Methoden der Ökonometrie

Universität Duisburg-Essen

Christoph Hanck

Winter 2022

Überblick

Was ist Ökonometrie und wofür brauchen wir sie?

Viele ökonomische Theorien spezifizieren Beziehungen zwischen Variablen. Interessant ist dabei oft eine quantitative Aussage. [Newton wäre auch nicht berühmt geworden für die Erkenntnis, *dass* ein Apfel *runter*fällt, wenn man ihn loslässt.]

Einige Beispiele:

- Um wie viel steigern Investitionen in das Humankapital (z.B. der Besuch der Universität) das Einkommen?
- Wie stark sinkt die Inflation, wenn die Zentralbank den Zinssatz anhebt?
- Um wie viel verringert sich die Kriminalität, wenn eine zusätzliche Million Euro für Polizisten auf den Straßen ausgegeben werden? `crime.R`
- Um wie viel erhöhen sich die Verkäufe, wenn Werbeausgaben erhöht werden?
- Wie viele zusätzliche Passagiere bringt eine Reduzierung der Zugticketpreise?
- ...

Allerdings sollen solche Aussagen typischerweise **ceteris paribus** sein, d.h., alle **anderen** wichtigen **Faktoren** (z.B. das Fähigkeitsniveau) **konstant** halten.

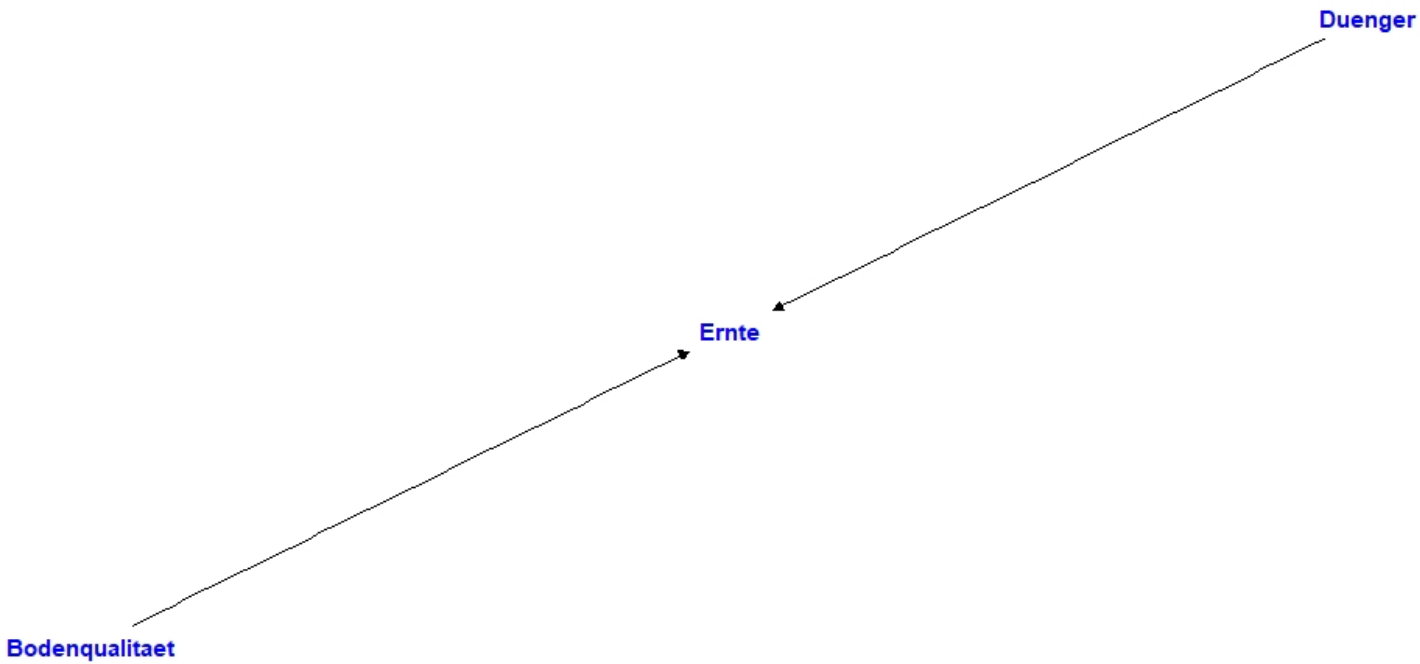
Anders formuliert handelt es sich um **kausale** Fragen, nicht nur Korrelationen. Erstere sind typischerweise interessanter. Siehe auch hier.

Ökonometrie ist die Kunst und Wissenschaft der Quantifizierung (Schätzung) solcher Beziehungen. Z.B., um *wie viel* erhöht sich das Einkommen *aufgrund* eines zusätzlichen Bildungsjahres?

Solche Schätzung ist recht einfach in **kontrollierten Experimenten**. Wir können z.B. die Einsatzmenge von Düngemitteln auf Äcker **zufällig zuteilen**, um den Effekt des Düngers auf den Ernteertrag zu messen.

In diesem Fall müssen andere wichtige Faktoren, wie Qualität des Bodens, *nicht* konstant gehalten werden, weil die Menge an Dünger den einzigen systematischen Unterschied zwischen den Äckern darstellt.

DAG Dünger



In den meisten interessanteren Szenarien (wie die o.g. Bildungsrenditen) sind solche Experimente aus unterschiedlichen Gründen nicht möglich.

Stattdessen müssen wir typischerweise mit **Beobachtungsdaten** arbeiten, die bereits bestehen werden und **nicht in einem Experiment gesammelt wurden**.

Dann müssen wir statistisch alle anderen wichtigen Faktoren konstant halten.

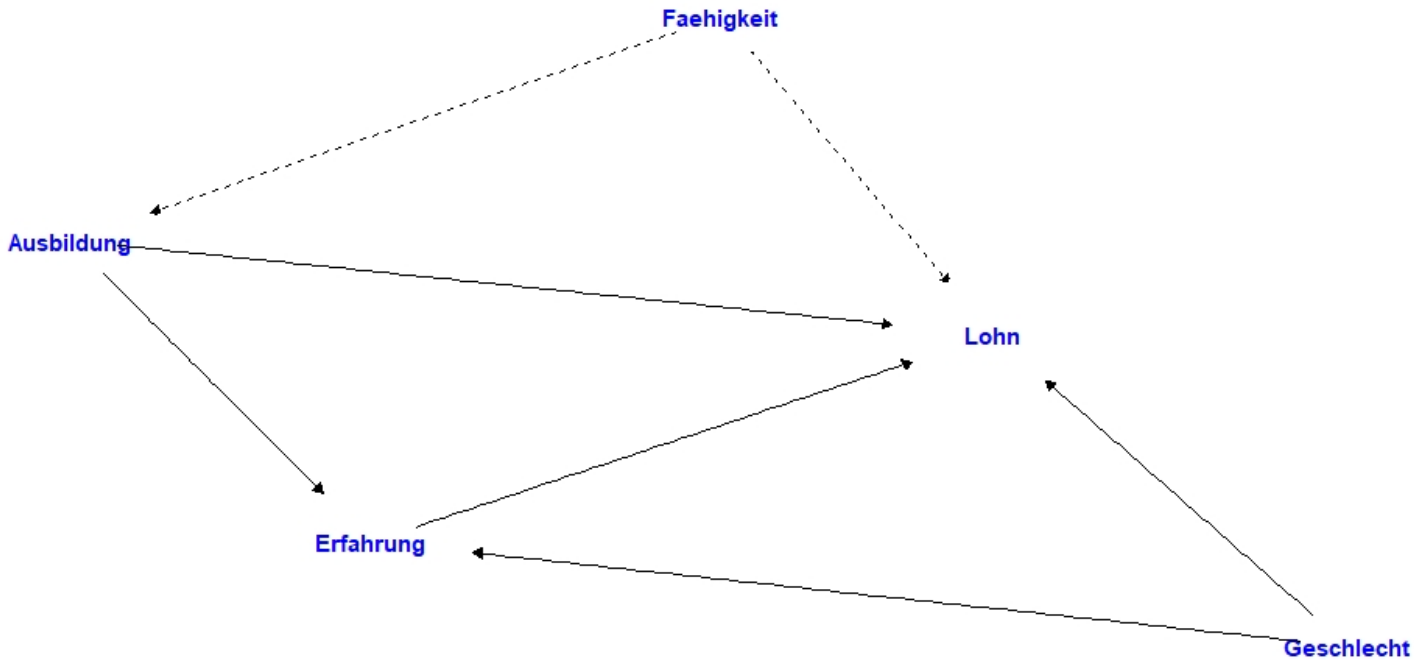
Beispielsweise werden fähigere Personen mehr verdienen, sich aber auch *entscheiden* mehr zu lernen. Höhergebildete haben zudem im Allgemeinen weniger Arbeitsmarkterfahrung, welche ebenfalls das Einkommen beeinflusst.

Die Arbeitsmarkterfahrung ist recht leicht zu beobachten, wohingegen „Fähigkeit“ schwer zu erfassen ist.

Wir werden sehen, dass das Berücksichtigen von Faktoren wie dem Ersteren leicht ist. Komplizierter wird es, wenn ein wichtiger Faktor nicht direkt beobachtet werden kann.

DAG Löhne

Die *angenommene* Problemstellung als DAG:



`DAGs.R`

Warum?

Siehe neben anderen:

http://blog.revolutionanalytics.com/2014/05/companies-using-r-in-2014.html
http://r4stats.com/articles/popularity
http://www.r-bloggers.com/r-passes-sas-in-scholarly-use-finally
https://www.r-bloggers.com/oreilly-data-scientist-salary-and-tools-survey-november-2014/

Das einfache Regressionsmodell

Der Einfachheit halber beginnen wir mit dem einfachen linearen Regressionsmodell.

Definition 1.1 (Einfaches **lineares** Regressionsmodell).

$$y_t = \beta_1 + \beta_2 X_t + u_t, \tag{1.1}$$

wobei

- y_t die **zu erklärende** oder **abhängige Variable**,
- X_t die **erklärende** oder **unabhängige Variable**,
- u_t der **Fehlerterm**, der alle unbeobachteten Einflüsse enthält,
- β_2 der **Steigungsparameter**, und β_1 der **Achsenabschnitt** ist.
- t ist der Index der Beobachtungen und läuft von 1 bis n.

y_t und X_t sind beobachtbar, der Rest nicht.

Das Modell ist linear, weil die Veränderung in y_t immer β_2 ist, wenn X_t sich um eine Einheit ändert, ungeachtet des Levels von X_t.

(1.1) ist interessant, wenn wir Annahmen bzgl. u_t treffen.

- u_t ist zufällig und unbeobachtbar—also modellieren wir unsere Unwissenheit über u_t.
- Besteht ein Zusammenhang zwischen u_t und X_t?

Zentrales Ziel dieses Kurses:

Wir wollen mehr erfahren über β_2 (und manchmal über β_1).

Wiederholung von etwas Wahrscheinlichkeitstheorie

Zufallsvariablen formalisieren Zufälligkeit.

Einfachster Fall: Skalare Zufallsvariable (ZV) X

- Menge der möglichen Werte x_i—$\mathbb{N}$, $\mathbb{R}$, $\mathbb{R}_+$,...
- Diskrete vs. stetige
- Wahrscheinlichkeiten
- Wahrscheinlichkeitsverteilung: weist x_i Wahrscheinlichkeit $p(x_i)$ zu
- Diskreter Fall:
  - ...$p(x_i)$ summiert sich zu eins, $\sum_{i=1}^{\infty} p(x_i) = 1$
  - ...Beispiel: binäre Daten, wie Zahlungsverzug/kein Zahlungsverzug

- Stetiger Fall:
  - ...jedes mögliche Ereignis x hat die Wahrscheinlichkeit null,
  - ...hat dennoch eine **Dichte** $f(x)$
  - ...sodass $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$
  - ...Beispiel: Dinge, die in Zeit oder Euro gemessen werden
  - ...können ebenfalls durch die **(kumulative) Verteilungsfunktion** $F(x)$ charakterisiert werden,
  - ...definiert als

$$F(x) = P(X \leqslant x), \tag{1.2}$$

    die Wahrscheinlichkeit, dass X gleich oder kleiner als ein Wert x ist.
  - ...hat welche Verbindung zu $f(x)$?

Regeln für Wahrscheinlichkeitsverteilungen:

- Wahrscheinlichkeiten liegen zwischen 0 und 1;
- Die leere Menge hat die Wahrscheinlichkeit 0, die Grundgesamtheit hat die Wahrscheinlichkeit 1;
- Die Vereinigung zweier disjunkter Ereignisse hat als Wahrscheinlichkeit die Summe der Wahrscheinlichkeiten dieser disjunkten Ereignisse.

Daraus folgen folgende Eigenschaften...

- $F(x)$ ist monoton steigend
- $F(x)$ konvergiert gegen 0 wenn $x \to -\infty$ und $F(x) \to 1$ wenn $x \to \infty$
- $P(a \leqslant X \leqslant b) = F(b) - F(a)$
- Was erhält man für $a = b$?
- Dies motiviert die Dichte $f(x) = F'(x)$.
- $P(a \leqslant X \leqslant b) = F(b) - F(a) = \int_a^b f(x)\,\mathrm{d}x$
- Diskrete ZV haben ebenfalls Verteilungsfunktionen.

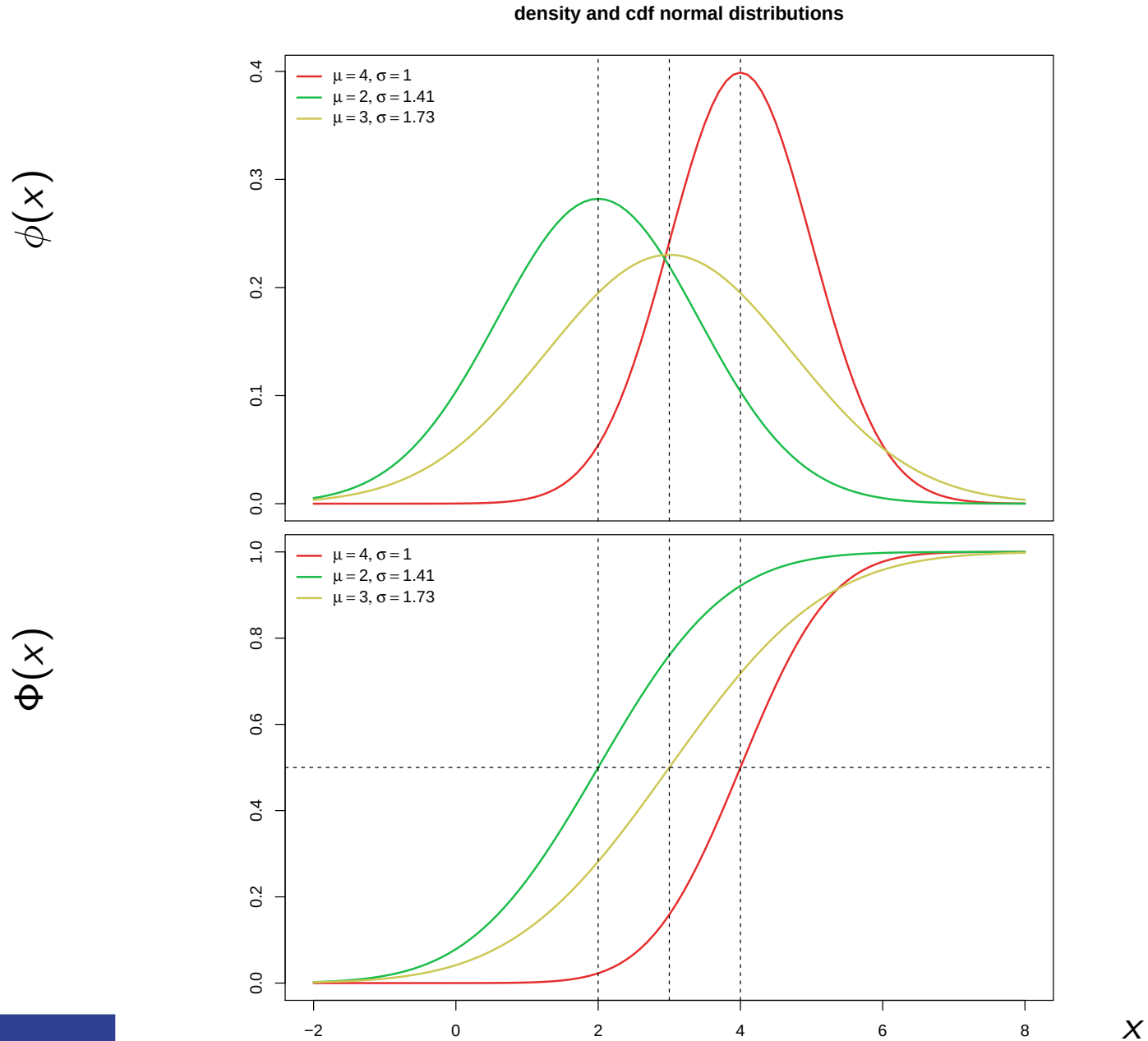
Beispiel 1.2 (Normalverteilung).

$$f(x) = \phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big\{-\frac{1}{2\sigma^2}(x-\mu)^2\Big\} \qquad \text{für } -\infty < x < \infty \tag{1.3}$$

Die Verteilungsfunktion (1.2) der Normalverteilung, Notation $\Phi(x)$, hat keine geschlossene Form. Alles, was wir sagen können, ist

$$\Phi(x) = \int_{-\infty}^{x} \phi(y)\,\mathrm{d}y \tag{1.4}$$

Einige Beispiele (cf. `pdfcdf.R`):



Momente

- Am bekanntesten: Der Erwartungswert
- Für eine diskrete ZV, die m mögliche Werte $x_1, x_2, \ldots, x_m$ annehmen kann, ist der Erwartungswert

$$E(X) = \sum_{i=1}^{m} p(x_i) x_i,$$

  die Summe jedes möglichen Wertes x_i multipliziert mit seiner Wahrscheinlichkeit.
- Für eine stetige ZV ist der Erwartungswert analog definiert als:

$$E(X) = \int_{-\infty}^{\infty} x f(x) \, \mathrm{d}x.$$

  Dennoch hat nicht jede ZV einen Erwartungswert!
- $E(X)$ (oft bezeichnet mit μ) bezieht sich auf die Grundgesamtheit und ist nicht zu verwechseln mit dem Stichprobenmittelwert.

- μ wird auch das erste Moment genannt...
- ...weil auch höhere Momente existieren (können). Im stetigen Fall:

$$m_k(X) = E(X^k) = \int_{-\infty}^{\infty} x^k f(x)\,\mathrm{d}x$$

- Solche Momente werden unzentrierte Momente genannt...
- ...da auch **zentrale Momente** existieren (können):

$$\mu_k(X) = E\big(X - E(X)\big)^k = \int_{-\infty}^{\infty} (x-\mu)^k f(x)\,\mathrm{d}x$$

- Was ist μ_2?
- ...die Quadratwurzel heißt **Standardabweichung**
- ...und wird zur Konstruktion von **Standardfehlern** wichtig sein

Multivariate Verteilungen

- **Mehrere** ZV (X_1, X_2) haben **gemeinsame** Verteilungen.
- Formal

$$F(x_1, x_2) = P\big((X_1 \leqslant x_1) \cap (X_2 \leqslant x_2)\big)$$

  oder einfach

$$F(x_1, x_2) = P(X_1 \leqslant x_1, X_2 \leqslant x_2)$$

- Im stetigen Fall ergibt sich eine **gemeinsame Dichte**

$$f(x_1, x_2) = \frac{\partial F(x_1, x_2)}{\partial x_1 \partial x_2} \tag{1.5}$$

  S. `BivariateNormalPlotly.R`

- Dann gilt

$$F(x_1, x_2) = \int_{-\infty}^{x_2} \int_{-\infty}^{x_1} f(y_1, y_2)\, \mathrm{d}y_1\, \mathrm{d}y_2$$

Unabhängigkeit

Wir bezeichnen zwei ZV als **statistisch unabhängig**, wenn wir die gemeinsame Verteilungsfunktion schreiben können als:

$$F(x_1, x_2) = F(x_1, \infty)F(\infty, x_2) \equiv F_1(x_1)F_2(x_2)$$

Kenntnis von X_1 liefert dann keine Information über X_2.

Unabhängigkeit definiert bezüglich der **Randdichten**

$$f_i(x_i) = \int_{-\infty}^{\infty} f(x_i, x_j)\,\mathrm{d}x_j, \qquad i \neq j \tag{1.6}$$

folgt analog als

$$f(x_1, x_2) = f_1(x_1)f_2(x_2)$$

Bedingte Wahrscheinlichkeiten

Für zwei beliebige Ereignisse A und B gilt,

$$P(A \cap B) = P(B)P(A|B) \qquad \text{oder} \qquad P(A|B) = \frac{P(A \cap B)}{P(B)},$$

gegeben $P(B) \neq 0$. Z.B.

- $A \cap B = \emptyset$ impliziert, dass $P(A|B) = 0$, genauso wie
- $B \subset A$ impliziert, dass $P(A \cap B) = P(B)$ und $P(A|B) = 1$.

Ein Venn-Diagramm hilft diese Notationen zu verstehen.

Beispiel: Was könnte die Wahrscheinlichkeit sein, dass eine Person über 2.000 Euro verdient, und wie hoch ist sie, wenn die Person älter als 30 ist?

Wir können ebenfalls mit **bedingten Dichten** arbeiten, welche auch funktionieren, wenn ein Ereignis $(X_2 = x_2)$ die Wahrscheinlichkeit 0 hat (sofern $f_2(x_2) > 0$):

$$f(x_1|x_2) = \frac{f(x_1, x_2)}{f_2(x_2)}$$

Dies hilft uns den **bedingten Erwartungswert** von X_1 gegeben X_2 zu definieren

$$E(X_1|X_2) = \int_{-\infty}^{\infty} x_1 f(x_1|x_2)\,\mathrm{d}x_1$$

Beachten Sie: Dies ist eine Funktion der ZV X_2.

Daher ist der bedingte Erwartungswert selbst eine ZV und wir können seinen Erwartungswert betrachten.

$$
\begin{aligned}
\underbrace{\int_{-\infty}^{\infty} E(X_1|X_2) f_2(x_2)\,\mathrm{d}x_2}_{=E[E(X_1|X_2)]} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 f(x_1|x_2)\,\mathrm{d}x_1 f_2(x_2)\,\mathrm{d}x_2 \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 f(x_1|x_2) f_2(x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 f(x_1, x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 \\
&= \int_{-\infty}^{\infty} x_1 \int_{-\infty}^{\infty} f(x_1, x_2)\,\mathrm{d}x_2\,\mathrm{d}x_1 \\
&= \int_{-\infty}^{\infty} x_1 f_1(x_1)\,\mathrm{d}x_1 \\
&= E(X_1) \qquad (1.7)
\end{aligned}
$$

Dieses Resultat ist bekannt als das **Gesetz der iterierten Erwartungen**. Es ist sowohl intuitiv als auch wichtig.

Eine wichtige Regel ist, dass für jede Funktion h gilt,

$$E(X_1 h(X_2)|X_2) = h(X_2)E(X_1|X_2). \tag{1.8}$$

Auch das macht intuitiv Sinn. Ein wichtiger Spezialfall ergibt sich, wenn $E(X_1|X_2) = 0$, was impliziert, dass

$$\begin{aligned} E(X_1 h(X_2)) &= E\big(E(X_1 h(X_2)|X_2)\big) \\ &= E\big(h(X_2)E(X_1|X_2)\big) \\ &= 0, \end{aligned} \tag{1.9}$$

bekannt als „(Erwartungswert-)Unabhängigkeit impliziert Unkorreliertheit.“

Zurück zur Ökonometrie, insbesondere (1.1). Nehme an, dass

$$E(u_t|X_t) = 0. \tag{1.10}$$

Dann ist, mit (1.1) und Nutzung von (1.8) und (1.10)

$$E(y_t|X_t) = \beta_1 + \beta_2 X_t + E(u_t|X_t) = \beta_1 + \beta_2 X_t$$

$E(u_t|X_t) = 0$ ist also letzten Endes die Annahme, dass unser Modell **korrekt spezifiziert** ist.

Betrachten wir

Beispiel 1.3 (Lineare Bildungsrenditen).

$$Lohn_t = \beta_1 + \beta_2 Ausbildung_t + u_t$$

Im Beispiel 1.3 enthält u_t angeborene Fähigkeiten. Für $X_t = Ausbildung_t$,

$$E(u_t|X_t) = 0$$

müssen die angeborenen Fähigkeiten immer gleich sein, unabhängig von den Bildungsjahren, z.B.

$$E(Fähigkeit_t|X_t = 8) = E(Fähigkeit_t|X_t = 16)$$

Wenn talentiertere Personen durchschnittlich mehr Bildungsjahre haben, ist die Annahme $E(u_t|X_t) = 0$ verletzt.

Wenn wir bei der Schätzung des Modells fälschlicherweise annehmen, dass $E(u_t|X_t) = 0$ ist, werden wir unser Ziel mehr über β_2 zu lernen nicht erreichen.

Für gewöhnlich wollen wir y_t **auf allgemeinere Informationssets** Ω_t bedingen, anstatt nur auf ein X_t. Z.B. $\Omega_t = \{X_t, X_t^2\}$ oder $\Omega_t = \{X_{t1}, X_{t2}\}$.

Wir wollen **exogene** Variablen in Ω_t einbinden, die außerhalb der Gleichung bestimmt werden. **Endogene** Variablen sind ebenfalls durch y_t bestimmt und dadurch auch durch u_t.

Neben der Spezifikation der Beziehung zu X_t müssen wir auch die Beziehung zwischen den u_t betrachten. Häufig treffen wir die Annahme, dass $\{y_t, X_t\}$ **unabhängig und identisch verteilt (u.i.v.)** sind.

Diese Annahme wäre bspw. verletzt, wenn **serielle Korrelation** vorläge, wenn z.B. auf große (kleine) Werte von u_t tendenziell große (kleine) Werte folgen. Mehr dazu in „Zeitreihenanalyse“.

Es liegt **Heteroskedastie** vor, wenn die bedingte Varianz verschiedener y_t unterschiedlich ist. `EVP.Rproj`

Lineare und nichtlineare Regressionsmodelle

Das lineare Modell erscheint sehr restriktiv, aber betrachten wir

$$
\begin{aligned}
y_t &= \beta_1 + \beta_2 X_t + \beta_3 X_t^2 + u_t \qquad (1.11)\\
y_t &= \gamma_1 + \gamma_2 \log X_t + u_t \\
y_t &= \delta_1 + \delta_2 \frac{1}{X_t} + u_t
\end{aligned}
$$

Alle diese Modelle sind linear **in den Parametern**. Ersteres ist ein Beispiel für ein **multiples** Regressionsmodell.

Das Folgende ist nicht linear

$$y_t = e^{\beta_1} X_{t2}^{\beta_2} X_{t3}^{\beta_3} + u_t$$

Aber selbst

$$y_t = e^{\beta_1} X_{t2}^{\beta_2} X_{t3}^{\beta_3} e^{v_t}$$

ist linear (oder präziser **log-linear**) für unserer Zwecke: Logarithmieren liefert

$$\log y_t = \beta_1 + \beta_2 \log X_{t2} + \beta_3 \log X_{t3} + v_t$$

Wir werden uns bald mit multiplen Regressionsmodellen beschäftigen. Hierfür benötigen wir Matrixalgebra. Siehe hierzu den Matrixalgebra-Reader.

Das lineare Modell in Matrixnotation

$$\underset{(n\times 1)}{\boldsymbol{y}} = \underset{(n\times 2)}{\boldsymbol{X}}\underset{(2\times 1)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\boldsymbol{u}} \tag{1.12}$$

wobei

$$\boldsymbol{X}_t = \begin{bmatrix} 1 & X_t \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix},$$

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \qquad \boldsymbol{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix},$$

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}.$$

Die Spalten von $\boldsymbol{X}$ werden **Regressoren** genannt. Beachten Sie, dass Matrixalgebra nur eine kompaktere Notation darstellt — eine typische Zeile von (1.12) ist das Gleiche wie (1.1).

Das Tolle an Matrixalgebra ist, dass wir uns keine Gedanken machen müssen, wenn wir mehr als eine X_t-Variable haben, wie in (1.11). Wir schreiben einfach

$$\underset{(n\times 1)}{\boldsymbol{y}} = \underset{(n\times k)}{\boldsymbol{X}}\underset{(k\times 1)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\boldsymbol{u}} \tag{1.13}$$

wobei

$$\boldsymbol{X}_t = \begin{bmatrix} X_{t1} & \cdots & X_{tk} \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix},$$

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \qquad \boldsymbol{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix},$$

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_n \end{bmatrix} = \begin{bmatrix} X_{11} & \cdots & X_{1k} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{nk} \end{bmatrix}.$$

Normalerweise, aber nicht zwingend, ist $X_{t1} = 1$.

