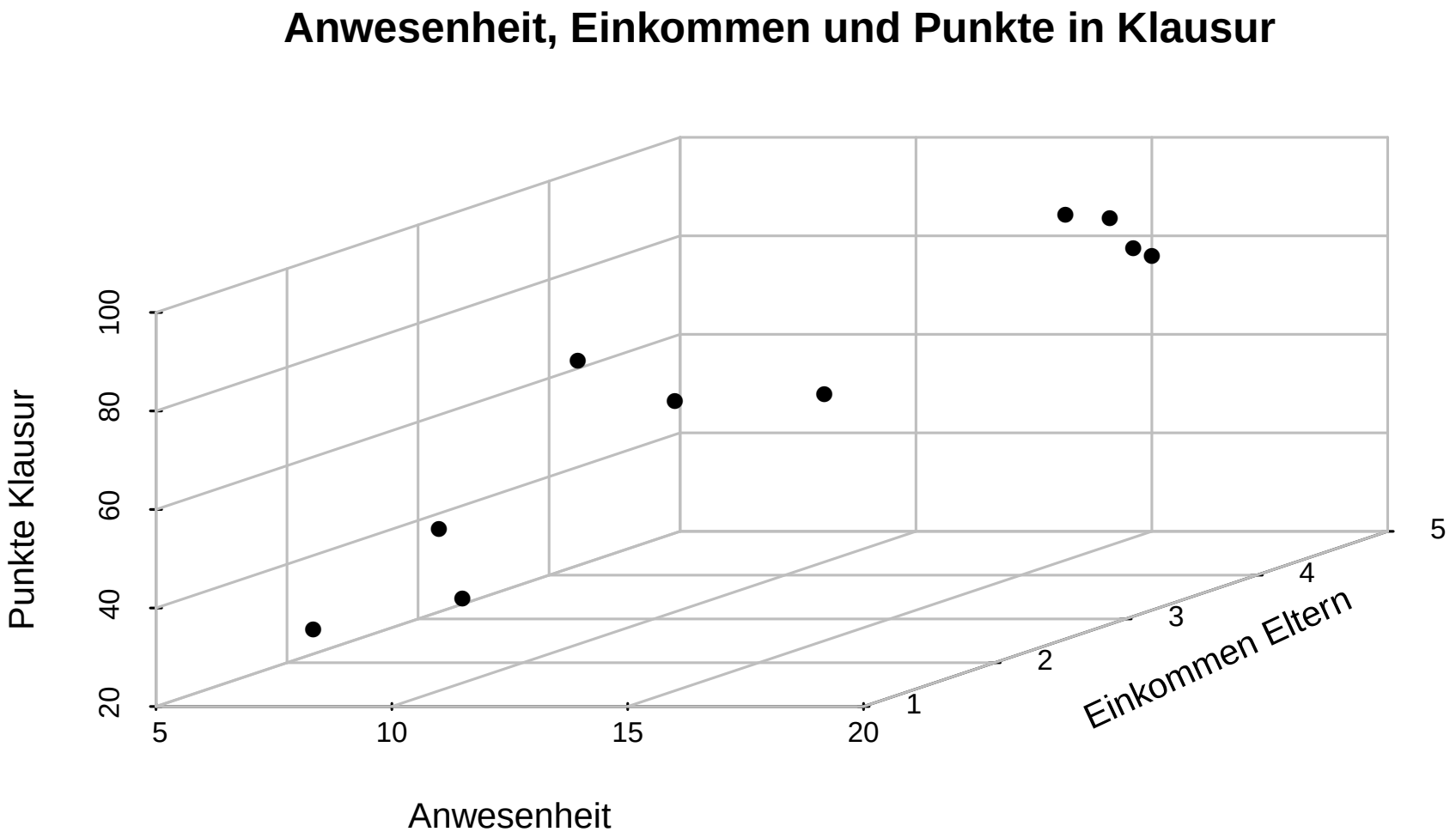
- Wir erklären y_i **linear** durch $x_{1i}, x_{2i}, \ldots, x_{Ki}$. Wir gehen davon aus, dass $x_{1i}, x_{2i}, \ldots, x_{Ki}$ auf y_i wirken (und nicht umgekehrt).
- y_i wird als **Regressand**, oder als **endogene**, **abhängige** oder **erklärte Variable** bezeichnet.
- $x_{1i}, x_{2i}, \ldots, x_{Ki}$ werden als **Regressoren**, oder als **exogene**, **unabhängige** oder **erklärende Variable** bezeichnet.
- α und $\beta_1, \beta_2, \ldots, \beta_K$ werden als **(Regressions)-Parameter** oder als **Koeffizienten** bezeichnet.
- u_i ist der **Fehler**, die **Störgröße** oder der **Störterm**.
- y_i und $x_{1i}, x_{2i}, \ldots, x_{Ki}$ werden beobachtet, α, $\beta_1, \beta_2, \ldots, \beta_K$ und u_i hingegen nicht.

Beispiel: Anwesenheit und Punkte bei Klausur

Für $n = 10$ Studierende liegen folgende Beobachtungen für die Teilnahme am Unterricht x_{1i}, monatliches elterliches Einkommen (in tausend Euro) x_{2i} und erreichte Punkte in der Abschlussklausur y_i vor:

i	x_{1i}	x_{2i}	y_i
1	15	2.5	70
2	12	1.7	84
3	20	3.2	92
4	9	1.9	34
5	11	1.0	56
6	16	4.5	82
7	18	3.6	96
8	5	2.2	25
9	11	2.8	66
10	14	4.7	87

Beispiel: Anwesenheit, Einkommen und Punkte in Klausur



Das Multivariate Regressionsmodell

- Ziel des multivariaten Regressionsmodell ist es, **eine lineare Schätzebene durch die Punktwolke zu legen**, so dass der **Abstand zwischen den Punkten und der Schätzebene am kleinsten** ist.
- In dem einfachen linearen Regressionsmodell entsprach die Schätzebene einer Geraden.
- Dadurch erhält man die **geschätzten Parameter** $\widehat{\alpha}$ und $\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_K$ für die wahren Parameter α und $\beta_1, \beta_2, \ldots, \beta_K$.
- Wie zuvor wird die Gleichung:

$$\widehat{y}_i = \widehat{\alpha} + \widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i} + \cdots + \widehat{\beta}_K x_{Ki}$$

  als geschätztes Modell bezeichnet.
- Wie erhalten wir Schätzer $\widehat{\alpha}$ und $\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_K$?

