SE Geographie und Ökonomie

Einheit 7:
Multivariates Lineares Regressionsmodell

Dieter Pennerstorfer
dieter.pennerstorfer@jku.at

Institut für Volkswirtschaftslehre
Johannes Kepler Universität Linz Linz

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 7

- 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** ist eine Generalisierung des einfachen linearen Regressionsmodells mit $K > 1$ erklärenden Variablen.
- Das multivariate lineare Regressionsmodell hat die Form:

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

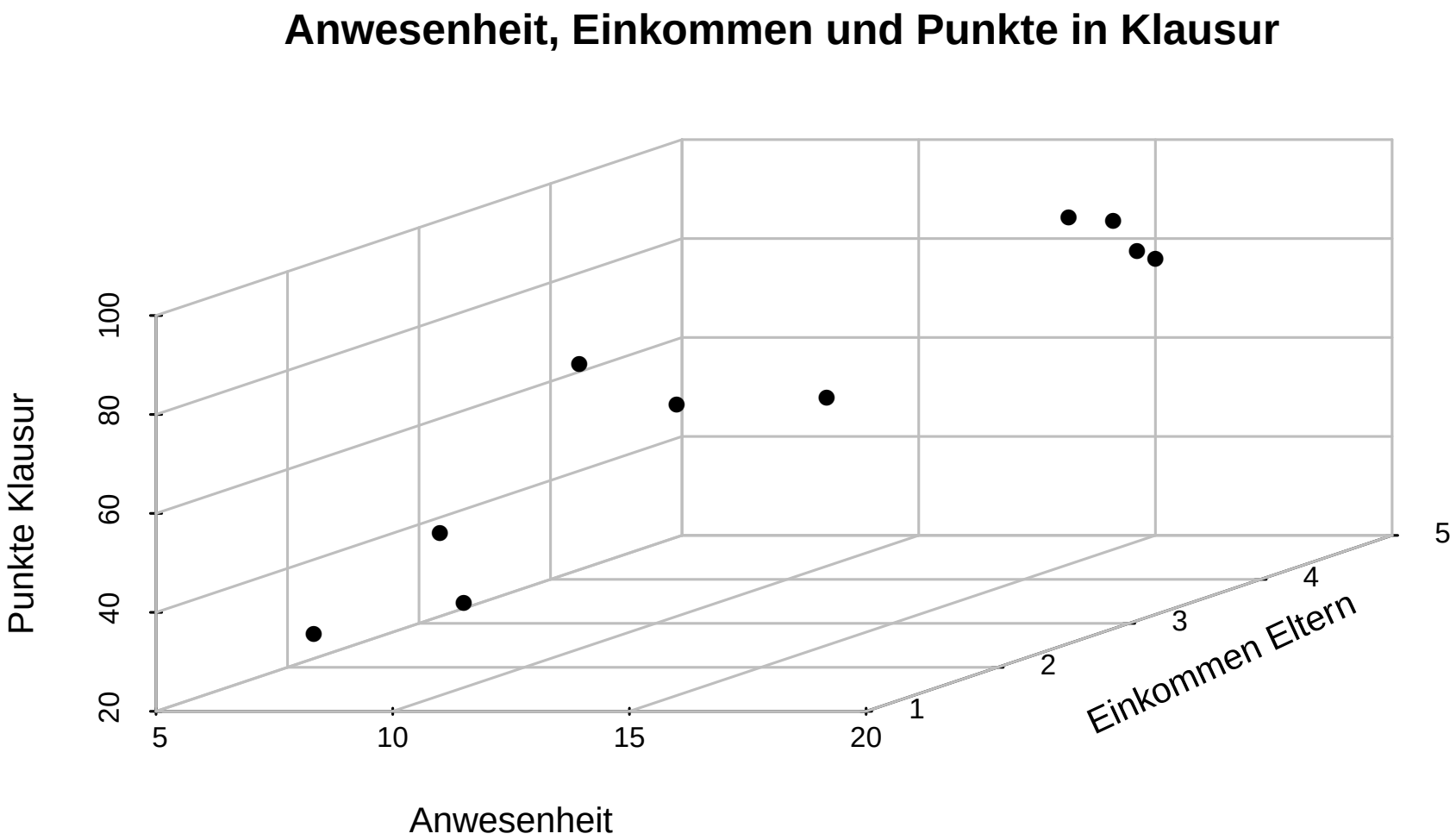
  für $i = 1, 2, ..., n$ Beobachtungen
- Wir erklären y_i **linear** durch $x_{1i}, x_{2i}, \ldots, x_{Ki}$. Wir gehen davon aus, dass $x_{1i}, x_{2i}, \ldots, x_{Ki}$ auf y_i wirken (und nicht umgekehrt).
- y_i wird als **Regressand**, oder als **endogene**, **abhängige** oder **erklärte Variable** bezeichnet.
- $x_{1i}, x_{2i}, \ldots, x_{Ki}$ 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:

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

- 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.
  - Wenn die entsprechenden A- und B-Annahmen (sie weiter unten) erfüllt sind, können die geschätzten Parameter als Wirkungszusammenhänge interpretiert werden.
  - $\widehat{\beta}_1$ 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.