Eine typische Zeile ist dann

$$y_t = \boldsymbol{X}_t\beta + u_t = \beta_1 + \sum_{i=2}^{k} \beta_i X_{ti} + u_t$$

Schätzung

Bisher haben wir uns nur mit der Frage des Modellaufbaus beschäftigt. In der Praxis kennen wir β nicht und müssen es aus unseren Daten $\boldsymbol{y}$, $\boldsymbol{X}$ **schätzen**.

Stellen wir uns vor, dass die Daten aus einer **Population** gezogen, oder durch einen **Datengenerierenden Prozess (DGP)** erzeugt wurden, wie in (1.13).

Wir stellen nun selbst ein Modell auf und versuchen unser Modell so gut wie möglich an den DGP anzugleichen. Danach schätzen wir die Parameter des Modells.

Eine einfache aber sehr nützliche Schätzmethode ist die **Momentenmethode**. Die Idee ist, die Momentenbedingungen der Population ($\mu = g(\theta)$) durch die Momentenbedingungen der Stichprobe ($\bar{X} = g(\theta)$) zu ersetzen und dann nach den Parametern $\hat{\theta}$ aufzulösen, die diese Gleichungen lösen.

Beispiel 1.4 (Einfache lineare Regression). (1.1) impliziert, dass

$$u_t = y_t - \beta_1 - \beta_2 X_t$$

Die Annahme (1.10) ($E(u_t|X_t) = 0$) liefert mit dem Gesetz des iterierten Erwartungswertes (1.7), dass

$$E(u_t) = E\big(E(u_t|X_t)\big) = 0$$

Daraus erhalten wir die **Momentenbedingung**

$$E(u_t) = E(y_t - \beta_1 - \beta_2 X_t) = 0$$

Ersetzen wir den Erwartungswert durch den Stichprobenmittelwert, erhalten wir

$$\frac{1}{n}\sum_{t=1}^{n} u_t = \frac{1}{n}\sum_{t=1}^{n}(y_t - \tilde{\beta}_1 - \tilde{\beta}_2 X_t) = 0 \tag{1.14}$$

Wir nehmen für den Moment der Einfachheit halber an, dass $\beta_2 = 0$. Dann

$$\frac{1}{n}\sum_{t=1}^{n}(y_t - \tilde{\beta}_1) = 0$$

Der Wert, der diese Gleichung löst, folgt aus

$$\frac{1}{n}\sum_{t=1}^{n} y_t = \frac{n}{n}\tilde{\beta}_1$$

so dass

$$\hat{\beta}_1 = \frac{1}{n}\sum_{t=1}^{n} y_t.$$

Was, wenn $\beta_2 \neq 0$? Dann haben wir zwei Unbekannte und benötigen daher eine weitere Momentenbedingung.

Wegen (1.9) impliziert $E(u_t|X_t) = 0$ auch, dass

$$E(X_t u_t) = 0,$$

da

$$E(X_t u_t) \overset{(1.7)}{=} E[E(X_t u_t|X_t)] \overset{(1.8)}{=} E[X_t E(u_t|X_t)] \overset{(1.10)}{=} E[X_t 0] = E[0] = 0$$

Daher ist

$$E\big(X_t(y_t - \beta_1 - \beta_2 X_t)\big) = 0$$

Ersetzen wir die Populationsmomente durch die Stichprobenmomente, erhalten wir

$$\frac{1}{n}\sum_{t=1}^{n}\big(X_t(y_t - \tilde{\beta}_1 - \tilde{\beta}_2 X_t)\big) = 0. \tag{1.15}$$

Durch Umschreiben von (1.14) und (1.15) ergibt sich (nachrechnen!):

$$\begin{aligned} \tilde{\beta}_1 + \left(\frac{1}{n}\sum_{t=1}^{n} x_t\right)\tilde{\beta}_2 &= \frac{1}{n}\sum_{t=1}^{n} y_t \\ \left(\frac{1}{n}\sum_{t=1}^{n} x_t\right)\tilde{\beta}_1 + \left(\frac{1}{n}\sum_{t=1}^{n} x_t^2\right)\tilde{\beta}_2 &= \frac{1}{n}\sum_{t=1}^{n} x_t y_t \end{aligned}$$

Multiplizieren mit n liefert

$$\begin{aligned} n\tilde{\beta}_1 + \left(\sum_{t=1}^{n} x_t\right)\tilde{\beta}_2 &= \sum_{t=1}^{n} y_t \\ \left(\sum_{t=1}^{n} x_t\right)\tilde{\beta}_1 + \left(\sum_{t=1}^{n} x_t^2\right)\tilde{\beta}_2 &= \sum_{t=1}^{n} x_t y_t \end{aligned}$$

In Matrixschreibweise,

$$\begin{bmatrix} n & \sum_{t=1}^{n} X_t \\ \sum_{t=1}^{n} X_t & \sum_{t=1}^{n} X_t^2 \end{bmatrix} \begin{bmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_2 \end{bmatrix} = \begin{bmatrix} \sum_{t=1}^{n} y_t \\ \sum_{t=1}^{n} X_t y_t \end{bmatrix}$$

oder noch kompakter

$$\boldsymbol{X}^\top \boldsymbol{X} \tilde{\beta} = \boldsymbol{X}^\top \boldsymbol{y} \tag{1.16}$$

Zum Auflösen von (1.16) nehmen wir die Inverse, um den **kleinsten Quadrate-Schätzer** zu erhalten.

$$\hat{\beta}_{\text{OLS}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y} \tag{1.17}$$

Formal ändert sich nicht viel in der multiplen Regression (1.13). Die Annahme, dass $E(u_t|X_t) = 0$ (siehe (1.10)) wird modifiziert zu

$$E(u_t|\boldsymbol{X}_t) = E(u_t|X_{t1}, \ldots, X_{tk}) = 0. \tag{1.18}$$

Das führt wieder über (1.9) zu $E(X_{ti}u_t) = 0$, $i = 1, \ldots, k$, und damit zu den k Momentenbedingungen

$$\frac{1}{n}\sum_{t=1}^{n}\left(X_{ti}(y_t - \boldsymbol{X}_t\tilde{\beta})\right) = 0 \qquad i = 1, \ldots, k \tag{1.19}$$

Zentral ist hierbei, dass (1.18) in Anwendungen wesentlich eher erfüllt sein wird als (1.10)!

(1.19) lautet in Matrixform (nachdem mit n multipliziert wurde)

$$\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\tilde{\beta}) = \boldsymbol{0},$$

wobei $\boldsymbol{0}$ ein Vektor mit k Nullen ist. Nun kann dieser Ausdruck umgeschrieben werden als (1.16), $\boldsymbol{X}^\top\boldsymbol{X}\tilde{\beta} = \boldsymbol{X}^\top\boldsymbol{y}$. Daher ist die Lösung ebenfalls (1.17).

Warum nennen wir

$$\hat{\beta}_{\text{OLS}} = \left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y}$$

dann eigentlich „OLS“-Schätzer?

Die Methode der kleinsten Quadrate

Für eine Stichprobe $(\boldsymbol{y}, \boldsymbol{X})$ sei $\tilde{\beta}$ ein hypothetischer Schätzer von β. $\tilde{\beta}$ definiert dann die **Residuen**

$$y_t - \boldsymbol{X}_t\tilde{\beta} \qquad (t = 1, \ldots, n).$$

Wir definieren die **Summe der quadrierten Residuen** als:

$$\begin{aligned} \mathrm{SSR}(\tilde{\beta}) &:= \sum_{t=1}^{n}(y_t - \boldsymbol{X}_t\tilde{\beta})^2 \\ &= (\boldsymbol{y} - \boldsymbol{X}\tilde{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\tilde{\beta}) \\ &\overset{\text{M20.4}}{=} \boldsymbol{y}^\top\boldsymbol{y} - 2\tilde{\beta}^\top\boldsymbol{X}^\top\boldsymbol{y} + \tilde{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}\tilde{\beta}. \end{aligned} \tag{1.20}$$

Der **Kleinste-Quadrate (OLS) Schätzer** von β, $\boldsymbol{b}$, ist definiert als

$$\boldsymbol{b} := \arg\min_{\tilde{\beta}} \ \mathrm{SSR}(\tilde{\beta}). \tag{1.21}$$

Die Matrixalgebra-Regeln 31, 1. und 2. sagen uns, dass

$$\frac{\partial(\tilde{\boldsymbol{\beta}}^\top \boldsymbol{X}^\top \boldsymbol{y})}{\partial \tilde{\boldsymbol{\beta}}} = \boldsymbol{X}^\top \boldsymbol{y}, \qquad \frac{\partial(\tilde{\boldsymbol{\beta}}^\top \boldsymbol{X}^\top \boldsymbol{X} \tilde{\boldsymbol{\beta}})}{\partial \tilde{\boldsymbol{\beta}}} = 2\boldsymbol{X}^\top \boldsymbol{X} \tilde{\boldsymbol{\beta}}.$$

Um SSR($\tilde{\boldsymbol{\beta}}$) zu minimieren, muss $\tilde{\boldsymbol{\beta}}$ die Bedingungen erster Ordnung erfüllen:

$$\frac{\partial \text{SSR}(\tilde{\boldsymbol{\beta}})}{\partial \tilde{\boldsymbol{\beta}}} = -2\boldsymbol{X}^\top \boldsymbol{y} + 2\boldsymbol{X}^\top \boldsymbol{X} \tilde{\boldsymbol{\beta}} \stackrel{!}{=} \boldsymbol{0}$$

oder

$$\boldsymbol{X}^\top \boldsymbol{X} \tilde{\boldsymbol{\beta}} = \boldsymbol{X}^\top \boldsymbol{y}.$$

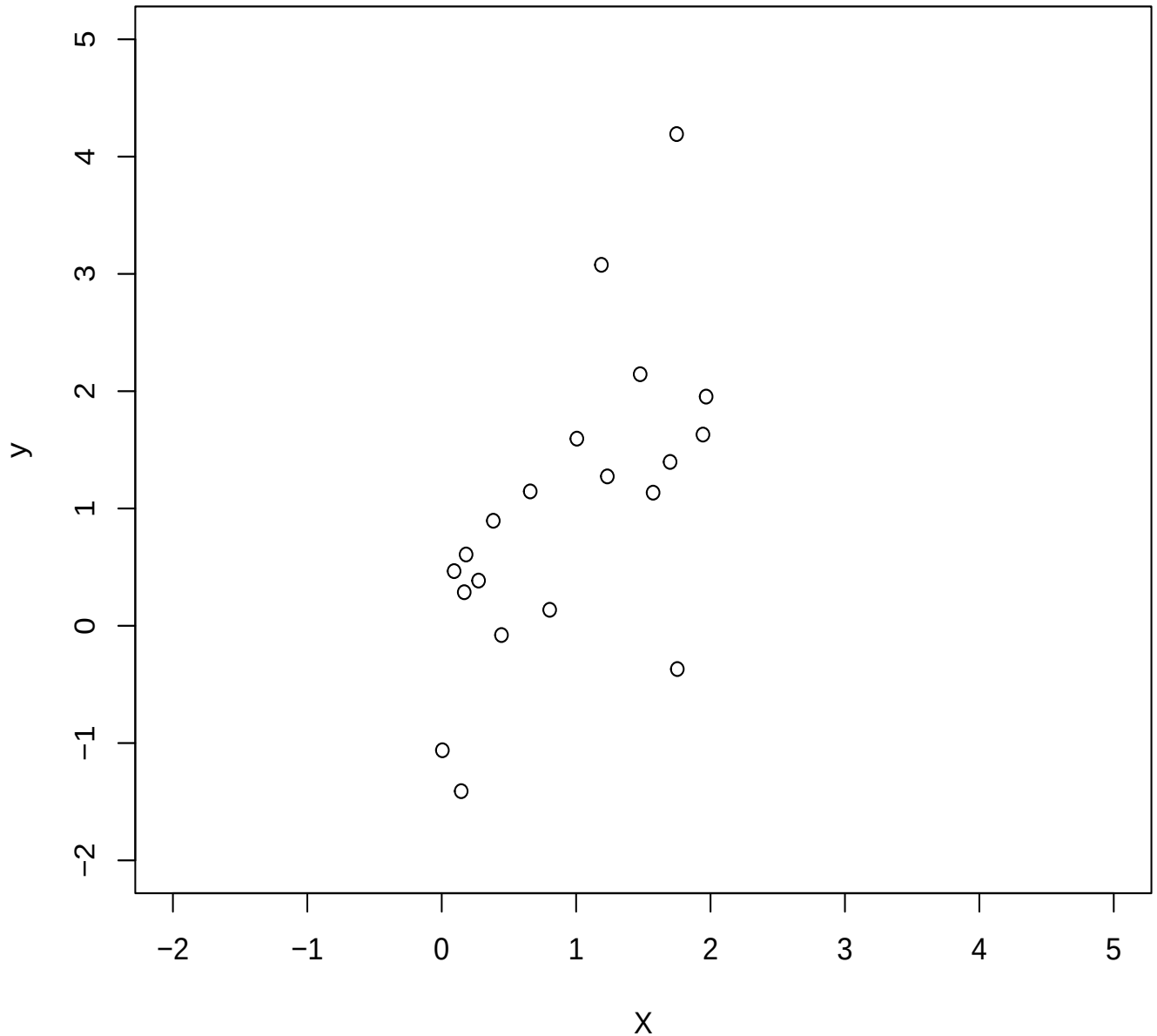
M6 liefert dann

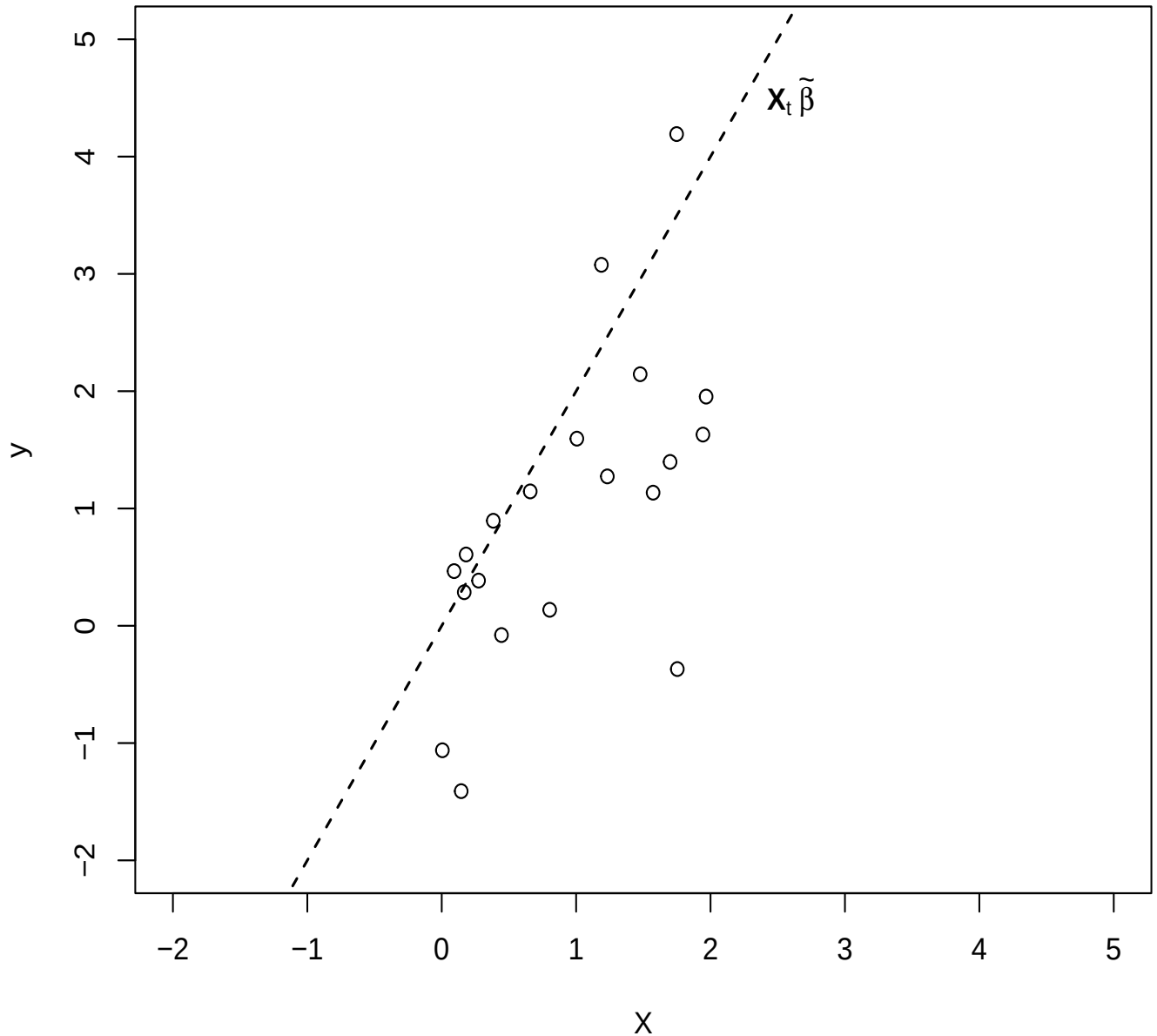
$$\boldsymbol{b} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

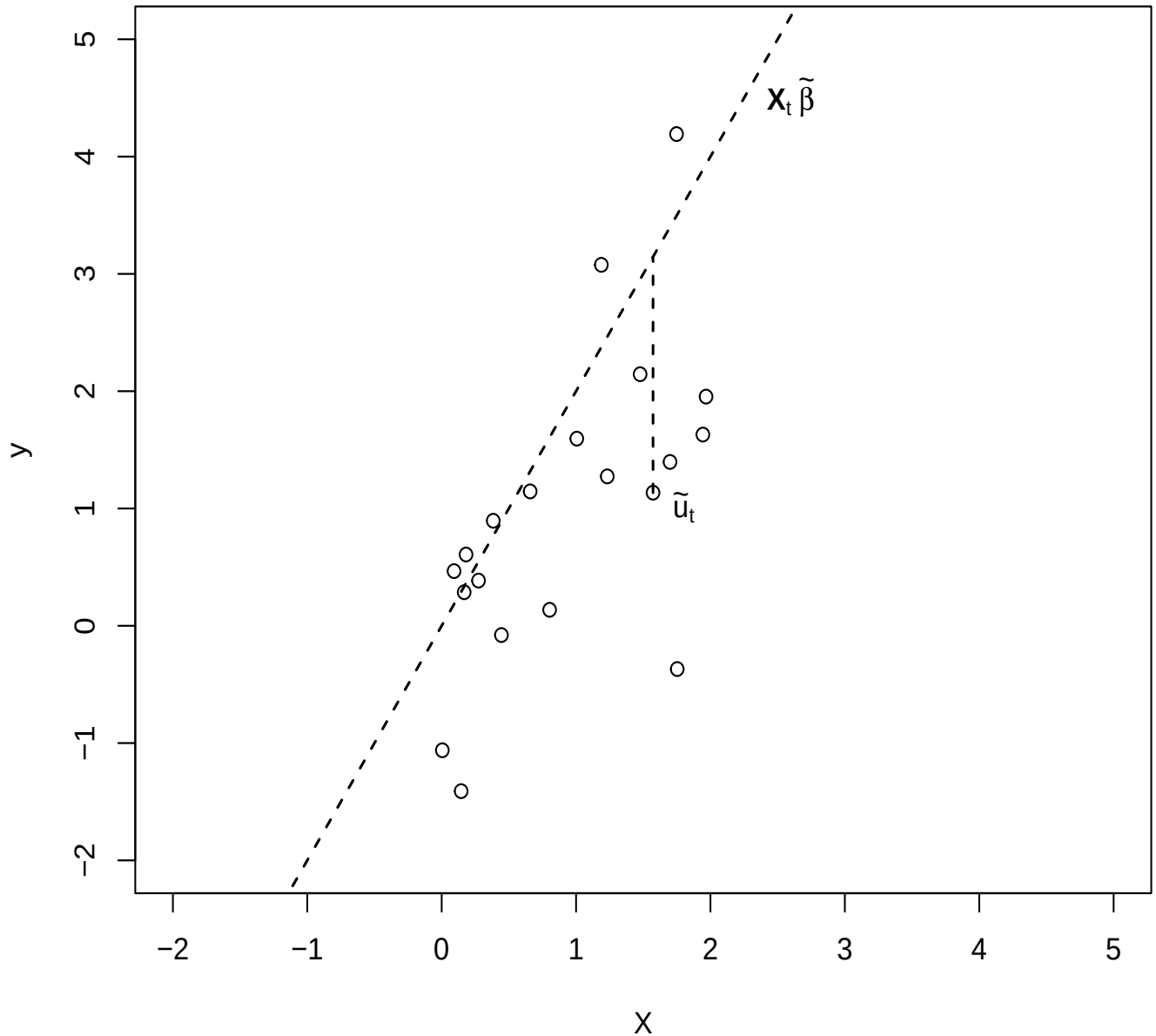
Das bedeutet, dass $\tilde{\boldsymbol{\beta}}$ muss (1.16) erfüllen muss—in anderen Worten

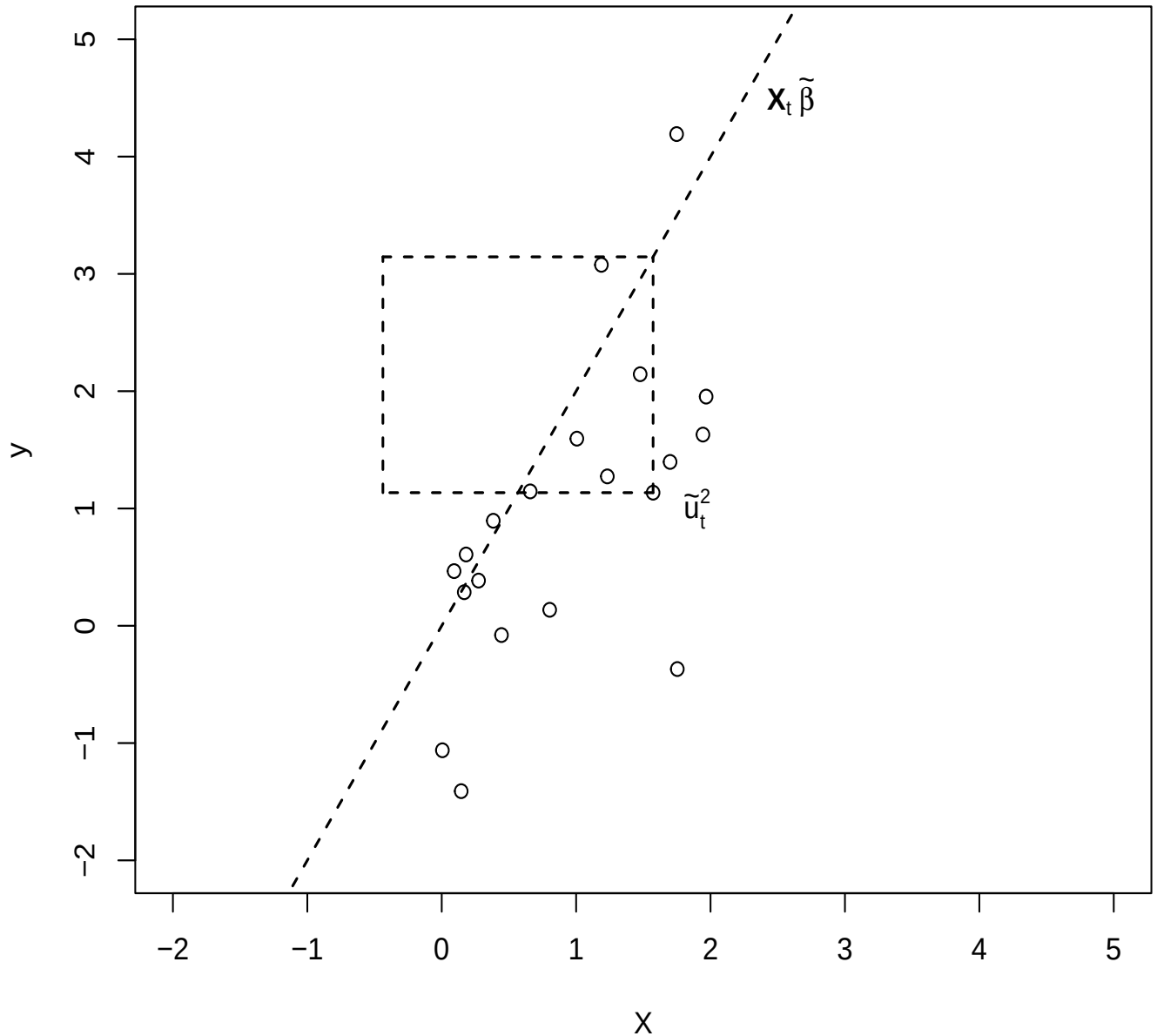
$$\boldsymbol{b} = \hat{\beta}_{\text{OLS}}!$$