Methode der Kleinsten Quadrate

- Ähnlich wie bei der linearen Einfachregression werden die geschätzten Parameter so gewählt, um die **Abweichung** der Residuen $\widehat{u}_i$ von der geschätzten Ebene zu **minimieren**.
- Es werden die quadrierten Abweichungen verwendet, da große Abweichungen besonders stark gewichtet werden sollen. Zudem ist die **Summe aller Quadrate** mathematisch **einfach zu minimieren**.
- Im Vergleich zu unserem Modell mit einer unabhängigen Variablen benötigen wir hierfür Matrixalgebra.
- Da die Schätzungen in EXCEL durchgeführt werden, wird an dieser Stelle nicht tiefer darauf eingegangen. Die Idee ist aber ident zur linearen Einfachregression.

Interpretation des OLS Schätzers

- Wenn wir die Kleinstquadrat-Methode (KQ bzw. OLS) bei dem erweiterten Datensatz anwenden, erhalten wir:

$$\widehat{y}_i = \underbrace{3.03}_{=\widehat{\alpha}} + \underbrace{4.50}_{=\widehat{\beta}_1} x_{1i} + \underbrace{2.55}_{=\widehat{\beta}_2} x_{2i}$$

- Wie zuvor können wir unsere Parameter und gegeben unser theoretisches Modell sinnvoll interpretieren

Interpretation des OLS Schätzers: $\hat{\beta}_1$ & $\hat{\beta}_2$

Interpretation der geschätzten Parameter $\widehat{\beta_1}$ und $\widehat{\beta_2}$:

- $\widehat{\beta_1}$ ist der Steigungsparameter für Anwesenheit.
- $\widehat{\beta_2}$ ist der Steigungsparameter für elter. Einkommen.
- $\widehat{\beta_1}$ gibt an, **um wie viele Einheiten sich y ändert, wenn wir x_1 um eine Einheit erhöhen** und alle anderen Variablen konstant lassen.
- Der Besuch einer zusätzlichen Einheit führt im Durchschnitt zu 4.50 mehr Punkten bei der Klausur.
- $\widehat{\beta_2}$ gibt an, **um wie viele Einheiten sich y ändert, wenn wir x_2 um eine Einheit erhöhen** und alles anderen Variablen konstant lassen.
- Wenn das elter. Einkommen um $1,000$ Euro steigt, erhöhen sich im Durchschnitt die Punkte bei der Klausur um 2.55.

Annahmen für das Multivariate Regressionsmodell

- In Vorlesung 5 wurde besprochen, unter welchen Annahmen unser Modell eine Ursachen-Wirkung Beziehung abbildet.
- Für das **multivariates Regressionsmodell** müssen die **A-Annahmen nur minimal anpasst** werden.
- Die **B-Annahmen ändern sich** hingegen **gar nicht**.

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 $x_1, x_2, \ldots, x_K$ relevant.

- Diese Annahme bleibt unverändert im Vergleich zu unserem univariaten Modell
- Der erste Teil von Annahme A1 (**Vollständigkeit**) sagt aus, dass wir all ökonomisch relevanten Variablen beobachten und auch in unserem ökonometrischen Modell verwenden.
- Der zweite Teil von Annahme A1 (**Relevanz**) sag aus, dass zwischen den erklärenden Variable $x_1, \ldots, x_K$ und der erklärten Variable y auch tatsächlich eine Ursachen-Wirkung-Beziehung existiert.

A-Annahme 2: Linearität

A2: Linearität

In unserem Modell ist der Zusammenhang zwischen x_{ik} und y_i linear.

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + u_i$$

- Im Gegensatz zu unserem einfachen univariaten Modell lässt sich Annahme A2 **nicht mehr** so einfach durch die **grafische Darstellung** einer Punktwolke **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.
- Wichtig ist hier, dass man die **funktionale Form der Variable x_k verändern** kann. Wenn der Zusammenhang zwischen dem Einkommen (als erklärte Variable y) und dem Alter nicht linear ist, dann kann für x_k auch **die quadrierten oder die logarithmierten Werte für Alter** verwendet werden.

A-Annahme 3: Konstante Paramter α & $\beta_1, \beta_2, \ldots, \beta_K$

A3: Konstante Parameter

Die Parameter α und $\beta_1, \beta_2, \ldots, \beta_K$ sind für alle n Beobachtungen von $x_{1i}, x_{2i}, \ldots, x_{Ki}$ und y_i konstant.

- Annahmen A3 schließt **Strukturbrüche** in unseren Daten aus.
- Wie bei Annahme A2 lässt sich A3 **nicht mehr** so einfach durch die **grafische Darstellung** einer Punktwolke **überprüfen**.
- Wenn wir wissen, wo der Strukturbruch entsteht, so können komplexere Modelle diesen Strukturbruch berücksichtigen. Dies ist oft in der sogenannten Zeitreihenanalyse der Fall.
- In der Realität geht man meistens von keinen unbekannten Strukturbrüchen aus, sondern von klar beobachtbaren und sehr wichtigen Ereignissen (Rezessionen, Covid-Krise, EU-Beitritt, ...).

B-Annahmen über die Störgrößen

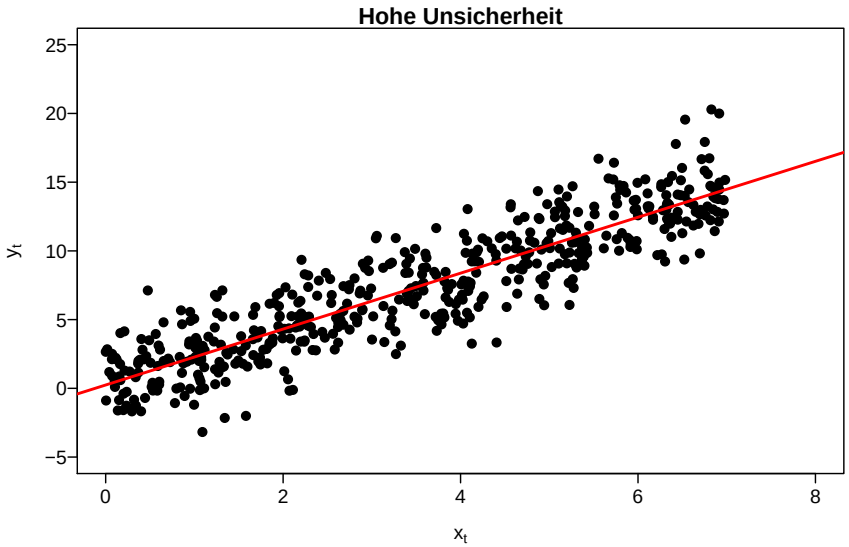
Die **B-Annahmen über die Störgrößen bleiben** im Vergleich zur linearen Einfachregression **unverändert**:

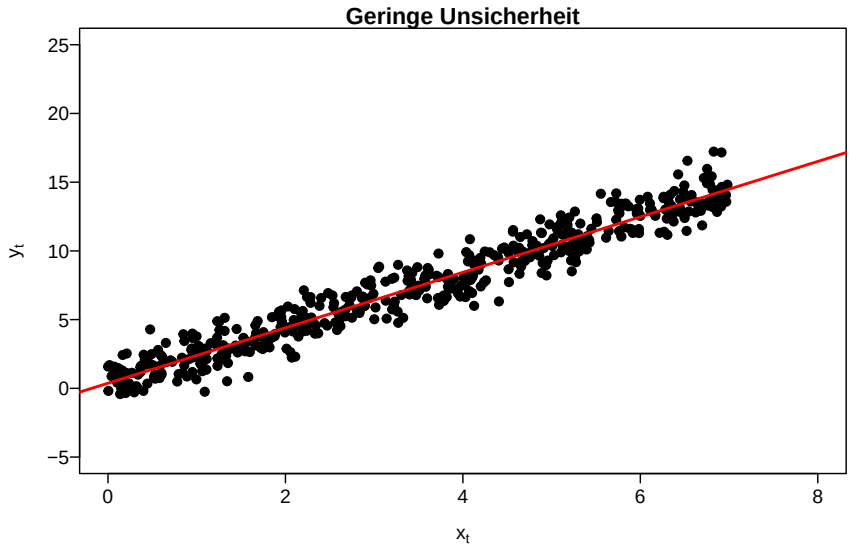
- B1: Erwartungswert von 0:
  - Die Störgröße u_i hat für alle Beobachtungen einen Erwartungswert von 0: $E[u_i] = 0$ für alle Beobachtungen $i = 1, ..., n$.
- B2: Homoskedastische Störgrößen:
  - Die Störgröße u_i hat für alle Beobachtungen i eine konstante Varianz: $var(u_i) = \sigma^2$ für alle Beobachtungen $i = 1, ..., n$.
- 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 $j = 1, ..., n$.
- B-Annahme 4: Normalverteilung der Störgrößen:
  - Die Störgrößen sind unabhängig und normalverteilt: $u_i \sim N(0, \sigma^2)$.

Unsicherheiten

- Ähnlich wenn wir einfache Tests durchführen sind unsere Schätzungen mit Unsicherheiten verbunden
- Wir können diese Unsicherheiten quantifizieren und dann
  a) Konfidenzintervalle um $\hat{beta}$ bilden
  b) Hypothesen testen
- Die Intuition hinter diesen Intervallen und Tests ist sehr ähnlich zu unseren einfachen Tests
- Die Schätzung der Unsicherheit verläuft hingegen anders
  - Die Unsicherheit wird druch den Störterm abgebildet

Punktschätzer & Unsicherheiten: Grafische Darstellung





- In beiden Datensätzen sind die Punktschätzer (die Lage der Regressionsgeraden) gleich und die Annahmen B1-B4 erfüllt.
- Im Datensatz der linken Grafik ist die Abweichung der Beobachtungen von der Regressionsgerade (also $|u_i|$) im Durchschnitt größer $\rightarrow$ **größere Unsicherheit** in der Schätzung des Modells.
- Wichtigkeit von **Stichprobenumfang** und **Streuung der erklärenden Variable** ist ebenfalls ersichtlich.

Schätzen der Unsicherheit (für Interessierte)

- Um die statistische Unsicherheit eines geschätzten Parameters $\widehat{\beta}_k$ zu bestimmen, muss (wie in der linearen Einfachregression) der **Standardfehler** (**se**) des geschätzten Parameters ($se(\widehat{\beta}_k)$) bestimmt werden. **EXCEL** macht dies für uns, und gibt die geschätzten Standardfehler für die Konstante ($\widehat{se}(\widehat{\alpha})$) sowie für jeden geschätzten β-Parameters ($\widehat{se}(\widehat{\beta}_k)$) an.
- Die Schätzung der Standardabweichung für $\widehat{\alpha}, \widehat{\beta}_1, \widehat{\beta}_2, \ldots \widehat{\beta}_K$ folgt ähnlichen Schritten wie bei der linearen Einfachregression (siehe Einheit 6). Da die Korrelation zwischen den Variablen berücksichtigt werden muss, muss die (sogenannte) **Varianz-Kovarianz Matrix** $\widehat{V}(\widehat{\alpha}, \widehat{\beta}_1, \widehat{\beta}_2, \ldots \widehat{\beta}_K) = \widehat{V}(\widehat{\beta})$ geschätzt werden.
- Formal ist die $\widehat{V}(\widehat{\beta})$ definiert als

$$\widehat{V}(\widehat{\beta}) = \widehat{\sigma}^2 (X'X)^{-1}$$

- Die unbekannte Stichprobenvarianz der Störterme σ^2 kann durch die geschätzte Stichprobenvarianz der Störterme $\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n} \widehat{u}_i^2}{n-K-1}$ ersetzt werden.
- Da die Varianz der Störgrößen σ^2 nicht bekannt ist, sondern geschätzt werden muss, **erhöht sich die Unsicherheit** der Aussage über den geschätzten Parameter $\widehat{\beta}_k$. Für den Intervallschätzer und für Hypothesentests verwendet man daher die Perzentile der **t-Verteilung** anstelle der Standardnormalverteilung.

Schätzen der Unsicherheit: Hintergrund

- Um die geschätzte **Varianz-Kovarianz Matrix** zu bestimmen, ist es hilfreich, die Daten der erklärenden Variablen mithilfe von **Matrixnotation** anzugeben:

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1K} \\ 1 & x_{21} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nK} \end{bmatrix}$$

- Die geschätzte **Varianz-Kovarianz Matrix** enthält dann folgende Einträge:

$$\widehat{V}(\widehat{\beta}) = \widehat{\sigma}^2(X'X)^{-1} = \begin{bmatrix} \widehat{var}(\widehat{\alpha}) & \widehat{cov}(\widehat{\alpha}, \widehat{\beta}_1) & \cdots & \widehat{cov}(\widehat{\alpha}, \widehat{\beta}_K) \\ \widehat{cov}(\widehat{\beta}_1, \widehat{\alpha}) & \widehat{var}(\widehat{\beta}_1) & \cdots & \widehat{cov}(\widehat{\beta}_1, \widehat{\beta}_K) \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{cov}(\widehat{\beta}_K, \widehat{\alpha}) & \widehat{cov}(\widehat{\beta}_K, \widehat{\beta}_1) & \cdots & \widehat{var}(\widehat{\beta}_K) \end{bmatrix}$$

- $(X'X)^{-1}$ ist eine Matrix der Dimension $(K+1) \times (K+1)$.
- $\widehat{V}(\widehat{\beta})$ berücksichtigt die mögliche Korrelation zwischen den unabhängigen Variablen.
- Um die **Standardfehler** der geschätzten Parameter zu erhalten, muss man die **Wurzel der Werte der Hauptdiagonalen** berechnen.