- 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.
- 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: Lineariät

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

Schätzen der Unsicherheit

- 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 in Multivariaten Regressionsmodellen

- Wir können daher für jeden Punktschätzer $\widehat{\beta}_k$ analog zu Einheit 6 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 einfachen Hypothesen

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: $H_0 : r_k\beta_k + r_m\beta_m = q$
  - Dies ist schwierig, da in EXCEL komplexe Hypothesentests nicht implementiert sind. Den Test “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.

Testen von mehrere Hypothesen

- 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 $\beta_2 \neq 0$
- Um mehrere Hypothesen auf einmal zu testen, muss ein sogenannter **F-Test** durchgeführt werden. Die Idee hinter dem Test ist, dass man die (größere) Summe der quadrierten **Residuen eines restringierten Modells** $S^0_{\widetilde{uu}}$ (unter H_0) mit der (kleineren) Summe der quadrierten **Residuen des vollständigen Modells** $S_{\widehat{uu}}$ (unter H_1) **vergleicht**.
- Wir lernen nur einen Spezialfall kennen, nämlich wenn:
  $H_0 : \beta_1 = \beta_2 = \cdots = \beta_K = 0.$
- Der entsprechende **F-Test berechnet sich als**

$$F = \frac{(S^0_{\widetilde{uu}} - S_{\widehat{uu}})/K}{S_{\widehat{uu}}/(n-K-1)}$$

  und muss mit dem **kritischen Wert** $F_{K,n-K-1}$ **der F-Verteilung** verglichen werden.

Testen von Hypothesen: Zusammenfassung

- Die Intuition der Hypothesentests des einfachen linearen Regressionsmodells ist relativ einfach auf multivariaten lineare Regressionsmodelle übertragbar.
  - Die geschätzten Standardfehler der geschätzten Parameter $\widehat{se}(\widehat{\beta}_k)$ sind schwieriger zu bestimmen als in der linearen Einfachregression, da die Parameter $\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_K$ einer **multivariaten Normalverteilung** folgen (wenn die A- und B-Annahmen erfüllt sind) und die geschätzten Standardfehler eines geschätzten Parameters **von der Korrelation mit anderen Variablen** abhängen.
  - In der **Anwendung** hat das für uns aber eine **geringe Relevanz**, da die geschätzten Standardfehler $\widehat{se}(\widehat{\beta}_k)$ ohnehin **von EXCEL berechnet und ausgewiesen** werden.
  - Sobald die geschätzten Standardfehler $\widehat{se}(\widehat{\beta}_k)$ bekannt sind, **können einfache Hypothesentests genau so durchgeführt werden, wie bei der linearen Einfachregression**.
- Das Testen von komplexen Hypothesen sowie das gleichzeitige Testen mehrerer Hypothesen wird hier nicht vertieft. Sie sollen lediglich den F-Test interpretieren können, der in EXCEL ausgewiesen wird.

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
  - $S_{\widehat{yy}} \equiv \sum_i (\widehat{y}_i - \bar{y})^2 =$ durch unser Modell **erklärte Variation** von y
  - $S_{\widehat{uu}} \equiv \sum_i \widehat{u}_i^2 =$ durch unser Modell **nicht erklärte Variation** von y
  - es lässt sich zeigen, dass $S_{yy} = S_{\widehat{yy}} + S_{\widehat{uu}}$
- **Definition**:
  - Das Bestimmtheitsmaß R^2 misst den **Anteil an der gesamten Variation von y, der durch unser Modell erklärt wird**.

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

$\longrightarrow$ 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 Studierenden, die 13 Mal anwesend waren und deren Eltern 2,500 Euro verdienen, kann man erwarten, dass sie auf die Klausur ca. 68 Punkte bekommen.

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.)
- **Gute Nachrichten**: Obwohl eine relevante Variable z nicht berücksichtigen wird (und daher Annahme A1 verletzt wird), ist der **Schätzer** $\widehat{\beta}_k$ der Variable x_k **unverzerrt**, wenn die ausgelassene Variable z und die Variable x_k **unkorreliert** sind.
  $\longrightarrow \widehat{\beta}_k$ kann daher als **kausaler Effekt** (als **Wirkungszusammenhang**) von der exogenen Variable x_k auf die abhängige Variable y interpretiert werden.
- Tatsächlich wird in der Ökonomie (um einen Wirkungszusammenhang aufzuzeigen) meist nicht versucht, eine vollständiges Schätzmodell zu formulieren (da die Zahl an relevanten Variablen sehr groß und zum Teil unbeobachtbar ist), sondern sicherzustellen, dass die Variable x_k, an der man hauptsächlich interessiert ist, mit den ausgelassenen Variablen unkorreliert ist.