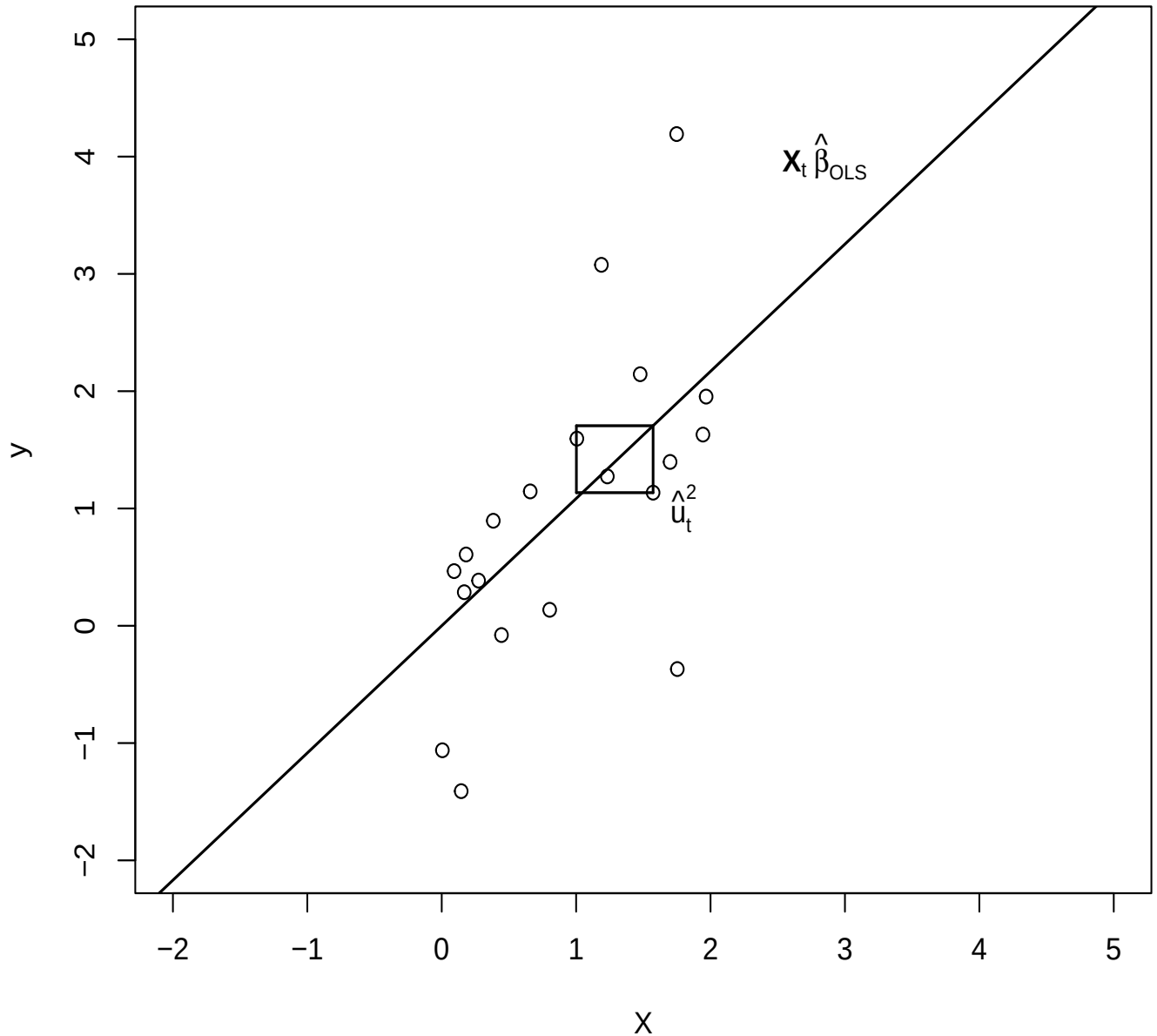
`BechdelTest.R`

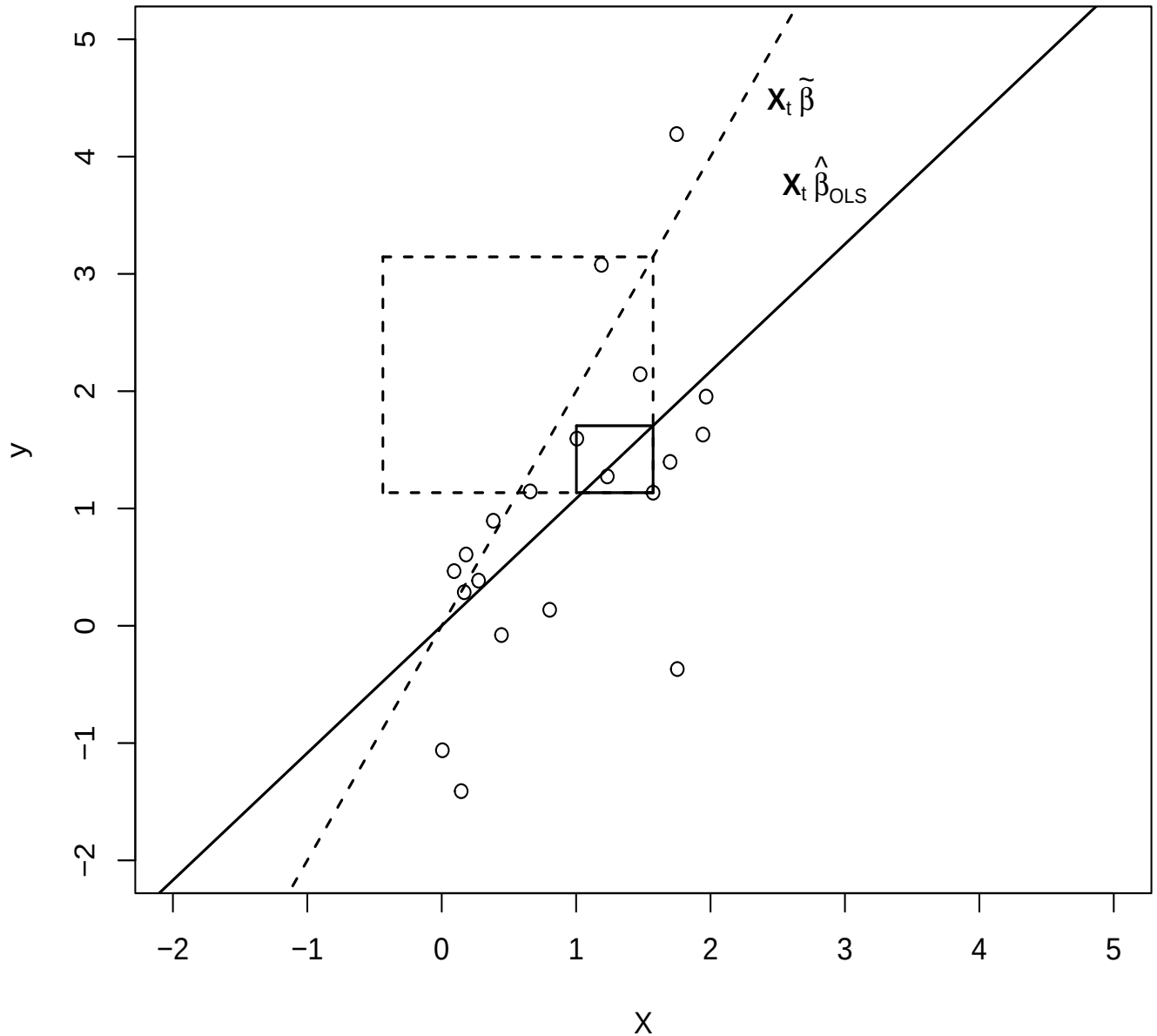

y
X
−2
−1
0
1
2
3
4
5


$\mathbf{X}_t\,\tilde{\beta}$
y
X
−2
−1
0
1
2
3
4
5


$\mathbf{X}_t\,\tilde{\beta}$
$\tilde{u}_t$
y
X
−2
−1
0
1
2
3
4
5


$\mathbf{X}_t\,\tilde{\beta}$
$\tilde{u}_t^2$
y
X
−2
−1
0
1
2
3
4
5


y
X
−2
−1
0
1
2
3
4
5
$\mathbf{x}_t \hat{\beta}_{OLS}$
$\hat{u}_t^2$


$\mathbf{X}_t\,\tilde{\beta}$
$\mathbf{X}_t\,\hat{\beta}_{OLS}$
y
X
−2
−1
0
1
2
3
4
5

Überblick

Bisher haben wir angenommen, dass die Inverse $\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}$ in (1.17) existiert. Dies ist jedoch nicht zwangsläufig der Fall. Die Inverse $\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}$ existiert, wenn $\boldsymbol{X}^\top \boldsymbol{X}$ den Rang k hat (Matrixalgebra-Reader Resultat 21.6).

Dies ist der Fall, wenn $\boldsymbol{X}$ Rang k hat (siehe Resultat 20.8—ausgenommen den Fall, dass $n < k$). $\boldsymbol{X}$ hat vollen Spaltenrang (= Rang k), wenn die Spalten $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ linear unabhängig sind (siehe Definition 12). Dies ist wiederum der Fall, wenn der einzige Weg

$$\sum_{i=1}^{k} c_i \boldsymbol{x}_i = \boldsymbol{0}$$

zu erhalten ist, für alle i $c_i = 0$ zu setzen. Es darf demnach nicht möglich sein einen Regressor als eine lineare Kombination eines anderen zu schreiben, wie in

$$\boldsymbol{x}_j = \sum_{i \neq j} \frac{c_i}{c_j} \boldsymbol{x}_i.$$

Falls dies doch möglich sein sollte, liegt **Multikollinearität** vor.

Beispiel 2.1 Ein volles **Dummy-Variablen** Modell mit Konstante liefert Multikollinearität.

$$\log(wage)_t = \beta_1 + \beta_2 \cdot man_t + \beta_3 \cdot woman_t + \boldsymbol{X}_t\boldsymbol{\beta}_4 + u_t$$

Wir haben $man_t + woman_t = 1$ für alle t (Bemerkung!). Eine Lösung ist es die Konstante oder einen der Dummies wegzulassen, z.B.

$$\log(wage)_t = \beta_1 \cdot man_t + \beta_2 \cdot woman_t + \boldsymbol{X}_t\boldsymbol{\beta}_3 + u_t \tag{2.1}$$

Die **Interpretation der Koeffizienten** ändert sich. Zum Beispiel ist für (2.1)

$$E(\log(wage)_t | woman_t = 1) = \beta_2 + \boldsymbol{X}_t\boldsymbol{\beta}_3$$

während $E(\log(wage)_t | woman_t = 1) = \beta_1 + \beta_2 + \boldsymbol{X}_t\boldsymbol{\beta}_3$ in

$$\log(wage)_t = \beta_1 + \beta_2 \cdot woman_t + \boldsymbol{X}_t\boldsymbol{\beta}_3 + u_t$$

`BechdelTest.R`

Definition 2.2 Die mit OLS **gefitteten Werte** sind

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{OLS}}.$$

Definition 2.3 Die **Projektionsmatrix** und die **Residuenmachermatrix** sind

$$\underset{(n\times n)}{\boldsymbol{P}_{\boldsymbol{X}}} := \boldsymbol{P} := \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top \tag{2.2}$$

$$\underset{(n\times n)}{\boldsymbol{M}_{\boldsymbol{X}}} := \boldsymbol{M} := \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top. \tag{2.3}$$

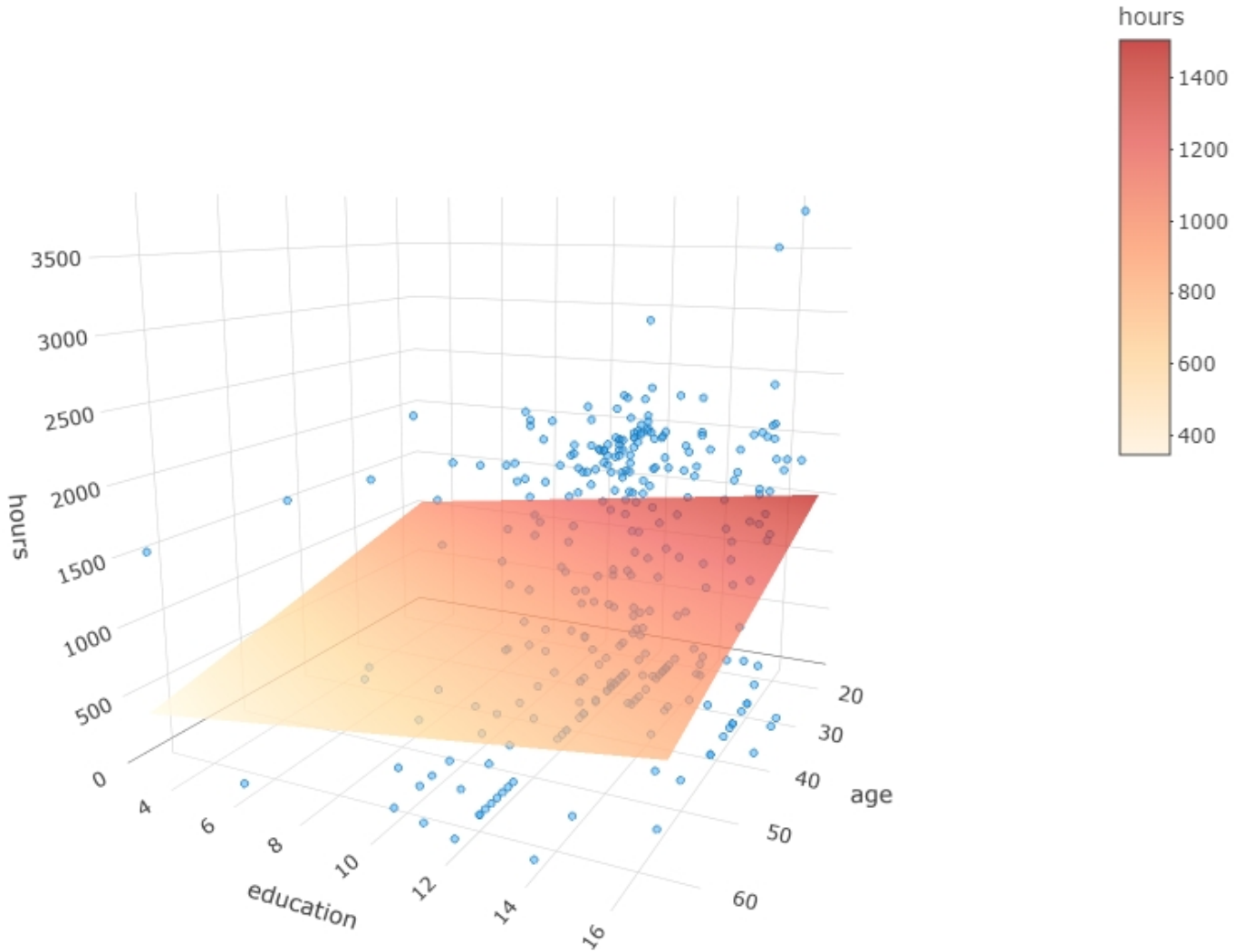
Diese Bezeichnungen kommen zustande, da

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{OLS}} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} = \boldsymbol{P}\boldsymbol{y}$$

und

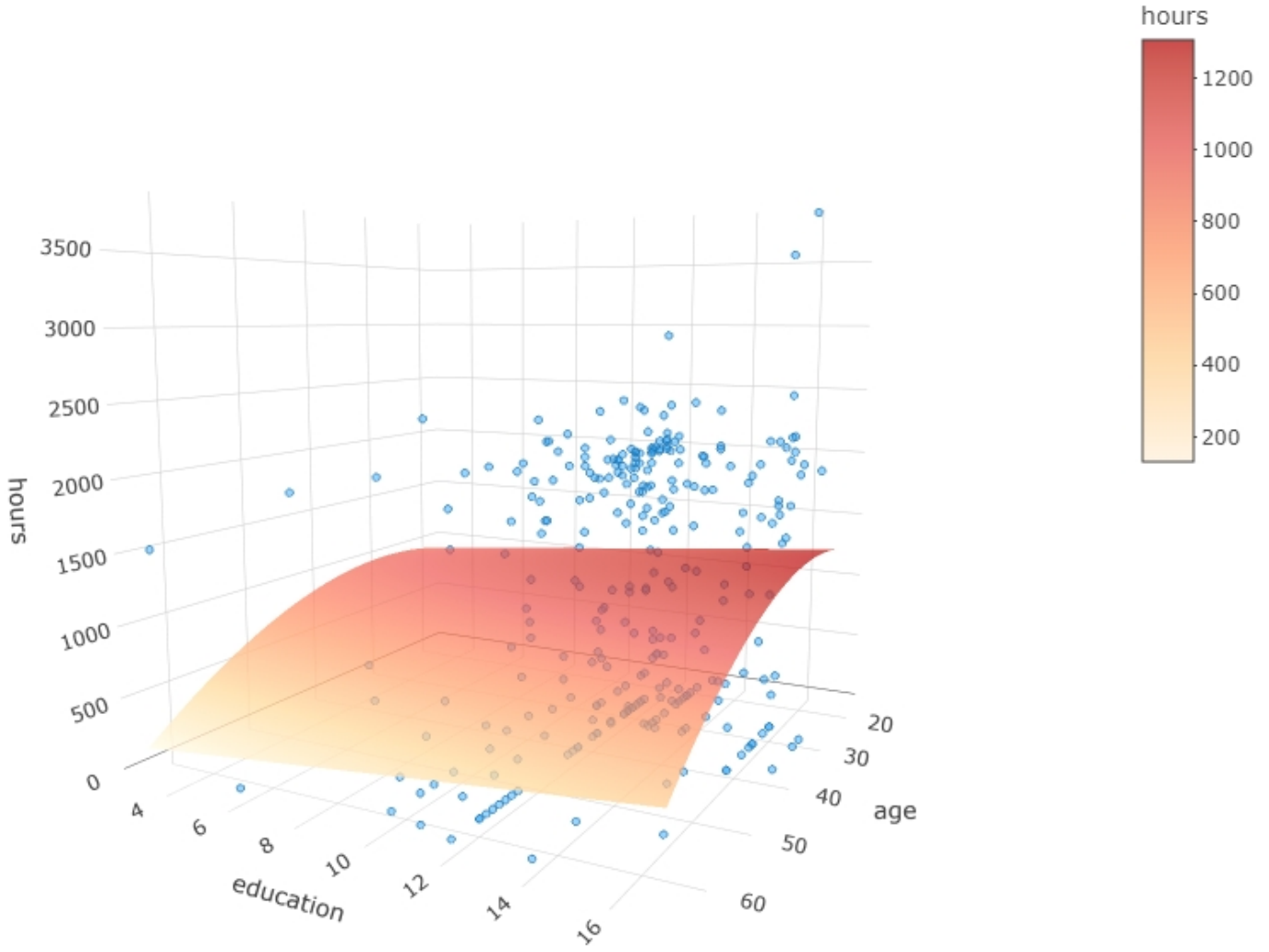
$$\hat{\boldsymbol{u}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y} = \boldsymbol{M}\boldsymbol{y}. \tag{2.4}$$

Mit zwei Regressoren kann Definition 2.2 noch visualisiert werden:



`RegPlane.R`. Erstellt in R mit dem Datensatz `Ecdat::workinghours`.

$\text{hours}_t = \hat{\beta}_1 + \hat{\beta}_2 \text{age}_t + \hat{\beta}_3 \text{age}_t^2 + \hat{\beta}_4 \text{education}_t$ (s. auch (1.11)):



Es ist nicht schwierig zu zeigen, dass

$$\boldsymbol{PX} = \boldsymbol{X} \tag{2.5}$$

und

$$\boldsymbol{MX} = \mathbf{O}. \tag{2.6}$$

Residuen und gefittete Werte sind **orthogonal** zueinander, da

$$\hat{\boldsymbol{u}}^\top \hat{\boldsymbol{y}} = 0 \tag{2.7}$$

`BechdelTest.R`

Dies ergibt sich aus einer Reihe von Eigenschaften. Zunächst ist $\boldsymbol{P}$ symmetrisch

$$\begin{aligned}
\boldsymbol{P}^\top = \left(\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\right)^\top &\overset{\text{M20.4}}{=} \left(\boldsymbol{X}^\top\right)^\top\left((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\right)^\top\boldsymbol{X}^\top \\
&\overset{\text{M21.3}}{=} \boldsymbol{X}\left((\boldsymbol{X}^\top\boldsymbol{X})^\top\right)^{-1}\boldsymbol{X}^\top \\
&= \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top = \boldsymbol{P}
\end{aligned}$$

Damit ist auch $\boldsymbol{M}$ offensichtlich symmetrisch. Daher ist

$$\begin{aligned}
\hat{\boldsymbol{u}}^\top\hat{\boldsymbol{y}} &= \boldsymbol{y}^\top\boldsymbol{M}^\top\boldsymbol{P}\boldsymbol{y} \\
&= \boldsymbol{y}^\top\boldsymbol{M}\boldsymbol{P}\boldsymbol{y}
\end{aligned}$$

Nutze nun Idempotenz von $\boldsymbol{P}$ ($\boldsymbol{P}\boldsymbol{P} = \boldsymbol{P}$, siehe Definition 9.2),

$$\begin{aligned}
\boldsymbol{M}\boldsymbol{P} &= \boldsymbol{P} - \boldsymbol{P}\boldsymbol{P} \\
&= \boldsymbol{P} - \boldsymbol{X}\underbrace{(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}}_{\boldsymbol{I}}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top \\
&= \boldsymbol{P} - \boldsymbol{P} = \boldsymbol{0} \qquad (2.8)
\end{aligned}$$

Dies impliziert demnach auch Orthogonalität von Residuen und Regressoren:

$$\boldsymbol{X}^\top \hat{\boldsymbol{u}} = \mathbf{0} \tag{2.9}$$

Übrigens ist auch $\boldsymbol{M}$ idempotent:

$$\begin{aligned} \boldsymbol{M}\boldsymbol{M} &= (\boldsymbol{I} - \boldsymbol{P})(\boldsymbol{I} - \boldsymbol{P}) \\ &= \boldsymbol{I} - 2\boldsymbol{P} + \boldsymbol{P}\boldsymbol{P} \\ &= \boldsymbol{I} - 2\boldsymbol{P} + \boldsymbol{P} \\ &= \boldsymbol{I} - \boldsymbol{P} \\ &= \boldsymbol{M} \end{aligned} \tag{2.10}$$

Die Definition der Residuen sagt uns, dass

$$\boldsymbol{y} = \boldsymbol{X}\hat{\beta}_{\text{OLS}} + \hat{\boldsymbol{u}} = \hat{\boldsymbol{y}} + \hat{\boldsymbol{u}}. \tag{2.11}$$

(2.7) erlaubt uns die **total sum of squares** TSS $:= \boldsymbol{y}^\top \boldsymbol{y}$ wie folgt zu zerlegen:

$$\begin{aligned}
\sum_{t=1}^{n} y_t^2 = \boldsymbol{y}^\top \boldsymbol{y} &= \left(\hat{\boldsymbol{y}} + \hat{\boldsymbol{u}}\right)^\top \left(\hat{\boldsymbol{y}} + \hat{\boldsymbol{u}}\right) \\
&= \hat{\boldsymbol{y}}^\top \hat{\boldsymbol{y}} + 2\hat{\boldsymbol{u}}^\top \hat{\boldsymbol{y}} + \hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}} \\
&\overset{(2.7)}{=} \hat{\boldsymbol{y}}^\top \hat{\boldsymbol{y}} + \hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}} \\
&=: \text{ESS} + \text{SSR}, \qquad (2.12)
\end{aligned}$$

wobei ESS die **explained sum of squares** und SSR die **sum of squared residuals** ist.

Dies ist eine fundamentale OLS-Eigenschaft: Wir zerlegen die Varianz von $\boldsymbol{y}$ in einen Teil, den wir erklären können und in einen, den wir nicht erklären können.

Solche Projektionsargumente sind nützlich, um einige Eigenschaften von OLS zu zeigen. Beispielsweise hat es keinen Einfluss auf die gefitteten Werte oder die Residuen, wenn wir die Maßeinheiten von $\boldsymbol{X}$ ändern.

Betrachten wir eine Transformation von $\boldsymbol{X}$ durch eine invertierbare $k \times k$ Matrix $\boldsymbol{A}$, $\boldsymbol{XA}$ (ändern wir z.B. die Einheit von Monaten zu Jahren oder von Metern zu Zentimetern).

Dann:

$$\begin{aligned} \boldsymbol{P_{XA}} &:= \boldsymbol{XA}\left((\boldsymbol{XA})^\top \boldsymbol{XA}\right)^{-1}(\boldsymbol{XA})^\top \\ &\overset{\text{M20.4}}{=} \boldsymbol{XA}\left(\boldsymbol{A}^\top \boldsymbol{X}^\top \boldsymbol{XA}\right)^{-1}\boldsymbol{A}^\top \boldsymbol{X}^\top \\ &\overset{\text{M20.5}}{=} \boldsymbol{XAA}^{-1}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}(\boldsymbol{A}^\top)^{-1}\boldsymbol{A}^\top \boldsymbol{X}^\top \\ &= \boldsymbol{P_X} \end{aligned} \tag{2.13}$$

Wie sieht es bei $\hat{\beta}_{\text{OLS}}$ aus? Betrachten wir

$$\begin{aligned}
\hat{\beta}^{\circ}_{\text{OLS}} &= (\underbrace{\boldsymbol{A}^\top \boldsymbol{X}^\top}_{\text{„}\boldsymbol{X}^\top\text{“}} \underbrace{\boldsymbol{X}\boldsymbol{A}}_{\text{„}\boldsymbol{X}\text{“}})^{-1} \underbrace{\boldsymbol{A}^\top \boldsymbol{X}^\top}_{\text{„}\boldsymbol{X}^\top\text{“}} \boldsymbol{y} \\
&\overset{\text{M20.5}}{=} \boldsymbol{A}^{-1}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}(\boldsymbol{A}^\top)^{-1}\boldsymbol{A}^\top \boldsymbol{X}^\top \boldsymbol{y} \\
&= \boldsymbol{A}^{-1}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y} \\
&= \boldsymbol{A}^{-1}\hat{\beta}_{\text{OLS}}
\end{aligned}$$

Wenn also folgende Werte im obigen Beispiel verwendet werden,

$$\boldsymbol{A} = \begin{pmatrix} 1/12 & 0 \\ 0 & 100 \end{pmatrix} \quad \text{so dass} \quad \boldsymbol{A}^{-1} = \begin{pmatrix} 12 & 0 \\ 0 & 1/100 \end{pmatrix}$$

wird der Effekt einer Änderung in den Regressoren—intuitiverweise—angepasst.

FWL

Das Frisch-Waugh-Lovell Theorem ist (nicht nur) ein nützliches Stück Regressionsalgebra. Betrachten wir das Modell

$$\boldsymbol{y} = \underbrace{\boldsymbol{X}_1}_{n \times k_1} \beta_1 + \underbrace{\boldsymbol{X}_2}_{n \times k_2} \beta_2 + \boldsymbol{u} \tag{2.14}$$

Wir interessieren uns für die Schätzung von β_2. Das FWL Theorem besagt, dass es zwei Möglichkeiten zur Berechnung von $\hat{\beta}_2$ gibt.

Theorem 2.4 *Folgende Berechnungen von $\hat{\beta}_2$ liefern numerisch identische Koeffizienten.*

1. *Regressiere $\boldsymbol{y}$ auf $\boldsymbol{X}_1$ und $\boldsymbol{X}_2$.*
2. - *Regressiere $\boldsymbol{y}$ auf $\boldsymbol{X}_1$ und speichere die Residuen, nenne sie $\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{y}$, wobei $\boldsymbol{M}_{\boldsymbol{X}_1} := \boldsymbol{I} - \boldsymbol{X}_1(\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top$.*
   - *Regressiere (jede Spalte in) $\boldsymbol{X}_2$ auf $\boldsymbol{X}_1$ und speichere die Residuen, nenne sie $\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2$.*
   - *Regressiere die Residuen aufeinander, also $\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{y}$ auf $\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2$, um den Schätzer $\hat{\beta}_2$ zu erhalten.*

Um dies zu zeigen, betrachten wir zunächst die Formel für $\hat{\beta}_2$ aus 2. Über die OLS-Formel (1.17) erhalten wir

$$\begin{aligned} \hat{\beta}_2 &= \big(\underbrace{\boldsymbol{X}_2^\top \boldsymbol{M}_{\boldsymbol{X}_1}^\top}_{„\boldsymbol{X}^\top“} \underbrace{\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2}_{„\boldsymbol{X}“}\big)^{-1} \underbrace{\boldsymbol{X}_2^\top \boldsymbol{M}_{\boldsymbol{X}_1}^\top}_{„\boldsymbol{X}^\top“} \underbrace{\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{y}}_{„\boldsymbol{y}“} \\ &\overset{(2.10)}{=} \big(\boldsymbol{X}_2^\top \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2\big)^{-1}\boldsymbol{X}_2^\top \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{y} \end{aligned} \tag{2.15}$$

Wenn wir die Regression wie in 1 durchführen, also direkt wie in (2.14), erhalten wir unter Verwendung von (2.4)

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\hat{\beta}_{\text{OLS}} + \boldsymbol{M}\boldsymbol{y} \\ &= \begin{pmatrix}\boldsymbol{X}_1 & \boldsymbol{X}_2\end{pmatrix}\begin{pmatrix}\hat{\beta}_1 \\ \hat{\beta}_2\end{pmatrix} + \boldsymbol{M}\boldsymbol{y} \\ &= \boldsymbol{X}_1\hat{\beta}_1 + \boldsymbol{X}_2\hat{\beta}_2 + \boldsymbol{M}\boldsymbol{y} \end{aligned} \tag{2.16}$$

Multiplizieren von (2.16) mit $\boldsymbol{X}_2^\top \boldsymbol{M}_{\boldsymbol{X}_1}$ ergibt

$$\boldsymbol{X}_2^\top \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{y} = \boldsymbol{X}_2^\top \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2\hat{\beta}_2 + \boldsymbol{X}_2^\top \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{M}\boldsymbol{y}$$

Hier fällt der erste Term in (2.16) weg, da $\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_1 = \mathbf{O}$.

$\boldsymbol{X}_2^\top \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{M}\boldsymbol{y}$ ist ebenfalls gleich null. Denn:

$$\begin{aligned}
\boldsymbol{M}\boldsymbol{M}_{\boldsymbol{X}_1} &= (\boldsymbol{I} - \boldsymbol{P})(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}_1}) \\
&= \boldsymbol{I} - \boldsymbol{P} - \boldsymbol{P}_{\boldsymbol{X}_1} + \boldsymbol{P}\boldsymbol{P}_{\boldsymbol{X}_1} \\
&= \boldsymbol{I} - \boldsymbol{P} - \boldsymbol{P}_{\boldsymbol{X}_1} + \boldsymbol{P}_{\boldsymbol{X}_1} \\
&= \boldsymbol{I} - \boldsymbol{P} \\
&= \boldsymbol{M}
\end{aligned}$$

Dass $\boldsymbol{P}\boldsymbol{P}_{\boldsymbol{X}_1} = \boldsymbol{P}_{\boldsymbol{X}_1}$ (dritte Gleichheit) gilt, geht auf die Tatsache zurück, dass die gefitteten Werte einer Regression einiger Spalten von $\boldsymbol{X}$ (hier also $\boldsymbol{X}_1$), $\boldsymbol{P}\boldsymbol{X}_1$, $\boldsymbol{X}_1$ entsprechen. Um dies besser zu sehen, schreiben wir $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$ als

$$\boldsymbol{P}(\boldsymbol{X}_1 : \boldsymbol{X}_2) = (\boldsymbol{X}_1 : \boldsymbol{X}_2),$$

so dass die ersten k_1 Spalten von $\boldsymbol{P}(\boldsymbol{X}_1 : \boldsymbol{X}_2)$ der Matrix $\boldsymbol{P}\boldsymbol{X}_1 = \boldsymbol{X}_1$ entsprechen. Daraus folgt, dass

$$\boldsymbol{P}\boldsymbol{P}_{\boldsymbol{X}_1} = \boldsymbol{P}\boldsymbol{X}_1(\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top = \boldsymbol{X}_1(\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top = \boldsymbol{P}_{\boldsymbol{X}_1}$$

Daraus folgt ebenfalls, dass

$$\boldsymbol{X}_2^\top \boldsymbol{M}_{\boldsymbol{X}_1} \boldsymbol{M} = \boldsymbol{X}_2^\top \boldsymbol{M} = \mathbf{O}, \tag{2.17}$$

da $\boldsymbol{M}\boldsymbol{X}_2$ den Residuen einer Regression von $\boldsymbol{X}_2$ auf $\boldsymbol{X}$ entspricht, was $\boldsymbol{X}_2$ enthält. Demzufolge hat die Regression perfekten Fit und die Residuen sind daher null.