  Zum Beispiel: $\widehat{se}(\widehat{\beta}_k) = \sqrt{\widehat{var}(\widehat{\beta}_k)}$

Intervallschätzer im Univariaten Modell: Beispiel

- Wie hilft uns der Intervallschätzer, die **statistische Unsicherheit des geschätzten Parameters** $\widehat{\beta}$ zu bestimmen?
- In unserem Anwesenheit-Studienerfolg Modell erhalten wir folgende Werte für $\widehat{\beta}$ und $\widehat{se}(\widehat{\beta})$:

$$\widehat{\beta} = 4.88$$
$$\widehat{se}(\widehat{\beta}) = 0.88$$

- Basierend auf diesen Schätzergebnissen liegt der **(unbeobachtete) Parameter** β
  - in **90 %** aller Fälle (d.h. $\alpha = 0.10$) im Intervall $[4.88 - 1.86 \cdot 0.88; 4.88 + 1.86 \cdot 0.88] = \mathbf{[3.25; 6.52]}$
  - in **95 %** aller Fälle (d.h. $\alpha = 0.05$) im Intervall $[4.88 - 2.31 \cdot 0.88; 4.88 + 2.31 \cdot 0.88] = \mathbf{[2.85; 6.91]}$
  - in **99 %** aller Fälle (d.h. $\alpha = 0.01$) im Intervall $[4.88 - 3.36 \cdot 0.88; 4.88 + 3.36 \cdot 0.88] = \mathbf{[1.93; 7.84]}$
  - **wenn** die Annahmen A1-A3 und B1-B4 erfüllt sind!!

Intervallschätzer: Anmerkungen

- Die Größe α legt die **Irrtumswahrscheinlichkeit** fest und wird auch als **Signifikanzniveau** bezeichnet.
- In der Praxis konzentrieren wir uns oft auf ein Signifikanzniveau von 5 %, d.h. in 95 % aller Fälle liegt der richtige Wert β innerhalb des entsprechenden Konfidenzintervalls.
- In **EXCEL** ist unter **Daten** $\rightarrow$ **Analyse** $\rightarrow$ **Datenanalyse** $\rightarrow$ **Regression** ein "Konfidenzniveau" von 95 % (also eine Irrtumswahrscheinlichkeit von $\alpha = 0.05$) voreingestellt (kann aber verändert werden).
- Der Satz "In X % aller Fälle..." ist darauf zurückzuführen, dass wir bei der (theoretischen) Konstruktion des Konfidenzintervalls von unendlich vielen Zufallsstichproben ausgehen.
- Oft hört/liest man auch "Das Konfidenzintervall umschließt den wahren Wert mit Wahrscheinlichkeit von X %". Diese Aussage erleichtert das Verständnis für Konfidenzintervalle, ist aber eigentlich nicht korrekt.
- Wenn die Anzahl der **Beobachtungen relative groß** ist, können die Perzentile der **t-Verteilung durch** die Perzentile einer **Standardnormalverteilung ersetzt** werden.

Intervallschätzer in Multivariaten Model

- Wir können für jeden Punktschätzer $\widehat{\beta}_k$ analog zu den vorherigen Folien einen Intervallschätzer konstruieren.
- Der unbeobachtete Parameter β liegt daher in $(1-\alpha)$ der Fälle in folgendem **Intervall**: $\left[\widehat{\beta}_k - t_{n-K-1;1-\frac{\alpha}{2}} \cdot \widehat{se}(\widehat{\beta}_k); \widehat{\beta}_k + t_{n-K-1;1-\frac{\alpha}{2}} \cdot \widehat{se}(\widehat{\beta}_k)\right]$

Regressionstabelle:

	Koeff.	*Stand.-f.*	*t-Stat.*	*P-Wert*	*Untere 95%*	*Obere 95%*
Schnittpunkt	3.032	13.073	0.232	0.823	-27.880	33.945
Anwesenheit	4.504	1.095	4.114	0.004	1.915	7.093
Eink. Eltern	2.550	4.037	0.632	0.548	-6.997	12.096

- Beispiel: 95 % Konfidenzintervall (d.h. Irrtumswahrscheinlichkeit $\alpha = 0.05$) für Anwesenheit:
  - $\widehat{\beta}_{Anwesenheit} = 4.504$ (siehe Tabelle)
  - $\widehat{se}(\widehat{\beta}_{Anwesenheit}) = 1.095$ (siehe Tabelle)
  - $t_{n-K-1;1-\frac{\alpha}{2}} = t_{10-2-1;1-\frac{0.05}{2}} = t_{7;0.975} = 2.365$ (siehe T.INV(0.975;7))
  - $[4.504 - 2.365 \cdot 1.095; 4.504 + 2.365 \cdot 1.095] = [1.915; 7.093]$

Testen von Hypothesen: Motivation

- Konfidenzintervalle sind sehr nützlich, aber manchmal möchten wir gewisse **Hypothesen** direkter **testen**.
- Sie als Studierende (aber auch mich als Lehrenden) könnten in unserem Anwesenheits-Klausur-Modell interessieren, ob
  - der geschätzter Effekt unterschiedlich von 0 ist ($\beta \neq 0$).
  - höhere Anwesenheit zu einer Punktsteigerung führt ($\beta > 0$).
  - die Anwesenheit bei einer zusätzlichen Einheit die Klausurpunkte um mehr als 3 Punkte steigert ($\beta > 3$).
- Wie diese Fragen zeigen, kann man an **einseitigen oder zweiseitigen Hypothesentests** interessiert sein.
- Wir können solche Hypothesen mit unseren bisher erlernten Mitteln einfach formulieren und testen!

Zweiseitiger Hypothesentest: Univariate Regression

- Vorbereitung: Allgemeine und spezielle **Voraussetzungen** des Tests **überprüfen** und **Irrtumswahrscheinlichkeit** (α-Fehler) **festlegen** (z.B. $\alpha = 0.01$ oder $\alpha = 0.05$).
- Formulierung der Hypothesen:

$$\text{Nullhypothese } H_0 : \beta = q$$
$$\text{Alternativhypothese } H_1 : \beta \neq q$$

- Mit dem Hypothesentest wird überprüft, ob es **wahrscheinlich** ist, dass der **(unbeobachtete) Parameter** β in der Nähe bzw. nicht in der Nähe von q liegt. Oft wird $q = 0$ gewählt und überprüft, ob der geschätzte Parameter $\widehat{\beta}$ signifikant von 0 verschieden ist (und sich β wahrscheinlich von 0 unterscheidet).
- H_0 wird abgelehnt, wenn $\widehat{\beta}$ signifikant **größer oder** signifikant **kleiner** als q ist $\rightarrow$ daher die Bezeichnung **"zweiseitiger" Hypothesentest**.