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

Korrelationsmatrix:

	Miete	*Größe*	*Einkommen*	*Frau*	*Baujahr*
Miete	1				
Größe	0.6381	1			
Einkommen	0.3071	0.2396	1		
Frau	−0.0366	−0.0238	−0.1555	1	
Baujahr	0.1201	−0.0221	0.0259	−0.0124	1

Verzerrung aufgrund von ausgelassenen Variablen: Beispiel

Welche Konsequenz hat das auslassen einer

- **irrelevanten Variable** (Geschlecht)?
  - Annahme A1 wird dadurch nicht verletzt.
- **relevanten Variable**, die aber mit anderen erklärenden Variablen **unkorreliert** ist (Baujahr)?
  - Annahme A1 wird dadurch verletzt.
- **relevanten Variable**, die mit anderen Variablen **korreliert** ist (Größe)?
  - Annahme A1 wird dadurch verletzt.

Auslassen einer irrelevanten 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	31.434	17.242	1.820	0.068	−2.380	65.247
Größe	0.561	0.016	35.000	0.000	0.530	0.593
Einkommen	1.300	0.141	9.250	0.000	1.024	1.576
~~Frau~~						
Baujahr	22.711	2.938	7.730	0.000	16.950	28.472

Auslassen einer irrelevanten 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	31.434	17.242	1.820	0.068	−2.380	65.247
Größe	0.561	0.016	35.000	0.000	0.530	0.593
Einkommen	1.300	0.141	9.250	0.000	1.024	1.576
~~Frau~~						
Baujahr	22.711	2.938	7.730	0.000	16.950	28.472

Schlussfolgerung: Das Auslassen einer (für das Modell) irrelevanten Variable hat **keine Auswirkungen** auf den Erwartungswert der geschätzten Parameter. Die Punktschätzer unterscheiden sich vom vollständigen Modell nur geringfügig.

Auslassen einer unkorrelierten 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	110.445	15.043	7.340	0.000	80.945	139.945
Größe	0.553	0.016	34.240	0.000	0.521	0.585
Einkommen	1.400	0.146	9.580	0.000	1.113	1.687
Frau	5.237	9.480	0.550	0.581	−13.355	23.829
~~Baujahr~~						

Auslassen einer unkorrelierten 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	110.445	15.043	7.340	0.000	80.945	139.945
Größe	0.553	0.016	34.240	0.000	0.521	0.585
Einkommen	1.400	0.146	9.580	0.000	1.113	1.687
Frau	5.237	9.480	0.550	0.581	−13.355	23.829
~~Baujahr~~						

Schlussfolgerung: Das Auslassen einer relevanten Variable, die mit den anderen erklärenden Variablen *Größe* und *Einkommen* unkorreliert ist, hat **keine Auswirkungen** auf den Erwartungswert der geschätzten Parameter der Variablen *Größe* und *Einkommen*. Die Punktschätzer unterscheiden sich vom vollständigen Modell nur geringfügig, **obwohl Annahme A1 verletzt** ist.

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

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

Schlussfolgerung: Das Auslassen einer relevanten Variable, die mit den anderen Variablen korreliert ist, führt dazu, dass die **geschätzten Parameter** der anderen Variablen **verzerrt** sind. Es ist daher im restringierten Modell **nicht zulässig**, den geschätzten Parameter für *Einkommen* als **Wirkungszusammenhang** zu interpretieren.

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.

Schätzen von Wirkungszusammenhängen in der Praxis

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.

- (erfundenes) **Beispiel für ein Quasi-Experiment**: Die Regierung hat im Oktober 2010 angekündigt, beginnend mit 1. Jänner 2011 die maximale Dauer der Elternkarenz zu verlängern. 10 Jahre später soll die Auswirkung dieser Maßnahme auf die Erwerbsbeteiligung bzw. die Erwerbskarrieren der Eltern untersucht werden.
- **Vorgehensweise**: Die Entscheidung, im Winter 2010/2011 ein Kind zu bekommen, ist bei der Ankündigung der Einführung der Maßnahme (im Oktober 2010) bereits getroffen. Es ist für die Eltern nicht (oder nur sehr eingeschränkt) möglich, den Geburtszeitpunkt zu bestimmen bzw. zu verändern. Ob die Eltern Anspruch auf die längere Elternkarenz haben (weil die Geburt im Jänner stattfindet) oder nicht (weil die Geburt im Dezember stattfindet) ist daher zufällig. Die Möglichkeit, die längere Elternkarenz in Anspruch zu nehmen, ist daher für Eltern, deren Kinder zwischen dem 1. Dezember 2010 und dem 31. Jänner 2011 zur Welt kommen, nicht mit anderen Variablen korreliert, die die Erwerbsbeteiligung beeinflussen. Es ist daher möglich, den Wirkungszusammenhang der Maßnahme abzuschätzen, obwohl das Modell unvollständig ist (und daher Annahme A1 verletzt ist).