Jetzt lösen wir

$$\boldsymbol{X}_2^\top \boldsymbol{M}_{\boldsymbol{X}_1} \boldsymbol{y} = \boldsymbol{X}_2^\top \boldsymbol{M}_{\boldsymbol{X}_1} \boldsymbol{X}_2 \hat{\beta}_2$$

nach $\hat{\beta}_2$ auf und sind fertig. Siehe (2.15).

`fwl.R`

Eine Anwendung

(Interessantere kommen später)

Beispiel 2.5 Nehmen wir $\boldsymbol{X}_1 = \boldsymbol{\iota} := (1, \ldots, 1)^\top$. Das Folgende ergibt identische $\hat{\beta}_2$, die Koeffizienten für $\boldsymbol{X}_2$.

- Regressiere $\boldsymbol{y}$ auf $\boldsymbol{\iota}$ (d.h. eine Konstante) und $\boldsymbol{X}_2$.
- Regressiere $\boldsymbol{y} - \bar{y}$ auf $\boldsymbol{X}_2 - \bar{\boldsymbol{X}}_2$, wobei letzteres die Spalten von $\boldsymbol{X}_2$ zentriert um deren jeweilige Spaltenmittelwerte bezeichnet.

Um dies zu sehen, betrachte

$$\boldsymbol{M}_\iota = \boldsymbol{I} - \boldsymbol{\iota}(\boldsymbol{\iota}^\top \boldsymbol{\iota})^{-1}\boldsymbol{\iota}^\top = \boldsymbol{I} - \frac{\boldsymbol{\iota}\boldsymbol{\iota}^\top}{n}, \tag{2.18}$$

so dass (wobei $\boldsymbol{x}_i$ der Spaltenvektor der n Beobachtungen einer Variable aus $\boldsymbol{X}_2$ sei)

$$\boldsymbol{M}_\iota \boldsymbol{x}_i = \boldsymbol{x}_i - \boldsymbol{\iota} n^{-1} \boldsymbol{\iota}^\top \boldsymbol{x}_i = \boldsymbol{x}_i - \boldsymbol{\iota}\bar{x}_i =: \boldsymbol{x}_i - \bar{\boldsymbol{x}}_i.$$

Daher sind die Residuen einer Regression von Variablen auf eine Konstante $\boldsymbol{M}_\iota \boldsymbol{x}_i$ lediglich die um den Mittelwert bereinigten Variablen.

Goodness of fit

Erinnern wir uns an die Zerlegung aus (2.12),

$$\text{TSS} = \boldsymbol{y}^\top \boldsymbol{y} = \hat{\boldsymbol{y}}^\top \hat{\boldsymbol{y}} + \hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}} = \text{ESS} + \text{SSR}$$

Dies liefert ein Maß für die Anpassungsgüte (goodness of fit), das **Bestimmtheitsmaß** oder auch R^2 genannt wird.

Definition 2.6 (unzentriertes R^2).

$$R_u^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{\text{TSS} - \text{SSR}}{\text{TSS}} = 1 - \frac{\text{SSR}}{\text{TSS}} = 1 - \frac{\hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}}}{\boldsymbol{y}^\top \boldsymbol{y}}$$

Wir interpretieren das R^2 als den Teil der Variation in den Daten, der durch das Modell erklärt wird.

- $R_u^2 = 1$, wenn die Regressionslinie perfekt passt.
- $R_u^2 = 0$, wenn $\hat{\beta}_{\text{OLS}} = \mathbf{0}$ (s. (2.11) mit dann $\boldsymbol{y} = \hat{\boldsymbol{u}}$).

Diese Version ist jedoch nicht robust gegen scheinbar harmlosen Transformationen von $\boldsymbol{y}$. Die folgende Variante, die in Anwendungen beliebter ist, vermeidet diesen Nachteil.

Definition 2.7 (zentriertes R^2).

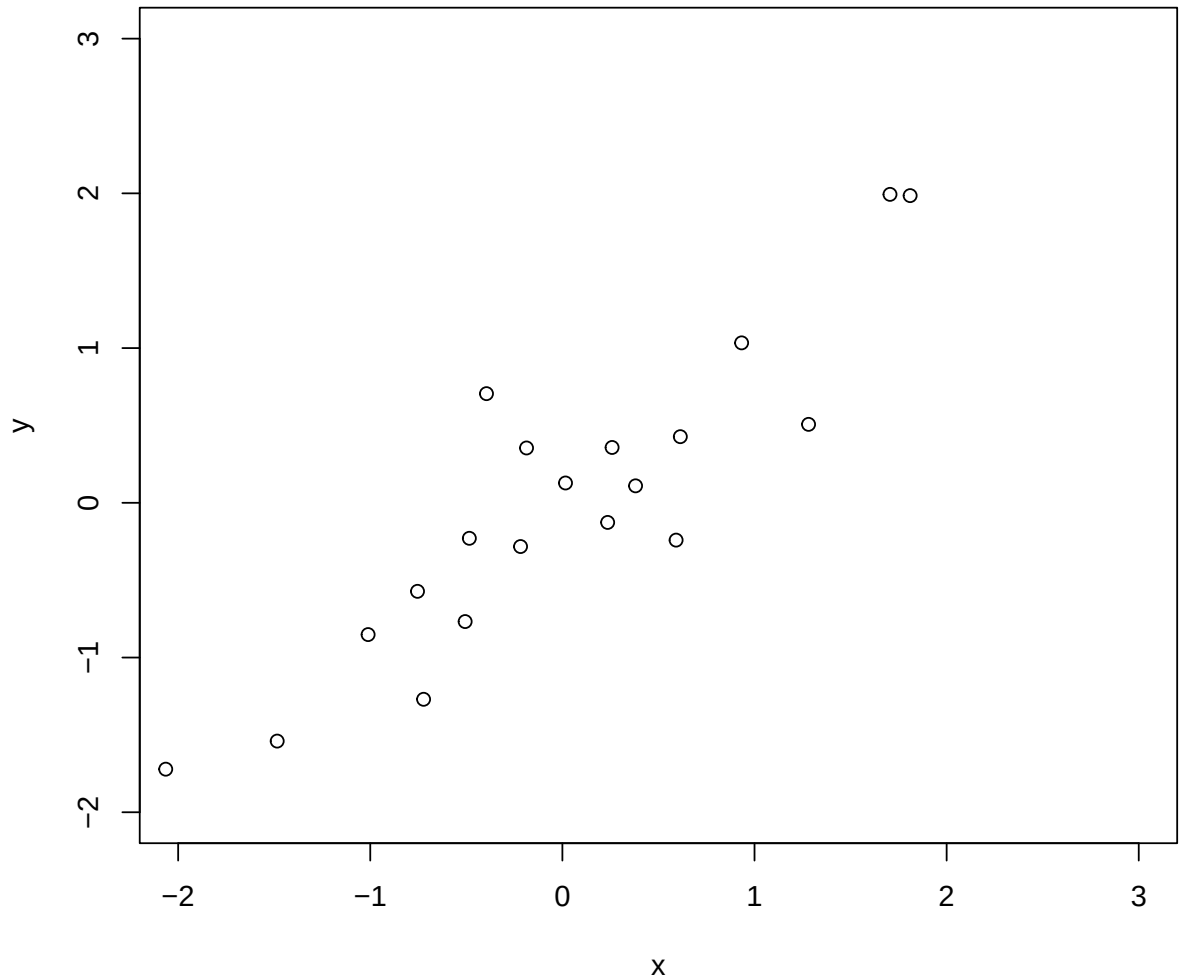
$$R^2 = 1 - \frac{\hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}}}{\boldsymbol{y}^\top \boldsymbol{M}_\iota \boldsymbol{y}} \tag{2.19}$$

$$\overset{\text{Ex. 2.5}}{=} 1 - \frac{\hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}}}{(\boldsymbol{y} - \bar{y})^\top (\boldsymbol{y} - \bar{y})}$$

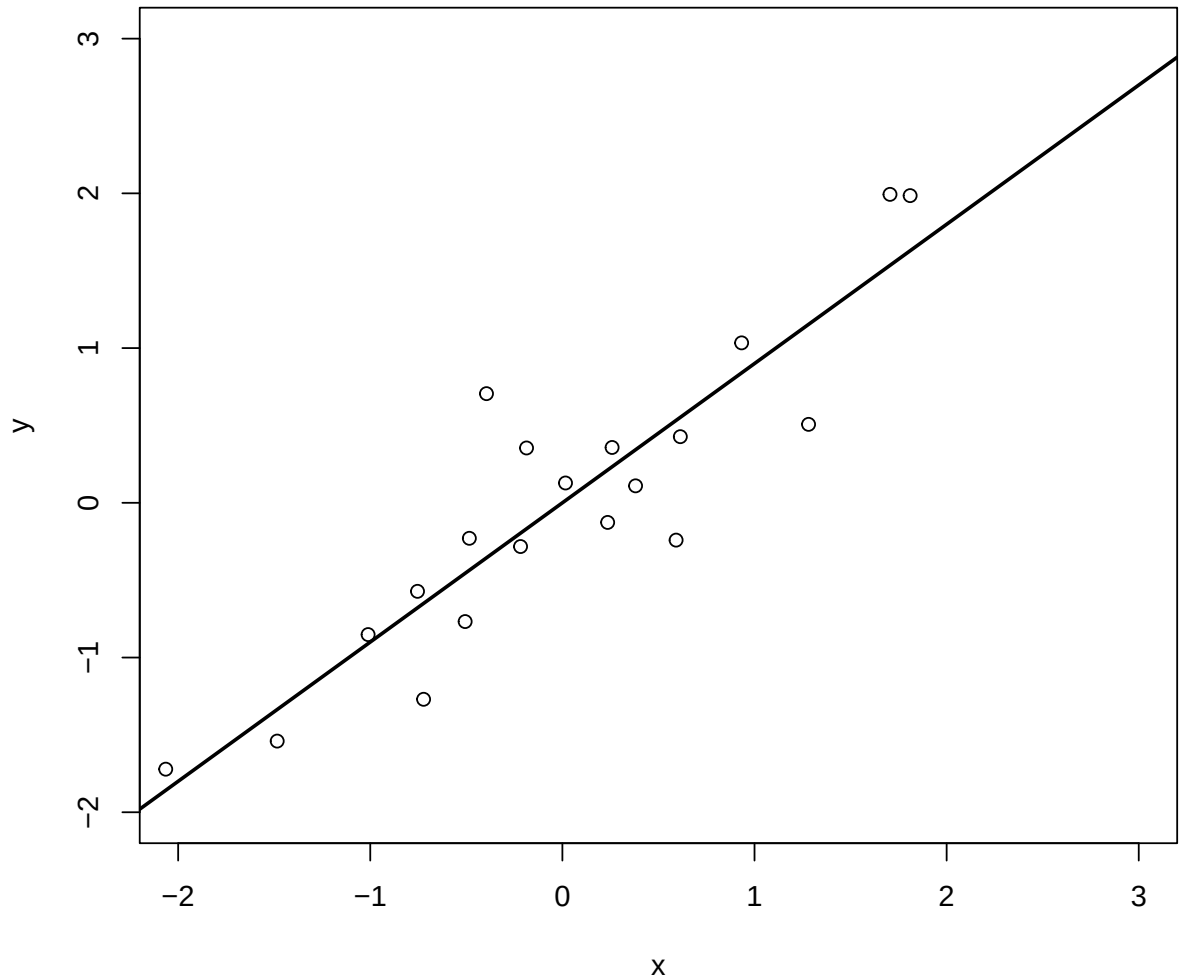
Dieses Maß macht jedoch nur Sinn, wenn eine Konstante in der Regression enthalten ist. Ansonsten ist ein $R^2 < 0$ möglich, was die Interpretation fragwürdig macht.

Wir müssen uns, unabhängig vom Wert für R^2, immer bewusst sein, was das R^2 für uns tun kann und was nicht.

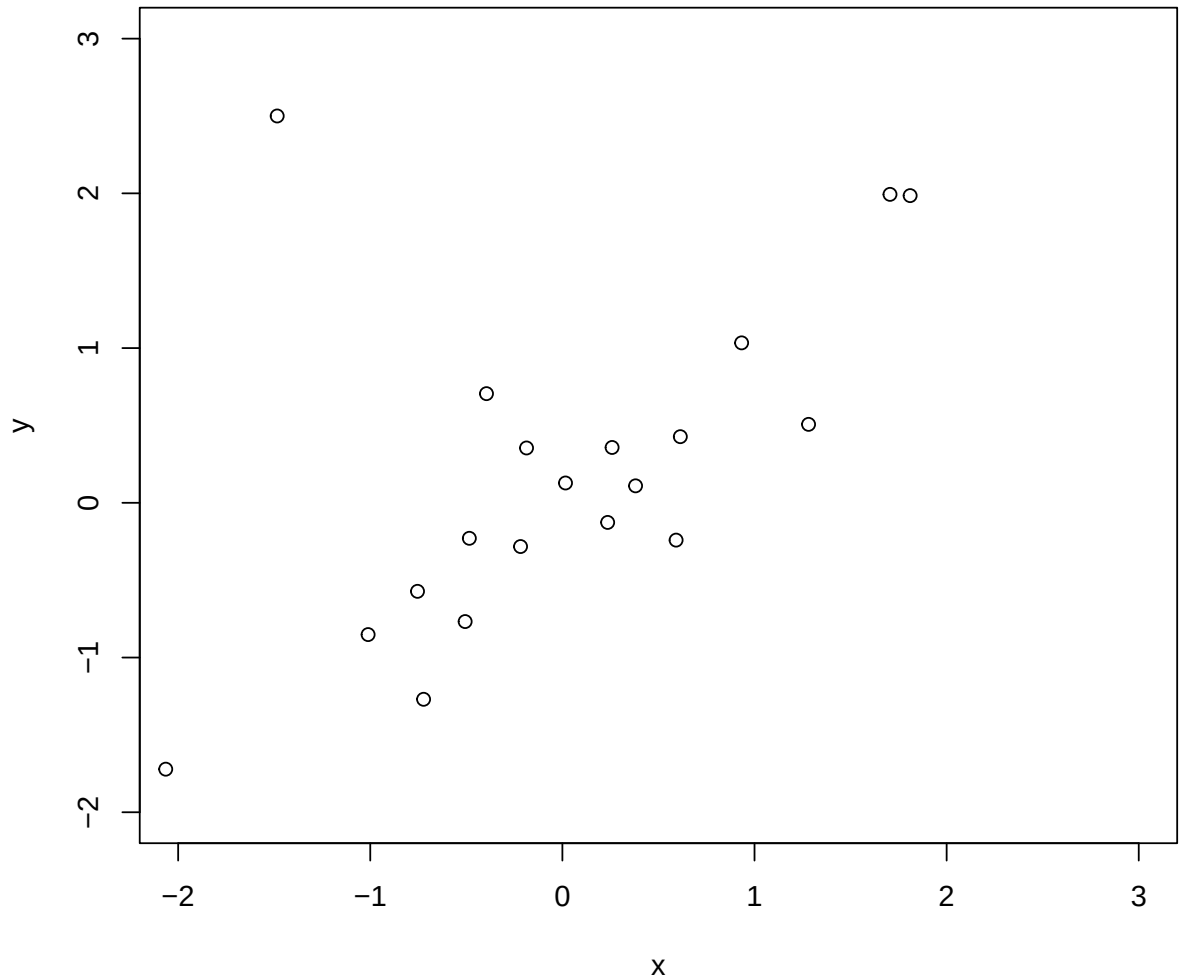
Die Minimierung der *quadrierten* Residuen in OLS impliziert, dass **Ausreißer (Outlier)** eine hohes Gewicht erhalten:



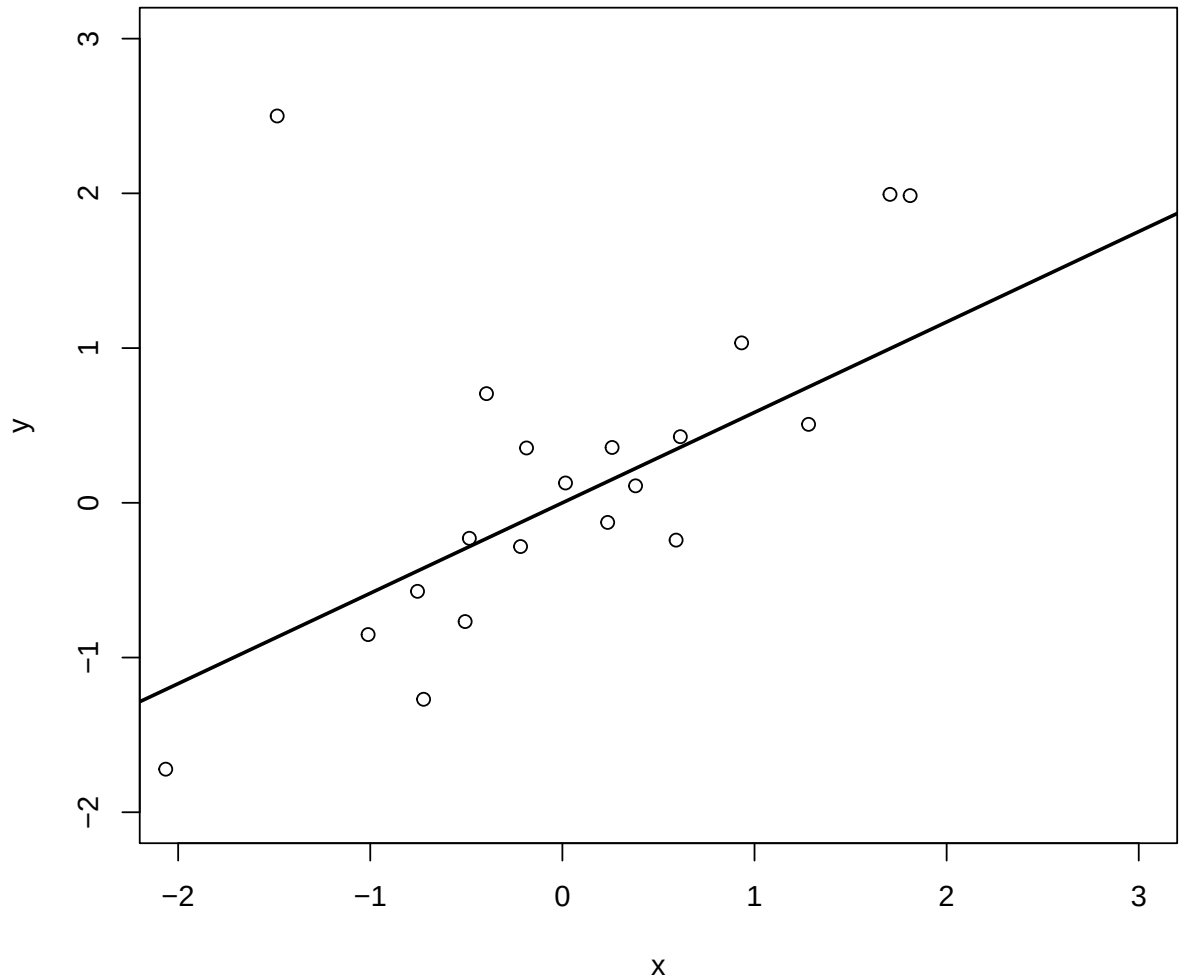
Die Minimierung der *quadrierten* Residuen in OLS impliziert, dass **Ausreißer (Outlier)** eine hohes Gewicht erhalten:



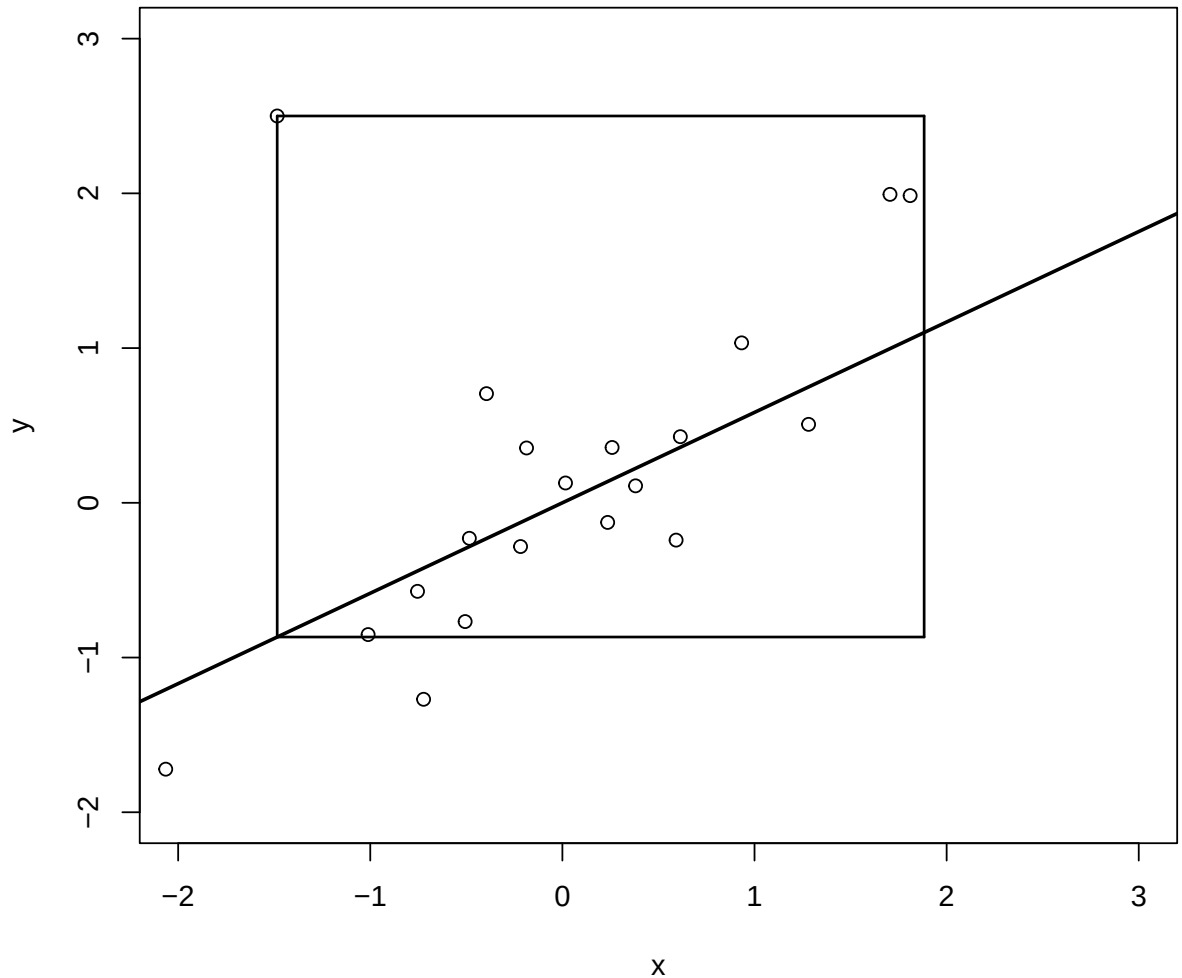
Die Minimierung der *quadrierten* Residuen in OLS impliziert, dass **Ausreißer (Outlier)** eine hohes Gewicht erhalten:



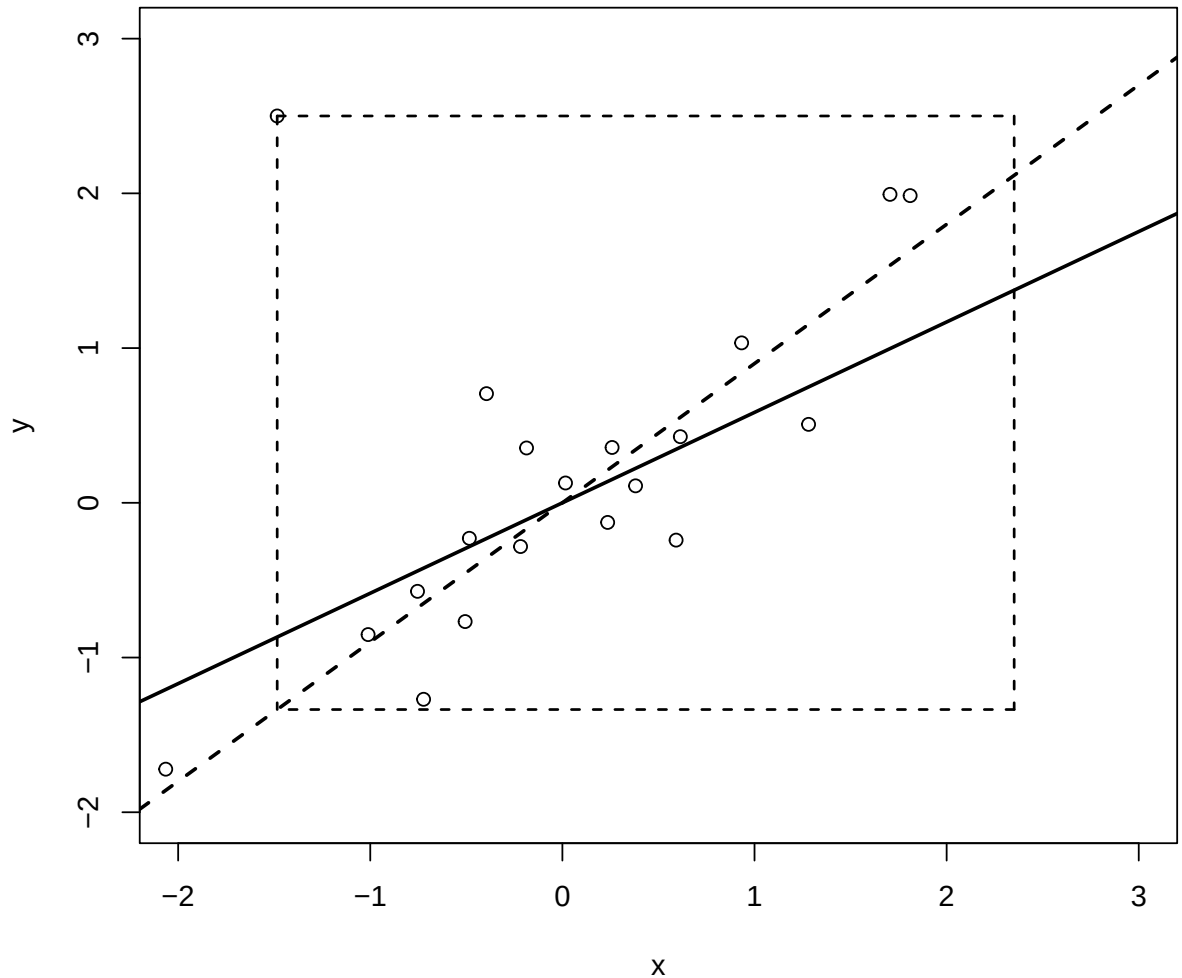
Die Minimierung der *quadrierten* Residuen in OLS impliziert, dass **Ausreißer (Outlier)** eine hohes Gewicht erhalten:



Die Minimierung der *quadrierten* Residuen in OLS impliziert, dass **Ausreißer (Outlier)** eine hohes Gewicht erhalten:



Die Minimierung der *quadrierten* Residuen in OLS impliziert, dass **Ausreißer (Outlier)** eine hohes Gewicht erhalten:



Daher ist es immer eine gute Idee die Daten erst einmal wie oben zu plotten.

Was wir hingegen mit Ausreißern machen sollen, ist weniger klar. Im einfachsten Fall wurde einfach falsch gemessen, oder fehlende Werte sind mit einem Zahlenwert codiert (oft -99), werden jedoch für eine echte Beobachtung gehalten. In diesem Fall können die Fehler korrigiert oder die Datenpunkte aus dem Datensatz entfernt werden.

Dennoch können auch ungewöhnlich große Realisationen vorkommen, wenn auch selten. Dann enthalten sie interessante Informationen über das Modell, gerade *weil* sie die Schätzungen so stark beeinflussen.