Standardisierte Teststatistik

1. **Hypothesen:**
   - $H_0: \ \beta = q; \ H_1: \ \beta \neq q$
2. **Entscheidungsregel**: mittels standardisierten Teststatistik
   - H_0 wird nicht verworfen, wenn t-Wert $= \left| \frac{\widehat{\beta} - q}{\widehat{se}(\widehat{\beta})} \right| \leq t_{n-2;1-\frac{\alpha}{2}}$
   - H_0 zugunsten von H_1 verworfen, wenn t-Wert $= \left| \frac{\widehat{\beta} - q}{\widehat{se}(\widehat{\beta})} \right| > t_{n-2;1-\frac{\alpha}{2}}$
   - Vorgehensweise:

     Berechnung einer **standardisierten Teststatistik (t-Wert)**:
     t-Wert $= \left| \frac{\widehat{\beta} - q}{\widehat{se}(\widehat{\beta})} \right|$

     Ermittlung des **kritischen Werts** c der Verteilung: $c = t_{n-2;1-\frac{\alpha}{2}}$

     H_0 **wird nicht verworfen**, wenn t-Wert $\leq c$

     H_0 **wird verworfen**, wenn t-Wert $> c$

Hinweis: Oft wird der t-Wert (in Lehrbüchern oder etwa auch in EXCEL) als $\frac{\widehat{\beta} - q}{\widehat{se}(\widehat{\beta})}$ berechnet und kann demnach auch negative Werte annehmen. Wichtig ist, dass Sie in diesem Fall den Betrag des ausgewiesenen t-Werts mit dem kritischen Wert c vergleichen!

Zweiseitiger Hypothesentest: Anleitung

Um einen **zweiseitigen Hypothesentest** durchzuführen, müssen Sie folgende Schritte durchführen:

- Aufstellen von H_0 ($H_0 : \beta = q$) und H_1 ($H_1 : \beta \neq q$)
- Festlegung des Signifikanzniveaus α
- Schätzung von $\widehat{se}(\widehat{\beta})$
- Berechnung des t-Wertes: $t = \left| \frac{\widehat{\beta} - q}{sd(\widehat{\beta})} \right|$
- Ermittlung des kritischen Wertes c: $c = t_{n-2;1-\frac{\alpha}{2}}$
- Vergleichen von c und t, **falls** $t > c$ **wird** H_0 **verworfen**.

Zweiseitiger Test: Beispiel

- Nehmen wir an, wir wurden beauftragt herauszufinden, ob es einen **Wirkungszusammenhang zwischen Anwesenheit und Prüfungserfolg** gibt. Wir wollen wissen, ob $\beta \neq 0$.
- Die Aussage ($\beta \neq 0$) wird daher als Alternativhypothese H_1 formuliert. Daher haben wir folgende Hypothesen: $H_0 : \beta = 0$ und $H_1 : \beta \neq 0$.
- Mit der Kleinstquadratschätzung haben wir für den Parameter β und den Standardfehler $se(\widehat{\beta})$ folgende Werte geschätzt:

$$\widehat{\beta} = 4.88$$

$$\widehat{se}(\widehat{\beta}) = 0.88$$

- Hinweis: Damit unser **Test valide** ist, müssen alle **Annahmen A1-A3 und B1-B4 erfüllt** sein. Wir nehmen an dieser Stelle an, dass diese Annahmen erfüllt sind.

Zweiseitiger Test: Beispiel

Wir testen die Hypothese, ob es einen **Wirkungszusammenhang zwischen Anwesenheit und Prüfungserfolg** gibt, in den folgenden Schritten:

1. Aufstellen von H_0 ($H_0 : \beta = 0$) und H_1 ($H_1 : \beta \neq 0$)
2. Festlegung des Signifikanzniveaus α: $\alpha = 0.10$
3. Schätzung von $se(\widehat{\beta})$: $\widehat{se}(\widehat{\beta}) = 0.88$
4. Berechnung des t-Wertes: $t = \left| \frac{\widehat{\beta} - q}{sd(\widehat{\beta})} \right| = \left| \frac{4.88 - 0}{0.88} \right| = 5.54$
5. Ermittlung des kritischen Wertes c für eine t-Verteilung mit 8 Freiheitsgraden: $c = t_{8;0.95} = 1.86$
6. Vergleichen von c und t: H_0 wird verworfen, da $|5.54| > 1.86$.

Einseitiger Hypothesentest: Univariate Regression

- Oft interessiert uns nicht nur, ob β unterschiedlich von q ist, sondern auch, ob der Parameter **größer oder kleiner** als q ist. Wir sprechen dann von einem **einseitigen Hypothesentest**.
- Wenn wir an der (*rechtsseitigen*) Hypothese, ob $\beta > q$ interessiert sind:

$$H_0 : \beta \leq q \ \ , H_1 : \beta > q$$

- Wenn wir an der (*linksseitigen*) Hypothese, ob $\beta < q$ interessiert sind:

$$H_0 : \beta \geq q \ \ , H_1 : \beta < q$$

- **Beachten Sie**:
  - Die Aussage, die wir zeigen möchten (an der wir interessiert sind), formulieren wir immer als Alternativhypothese H_1.
  - $\beta = q$ ist immer Teil der Nullhypothese H_0.

Einseitiger (rechtsseitiger) Hypothesentest: Anleitung

Um einen **einseitigen (rechtsseitiger) Hypothesentest** durchzuführen (wenn Sie interessiert sind, ob $\beta > q$), müssen Sie folgende Schritte durchführen:

1. Aufstellen von H_0 ($H_0 : \beta \leq q$) und H_1 ($H_1 : \beta > q$)
2. Festlegung des Signifikanzniveaus α
3. Schätzung von $\widehat{se}(\widehat{\beta})$
4. Berechnung des t-Wertes: $t = \left| \frac{\widehat{\beta} - q}{sd(\widehat{\beta})} \right|$
5. Ermittlung des kritischen Wertes c: $c = t_{n-2;1-\alpha}$
6. Vergleichen von c und t: falls $t > c$ und $\widehat{\beta} > q$ wird H_0 verworfen

Einseitiger (linksseitiger) Hypothesentest: Anleitung

Um einen **einseitigen (linksseitigen) Hypothesentest** durchzuführen (wenn Sie interessiert sind, ob $\beta < q$), müssen Sie folgende Schritte durchführen:

1. Aufstellen von H_0 ($H_0 : \beta \geq q$) und H_1 ($H_1 : \beta < q$)
2. Festlegung des Signifikanzniveaus α
3. Schätzung von $\widehat{se}(\widehat{\beta})$
4. Berechnung des t-Wertes: $t = \left| \frac{\widehat{\beta} - q}{sd(\widehat{\beta})} \right|$
5. Ermittlung des kritischen Wertes c: $c = t_{n-2;1-\alpha}$
6. Vergleichen von c und t: falls $t > c$ und $\widehat{\beta} < q$ wird H_0 verworfen