Meist ist es eine gute Strategie die Ergebnisse mit und ohne Ausreißer darzustellen und die Unterschiede zu interpretieren.

`MinimizeSumPowerResiduals.R`

Um Ausreißer etwas formaler zu analysieren, betrachten wir die Schätzung von β, wenn wir die t-te Beobachtung herauslassen. Wenn

$$\boldsymbol{y}_{(t)} = \begin{pmatrix} y_1 \\ \vdots \\ y_{t-1} \\ y_{t+1} \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{X}_{(t)} = \begin{pmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_{t-1} \\ \boldsymbol{X}_{t+1} \\ \vdots \\ \boldsymbol{X}_n \end{pmatrix}, \quad \boldsymbol{u}_{(t)} = \begin{pmatrix} u_1 \\ \vdots \\ u_{t-1} \\ u_{t+1} \\ \vdots \\ u_n \end{pmatrix},$$

dann ist

$$\boldsymbol{y}_{(t)} = \boldsymbol{X}_{(t)}\beta + \boldsymbol{u}_{(t)}.$$

Eine OLS-Regression liefert

$$\hat{\beta}_{(t)} = (\boldsymbol{X}_{(t)}^\top \boldsymbol{X}_{(t)})^{-1} \boldsymbol{X}_{(t)}^\top \boldsymbol{y}_{(t)}.$$

Wenn die Differenz zwischen $\hat{\beta}$ und $\hat{\beta}_{(t)}$ „groß“ ist (die Differenz ist ein Vektor), dann bezeichnen wir die t-te Beobachtung als **einflussreich**.

Die Diagonalelemente der **Projektionsmatrix $\boldsymbol{P_X}$** sind gegeben durch $h_t = \boldsymbol{e}_t^\top \boldsymbol{P_X} \boldsymbol{e}_t$, $t = 1, \ldots, n$, wobei die $\boldsymbol{e}_t$ Einheitsvektoren sind, welche die Spalten der Einheitsmatrix bilden. Man kann zeigen, dass

$$0 \leq h_t \leq 1, \qquad t = 1, \ldots, n,$$
$$\sum_{t=1}^{n} h_t = k.$$

Um den Einfluss der t-ten Beobachtung zu ermitteln, betrachten wir das folgende Resultat. Erinnern wir uns an $\hat{\boldsymbol{u}} = \boldsymbol{y} - \hat{\boldsymbol{y}}$.

Resultat 2.8

$$\hat{\beta} - \hat{\beta}_{(t)} = \left(\frac{\hat{u}_t}{1 - h_t} \right) (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}_t^\top \tag{2.20}$$

Wenn h_t groß ist, z.B. nah bei eins, sagen wir, dass $\boldsymbol{X}_t$ starke **Hebelwirkung (Leverage)** hat. Wenn $|\hat{u}_t|$ ebenfalls groß ist, ist die t-te Beobachtung einflussreich.

Um das Resultat 2.8 auf eine andere Weise als im Lehrbuch zu zeigen, betrachten wir das folgende Matrixalgebra-Resultat.

Resultat 2.9 Gegeben sei eine nicht-singuläre Matrix $\boldsymbol{A}$, ein Vektor $\boldsymbol{b}$ und ein Skalar λ. Wenn

$$\lambda \neq -\frac{1}{\boldsymbol{b}^\top \boldsymbol{A}^{-1} \boldsymbol{b}}, \tag{2.21}$$

dann ist

$$(\boldsymbol{A} + \lambda \boldsymbol{b}\boldsymbol{b}^\top)^{-1} = \boldsymbol{A}^{-1} - \left(\frac{\lambda}{1 + \lambda \boldsymbol{b}^\top \boldsymbol{A}^{-1} \boldsymbol{b}} \right) \boldsymbol{A}^{-1} \boldsymbol{b}\boldsymbol{b}^\top \boldsymbol{A}^{-1}. \tag{2.22}$$

Beachte, dass $\boldsymbol{A} + \lambda \boldsymbol{b}\boldsymbol{b}^\top$ singulär ist, wenn (2.21) nicht erfüllt ist. Der **Beweis für Resultat 2.9** folgt direkt aus der Überprüfung von

$$\left\{ \boldsymbol{A}^{-1} - \left(\frac{\lambda}{1 + \lambda \boldsymbol{b}^\top \boldsymbol{A}^{-1} \boldsymbol{b}} \right) \boldsymbol{A}^{-1} \boldsymbol{b}\boldsymbol{b}^\top \boldsymbol{A}^{-1} \right\} (\boldsymbol{A} + \lambda \boldsymbol{b}\boldsymbol{b}^\top) = \boldsymbol{I}.$$

Das folgende Resultat ist hilfreich beim Beweis von 2.8.

Resultat 2.10

$$(\boldsymbol{X}_{(t)}^\top \boldsymbol{X}_{(t)})^{-1} \boldsymbol{x}_t^\top = \left(\frac{1}{1-h_t}\right) (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_t^\top. \tag{2.23}$$

Beweis von Resultat 2.10: Durch Resultat 2.9 erhalten wir

$$\begin{aligned}(\boldsymbol{X}_{(t)}^\top \boldsymbol{X}_{(t)})^{-1} &= (\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{x}_t^\top \boldsymbol{x}_t)^{-1} \\ &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} + \frac{(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_t^\top \boldsymbol{x}_t (\boldsymbol{X}^\top \boldsymbol{X})^{-1}}{1 - \boldsymbol{x}_t (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_t^\top}.\end{aligned}$$

Daher erhalten wir

$$\begin{aligned}(\boldsymbol{X}_{(t)}^\top \boldsymbol{X}_{(t)})^{-1} \boldsymbol{x}_t^\top &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_t^\top + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_t^\top \left(\frac{\boldsymbol{x}_t (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_t^\top}{1 - \boldsymbol{x}_t (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_t^\top}\right) \\ &= \left(\frac{1}{1-h_t}\right) (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_t^\top.\end{aligned}$$

q.e.d.

Der Beweis von Resultat 2.8 folgt nun aus Resultat 2.10.

Beweis von Resultat 2.8: Da

$$\boldsymbol{X}^\top \boldsymbol{X}\hat{\beta} = \boldsymbol{X}^\top \boldsymbol{y},$$

erhalten wir

$$(\boldsymbol{X}_{(t)}^\top \boldsymbol{X}_{(t)} + \boldsymbol{X}_t^\top \boldsymbol{X}_t)\hat{\beta} = \boldsymbol{X}_{(t)}^\top \boldsymbol{y}_{(t)} + \boldsymbol{X}_t^\top y_t,$$

oder

$$\left\{\boldsymbol{I}_k + (\boldsymbol{X}_{(t)}^\top \boldsymbol{X}_{(t)})^{-1}\boldsymbol{X}_t^\top \boldsymbol{X}_t\right\}\hat{\beta} = \hat{\beta}_{(t)} + (\boldsymbol{X}_{(t)}^\top \boldsymbol{X}_{(t)})^{-1}\boldsymbol{X}_t^\top(\boldsymbol{X}_t\hat{\beta} + \hat{u}_t).$$

Daher

$$\begin{aligned}\hat{\beta} &= \hat{\beta}_{(t)} + (\boldsymbol{X}_{(t)}^\top \boldsymbol{X}_{(t)})^{-1}\boldsymbol{X}_t^\top \hat{u}_t \\ &= \hat{\beta}_{(t)} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}_t^\top \frac{\hat{u}_t}{1-h_t},\end{aligned}$$

wobei die letzte Gleichung aus Resultat 2.10 folgt q.e.d.

Überblick

Das klassische lineare Modell ist gegeben durch

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{u}, \tag{3.1}$$

wobei wir den n-Vektor $\boldsymbol{y}$, die abhängige Variable, beobachten und die $n \times k$ Matrix $\boldsymbol{X}$, die erklärenden Variablen enthält.

Der n-Vektor der Störgrößen $\boldsymbol{u}$ wird nicht beobachtet. Wir treffen Annahmen über seine Verteilung (wie beispielsweise (1.18)).

Auch $\boldsymbol{X}$ ist (außer z.B. in Experimenten) stochastisch (denken wir z.B. an Erfahrung). Wir nehmen an, dass der k-Vektor der Koeffizienten β nicht zufällig aber unbekannt ist.

Wir nehmen zudem **Exogenität** an:

Annahme 3.1

$$E(\boldsymbol{u}|\boldsymbol{X}) = \boldsymbol{0}. \tag{3.2}$$

Dies ist für eine für eine u.i.v. Stichprobe dasselbe wie $E(u_t|\boldsymbol{X}_t) = 0$ in (1.18).

Unter der Annahme 3.1 ist der OLS-Schätzer bedingt unverzerrt: Wir haben mit (3.1)

$$\hat{\beta}_{\text{OLS}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} = \beta + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u}, \tag{3.3}$$

und daher ist

$$E(\hat{\beta}_{\text{OLS}} | \boldsymbol{X}) \overset{(1.9)}{=} \beta + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top E(\boldsymbol{u} | \boldsymbol{X}) = \beta. \tag{3.4}$$

Aufgrund des Gesetzes der iterierten Erwartungen (1.7) ist $\hat{\beta}_{\text{OLS}}$ auch unbedingt erwartungstreu.

Siehe `unbiasedness.R` für einige graphische Darstellungen.

Der Begriff einer **Varianz-Kovarianzmatrix** eines $m \times 1$ Vektors $\boldsymbol{x}$,

$$\boldsymbol{x} = (X_1, \ldots, X_m)^\top,$$

ist gleich hilfreich:

$$\begin{aligned} Var(\boldsymbol{x}) &= E\Big((\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^\top\Big) \qquad (3.5) \\ &= \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & \ldots & Cov(X_1, X_m) \\ Cov(X_1, X_2) & Var(X_2) & \ldots & Cov(X_2, X_m) \\ \vdots & \ddots & \ldots & \vdots \\ \vdots & & \ddots & \vdots \\ Cov(X_1, X_m) & \ldots & Cov(X_{m-1}, X_m) & Var(X_m) \end{pmatrix} \end{aligned}$$

Für

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$$

erhalten wir außerdem

$$\begin{aligned}
Var(\boldsymbol{y}) &= E\big((\boldsymbol{y} - E(\boldsymbol{y}))(\boldsymbol{y} - E(\boldsymbol{y}))^\top\big) \\
&= E\big((\boldsymbol{A}\boldsymbol{x} - E(\boldsymbol{A}\boldsymbol{x}))(\boldsymbol{A}\boldsymbol{x} - E(\boldsymbol{A}\boldsymbol{x}))^\top\big) \\
&= E\big((\boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}E(\boldsymbol{x}))(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}E(\boldsymbol{x}))^\top\big) \\
&= E\big(\boldsymbol{A}(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{A}(\boldsymbol{x} - E(\boldsymbol{x})))^\top\big) \\
&\overset{\text{M20.4}}{=} E\big(\boldsymbol{A}(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^\top\boldsymbol{A}^\top\big) \\
&= \boldsymbol{A}E\big((\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^\top\big)\boldsymbol{A}^\top \\
&= \boldsymbol{A}Var(\boldsymbol{x})\boldsymbol{A}^\top \qquad (3.6)
\end{aligned}$$

Nehmen wir zusätzlich zu 3.1 ebenfalls an, dass

Annahme 3.2 $Var(\boldsymbol{u}|\boldsymbol{X}) = \sigma^2 \boldsymbol{I}_n$, wobei σ eine unbekannte Konstante ist, also ein zusätzlicher Parameter.

Aus 3.1 folgt $E(\boldsymbol{u}\boldsymbol{u}^\top|\boldsymbol{X}) = Var(\boldsymbol{u}|\boldsymbol{X})$, da

$$Var(\boldsymbol{u}|\boldsymbol{X}) = E\big((\boldsymbol{u} - \underbrace{E(\boldsymbol{u}|\boldsymbol{X})}_{=\mathbf{0}})((\boldsymbol{u} - E(\boldsymbol{u}|\boldsymbol{X}))^\top|\boldsymbol{X}\big) = E(\boldsymbol{u}\boldsymbol{u}^\top|\boldsymbol{X})$$

Was bedeutet diese Annahme? `educearn.R`

Die Kovarianzmatrix stellt sich dann wie folgt dar:

$$
\begin{aligned}
Var(\hat{\beta}_{\text{OLS}}|\boldsymbol{X}) &= E\{(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}|\boldsymbol{X}\} \\
&\overset{(3.5)}{=} (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top Var(\boldsymbol{u}|\boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \\
&= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\{\sigma^2\boldsymbol{I}_n\}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \\
&= \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\underbrace{\boldsymbol{X}^\top\boldsymbol{I}_n\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}}_{=\boldsymbol{I}} \\
&= \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}. \qquad (3.7)
\end{aligned}
$$

$Var(\hat{\beta}_{\text{OLS}}|\boldsymbol{X})$ ist unbekannt, da σ^2 unbekannt ist.

σ^2 aus 3.2 kann unverzerrt geschätzt werden durch $\hat{\sigma}^2$. $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ wird der **Standardfehler der Regression** genannt.

$$
\begin{aligned}
\hat{\sigma}^2 &= \frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}_{\text{OLS}})^\top (\boldsymbol{y} - \boldsymbol{X}\hat{\beta}_{\text{OLS}})}{n-k} \qquad (3.8)\\
&= \frac{\hat{\boldsymbol{u}}^\top \hat{\boldsymbol{u}}}{n-k} \overset{(2.4)}{=} \frac{\boldsymbol{y}^\top \boldsymbol{M}^\top \boldsymbol{M}\boldsymbol{y}}{n-k}\\
&\overset{(2.10)}{=} \frac{\boldsymbol{y}^\top \boldsymbol{M}\boldsymbol{y}}{n-k}\\
&= \frac{(\boldsymbol{X}\beta + \boldsymbol{u})^\top \boldsymbol{M}(\boldsymbol{X}\beta + \boldsymbol{u})}{n-k}\\
&\overset{(2.6)}{=} \frac{\boldsymbol{u}^\top \boldsymbol{M}\boldsymbol{u}}{n-k}\\
&= \frac{\text{tr}\left(\boldsymbol{u}^\top \boldsymbol{M}\boldsymbol{u}\right)}{n-k}\\
&\overset{\text{M20.6}}{=} \frac{\text{tr}\left(\boldsymbol{M}\boldsymbol{u}\boldsymbol{u}^\top\right)}{n-k}\\
&= \text{tr}\left\{\frac{(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{u}\boldsymbol{u}^\top}{n-k}\right\}.
\end{aligned}
$$

Daher erhalten wir

$$
\begin{aligned}
E(\hat{\sigma}^2|\boldsymbol{X}) &= E\left[\operatorname{tr}\left\{\frac{(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{u}\boldsymbol{u}^\top}{n-k}\right\}\middle|\,\boldsymbol{X}\right] \\
&= \operatorname{tr}\left[E\left\{\frac{(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{u}\boldsymbol{u}^\top}{n-k}\middle|\,\boldsymbol{X}\right\}\right] \\
&= \operatorname{tr}\left[\frac{(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})E(\boldsymbol{u}\boldsymbol{u}^\top|\boldsymbol{X})}{n-k}\right] \\
&\overset{3.2}{=} \operatorname{tr}\left[\frac{(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\sigma^2\boldsymbol{I}}{n-k}\right] \\
&= \sigma^2\frac{\operatorname{tr}(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})}{n-k}
\end{aligned}
$$

Außerdem ist

$$\begin{aligned}
\operatorname{tr}(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}}) &\overset{\text{M19.5}}{=} \operatorname{tr}(\boldsymbol{I}_n) - \operatorname{tr}(\boldsymbol{P}_{\boldsymbol{X}}) \\
&= n - \operatorname{tr}(\boldsymbol{P}_{\boldsymbol{X}}) \\
&= n - \operatorname{tr}\big(\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top\big) \\
&= n - \operatorname{tr}\big(\boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\big) \\
&= n - \operatorname{tr}(\boldsymbol{I}_k) \\
&= n - k
\end{aligned} \tag{3.9}$$

Also ist $\hat{\sigma}^2$ ein unverzerrter Schätzer von σ^2.

Frage 3.3 Was ist mit $E(\hat{\sigma})$? Ist es gleich σ, ist es kleiner oder ist es größer?

Gauss-Markov

Unter den Annahmen 3.1 und 3.2 können wir zeigen, dass der OLS-Schätzer $\hat{\beta}_{\text{OLS}}$ BLUE („Best Linear Unbiased Estimator“) ist.

Dies ist die Aussage des **Gauss-Markov-Theorems**.

Betrachten wir einen anderen linearen Schätzer

$$\hat{\tilde{\beta}} = \boldsymbol{C}\boldsymbol{y} = \boldsymbol{C}\boldsymbol{X}\beta + \boldsymbol{C}\boldsymbol{u}.$$

Damit $\hat{\tilde{\beta}}$ unverzerrt ist, muss gelten, dass

$$E(\hat{\tilde{\beta}}|\boldsymbol{X}) = \boldsymbol{C}\boldsymbol{X}\beta = \beta.$$

Dies sollte ungeachtet des Wertes von β gelten, das heißt es sollte für alle $\beta \in \mathbb{R}^k$ gelten. Damit $\hat{\tilde{\beta}}$ unverzerrt ist, benötigen wir muss also

$$\boldsymbol{C}\boldsymbol{X} = \boldsymbol{I}_k. \tag{3.10}$$

Die Kovarianzmatrix von $\hat{\hat{\beta}}$ ist gegeben durch

$$\begin{aligned} Var(\hat{\hat{\beta}}|\boldsymbol{X}) &= Var(\boldsymbol{C}\boldsymbol{u}|\boldsymbol{X}) \overset{3.2\&(3.6)}{=} \boldsymbol{C}\{\sigma^2 \boldsymbol{I}_n\}\boldsymbol{C}^\top \\ &= \sigma^2 \boldsymbol{C}\boldsymbol{C}^\top. \end{aligned}$$

Die kleinste Kovarianzmatrix (im Matrixsinn) wird über die Matrix $\boldsymbol{C}$ gefunden, welche (3.10) erfüllt und $\boldsymbol{C}\boldsymbol{C}^\top$ minimiert.

Die Lösung ist einfach: Erinnern wir uns, dass $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$. Dann gilt

$$\boldsymbol{C}\boldsymbol{C}^\top \geq \boldsymbol{C}\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{C}^\top.$$

Um dies zu sehen, müssen wir zeigen, dass

$$\boldsymbol{C}\boldsymbol{C}^\top - \boldsymbol{C}\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{C}^\top = \boldsymbol{C}(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{C}^\top = \boldsymbol{C}\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{C}^\top \geq \boldsymbol{0}, \tag{3.11}$$

das heißt, dass $\boldsymbol{C}\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{C}^\top$ p.s.d. (positiv semi-definit) ist, also dass $\boldsymbol{a}^\top \boldsymbol{C}\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{C}^\top \boldsymbol{a} \geq 0$ für jeden Vektor $\boldsymbol{a}$ gilt.

Da wir einfach $\boldsymbol{d} = \boldsymbol{C}^\top \boldsymbol{a}$ definieren können, ist die Bedingung erfüllt, wenn $\boldsymbol{d}^\top \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{d} \geq 0$, das heißt, wenn $\boldsymbol{M}_{\boldsymbol{X}}$ p.s.d. ist. Wir wissen bereits, dass $\boldsymbol{M}_{\boldsymbol{X}}$ symmetrisch und idempotent ist (vgl. (2.10)).

Daher sagt uns Resultat 27.3, dass alle Eigenwerte 0 oder 1 sind, also nicht negativ. Resultat 27.7 sagt uns dann, dass $\boldsymbol{M}_{\boldsymbol{X}}$ p.s.d. ist.

Wir haben dann

$$\boldsymbol{C}\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{C}^\top = \underbrace{\boldsymbol{C}\boldsymbol{X}}_{\boldsymbol{I}}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\underbrace{\boldsymbol{X}^\top\boldsymbol{C}^\top}_{\boldsymbol{I}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}$$

Wenn (3.10) erfüllt ist, sehen wir, dass die Kovarianzmatrix von $\hat{\tilde{\beta}}$ nicht kleiner sein kann als $\sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$. Daher gilt

$$Var(\hat{\tilde{\beta}}|\boldsymbol{X}) \geq Var(\hat{\beta}_{\text{OLS}}|\boldsymbol{X}).$$

Beachten Sie, dass die Untergrenze durch $\hat{\beta}_{\text{OLS}}$ erreicht wird, wobei

$$\boldsymbol{C} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top,$$

da

$$\begin{aligned}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{X}} \left((\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top\right)^\top &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\end{aligned}$$

Daher bezeichnen wir $\hat{\beta}_{\text{OLS}}$ als **Best Linear Unbiased Estimator**: BLUE.

Dies ist ein starkes Resultat, aber wir müssen uns vor Augen halten, dass es nur unter sehr starken Annahmen erfüllt ist.

Siehe hier für eine Illustration.

Ausschluss und Inklusion von Variablen

Nehmen wir an, dass das wahre Modell (mit $E(u|\boldsymbol{X}) = \mathbf{0}$; definiere $\boldsymbol{X} = (\boldsymbol{X}_1 : \boldsymbol{X}_2)$) gegeben ist durch

$$\boldsymbol{y} = \boldsymbol{X}_1\beta + \boldsymbol{X}_2\gamma + \boldsymbol{u}, \tag{3.12}$$

jedoch führen wir die Schätzung basierend auf

$$\boldsymbol{y} = \boldsymbol{X}_1\beta + \boldsymbol{v} \tag{3.13}$$

durch. Das bedeutet, dass wir β mit

$$\hat{\beta}_{\text{OLS,K}} = (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top \boldsymbol{y},$$

schätzen, während Modell (3.12) das wahre Modell ist. Das heißt, dass wir **relevante Variablen** auslassen (wenn $\gamma \neq \mathbf{0}$).

Die Konsequenz dieses Auslassens ist eine Verzerrung in der Schätzung von β.

$$\begin{aligned} E(\hat{\beta}_{\text{OLS,K}}|\boldsymbol{X}) &= (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top E(\boldsymbol{y}|\boldsymbol{X}) \\ &\overset{(3.12)}{=} (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top E(\boldsymbol{X}_1\beta + \boldsymbol{X}_2\gamma + \boldsymbol{u}|\boldsymbol{X}) \\ &= (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top (\boldsymbol{X}_1\beta + \boldsymbol{X}_2\gamma) \\ &= \beta + (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top \boldsymbol{X}_2\gamma, \end{aligned} \tag{3.14}$$

Dies ist ungleich β, außer wenn $\gamma = \mathbf{0}$ (also wenn die zusätzlichen Regressoren irrelevant sind) oder $\boldsymbol{X}_1^\top \boldsymbol{X}_2 = \mathbf{O}$ (also wenn $\boldsymbol{X}_1$ und $\boldsymbol{X}_2$ orthogonal sind).

Betrachten wir nochmal `unbiasedness.R`.

Nehme jetzt an, dass (3.13) das wahre Modell ist (mit dann $E(\boldsymbol{v}|\boldsymbol{X}) = \mathbf{0}$), wir jedoch auf Basis von Modell (3.12) schätzen.

Dies bedeutet **Inklusion irrelevanter Variablen**, da

$$\gamma = \mathbf{0}.$$

Beide Modelle (3.12) und (3.13) sind also korrekt, jedoch verwendet (3.13) die Information, dass $\gamma = \mathbf{0}$, (3.12) hingegen nicht.

Als Konsequenz folgt daraus, dass der Schätzer (siehe (2.15))

$$\hat{\beta}_{\text{OLS,L}} = \left\{ \boldsymbol{X}_1^\top (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}_2}) \boldsymbol{X}_1 \right\}^{-1} \boldsymbol{X}_1^\top (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}_2}) \boldsymbol{y},$$

unverzerrt ist, da

$$\begin{aligned} E(\hat{\beta}_{\text{OLS,L}} | \boldsymbol{X}) &= \left\{ \boldsymbol{X}_1^\top (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}_2}) \boldsymbol{X}_1 \right\}^{-1} \boldsymbol{X}_1^\top (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}_2}) E(\boldsymbol{X}_1 \beta + \boldsymbol{v} | \boldsymbol{X}) \\ &= \left\{ \boldsymbol{X}_1^\top (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}_2}) \boldsymbol{X}_1 \right\}^{-1} \boldsymbol{X}_1^\top (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}_2}) \boldsymbol{X}_1 \beta = \beta. \end{aligned}$$

Er hat jedoch eine größere Kovarianzmatrix (die man analog zu (3.7) herleitet) als die des OLS-Schätzers von (3.13) (vgl. (3.7)):

$$Var(\hat{\beta}_{\text{OLS,L}} | \boldsymbol{X}) = \sigma^2 \left\{ \boldsymbol{X}_1^\top (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}_2}) \boldsymbol{X}_1 \right\}^{-1} \geq \sigma^2 (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} = Var(\hat{\beta}_{\text{OLS,K}} | \boldsymbol{X}).$$

Dies gilt, da diese Bedingung äquivalent ist zu (siehe z.B. Magnus and Neudecker [1988, Thm. 1.24] oder hier)

$$\frac{1}{\sigma^2}(\boldsymbol{X}_1^\top \boldsymbol{X}_1) \geq \frac{1}{\sigma^2}\left\{\boldsymbol{X}_1^\top (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}_2})\boldsymbol{X}_1\right\}$$

Dies wiederum ist der Fall, wenn

$$\boldsymbol{X}_1^\top \boldsymbol{X}_1 - \left\{\boldsymbol{X}_1^\top (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}_2})\boldsymbol{X}_1\right\} = \boldsymbol{X}_1^\top \boldsymbol{X}_1 - \boldsymbol{X}_1^\top \boldsymbol{X}_1 + \boldsymbol{X}_1^\top \boldsymbol{P}_{\boldsymbol{X}_2}\boldsymbol{X}_1 = \boldsymbol{X}_1^\top \boldsymbol{P}_{\boldsymbol{X}_2}\boldsymbol{X}_1 \geq 0$$

Dass dies gilt, folgt mit dem Argument, das (3.11) gezeigt hat.

Zusätzliche Regressoren reduzieren daher das Risiko des **omitted variable bias**, was (sehr) vorteilhaft ist.

Auf der anderen Seite verringern sie die Präzision des Schätzers (außer im Fall von Orthogonalität $\boldsymbol{P}_{\boldsymbol{X}_2}\boldsymbol{X}_1 = \boldsymbol{O} \Leftrightarrow \boldsymbol{X}_2^\top \boldsymbol{X}_1 = \boldsymbol{O}$), was weniger gut ist.

`long short reg.R`

Was machen wir also?

Wenn alle Stichproben extrem groß wären, erscheint Überspezifizierung naheliegend. Der Bias durch Unterspezifizierung verschwindet nicht mit zunehmender Stichprobengröße, jedoch konvergiert die Varianz „konsistenter“ (s.u., bspw. (4.32)) Schätzer gegen null.

Um dies zu sehen, betrachten wir ein einfaches Beispiel einer Regression auf eine Konstante. Nach Einsetzen von (1.17) und (3.7) erhalten wir:

$$\hat{\beta} = \bar{y}$$

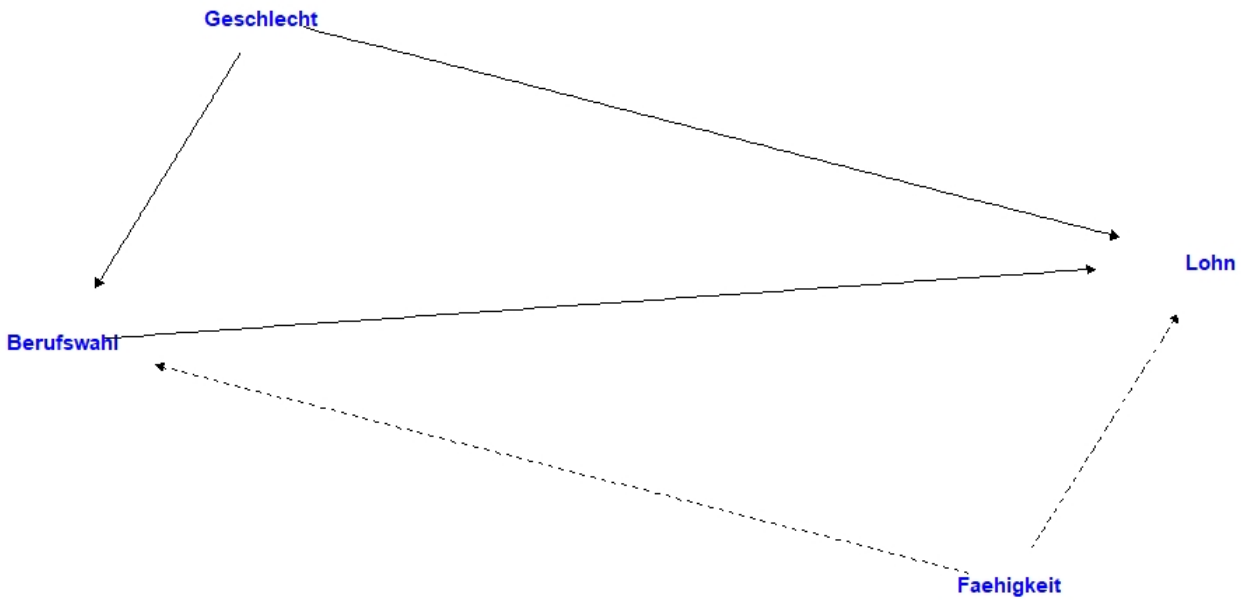
und für eine u.i.v.-Stichprobe mit der Varianz σ^2,

$$Var(\bar{y}) = \frac{\sigma^2}{n} \to 0$$

Auf der anderen Seite mag bei kleinem n der Effizienzverlust durch die Hinzunahme von zu vielen Variablen schwerer wiegen als die Konsequenzen des Omitted Variable Bias.

Bedeutet dies dann, dass man bei großem n alle auffindbaren Regressoren verwenden sollte? Nein!

Beispiel: Berufswahl und Lohndiskriminierung



Siehe hier bzw. hier.

Überblick

Die Idee

Es wäre ein Glückstreffer, wenn $\hat{\beta}_{\text{OLS}}$ dem wahren Vektor β in einer endlichen Stichprobe entspräche. Daher müssen wir die Zufälligkeit von $\hat{\beta}_{\text{OLS}}$ miteinbeziehen, wenn wir Schlussfolgerungen über β treffen wollen.

Im einfachsten Fall: Der Erwartungswert einer Population, aus der eine Zufallsstichprobe gezogen wurde. Nehmen wir den folgenden DGP an

$$y_t = \beta + u_t, \qquad u_t \overset{\text{u.i.v.}}{\sim} (0, \sigma^2)$$

wobei β der Erwartungswert ist. Der kleinste-Quadrate-Schätzer von β und seine Varianz sind gegeben durch (s.o.)

$$\hat{\beta} = \bar{y}$$

und

$$Var(\bar{y}) = \frac{\sigma^2}{n}$$

Wir wollen folgende **Nullhypothese** testen:

$$H_0 : \beta = \beta_{H0}, \tag{4.1}$$

wobei β_{H0} einen spezifischen Wert von β darstellt.

Wir brauchen eine **Teststatistik** z, die eine Zufallsvariable (ZV) ist, die (zumindest asymptotisch) eine bekannte Verteilung hat, wenn H_0 zutrifft.

Wenn der realisierte Wert von z nicht „ungewöhnlich“ ist, wenn H_0 zutrifft, dann liefert der Test keine Evidenz gegen H_0.

Wenn der Wert jedoch nur selten unter H_0 (ökonometrische Abkürzung für „falls die Nullhypothese wahr ist“) erreicht würde, ist dies ein Anhaltspunkt gegen H_0.

Wir entscheiden uns dann, H_0 **abzulehnen**.

Der Einfachheit halber treffen wir zunächst zwei sehr starke Annahmen:

1. Die u_t aus unserem Regressionsmodell sind normalverteilt.
2. Wir wissen, dass ihre Varianz σ^2 ist.

Unter diesen Annahmen können wir folgende Teststatistik benutzen

$$z = \frac{\hat{\beta} - \beta_{H0}}{\left(Var(\hat{\beta})\right)^{1/2}} = \frac{n^{1/2}}{\sigma}(\hat{\beta} - \beta_{H0}).$$

Unter der Nullhypothese (4.1), $\beta = \beta_{H0}$. Des Weiteren gilt $E(\hat{\beta}) = E(\bar{y}) = \beta$, so dass

$$E(z) = E\left(\frac{n^{1/2}}{\sigma}(\hat{\beta} - \beta_{H0})\right) = \frac{n^{1/2}}{\sigma}[E(\hat{\beta}) - \beta_{H0}] = \frac{n^{1/2}}{\sigma}[\beta - \beta] = 0$$

und

$$Var(z) = Var\left(\frac{n^{1/2}}{\sigma}\hat{\beta}\right) = \frac{n}{\sigma^2}\frac{\sigma^2}{n} = 1$$

Daraus folgt $z \sim \mathcal{N}(0,1)$—z hat einen Erwartungswert von 0, eine Varianz von 1 und ist eine lineare Kombination von normalverteilten Zufallsvariablen, die ebenfalls normalverteilt ist. Daher hat z unter H_0 eine bekannte Verteilung.

Es gibt außerdem (zumindest implizit) eine **Alternativhypothese**, meist als H_1 bezeichnet. Wir testen H_0 gegen H_1. Nehmen Sie an, dass

$$H_1 : \beta \neq \beta_{H0}.$$

Unter H_1 ist z nicht $\mathcal{N}(0,1)$-verteilt. Wenn $\beta = \beta_1$, dann gilt $E(z) \neq 0$ und kann daher nicht $\mathcal{N}(0,1)$-verteilt sein.

Es ist nicht schwierig zu zeigen, dass

$$z \sim \mathcal{N}\left(\frac{n^{1/2}}{\sigma}(\beta_1 - \beta_{H0}), 1\right) \tag{4.2}$$

Wenn n groß oder die Differenz $(\beta_1 - \beta_{H0})$ absolut groß ist, ist auch $E(z)$ absolut groß. Wir lehnen H_0 ab, wenn z „ausreichend“ (siehe unten) weit entfernt von 0 ist—egal in welche Richtung: Da wir gegen

$$H_1 : \beta \neq \beta_{H0}$$

testen, führen wir einen **zweiseitigen Test** durch und verwerfen H_0, wenn $|z|$ groß ist.

Wenn wir gegen $H_1 : \beta > \beta_{H0}$ testen wollen, führen wir einen **einseitigen Test** durch und verwerfen H_0, wenn z ausreichend groß (im Sinne von positiv) ist.

Da z jeden Wert annehmen kann, auch unter H_0, ist kein Wert z absolut inkompatibel mit H_0. Daher können wir uns nie sicher sein, dass H_0 falsch ist. Wir brauchen also eine **Ablehnungsregel**, die ablehnt, wenn z in eine bestimmte Ablehnungsregion fällt.

Wenn wir eine wahre H_0 verwerfen, machen wir einen **Fehler erster Art**. Die Wahrscheinlichkeit eines solchen Fehlers ist demnach die Wahrscheinlichkeit unter H_0, dass z in den Ablehnungsbereich fällt. Diese Wahrscheinlichkeit nennen wir **Signifikanzniveau**, bezeichnet als α. Wir *konstruieren* Ablehnungsregionen so, dass α kleine Werte wie .05 oder .01 annimmt.

Die Wahrscheinlichkeit, dass ein Test eine falsche H_0 ablehnt, heißt **Macht**. Die Macht ist abhängig von der Datengenerierung und von n. So steigt die Macht mit der Distanz $|\beta_1 - \beta_{H0}|$ in (4.2).

Eine falsche H_0 fälschlicherweise nicht abzulehnen wird als **Fehler zweiter Art** bezeichnet. Die Wahrscheinlichkeit dieses Fehlers ist 1 – Macht.

Um eine Ablehnungsregion zum Niveau α zu konstruieren, berechnen wir zunächst den zu α gehörigen **kritischen Wert** c_α. Für eine $\mathcal{N}(0,1)$ Teststatistik eines zweiseitigen Tests unter H_0 ist der kritische Wert implizit gegeben durch

$$1 - \alpha/2 = \Phi(c_\alpha) \tag{4.3}$$

wobei Φ die Verteilungsfunktion (engl. CDF) der $\mathcal{N}(0,1)$-Verteilung (1.4) bezeichnet. Daher ist

$$\Phi^{-1}(1 - \alpha/2) = c_\alpha,$$

wobei Φ^{-1} die **Quantilsfunktion** bezeichnet.

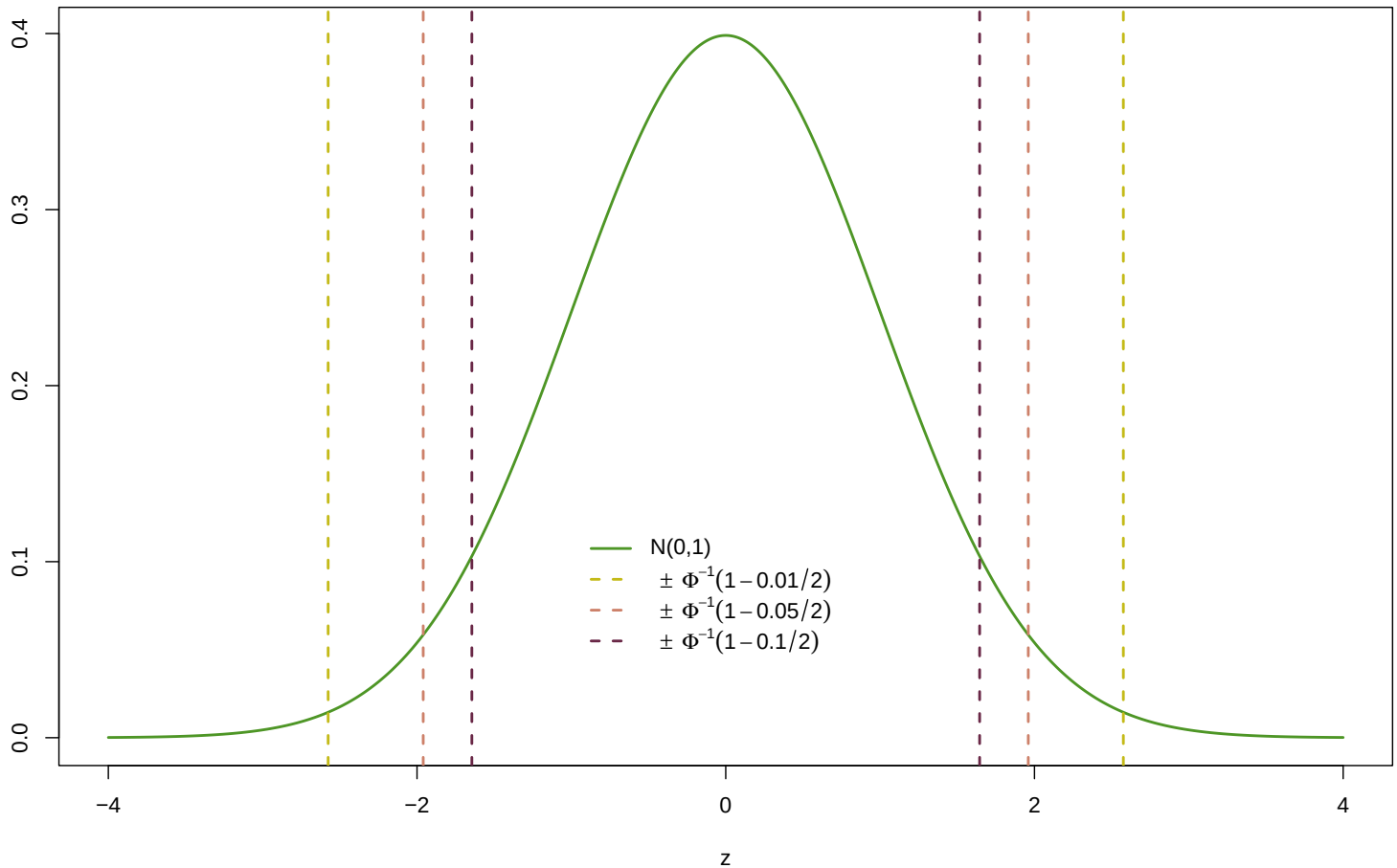
Die Wahrscheinlichkeit, dass $z > c_\alpha$, ist $1 - (1 - \alpha/2) = \alpha/2$ und durch Symmetrie auch $P(z < -c_\alpha) = \alpha/2$. Daher ist

$$P(|z| > c_\alpha) = \alpha,$$

wie gewünscht. Bei $\alpha = 0.05$ erhalten wir beispielsweise

$$\Phi^{-1}(1 - \alpha/2) = 1.96.$$

Graphisch können wir die jeweiligen Elemente auch von der Normalverteilungsdichte (1.3) ablesen.

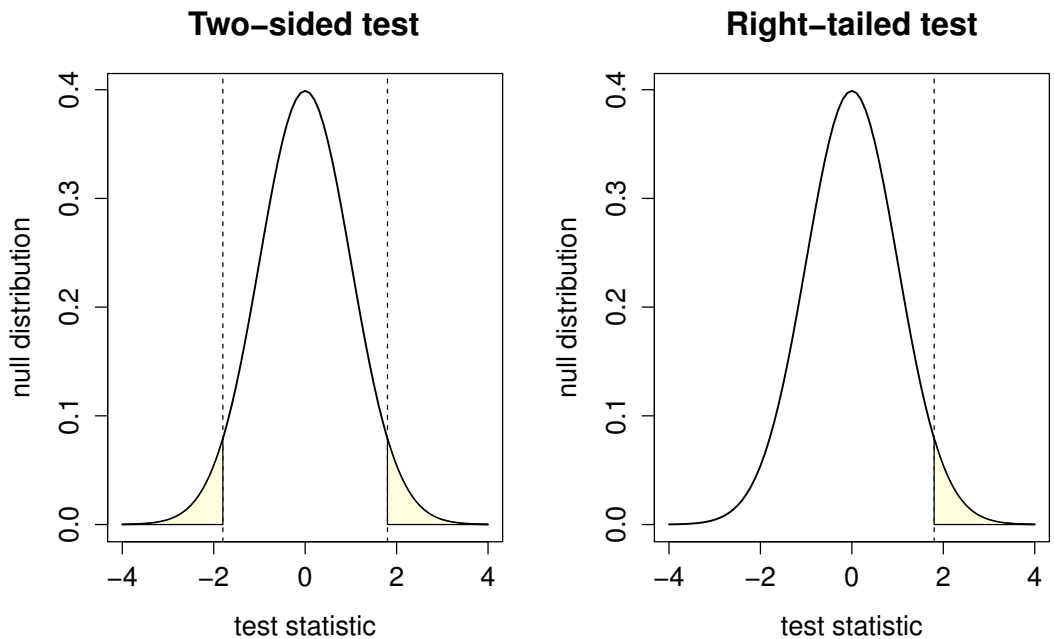


Das Ergebnis eines Tests ist Ja oder Nein: Ablehnen oder nicht ablehnen.

Ein ausgefeilterer Ansatz ist es den ***p*-Wert**, auch bezeichnet als das marginale Signifikanzniveau, der beobachteten Teststatistik $\hat{z}$ zu berechnen.

Der p-Wert ist definiert als das kleinste Niveau, zu dem ein Test basierend auf $\hat{z}$ gerade noch abgelehnt wird. Alternativ ist er die Wahrscheinlichkeit unter H_0 genau so oder extremere Statistiken als die beobachtete zu erhalten.

Siehe `pValues.R`.



Für einen zweiseitigen Test in unserem vorherigen Beispiel gilt

$$p(\hat{z}) = 2(1 - \Phi(|\hat{z}|))$$

Um dies zu sehen, erinnern wir uns, dass der Test basierend auf $\hat{z}$ ablehnt, wenn

$$|\hat{z}| > c_\alpha$$

Dies ist äquivalent zu

$$\Phi(|\hat{z}|) > \Phi(c_\alpha),$$

da Φ strikt steigt.

Weiter gilt mit (4.3)

$$\Phi(c_\alpha) = 1 - \alpha/2$$

Der kleinste Wert α, für den obige Ungleichung erfüllt ist, wird über die Lösung der folgenden Gleichung nach α ermittelt:

$$\Phi(|\hat{z}|) = 1 - \alpha/2$$

Wir erhalten $2(1 - \Phi(|\hat{z}|))$.

Der *p*-Wert präsentiert die Information direkter. Beispielsweise führen Teststatistiken 2.02 und 5.77 beide zur Ablehnung bei $\alpha = 0.05$. 5.77 liefert jedoch eine „deutlichere“ Ablehnung der Nullhypothese als 2.02.

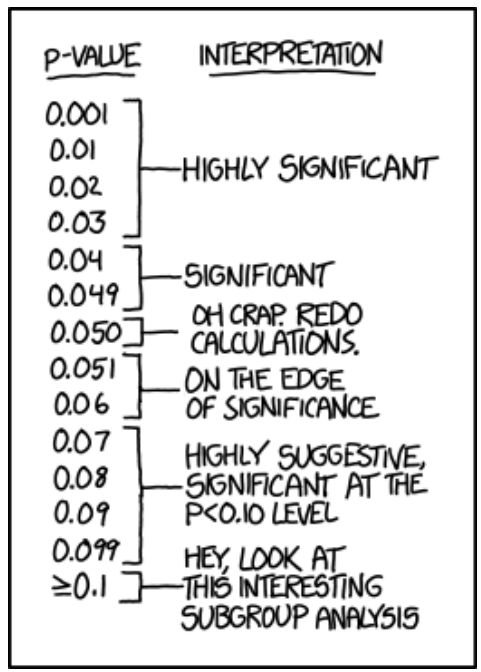
p-Werte verdeutlichen das Ausmaß der Differenz. Der *p*-Wert zu 2.02 ist .0434, während der *p*-Wert zu 5.77 gleich 7.93×10^{-9} und damit deutlich kleiner ist.

Dennoch sollte der *p*-value nicht überschätzt werden! Siehe hierfür hier und hier oder detaillierter hier.

Oder wie andere gesagt haben:

> *„... surely, God loves the .06 nearly as much as the .05. Can there be any doubt that God views the strength of evidence for or against the null as a fairly continuous function of the magnitude of p?“* —Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. American Psychologist, 44(10), 1276-1284.

Ein verwandtes Problem:



Nun werden wir die starken Annahmen, dass Fehlerterme normalverteilt sind, und dass die Varianz der Fehlerterme bekannt ist, lockern. Zunächst beschäftigen wir uns mit der letzteren Annahme und kommen dann später auf erstere zurück.

Zusätzlich haben wir unsere Aufmerksamkeit auf eine einzige Restriktion bezüglich eines einzigen Parameters beschränkt.

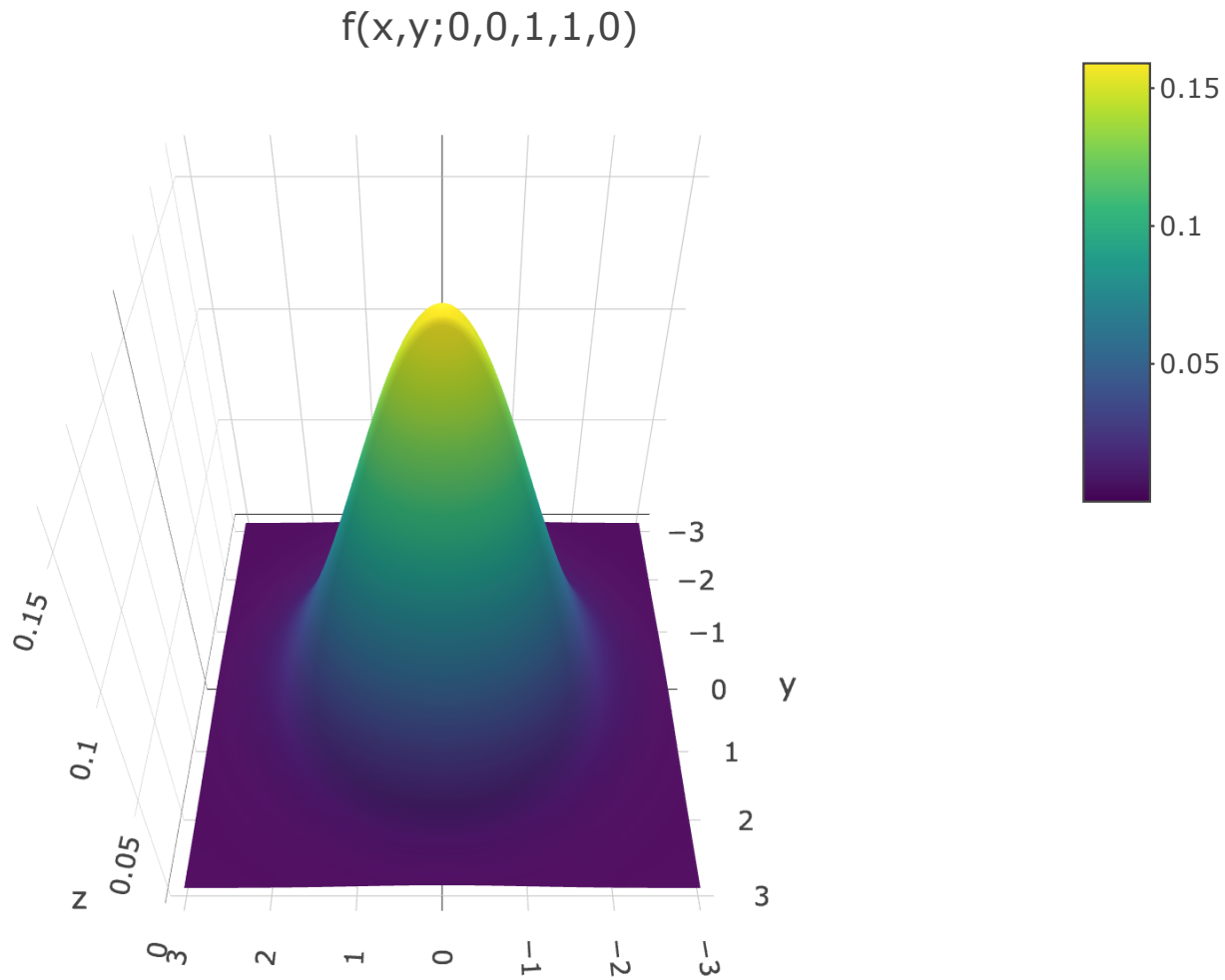
Wir werden uns einige Verteilungen ansehen müssen.

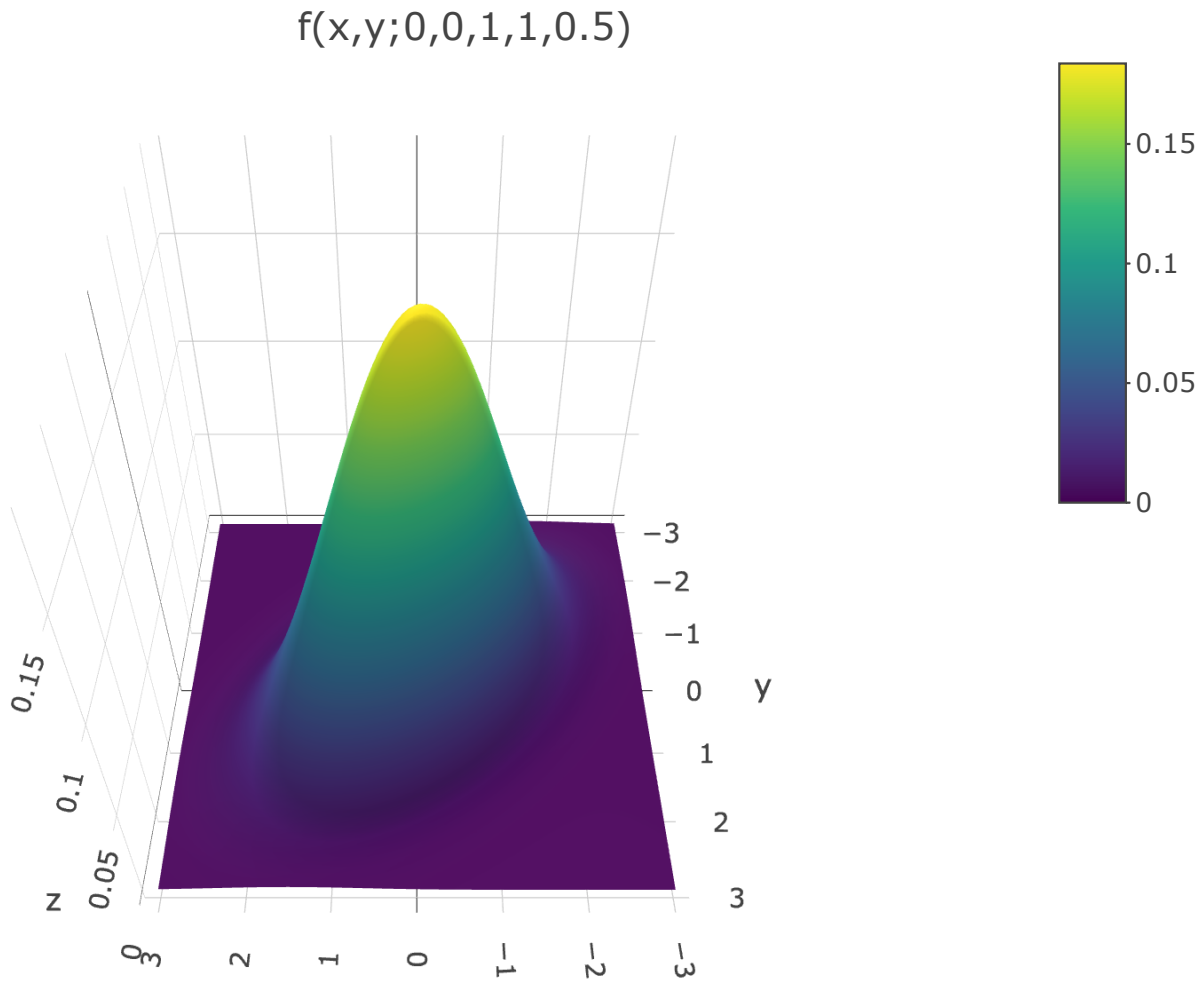
Die bivariate Normalverteilung

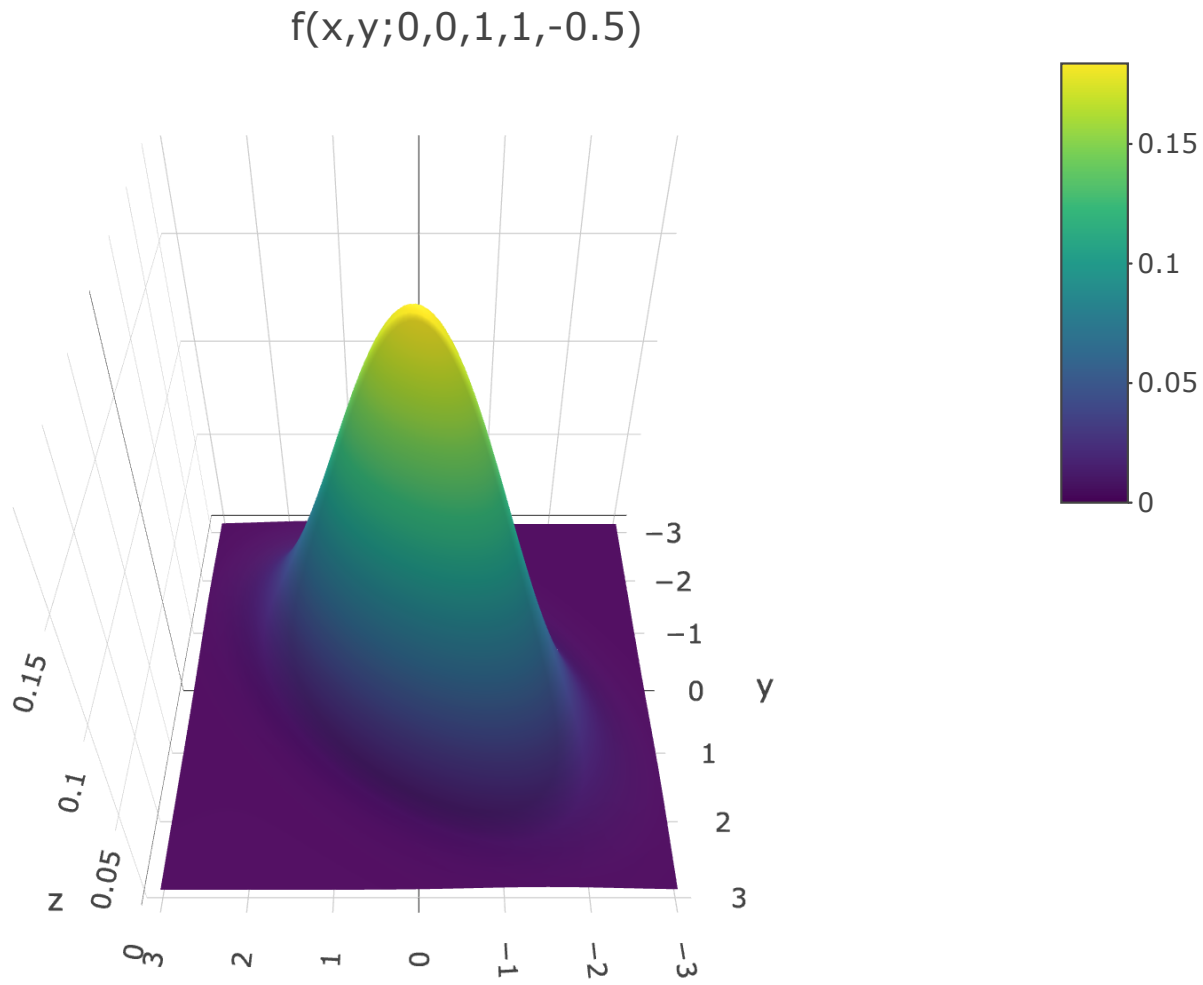
Definition 4.1 Ein Paar von Zufallsvariablen X und Y hat die **bivariate Normalverteilung**, wenn deren Dichte (1.5) gegeben ist durch