Einseitiger (rechtsseitiger) Hypothesentest: Beispiel I

- Es wird überlegt, eine **Anwesenheitspflicht** einzuführen. Dies sollte aber nur geschehen, wenn es genug **Evidenz** dafür gibt, dass **Anwesenheit die Klausurergebnisse verbessert**. Wir sind daher interessiert, ob es sehr wahrscheinlich ist, dass $\beta > 0$.
- Um eine **Empfehlung** abgeben zu können, ob eine Anwesenheitspflicht die Klausurergebnisse verbessert, müssen wir die entsprechenden Hypothesen formulieren:

$$H_0 : \beta \leq 0; \quad H_1 : \beta > 0$$

- Von unserer Hypothese sehen wir, dass wir einen **einseitiger Hypothesentest** durchführen müssen. Wir definieren die interessante Hypothese als H_1.
- Um einen Hypothesentest durchführen zu können, müssen wir ein **Signifikanzniveau** festlegen. Da die Einführung einer Anwesenheitspflicht einen starken Eingriff in die Autonomie der Studierenden darstellt,legen wir daher eine **sehr niedrige Irrtumswahrscheinlichkeit** von 1 % fest: $\alpha = 0.01$

Einseitiger (rechtsseitiger) Hypothesentest: Beispiel I

Um einen **einseitigen (rechtsseitiger) Hypothesentest** durchzuführen, müssen Sie folgende Schritte durchführen:

1. Aufstellen von H_0 ($H_0 : \beta \leq 0$) und H_1 ($H_1 : \beta > 0$)
2. Festlegung des Signifikanzniveaus α: $\alpha = 0.01$
3. Schätzung von $se(\widehat{\beta})$: $\widehat{se}(\widehat{\beta}) = 0.88$
4. Berechnung des t-Wertes: $t = \left| \frac{\widehat{\beta} - q}{sd(\widehat{\beta})} \right| = \left| \frac{4.88 - 0}{0.88} \right| = 5.54$
5. Ermittlung des kritischen Wertes c: $c = t_{n-2;1-\alpha} = t_{8;0.99} = 2.90$
6. Vergleichen von c und t: Da $t > c$ und $\widehat{\beta} > 0$ wird H_0 verworfen.

Schlussfolgerung: Es ist **sehr wahrscheinlich** (zu mehr als 99 %), dass es in der Grundgesamtheit einen **positiven Effekt** von Anwesenheit auf den Klausurerfolg gibt. Man könnte daher eine **entsprechende Empfehlung abgeben**.

Umsetzung in EXCEL

- **Daten** $\rightarrow$ **Analyse** $\rightarrow$ **Datenanalyse** $\rightarrow$ **Regression**

	Koeff.	*Stand.-f.*	*t-Stat.*	*P-Wert*	*Untere 95%*	*Obere 95%*
Schnittpunkt	5.233	12.117	0.432	0.6772	-22.710	33.175
Anwesenheit	4.883	0.881	5.544	0.0005	2.852	6.914

- **EXCEL** berechnet **automatisch** folgende Maßzahlen:
  - *Koeffizient* $= \widehat{\beta}$ (Punktschätzer)
  - *Standardfehler* $= \widehat{se}(\widehat{\beta})$
  - *t-Statistik* $= \frac{\widehat{\beta}-0}{\widehat{se}} = \frac{\widehat{\beta}}{\widehat{se}}$ ist der t-Wert für den zweiseitigen Hypothesentest mit $H_0 : \beta = 0$. (Falls t-Statistik < 0 muss der Betrag genommen werden.)
  - *P-Wert* ist die dazugehörige Wahrscheinlichkeit.
  - *Untere* bzw. *Obere 95 %* gibt den Intervallschätzer für das 95 %-Konfidenzintervall an. Es kann wahlweise noch ein weiteres (z.B. 90 % oder 99 %) Intervall ausgewiesen werden.
  - Die Nullhypothese $H_0 : \beta = 0$ wird verworfen wen der *P-Wert* kleiner als das Signifikanzniveau α ist.

Umsetzung in EXCEL

- **Daten** $\rightarrow$ **Analyse** $\rightarrow$ **Datenanalyse** $\rightarrow$ **Regression**

	Koeff.	*Stand.-f.*	*t-Stat.*	*P-Wert*	*Untere 95%*	*Obere 95%*
Schnittpunkt	5.233	12.117	0.432	0.6772	-22.710	33.175
Anwesenheit	4.883	0.881	5.544	0.0005	2.852	6.914

- Um **andere Hypothesen** zu testen, muss:
  - selbständig der t-Wert mit $t = \left| \frac{\widehat{\beta} - q}{\widehat{se}} \right|$ berechnet werden.
  - der kritische Wert c der t-Verteilung bestimmt werden:
    - mit T.INV$(1 - \alpha/2;\ n-2)$ oder T.INV.2S$(\alpha;\ n-2)$ bei zweiseitigem Hypothesentest
    - mit T.INV$(1 - \alpha;\ n-2)$ bei einseitigem Hypothesentest
  - Alternativ dazu kann direkt der p-Wert berechnet werden:
    - T.VERT.2S(|t-Wert|; $n-2$): p-Wert bei zweiseitigem Hypothesentest; wenn p-Wert$< \alpha$ wird H_0 verworfen
    - T.VERT.RE(|t-Wert|; $n-2$): p-Wert bei einseitigem Hypothesentest; wenn p-Wert$< \alpha$ wird H_0 verworfen (und $\widehat{\beta} > q$ bei einem rechtsseitigen Test bzw. $\widehat{\beta} < q$ bei einem linksseitigen Test)

Testen von einfachen Hypothesen: Multivariate Regression

Das **Testen einfacher Hypothesen** verläuft **genauso wie beim einfachen linearen Regressionsmodell**. Mit einfachen Hypothesen sind gemeint:

- Zweiseitiger Hypothesentest: $H_0 : \beta_k = q$, $H_1 : \beta_k \neq q$
- Einseitiger rechtsseitiger Hypothesentest: $H_0 : \beta_k \leq q$, $H_1 : \beta_k > q$
- Einseitiger linksseitiger Hypothesentest: $H_0 : \beta_k \geq q$, $H_1 : \beta_k < q$
- Bei Schwierigkeiten bitte bei Einheit 6 nachlesen!

Testen von einfachen Hypothesen: Regressionstabelle in EXCEL

Regressionstabelle:

	Koeff.	*Stand.-f.*	*t-Stat.*	*P-Wert*	*Untere 95%*	*Obere 95%*
Schnittpunkt	3.032	13.073	0.232	0.823	-27.880	33.945
Anwesenheit	4.504	1.095	4.114	0.004	1.915	7.093
Eink. Eltern	2.550	4.037	0.632	0.548	-6.997	12.096

- Bei der ausgewiesenen Regressionstabelle wird für jeden geschätzten Koeffizienten der **zweiseitige Hypothesentest** $H_0 : \beta_k = 0$ und $H_1 : \beta_k \neq 0$ durchgeführt.
- Wenn die $|t - Statistik| > t_{n-K-1;1-\frac{\alpha}{2}}$, dann wird H_0 verworfen.
- Wenn der $P - Wert < \alpha$, dann wird H_0 verworfen.
- Beide Möglichkeiten müssen **immer zum gleichen Ergebnis** führen.