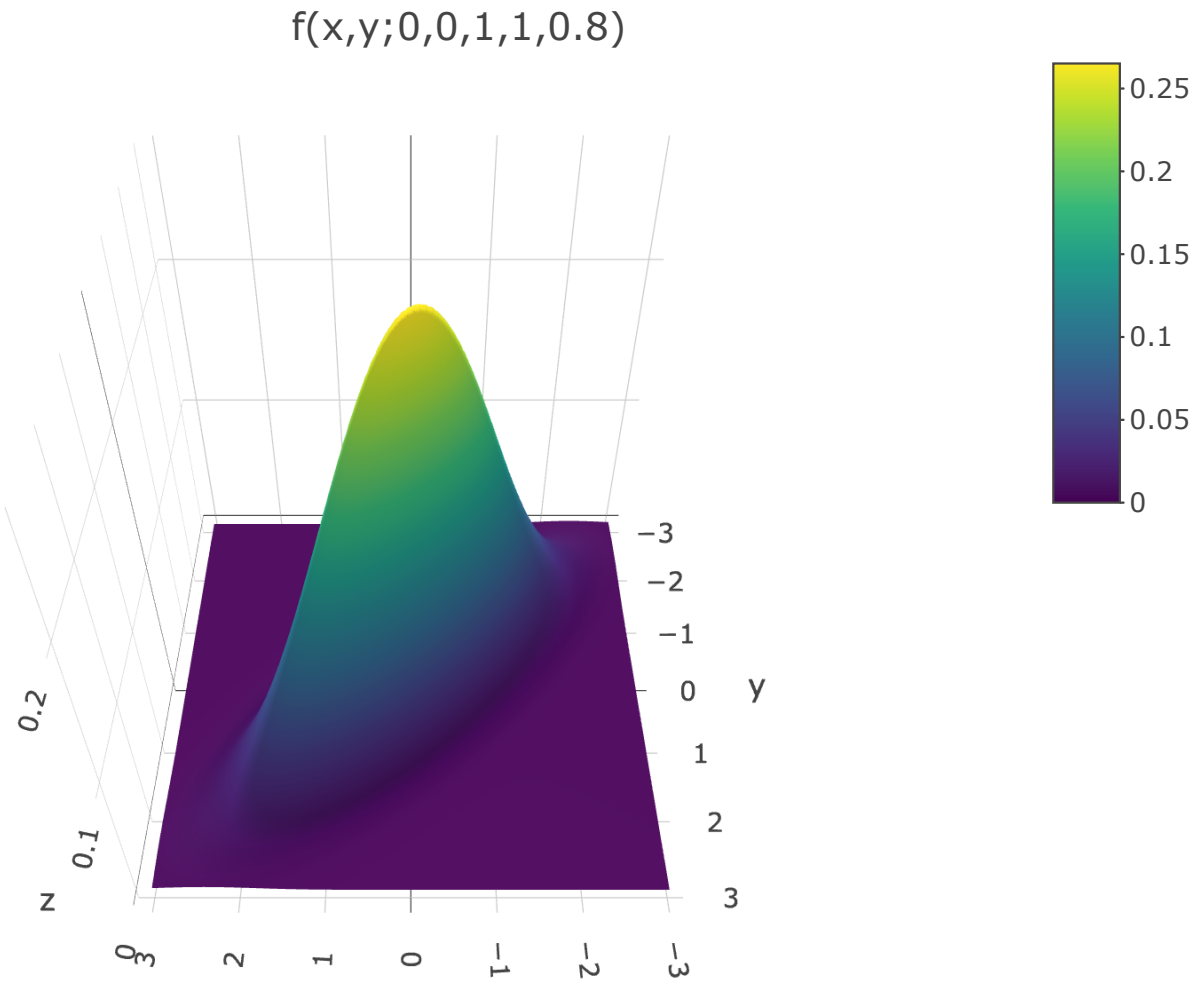
$$
\begin{aligned}
f(x,y;\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho) \quad = \quad & \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\Bigg\{ -\frac{1}{2(1-\rho^2)} \times \\
& \Bigg[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \left(\frac{y-\mu_2}{\sigma_2}\right)^2 \\
& -2\rho \left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) \Bigg] \Bigg\}
\end{aligned}
$$

mit $\sigma_1, \sigma_2 > 0$, $-1 < \rho < 1$ und $-\infty < x, y < \infty$.


f(x,y;0,0,1,1,0)
0.15
0.1
0.05
z
0.15
0.1
0.05
3
2
1
0
−1
−2
−3
x
−3
−2
−1
0
1
2
3
y


f(x,y;0,0,1,1,0.5)
0.15
0.1
0.05
0
z
0.15
0.1
0.05
0
3
2
1
0
−1
−2
−3
y
−3
−2
−1
0
1
2
3


f(x,y;0,0,1,1,-0.5)
0.15
0.1
0.05
0
z
0.15
0.1
0.05
0
3
2
1
0
−1
−2
−3
y
−3
−2
−1
0
1
2
3


f(x,y;0,0,1,1,0.8)
0.25
0.2
0.15
0.1
0.05
0
y
−3
−2
−1
0
1
2
3
z
0.2
0.1
0
3
2
1
0
−1
−2
−3

Der Parameter ρ ist der **Korrelationskoeffizient** zwischen X und Y, das heißt

$$\rho = \frac{Cov(X, Y)}{\sigma_1 \sigma_2}$$

Die Notation μ_1, μ_2 genauso wie σ_1^2, σ_2^2 deutet marginale Erwartungswerte und Varianzen von X and Y an. Dies ist auch der Fall. Herausintegrieren von Y wie in (1.6) (was ein bisschen umständlich ist), d.h.

$$\int_{-\infty}^{\infty} f(x, y; \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)\, \mathrm{d}y,$$

liefert die marginale Dichte von X,

$$g(x) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\Big\{-\frac{1}{2\sigma_1^2}(x - \mu_1)^2\Big\}, \tag{4.4}$$

aus der wir ablesen können, dass $E(X) = \mu_1$ und $Var(X) = \sigma_1^2$. Durch Symmetrie gilt das Gleiche für Y.

Hier finden Sie eine Heatmap mit einer bivariaten Normalverteilung: `BivariateNormal.R`. Siehe auch `Regression to the mean.R` für eine Anwendung.

Diese Ergebnisse können zur **multivariaten Normalverteilung** generalisiert werden.

Erinnern wir uns an (3.5) und betrachten zunächst einen Vektor aus Zufallsvariablen $\boldsymbol{z} = (z_1, \ldots, z_m)^\top$, wobei jedes $z_i \sim \mathcal{N}(0, 1)$ unabhängig von den anderen z_j sei. Dies bedeutet, dass $Var(\boldsymbol{z}) = \boldsymbol{I}$. Transformieren wir $\boldsymbol{z}$ mit der $m \times m$ Matrix $\boldsymbol{A}$ mit vollem Rang

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{z},$$

erhalten wir

$$E(\boldsymbol{x}) = \boldsymbol{A}E(\boldsymbol{z}) = \boldsymbol{0}$$

Des Weiteren gilt

$$\begin{aligned} Var(\boldsymbol{x}) &\overset{(3.6)}{=} \boldsymbol{A}E\left(\boldsymbol{z}\boldsymbol{z}^\top\right)\boldsymbol{A}^\top \\ &= \boldsymbol{A}Var(\boldsymbol{z})\boldsymbol{A}^\top \\ &= \boldsymbol{A}\boldsymbol{A}^\top \\ &=: \boldsymbol{\Omega} \end{aligned}$$

Studentsche t-, F- and Chi-Quadrat-Verteilungen

Verwandt zu der Normalverteilung haben wir:

- Die Chi-Quadrat-Verteilung:
  Wenn $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, dann

$$\boldsymbol{z}^\top \boldsymbol{z} \sim \chi^2_n, \tag{4.5}$$

- Die Studentsche t-Verteilung:
  Wenn $z \sim \mathcal{N}(0,1)$ und davon unabhängig $y \sim \chi^2_n$, dann

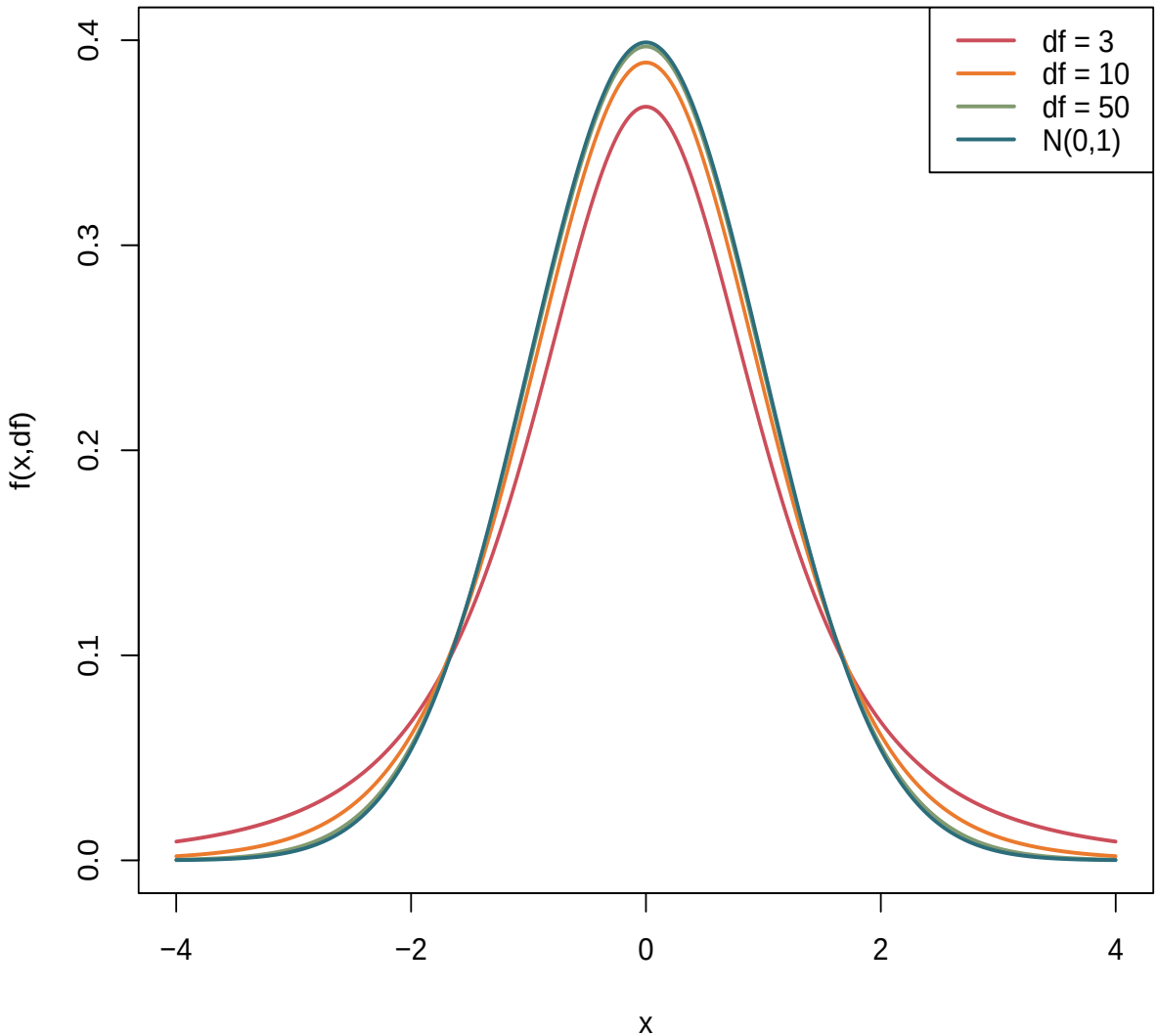
$$\frac{z}{\sqrt{y/n}} \sim t_n, \tag{4.6}$$

- (Fishers) F-Verteilung
  Wenn $y_1 \sim \chi^2_{n_1}$ und unabhängig, $y_2 \sim \chi^2_{n_2}$, dann

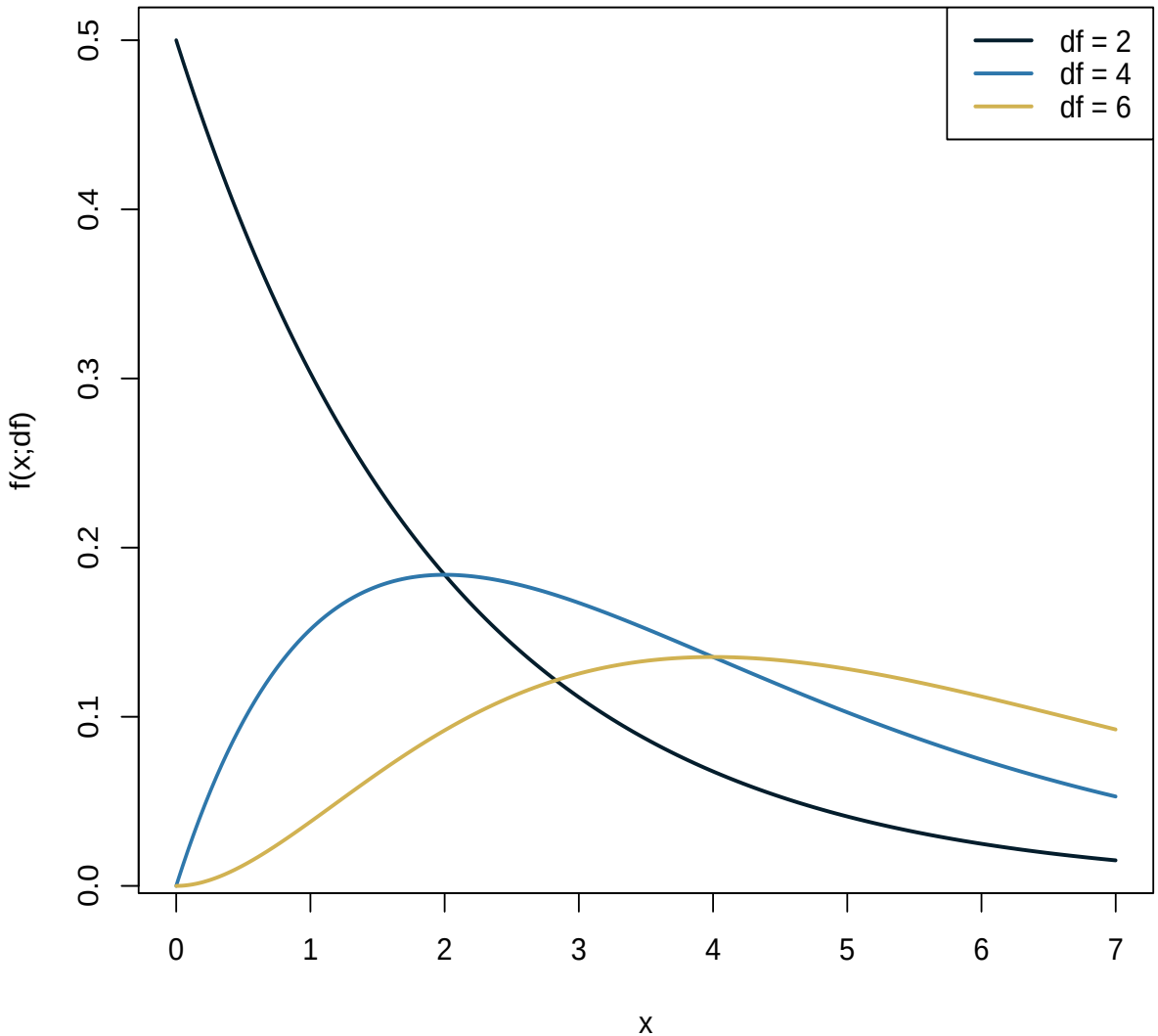
$$\frac{y_1/n_1}{y_2/n_2} \sim F_{n_1,n_2}. \tag{4.7}$$

Hier sind ein paar Bilder.

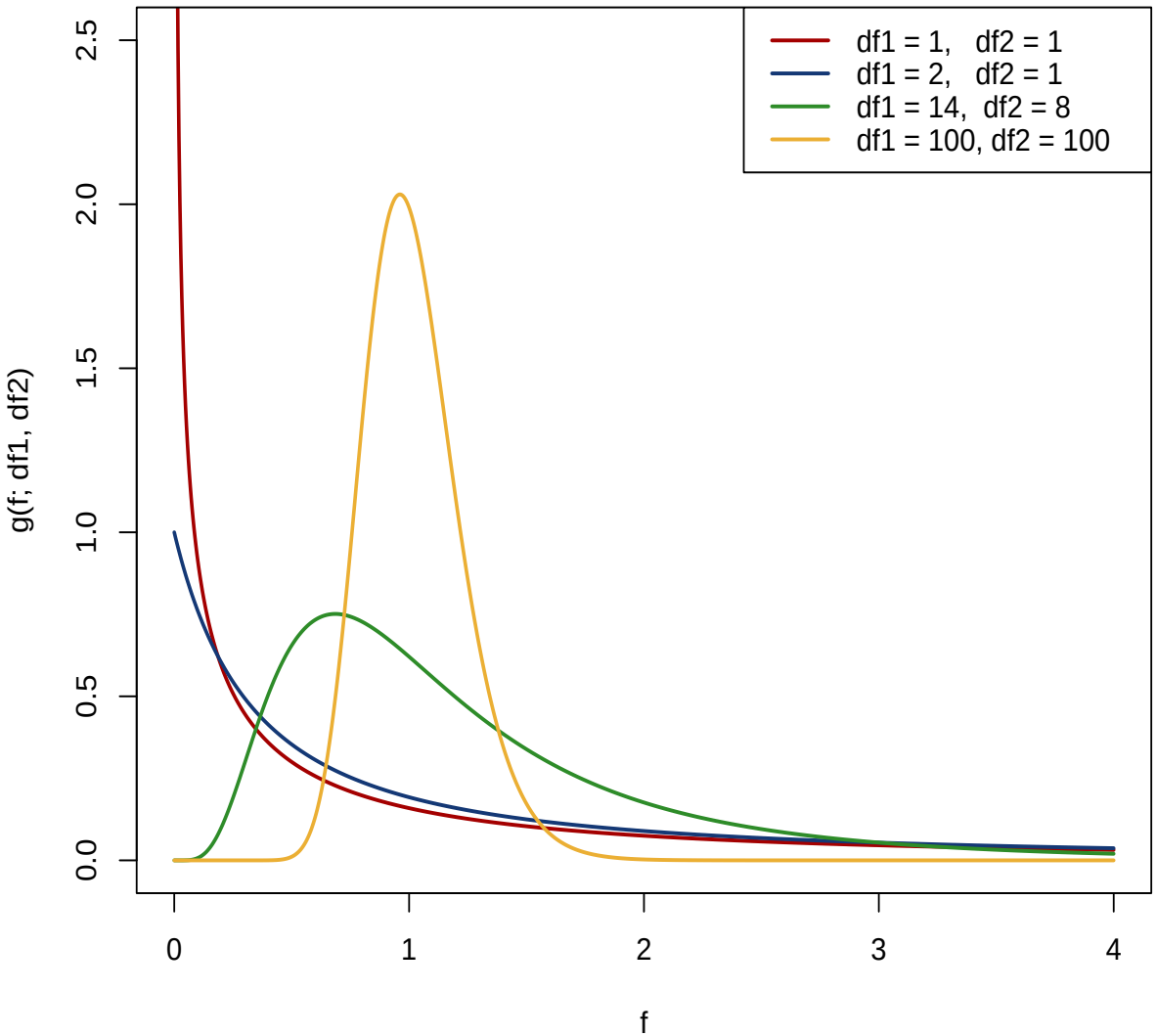
Die t-Verteilung:



Die χ^2-Verteilung:



Die F-Verteilung:



Wir betrachten weiter das lineare Modell (3.1)

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{u}.$$

Wir erweitern nun Annahme 3.2 zu

$$\boldsymbol{u}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n). \tag{4.8}$$

In diesem Fall überzeugen wir uns, dass

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{X} &\sim \mathcal{N}(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I}_n), \\ \hat{\beta}_{\text{OLS}}|\boldsymbol{X} &\sim \mathcal{N}\left(\beta, \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\right). \end{aligned} \tag{4.9}$$

Um Tests ohne die Annahme, dass σ^2 bekannt ist, entwickeln zu können, brauchen wir

Resultat 4.2 Die Zufallsvektoren

$$(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{u} \text{ und } \boldsymbol{X}^\top \boldsymbol{u} \tag{4.10}$$

sind bedingt auf $\boldsymbol{X}$ unabhängig und gemeinsam normalverteilt.

Beweis von Resultat 4.2

$$\begin{pmatrix} \boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}} \\ \boldsymbol{X}^\top \end{pmatrix} \boldsymbol{u}|\boldsymbol{X} \tag{4.11}$$

hat eine multivariate Normalverteilung, da die Elemente lineare Funktionen von $\boldsymbol{u}$ sind. Der Erwartungswert ist null und die Kovarianzmatrix block-diagonal, da

$$E\big((\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{u}\boldsymbol{u}^\top \boldsymbol{X}|\boldsymbol{X}\big) = \sigma^2(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{X} \overset{(2.6)}{=} \mathbf{O}. \tag{4.12}$$

Die zwei Terme in (4.10) sind unabhängig, da sie eine multivariate Normalverteilung und eine Korrelation von null haben. q.e.d.

Erinnern wir uns, dass

$$\boldsymbol{y} - \boldsymbol{X}\hat{\beta}_{\text{OLS}} = (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y} \overset{(2.6)}{=} (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{u}.$$

Daraus folgt, dass

$$\hat{\sigma}^2 = \frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}_{\text{OLS}})^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}_{\text{OLS}})}{n-k}$$

aus (3.8) eine Funktion von $(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{u}$ ist.

Des Weiteren ist $\hat{\boldsymbol{\beta}}_{\text{OLS}} = \boldsymbol{\beta} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u}$ eine Funktion von $\boldsymbol{X}^\top \boldsymbol{u}$. Daher sind $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ und $\hat{\sigma}^2$ unabhängig. Außerdem ist

$$(n-k)\frac{\hat{\sigma}^2}{\sigma^2}\Big| \boldsymbol{X} \sim \chi^2_{n-k} \tag{4.13}$$

Warum? Beachten Sie, dass durch Symmetrie und Idempotenz gilt

$$(n-k)\hat{\sigma}^2 = \boldsymbol{u}^\top (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})^\top (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{u} = \boldsymbol{u}^\top (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{u}$$

Dividieren durch σ^2 liefert

$$(n-k)\frac{\hat{\sigma}^2}{\sigma^2} = \frac{\boldsymbol{u}}{\sigma}^\top (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\frac{\boldsymbol{u}}{\sigma},$$

wobei (4.8) impliziert, dass $\frac{\boldsymbol{u}}{\sigma}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.

Dies ist eine quadratische Form (siehe Definition 14 im Matrixalgebrareader) von Standardnormal-Zufallsvektoren in einer idempotenten Matrix des Rangs $n - k$, siehe (3.9) und Resultat 26.2.

Warum ist das interessant? Theorem 4.1 des Buches sagt uns, dass eine quadratische Form in einem standardnormalverteilten n-Vektor und einer idempotenten Matrix $\boldsymbol{J}$ des Rangs m χ^2_m ist. Daher gilt (4.13).

Ein alternativer Beweis nutzt die Spektralzerlegung (Resultat 29):

$$\boldsymbol{x}^\top \boldsymbol{J}\boldsymbol{x} = \boldsymbol{x}^\top \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top \boldsymbol{x}$$

Wir wissen, dass die Spur die Summe der Eigenwerte ist (Resultat 27.8), welche entweder eins oder null (Resultat 27.3) sind.

Daher hat $\boldsymbol{\Lambda}$ m Eigenwerte, die gleich 1 sind und $n - m$, die gleich 0 sind. Da $\boldsymbol{U}$ orthogonal ist, gilt

$$E\left(\boldsymbol{U}^\top \boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{U}\right) = \boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}$$

so dass auch

$$\boldsymbol{z} := \boldsymbol{U}^\top \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}).$$

Dann gilt

$$\boldsymbol{z}^\top \boldsymbol{\Lambda}\boldsymbol{z} = \sum_{i=1}^{m} z_i^2 \sim \chi_m^2$$

Wir erhalten insgesamt

(i) (4.9): $\hat{\beta}_{\text{OLS}} - \beta | \boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1})$,

(ii) (4.13): $(n-k)\frac{\hat{\sigma}^2}{\sigma^2} \Big| \boldsymbol{X} \sim \chi^2_{n-k}$,

(iii) Resultat 4.2: die Terme (i) und (ii) sind unabhängig.

Diese Bausteine ermöglichen es uns t-, F- and Chi-Quadrat-Teststatistiken zu konstruieren.

t-Tests

Betrachten wir eine einzige lineare Restriktion

$$H_0 : \boldsymbol{a}^\top \beta = r \tag{4.14}$$

wobei $\boldsymbol{a} : k \times 1$ ein bekannter Vektor und r ein bekanntes Skalar ist.

Der häufigste Fall einer Nullhypothese ist $H_0 : \beta_i = 0$, was dem Standardoutput der meisten Regressionspakete entspricht.

Wir haben

$$\boldsymbol{a}^\top(\hat{\beta}_{\text{OLS}} - \beta)|\boldsymbol{X} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{a}^\top(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{a}).$$

Daher ist

$$\frac{\boldsymbol{a}^\top(\hat{\beta}_{\text{OLS}} - \beta)}{\sqrt{\sigma^2 \boldsymbol{a}^\top(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{a}}} \sim \mathcal{N}(0, 1).$$

Dieser Ausdruck ist (Resultat 4.2) unabhängig von

$$(n-k)\frac{\hat{\sigma}^2}{\sigma^2}\,\Big|\,\boldsymbol{X} \sim \chi^2_{n-k}.$$

Daher gilt

$$\underbrace{\frac{\boldsymbol{a}^\top(\hat{\boldsymbol{\beta}}_{\text{OLS}}-\boldsymbol{\beta})}{\sqrt{\sigma^2\boldsymbol{a}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{a}}}}_{\sim\mathcal{N}(0,1)} \Bigg/ \underbrace{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}}}_{\overset{\text{bed. auf } \boldsymbol{X}}{\sim}\ \sqrt{\chi^2_{n-k}/(n-k)}} = \frac{\boldsymbol{a}^\top(\hat{\boldsymbol{\beta}}_{\text{OLS}}-\boldsymbol{\beta})}{\sqrt{\hat{\sigma}^2\boldsymbol{a}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{a}}}$$

$$\overset{(4.6)}{\sim}\ t_{n-k}. \tag{4.15}$$

Mit (4.14), d.h. $H_0 : \boldsymbol{a}^\top\boldsymbol{\beta} = r$, haben wir eine Teststatistik (d.h. etwas, das nicht von etwas Unbekannten abhängt)

$$t = \frac{\boldsymbol{a}^\top\hat{\boldsymbol{\beta}}_{\text{OLS}} - r}{\sqrt{\hat{\sigma}^2\boldsymbol{a}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{a}}}. \tag{4.16}$$

t hat unter der Nullhypothese eine t_{n-k}-Verteilung.

t-Werte werden standardmäßig für $H_0 : \beta_i = 0$ ausgegeben. Diesen Fall erhalten wir aus (4.16) für $\boldsymbol{a} = \boldsymbol{e}_i$ (einen Vektor mit einer 1 an der Position i und sonst null) und $r = 0$. Daher ist der t-Wert für $H_0 : \beta_i = 0$ gegeben durch

$$t = \frac{\hat{\beta}_i - 0}{\text{se}(\hat{\beta}_i)} = \frac{\hat{\beta}_i}{\text{se}(\hat{\beta}_i)}, \tag{4.17}$$

wobei der Standardfehler von $\hat{\beta}_i$ gegeben ist durch

$$\text{se}(\hat{\beta}_i) = \sqrt{\widehat{Var}(\hat{\beta}_i)} = \sqrt{\hat{\sigma}^2 \boldsymbol{e}_i^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{e}_i} = \sqrt{\hat{\sigma}^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}_{ii}}.$$

Siehe `Testing.R`.