Testen komplexer Hypothesen

- Man kann auch **komplexe Nullhypothese** testen, wie z.B.:
  - $H_0 : \frac{\beta_1}{3} + \frac{\beta_2}{2} = 0$
  - oder allgemein: $H_0 : r_k \beta_k + r_m \beta_m = q$
  - Man kann **mehrere Hypothesen auf einmal** testen. Allgemein spricht man vom **simultanen Test mehrerer Linearkombinationen von Parametern**.
    - ⋆ Zum Beispiel:
      $H_0 : \beta_1 = 0$ und gleichzeitig $\beta_2 = 0$
      $H_1 : \beta_1 \neq 0$ und/oder $\beta_2 \neq 0$
  - Diese Tests sind relativ schwierig, da in EXCEL komplexe Hypothesentests nicht implementiert sind. Die Tests "händisch" zu berechnen erfordert die Schätzung der Kovarianz der beiden geschätzten Parameter $\widehat{\beta}_k$ und $\widehat{\beta}_m$, was die Möglichkeiten in unserem Kurs allerdings übersteigt.

Hypothesentest: Zusammenfassung

- Um **Hypothesen für die Grundgesamtheit** zu testen, muss die **statistische Unsicherheit**, die mit der **Schätzungen der Modellparameter** verknüpft ist, berücksichtigt werden.
- Die Hypothese, an der wir interessiert sind, wird immer als Alternativhypothese H_1 formuliert.
- Wenn man daran interessiert ist, ob sich der Parameter von einem bestimmten Wert unterscheidet (d.h. $H_1 : \beta_k \neq q$), muss ein **zweiseitiger Hypothesentest** durchgeführt werden.
- Wenn man daran interessiert ist, ob der Parameter größer als ein bestimmter Wert ist (d.h. $H_1 : \beta_k > q$), muss ein **einseitiger (rechtsseitiger) Hypothesentest** durchgeführt werden.
- Wenn man daran interessiert ist, ob der Parameter kleiner als ein bestimmter Wert ist (d.h. $H_1 : \beta_k < q$), muss ein **einseitiger (linksseitiger) Hypothesentest** durchgeführt werden.
- Die Hypothesentests liefern nur dann gültige Ergebnisse, wenn alle **Annahmen A1-A3 und B1-B4 erfüllt** sind.

Interpretation des Linearen Regressionsmodell

Wir wollen noch drei Aspekte besprechen, die sowohl das einfache als auch das multivariate lineare Regressionsmodell betreffen:

1) **Bestimmtheitsmaß** R^2:
   - Wie viel der Variation der zu erklärenden Variable y kann durch unser Modell erklärt werden?
   - Gütekriterium für die Regression.
   - Damit können wir alle **Maßzahlen**, die EXCEL bei einer **Regression** ausgibt, **verstehen und interpretieren**.
2) **Prognose**:
   - Welchen Wert für die erklärte Variable y kann man erwarten, wenn man die Werte der erklärenden Variablen kennt (**bedingte Prognose**)?
3) **Wirkungszusammenhang**:
   - Wann ist die Interpretation eines Wirkungszusammenhangs (d.h. eine **kausale Interpretation** des geschätzten Parameters) zulässig, selbst wenn die Annahme A1 (Vollständigkeit und Relevanz) nicht erfüllt ist?
   - **Verzerrung aufgrund von ausgelassenen Variablen** (englisch: *omitted variable bias*).

Das Bestimmtheitsmaß R^2

- Wie viel der **Variation der zu erklärenden Variable** y kann **durch unser Modell erklärt** werden?
- Vorbemerkungen:
  - Regressionsmodell: $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + u_i$
  - geschätztes Modell: $\widehat{y}_i = \widehat{\alpha} + \widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i} + \cdots + \widehat{\beta}_K x_{Ki}$
  - geschätzte Störgrößen (Residuen):
    $\widehat{u}_i = y_i - \widehat{y}_i = y_i - (\widehat{\alpha} + \widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i} + \cdots + \widehat{\beta}_K x_{Ki})$
  - daraus ergibt sich: $y_i = (\widehat{\alpha} + \widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i} + \cdots + \widehat{\beta}_K x_{Ki}) + \widehat{u}_i$
  - ein Teil der Schwankungen von y_i kann durch das geschätzte Modell $(\widehat{\alpha} + \widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i} + \cdots + \widehat{\beta}_K x_{Ki})$ erklärt werden, ein Teil muss unerklärt blieben $(\widehat{u}_i)$.
  - das Modell ist **umso besser, je größer der Teil ist, der erklärt werden kann**.

Das Bestimmtheitsmaß R^2

- Wie viel der Variation der zu erklärenden Variable y kann durch unser Modell erklärt werden?
- **Begriffsbestimmung**:
  - $y_i = (\widehat{\alpha} + \widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i} + \cdots + \widehat{\beta}_K x_{Ki}) + \widehat{u}_i$
  - $S_{yy} \equiv \sum_i (y_i - \bar{y})^2 =$ **Variation der abhängigen Variable** y
  - $S_{\widehat{y}\widehat{y}} \equiv \sum_i (\widehat{y}_i - \bar{y})^2 =$ durch unser Modell **erklärte Variation** von y
  - $S_{\widehat{u}\widehat{u}} \equiv \sum_i \widehat{u}_i^2 =$ durch unser Modell **nicht erklärte Variation** von y
  - es lässt sich zeigen, dass $S_{yy} = S_{\widehat{y}\widehat{y}} + S_{\widehat{u}\widehat{u}}$
- **Definition**:
  - Das Bestimmtheitsmaß R^2 misst den **Anteil an der gesamten Variation von y, der durch unser Modell erklärt wird**.

$$R^2 = \frac{\text{erklärte Variation}}{\text{gesamte Variation}} = \frac{S_{yy} - S_{\widehat{u}\widehat{u}}}{S_{yy}} = \frac{S_{\widehat{y}\widehat{y}}}{S_{yy}}$$

$\longrightarrow$ Damit können wir alle Maßzahlen erklären, die EXCEL ausgibt!

Prognose

Um eine **bedingte Prognose** zu erhalten, setzt man in das geschätzte Modell $\widehat{y}_i = \widehat{\alpha} + \widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i} + \cdots + \widehat{\beta}_K x_{Ki}$ anstelle von $x_{1i}, x_{2i}, \ldots, x_{Ki}$ bestimmte Werte $x_{10}, x_{20}, \ldots, x_{K0}$ ein, um die **Punktprognose** $\widehat{y}_0$ zu erhalten. Diese Prognose ist natürlich auch mit **Unsicherheit** behaftet (was wir hier aber nicht vertiefen können).

Regressionstabelle:

	Koeff.	*Stand.-f.*	*t-Stat.*	*P-Wert*	*Untere 95%*	*Obere 95%*
Schnittpunkt	3.032	13.073	0.232	0.823	-27.880	33.945
Anwesenheit	4.504	1.095	4.114	0.004	1.915	7.093
Eink. Eltern	2.550	4.037	0.632	0.548	-6.997	12.096

Beispiel:

- geschätzte Gleichung: $\widehat{y}_{Punkte} = 3.032 + 4.504 \cdot x_{Anwesenheit} + 2.550 \cdot x_{Eink.Eltern}$
- gesucht: Punktprognose für eine Studierende, die 13 Mal anwesend war und deren Eltern 2,500 Euro verdienen.
- $\widehat{y}_{Punkte} = 3.032 + 4.504 \cdot 13 + 2.550 \cdot 2.5 = 67.959$
- **Interpretation**: Bei einer Studierenden, die 13 Mal anwesend war und deren Eltern 2,500 Euro verdienen, kann man erwarten, dass sie auf die Klausur etwa 68 Punkte bekommt.

Wirkungszusammenhang: Verzerrung aufgrund von ausgelassenen Variablen

- Die Annahme A1 besagt, dass **alle relevanten exogenen Variablen** im Modell **berücksichtigt** werden bzw. (umgekehrt formuliert) keine relevante exogene Variable unberücksichtigt bleibt (ausgelassen wird).
- Die **Verzerrung aufgrund von ausgelassenen Variablen** (englisch: *omitted variable bias*) bedeutet, dass ausgelassene Variablen zu verzerrten Schätzern führen, wenn die ausgelassene Variable z (i) relevant ist und (ii) mit einer Variable x_k korreliert ist. (Würden wir z kennen, können wir sogar die Richtung und das Ausmaß der Verzerrung bestimmen.)

Verzerrung aufgrund von ausgelassenen Variablen: Beispiel

Beispiel: Schätzmodell zur Erklärung der Höhe der Miete:

Regressionstabelle:

	Koeff.	*Stand.-f.*	*t-Stat.*	*P-Wert*	*Untere 95%*	*Obere 95%*
Schnittpunkt	30.088	18.056	1.670	0.096	-5.322	65.498
Größe	0.561	0.016	34.980	0.000	0.530	0.593
Einkommen	1.306	0.142	9.170	0.000	1.026	1.585
Frau	2.352	9.347	0.250	0.801	-15.979	20.683
Baujahr	22.717	2.938	7.730	0.000	16.954	28.480

Was passiert, wenn wir eine **relevanten Variable**, die mit anderen Variablen **korreliert** ist auslasse (nicht beobachte)?

Auslassen einer relevanten und korrelierten Variable

Regressionstabelle: (vollständiges Modell)

	Koeff.	*Stand.-f.*	*t-Stat.*	*P-Wert*	*Untere 95%*	*Obere 95%*
Schnittpunkt	30.088	18.056	1.670	0.096	-5.322	65.498
Größe	0.561	0.016	34.980	0.000	0.530	0.593
Einkommen	1.306	0.142	9.170	0.000	1.026	1.585
Frau	2.352	9.347	0.250	0.801	-15.979	20.683
Baujahr	22.717	2.938	7.730	0.000	16.954	28.480

Regressionstabelle: (restringiertes Modell)

	Koeff.	*Stand.-f.*	*t-Stat.*	*P-Wert*	*Untere 95%*	*Obere 95%*
Schnittpunkt	489.440	15.775	31.030	0.000	458.502	520.378
~~Größe~~						
Einkommen	2.498	0.176	14.210	0.000	2.153	2.843
Frau	6.867	11.896	0.580	0.564	-16.464	30.197
Baujahr	19.727	3.739	5.280	0.000	12.395	27.059

Abschließende Betrachtung des Anwesenheits-Klausur-Modells

In unserem einfachen Anwesenheits-Klausur-Modells ist die **Interpretation** des geschätzten Parameters $\widehat{\beta}_{Anwesenheit}$ a**ls Wirkungszusammenhang nicht zulässig**, da es eine Vielzahl an Variablen geben kann (und geben wird), die nicht berücksichtigt werden können (Einkommen und Bildung der Eltern, Motivation, Erfahrung, Fähigkeiten der Selbstorganisation,...), obwohl diese Variablen

(i) relevant sind (und ein Auslassen Annahme A1 verletzt), und

(ii) mit der Anwesenheit korreliert sind (zumindest teilweise).

Es ist daher eine **Verzerrung aufgrund von ausgelassenen Variablen** (*omitted variable bias*) zu erwarten.

Abschließende Betrachtung des Anwesenheits-Klausur-Modells

Wäre unter folgenden Bedingungen die Interpretation als Wirkungszusammenhang möglich?

- Gedankenexperiment: Aufgrund der vorgeschriebenen Verpflichtung, ausreichend Abstand zu halten, darf der Hörsaal nur zu 50 % ausgelastet werden. Daher wird der Unterricht hybrid angeboten. Die Ausgestaltung des hybriden Unterrichts erfolgt derart, dass für jede Woche 50 % der Studierenden zufällig ausgewählt werden, die in Präsenz teilnehmen dürfen. Für diese Studierenden ist die Teilnahme verpflichtend. Die anderen Studierenden müssen sich den Stoff selbständig erarbeiten (unter den gleichen Bedingungen wie früher, d.h. sie bekommen außer Folien und Lehrbücher keine zusätzlichen Unterrichtsmaterialien). Die Zufallsauswahl der Studierenden für jede Einheit führt dazu, dass manche Studierende öfters teilnehmen dürfen/müssen als andere Studierende.

Abschließende Betrachtung des Anwesenheits-Klausur-Modells

Wäre unter folgenden Bedingungen die Interpretation als Wirkungszusammenhang möglich?

- Ja, unter diesen Bedingungen ist eine **kausale Interpretation** des geschätzten Parameters $\widehat{\beta}_{Anwesenheit}$ (die Interpretation als Wirkungszusammenhang) **zulässig**. Das Modell ist zwar **unvollständig** (d.h. **Annahme A1 ist verletzt**), aber die Variable, die uns interessiert (die Anwesenheit) ist mit den unbeobachteten (und daher ausgelassenen) Variablen **nicht korreliert**! Das liegt daran, dass die **Anwesenheit** nicht von den Studierenden gewählt, sondern **zufällig bestimmt** wird.

Um den Wirkungsmechanismen zu bestimmen, werden oft **Experimente** durchgeführt, in denen **zufällig bestimmt** wird, welche Beobachtungen (hier: Studierende) eine **"Behandlung"** (englisch *treatment*; hier: die Anwesenheit) bekommen. Wenn ein Experiment nicht durchführbar ist, kann durch ein **Quasi-Experiment** ein Wirkungszusammenhang geschätzt werden.