F-Test

Oft wollen wir mehrere Restriktionen auf einmal testen.

Sei $H_0 : \boldsymbol{R}^\top \beta = \boldsymbol{r}$ mit $\boldsymbol{R}$ und $\boldsymbol{r}$ bekannt, nicht zufällig und $\boldsymbol{R} : k \times q$ hat vollen Spaltenrang q. Hiermit werden q lineare Restriktionen abgebildet. Ein Beispiel findet sich unten in (4.22).

Aus (4.9) und (3.6) erhalten wir

$$\boldsymbol{R}^\top(\hat{\beta}_{\text{OLS}} - \beta)|\boldsymbol{X} \sim \mathcal{N}\left(\boldsymbol{0}, \sigma^2 \boldsymbol{R}^\top(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{R}\right)$$

Definiere

$$\boldsymbol{B} = \boldsymbol{R}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{R}$$

und sei $\boldsymbol{B}^{-1/2}$ eine „Matrixwurzel“ von $\boldsymbol{B}^{-1}$ (siehe z.B. Resultat 30). Für

$$\boldsymbol{g} := \frac{\boldsymbol{B}^{-1/2}}{\sigma} \boldsymbol{R}^\top (\hat{\beta}_{\text{OLS}} - \beta)$$

gilt dann, dass

$$\boldsymbol{g} | \boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_q), \tag{4.18}$$

da

$$\begin{aligned} Var(\boldsymbol{g}|\boldsymbol{X}) &= \frac{\boldsymbol{B}^{-1/2}}{\sigma} \boldsymbol{R}^\top Var\left(\hat{\beta}_{\text{OLS}}|\boldsymbol{X}\right) \boldsymbol{R} \frac{\boldsymbol{B}^{-1/2}}{\sigma} \\ &\overset{(3.7)}{=} \frac{\boldsymbol{B}^{-1/2}}{\sigma} \sigma^2 \boldsymbol{B} \frac{\boldsymbol{B}^{-1/2}}{\sigma} = \boldsymbol{I} \end{aligned} \tag{4.19}$$

Dies ist unabhängig von

$$d := (n-k)\frac{\hat{\sigma}^2}{\sigma^2}, \qquad d|\boldsymbol{X} \overset{(4.13)}{\sim} \chi^2_{n-k}.$$

Daher gilt

$$\frac{\overbrace{\boldsymbol{g}^\top \boldsymbol{g}}^{\sim\chi^2_q, \text{vgl. } (4.5)} /q}{d/(n-k)} = \frac{(\hat{\beta}_{\text{OLS}} - \beta)^\top \boldsymbol{R} \left\{\boldsymbol{R}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{R}\right\}^{-1} \boldsymbol{R}^\top (\hat{\beta}_{\text{OLS}} - \beta)/q}{\hat{\sigma}^2} \sim F_{q,n-k}.$$

D.h., unter $H_0 : \boldsymbol{R}^\top \beta = \boldsymbol{r}$ erhalten wir

$$F = \frac{(\boldsymbol{R}^\top \hat{\beta}_{\text{OLS}} - \boldsymbol{r})^\top \left\{\boldsymbol{R}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{R}\right\}^{-1} (\boldsymbol{R}^\top \hat{\beta}_{\text{OLS}} - \boldsymbol{r})/q}{\hat{\sigma}^2} \sim F_{q,n-k}. \qquad (4.20)$$

Zur Illustration betrachten wir den Spezialfall $\boldsymbol{R}^\top = \boldsymbol{I}$, $\boldsymbol{r} = \boldsymbol{0}$, $q = 2$, $\hat{\sigma}^2 = 1$ und $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{I}$. Dann gilt

$$F = \hat{\boldsymbol{\beta}}_{\text{OLS}}^\top \hat{\boldsymbol{\beta}}_{\text{OLS}}/2 = \frac{\hat{\beta}_{\text{OLS},1}^2 + \hat{\beta}_{\text{OLS},2}^2}{2}, \tag{4.21}$$

dies ist die quadrierte euklidische Distanz zwischen dem OLS-Schätzer und dem Ursprung, standardisiert durch die Anzahl der Elemente.

Der F-Test kann in speziellen Fällen eine aufschlussreichere Form annehmen. Betrachten wir den Test

$$H_0 : \beta_2 = \mathbf{0} \tag{4.22}$$

in

$$\boldsymbol{y} = \boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2 + \boldsymbol{u} \tag{4.23}$$

Wir zeigen nun, dass die F-Statistik sich auch durch den intuitiven Ausdruck

$$F_{\beta_2} = \frac{(\text{RSSR} - \text{USSR})/r}{\text{USSR}/(n-k)} \tag{4.24}$$

darstellen lässt. Hierbei ist

- r die Spaltendimension von $\boldsymbol{X}_2$,
- $\text{USSR} = \boldsymbol{y}^\top \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{y}$ die Summe der quadrierten Residuen einer vollen Regression (4.23),
- $\text{RSSR} = \boldsymbol{y}^\top \boldsymbol{M}_{\boldsymbol{X}_1} \boldsymbol{y}$ die Summe der quadrierten Residuen einer Regression von $\boldsymbol{y}$ auf $\boldsymbol{X}_1$, d.h. unter Auferlegung von H_0.

Diese Nullhypothese entspricht den Restriktionen $\boldsymbol{R}^\top = [\boldsymbol{O} \;\; \boldsymbol{I}]$ und $\boldsymbol{r} = \boldsymbol{0}$.

Definiere den partitionierten KQ-Schätzer $\hat{\beta}_{\text{OLS}} = (\hat{\beta}_{\text{OLS},1}^\top, \hat{\beta}_{\text{OLS},2}^\top)^\top$.

Dann folgt

$$\boldsymbol{R}^\top \hat{\beta}_{\text{OLS}} = \hat{\beta}_{\text{OLS},2}$$

und

$$\boldsymbol{R}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{R} =: \tilde{\boldsymbol{D}},$$

dem südöstlichen Block von

$$\begin{aligned}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} &= \begin{pmatrix} \boldsymbol{X}_1^\top \boldsymbol{X}_1 & \boldsymbol{X}_1^\top \boldsymbol{X}_2 \\ \boldsymbol{X}_2^\top \boldsymbol{X}_1 & \boldsymbol{X}_2^\top \boldsymbol{X}_2 \end{pmatrix}^{-1} \\ &=: \begin{pmatrix} \tilde{\boldsymbol{A}} & \tilde{\boldsymbol{B}} \\ \tilde{\boldsymbol{C}} & \tilde{\boldsymbol{D}} \end{pmatrix}\end{aligned}$$

Ergebnisse für partitionierte Inversen (vgl. Matrixreader, Resultat 23) liefern

$$\tilde{\boldsymbol{D}} = (\boldsymbol{X}_2^\top \boldsymbol{X}_2 - \boldsymbol{X}_2^\top \boldsymbol{X}_1 (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{X}_2)^{-1} = (\boldsymbol{X}_2^\top \boldsymbol{M}_{\boldsymbol{X}_1} \boldsymbol{X}_2)^{-1}$$

wobei $\boldsymbol{M}_{\boldsymbol{X}_1} = \boldsymbol{I} - \boldsymbol{X}_1 (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top$.

Der Zähler der F-Statistik ist also (ohne Teilen durch r)

$$F_{\text{num}} = \hat{\beta}_{\text{OLS},2}^{\top}(\boldsymbol{X}_2^{\top}\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2)\hat{\beta}_{\text{OLS},2}$$

Das FWL-Theorem 2.4 liefert

$$\hat{\beta}_{\text{OLS},2} \overset{(2.15)}{=} (\boldsymbol{X}_2^{\top}\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^{\top}\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{y},$$

so dass

$$\begin{aligned} F_{\text{num}} &= \boldsymbol{y}^{\top}\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2(\boldsymbol{X}_2^{\top}\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2)^{-1}(\boldsymbol{X}_2^{\top}\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2)(\boldsymbol{X}_2^{\top}\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^{\top}\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{y} \\ &= \boldsymbol{y}^{\top}\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2(\boldsymbol{X}_2^{\top}\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^{\top}\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{y} \qquad (4.25) \end{aligned}$$

Wir zeigen nun, dass $F_{\text{num}} = \text{RSSR} - \text{USSR}$. Mit FWL (das auch zeigt, dass die Residuen der beiden Vorgehensweisen identisch sind) können wir USSR als die SSR der Regression

$$\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{y} \quad \text{auf} \quad \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2$$

schreiben. D.h.,

$$\begin{aligned}
\text{USSR} &= \boldsymbol{y}^\top \boldsymbol{M}_{\boldsymbol{X}_1}^\top \boldsymbol{M}_{\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{x}_2} \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{y} \\
&= \boldsymbol{y}^\top \boldsymbol{M}_{\boldsymbol{X}_1}^\top (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{x}_2}) \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{y} \\
&= \boldsymbol{y}^\top \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{M}_{\boldsymbol{X}_1} \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2 ((\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2)^\top \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2)^{-1} (\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2)^\top \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{y} \\
&= \boldsymbol{y}^\top \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{M}_{\boldsymbol{X}_1} \boldsymbol{X}_2 (\boldsymbol{X}_2^\top \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2)^{-1} \boldsymbol{X}_2^\top \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{y}
\end{aligned}$$

Also ist RSSR – USSR gleich (4.25):

$$\begin{aligned}
\text{RSSR} - \text{USSR} &= \boldsymbol{y}^\top \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{y} - (\boldsymbol{y}^\top \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{M}_{\boldsymbol{X}_1} \boldsymbol{X}_2 (\boldsymbol{X}_2^\top \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2)^{-1} \boldsymbol{X}_2^\top \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{y}) \\
&= \boldsymbol{y}^\top \boldsymbol{M}_{\boldsymbol{X}_1} \boldsymbol{X}_2 (\boldsymbol{X}_2^\top \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2)^{-1} \boldsymbol{X}_2^\top \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{y}
\end{aligned}$$

Der Nenner von (4.20) ist offenbar

$$\text{USSR}/(n-k).$$

`SimFstat.R`

Wenn $\boldsymbol{X}_1 = \boldsymbol{\iota}$, wir also in (4.22) testen, ob alle Steigungskoeffizienten gleich null sind, ist

$$\text{RSSR} = \boldsymbol{y}^\top \boldsymbol{M}_\iota \boldsymbol{y} \tag{4.26}$$

Damit gilt dann wegen (2.19), dass

$$R^2 = 1 - \frac{\text{USSR}}{\text{RSSR}}$$

oder

$$\text{USSR} = (1 - R^2)\text{RSSR}$$

und daher

$$\frac{\text{RSSR} - \text{USSR}}{\text{USSR}} = \frac{R^2}{1 - R^2}$$

Chow-Test

Der Chow-Test auf Strukturbruch testet

$$H_0 : \beta_1 = \beta_2 \tag{4.27}$$

für zwei unabhängige Stichproben $\boldsymbol{y}_1 = \boldsymbol{X}_1\beta_1 + \boldsymbol{u}_1$, and $\boldsymbol{y}_2 = \boldsymbol{X}_2\beta_2 + \boldsymbol{u}_2$, wobei $\boldsymbol{u}_1|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_{n_1})$ und $\boldsymbol{u}_2|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_{n_2})$.

Beachten Sie, dass wir die Störtermvarianz als gleich voraussetzen und die Zahl der Spalten von $\boldsymbol{X}_1$ und $\boldsymbol{X}_2$ gleich ist, da anderenfalls H_0 nicht wahr sein kann. Siehe `Chow.R`.

Wenn $\boldsymbol{X}_1$ und $\boldsymbol{X}_2$ vollen Spaltenrang k haben, können wir einen Test für H_0 formulieren, der auf dem klassischen F-Test basiert.

Wir reduzieren dazu ein System mehrerer Gleichungen auf ein Modell mit nur einer Gleichung.

Im gegebenen Fall haben wir als unrestringiertes Modell

$$\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1 & \mathbf{O} \\ \mathbf{O} & \boldsymbol{X}_2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \end{pmatrix} = \boldsymbol{X}\beta + \boldsymbol{u}, \tag{4.28}$$

wobei $\boldsymbol{X}$, β und $\boldsymbol{u}$ implizit definiert sind, mit $\boldsymbol{u}|\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_{n_1+n_2})$,
$\boldsymbol{X}$: $(n_1 + n_2) \times 2k$ und β: $2k \times 1$.

Die daraus folgende Summe der quadrierten Residuen des unrestringierten Modells ist dabei

$$\hat{\boldsymbol{u}}_{\text{OLS}}^\top \hat{\boldsymbol{u}}_{\text{OLS}} = \hat{\boldsymbol{u}}_1^\top \hat{\boldsymbol{u}}_1 + \hat{\boldsymbol{u}}_2^\top \hat{\boldsymbol{u}}_2 \tag{4.29}$$

wobei $\hat{\boldsymbol{u}}_1$ und $\hat{\boldsymbol{u}}_2$ die OLS-Residuen der Regressionen $\boldsymbol{y}_i = \boldsymbol{X}_i \hat{\beta}_i + \hat{\boldsymbol{u}}_i$ sind, mit

$$\hat{\beta}_i = (\boldsymbol{X}_i^\top \boldsymbol{X}_i)^{-1} \boldsymbol{X}_i^\top \boldsymbol{y}_i$$

für $i = 1, 2$.

Das restringierte Modell ist

$$\boldsymbol{y} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} \beta + \boldsymbol{u}. \tag{4.30}$$

Die Residuen des restringierten Modells $\hat{\boldsymbol{u}}_{\text{RLS}}$ erhalten wir durch eine Regression von $\boldsymbol{y}$ auf $(\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top)^\top$, das k Regressoren hat. Die SSR der hieraus resultierenden OLS Residuen ist $\hat{\boldsymbol{u}}_{\text{RLS}}^\top \hat{\boldsymbol{u}}_{\text{RLS}}$.

Die Chow-Teststatistik ist nun (4.24), die Version des F-Tests, in welcher die SSR des restringierten und des unrestringierten Modells verglichen werden:

$$F_{Chow} = \frac{(\hat{\boldsymbol{u}}_{\text{RLS}}^\top \hat{\boldsymbol{u}}_{\text{RLS}} - \hat{\boldsymbol{u}}_1^\top \hat{\boldsymbol{u}}_1 - \hat{\boldsymbol{u}}_2^\top \hat{\boldsymbol{u}}_2)/k}{(\hat{\boldsymbol{u}}_1^\top \hat{\boldsymbol{u}}_1 + \hat{\boldsymbol{u}}_2^\top \hat{\boldsymbol{u}}_2)/(n_1 + n_2 - 2k)} \sim F_{k, n_1+n_2-2k}. \tag{4.31}$$

`okcupid.R`

Konfidenzintervalle

Ein Konfidenzintervall zum Konfidenzniveau $(1-\alpha)100\%$ ist die Menge von Parameterwerten θ^* für welche die Nullhypothese $H_0 : \theta = \theta^*$ zum Niveau α nicht abgelehnt würde. Nehmen wir an, dass H_0 abgelehnt wird, wenn

$$\tau(\boldsymbol{y}, \theta^*) \in \mathcal{A}(\boldsymbol{y}, \theta^*).$$

Das $(1-\alpha)100\%$ Konfidenzintervall für θ ist dann

$$\{\theta^* | \tau(\boldsymbol{y}, \theta^*) \notin \mathcal{A}(\boldsymbol{y}, \theta^*)\}.$$

In der Praxis werden Konfidenzintervalle oft als eine Spanne „plausibler Werte“ für einen Parameter interpretiert.

Der einfachste Weg ein Konfidenzintervall zu konstruieren ist über ein **Pivot**, eine Funktion von Beobachtungen und Parametern, deren *Verteilung* nicht von den unbekannten Parametern abhängt.

Beispiel 4.3 Betrachte eine Stichprobe $y_1, \ldots, y_n$ aus einer $\mathcal{N}(\mu, 1)$-Verteilung. Wir wissen, dass $\bar{Y} - \mu \sim \mathcal{N}(0, 1/n)$. Es sei wieder c_α das $\mathcal{N}(0,1)$-$(1-\alpha/2)$-Quantil. Daher gilt

$$\begin{aligned}
1-\alpha &= \mathrm{P}\{(\bar{Y}-\mu)/(1/\sqrt{n}) \in (-c_\alpha, c_\alpha)\} \\
&= \mathrm{P}\{-c_\alpha \leqslant (\bar{Y}-\mu)\sqrt{n} \leqslant c_\alpha\} \\
&= \mathrm{P}\{c_\alpha \geqslant (\mu-\bar{Y})\sqrt{n} \geqslant -c_\alpha\} \\
&= \mathrm{P}\{-c_\alpha/\sqrt{n} \leqslant \mu-\bar{Y} \leqslant c_\alpha/\sqrt{n}\} \\
&= \mathrm{P}\{\bar{Y}-c_\alpha/\sqrt{n} \leqslant \mu \leqslant \bar{Y}+c_\alpha/\sqrt{n}\} \\
&= \mathrm{P}\{(\bar{Y}-c_\alpha/\sqrt{n}, \bar{Y}+c_\alpha/\sqrt{n}) \ni \mu\}
\end{aligned}$$

ist ein Konfidenzintervall zum Niveau $1-\alpha$. Die *Breite* des Intervalls hängt hier nicht von den Daten ab, sondern nur von α und n. Dies kommt daher, dass wir σ^2 als bekannt angenommen haben (und der Einfachheit halber gleich eins).

Setzen Sie einen Randwert des Konfidenzintervalls ein, um zu überprüfen, dass er tatsächlich ein μ^* repräsentiert, für welches gerade $H_0 : \mu = \mu^*$ nicht abgelehnt werden würde!

Es ist wichtig zu verstehen, was ein Konfidenzintervall aussagt und was nicht:

Das Intervall überdeckt den wahren Wert mit Wahrscheinlichkeit $1 - \alpha$. Was bedeutet diese Wahrscheinlichkeit?

Die Wahrscheinlichkeit ist eine Wahrscheinlichkeit über wiederholte Stichproben.

Sie besagt nicht, dass der wahre Parameter für ein konkret berechnetes Intervall mit Wahrscheinlichkeit 95% in dem Intervall liegt. Das tut er oder er tut es nicht. Welcher Fall eingetreten ist, wissen wir aber leider nicht, ähnlich wie wir auch nicht wissen, ob eine konkrete Verwerfung eines Tests eine korrekte Entscheidung oder aber ein Fehler erster Art ist.

Ein 95%-Konfidenzintervall überdeckt demnach den wahren Wert in etwa 95% der Fälle. Siehe `ConfidenceIntervals.R`.

Konfidenzintervalle

Betrachte etwas allgemeiner ein Pivot für einen Regressionskoeffizienten β_i, siehe (4.17)

$$\tau(\boldsymbol{y}, \beta_i) = \frac{\hat{\beta}_i - \beta_i}{\mathsf{se}(\hat{\beta}_i)}.$$

Mit dem $1 - \alpha/2$-Quantil der t_{n-k}-Verteilung $t_{n-k,\alpha}$ erhalten wir aus (4.15) das Konfidenzintervall

$$\{\beta_i \mid -t_{n-k,\alpha} \leq \tau(\boldsymbol{y}, \beta_i) \leq t_{n-k,\alpha}\} = \left[\hat{\beta}_i - \mathsf{se}(\hat{\beta}_i) t_{n-k,\alpha}, \hat{\beta}_i + \mathsf{se}(\hat{\beta}_i) t_{n-k,\alpha}\right].$$

Wir nennen diese Prozedur die **Invertierung von t-Tests**.

Asymptotische Theorie

„Everybody believes in the [Normal] distribution: the experimenters, because they think it can be proved by mathematics; and the mathematicians, because they believe it has been established by observation.“

—Whittaker, E. T. and Robinson, G. „Normal Frequency Distribution.“ Ch. 8 in The Calculus of Observations: A Treatise on Numerical Mathematics, 4th ed. New York: Dover, pp. 164-208, 1967. p. 179.

Asymptotische Theorie

Wir beschäftigen uns nun mit der Lockerung der Annahme von normalverteilten Fehlern. Dafür wiederholen wir einige Konzepte der asymptotischen Theorie.

Für eine Folge von deterministischen Variablen gilt $(x_i)_{i=1}^{\infty} = x_1, x_2, x_3, \ldots$, $x_n \to a$ oder $\lim_{n\to\infty} x_n = a$, wenn x_n beliebig nah an a für ausreichend große n liegt.

Bei Zufallsvariablen/-vektoren müssen wir hingegen verschiedene Arten von Konvergenz unterscheiden:

(i) Konvergenz in quadratischem Mittel (q.m.)

(ii) Konvergenz in Wahrscheinlichkeit (plim)

(iii) Konvergenz in Verteilung.

Wir formulieren Gesetze der Großen Zahlen (Laws of Large Numbers/LLN) und zentrale Grenzwertsätze (Central Limit Theorems/CLT).

1. **(schwaches) Gesetz der großen Zahlen**: Für unabhängige und identisch verteilte (u.i.v.) Zufallsvariablen $X_1, \ldots, X_n$ gilt

$$\overline{X}_n = \frac{\sum_{i=1}^{n} X_i}{n} \xrightarrow{p} \mu = E(X_i), \tag{4.32}$$

   vorausgesetzt dieser Erwartungswert existiert. `ShinyApp:WLLN`

Was bedeutet das? Das LLN ist ein Beispiel für **Konvergenz in Wahrscheinlichkeit**. Betrachten wir eine Folge von Zufallsvariablen $Y_1, Y_2, \ldots$.

2. **Konvergenz in Wahrscheinlichkeit**: Wir sagen, dass die Folge $\{Y_n\}$ in Wahrscheinlichkeit gegen c konvergiert, wenn für alle $\epsilon > 0$

$$\lim_{n\to\infty} \mathrm{P}\left(|Y_n - c| < \epsilon\right) = 1. \tag{4.33}$$

   Notation: $Y_n \xrightarrow{p} c$ oder $\mathrm{plim}_{n\to\infty}(Y_n) = c$.

3. **Konvergenz im quadratischen Mittel**: Ein weiteres Konzept verwendet Momente (die nicht immer existieren): wenn $E(Y_n) \to c$ und $Var(Y_n) \to 0$, dann konvergiert $\{Y_n\}$ im quadratischen Mittel gegen c. Notation: $Y_n \xrightarrow{q.m.} c$.

4. Wenn die Varianz existiert, gilt

$$Y_n \xrightarrow{q.m.} c \Rightarrow Y_n \xrightarrow{p} c \tag{4.34}$$

   Dies ist oft der einfachste Weg, um 2. zu zeigen.

5. Für stetige Funktionen $g(\cdot)$, die nicht von n abhängen, gilt

$$\mathrm{plim}_{n\to\infty}\left\{g(Y_n)\right\} = g\left(\mathrm{plim}_{n\to\infty}(Y_n)\right) \tag{4.35}$$

6. **Konvergenz in Verteilung**: Wenn die Verteilungsfunktionen $F_{Y_n}(x)$ gegen die Verteilungsfunktion der Zufallsvariable X konvergiert, also $F_{Y_n}(x) \to F_X(x)$ für alle Stetigkeitsstellen x gilt, sagen wir

$$Y_n \xrightarrow{d} X, \qquad \text{oder} \qquad Y_n \overset{a}{\sim} F_X, \tag{4.36}$$

d.h. Y_n hat die asymptotische Verteilung F_X.

7. **Satz von Slutzky**: Nehmen wir an, es liegen Zufallsfolgen $X_n \xrightarrow{p} a$ und $Z_n \xrightarrow{d} Z$ vor, wobei a nicht zufällig ist. Dann gilt

$$X_n Z_n \xrightarrow{d} aZ, \tag{4.37}$$

$$X_n + Z_n \xrightarrow{d} a + Z. \tag{4.38}$$

Diese Resultate erlauben es asymptotische Verteilungen abzuleiten, die wiederum verwendet werden können, um Verteilungen in endlichen Stichproben zu approximieren.

Hierbei sind zentrale Grenzwertsätze zentraler Baustein.

8. **Zentraler Grenzwertsatz**: Für u.i.v. Beobachtungen $X_1, \ldots, X_n$ mit Erwartungswert μ und Varianz σ^2 gilt

$$n^{1/2}(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2). \tag{4.39}$$

Anders ausgedrückt, gilt für $Y_n = n^{1/2}(\overline{X}_n - \mu)/\sigma$, dass

$$F_{Y_n}(x) \to \Phi(x)$$

für alle x, mit Φ der CDF der $\mathcal{N}(0,1)$-Verteilung (1.4).

Siehe `clt.R`.

Wir sagen oft etwas salopp, dass $\overline{X}_n$ eine asymptotische Normalverteilung hat, obwohl eigentlich gilt $\overline{X}_n \xrightarrow{p} \mu$. Wir meinen, dass die Verteilung von

$$n^{1/2}(\overline{X}_n - \mu)$$

gegen eine Normalverteilung konvergiert.

9. Wenn $Y_n \xrightarrow{d} 0$, dann $Y_n \xrightarrow{p} 0$.
   Wenn also, basierend auf Resultat 7 oben, $Y_n \xrightarrow{p} 0$ und $Z_n \xrightarrow{d} Z$, dann

$$Y_n Z_n \xrightarrow{p} 0. \tag{4.40}$$

   Folglich gilt für

$$n^{1/2}(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

   dass $\text{plim}_{n\to\infty} \overline{X}_n = \mu$.

   Dies ist der Fall, da $n^{-1/2} \to 0$ (was gelesen werden kann als $n^{-1/2} \xrightarrow{p} 0$) und $n^{1/2}(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ daher gilt

$$\overline{X}_n - \mu = n^{-1/2} \left\{ n^{1/2}(\overline{X}_n - \mu) \right\} \xrightarrow{p} 0, \tag{4.41}$$

10. Die obigen Resultate wurden für skalarwertige Zufallsvariablen formuliert, es gelten jedoch sehr ähnliche Resultate für Zufallsvektoren und -Matrizen.

Asymptotik für Regressionen

Erinnern wir uns an die Darstellung

$$\hat{\beta}_{\text{OLS}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \overset{(3.3)}{=} \beta + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u}.$$

Beachten Sie

$$\boldsymbol{X}^\top \boldsymbol{X} = \sum_{t=1}^{n} \boldsymbol{X}_t^\top \boldsymbol{X}_t,$$

wobei $\boldsymbol{X}_t$ die t-te Zeile von $\boldsymbol{X}$ ist. Also ist

$$\frac{\boldsymbol{X}^\top \boldsymbol{X}}{n} = \frac{1}{n} \sum_{t=1}^{n} \boldsymbol{X}_t^\top \boldsymbol{X}_t,$$

ein Mittelwert.

Wir treffen die folgenden Annahmen.

Annahme 4.4 Die Folge $\{y_t, \boldsymbol{X}_t\}$ ist gemeinsam u.i.v.

Annahme 4.4 impliziert mit dem Gesetz der großen Zahlen (vgl. (4.32)), dass

$$\frac{\boldsymbol{X}^\top \boldsymbol{X}}{n} \xrightarrow{p} E(\boldsymbol{X}_t^\top \boldsymbol{X}_t) =: \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}} > \mathbf{O}, \tag{4.42}$$

sofern die Momente existieren.

Analog ist

$$\frac{\boldsymbol{X}^\top \boldsymbol{u}}{n} = \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{X}_t^\top u_t \tag{4.43}$$

ein Mittelwert. Wenn wir bei der Aufstellung unseres Modells einen guten Job gemacht haben, wird

Annahme 4.5

$$E(\boldsymbol{X}_t^\top u_t) = \mathbf{0} \tag{4.44}$$

erfüllt sein. Wie motiviert sich diese Annahme?

Unter Annahme 4.4 wird (3.2) zu $E(u_t|\boldsymbol{X}_t) = 0$ und wir erhalten durch das Gesetz der iterierten Erwartungen (LIE) (1.7)

$$\begin{aligned} E(\boldsymbol{X}_t^\top u_t) &= E(E(\boldsymbol{X}_t^\top u_t|\boldsymbol{X}_t)) \\ &= E(\boldsymbol{X}_t^\top E(u_t|\boldsymbol{X}_t)) \\ &= \mathbf{0} \end{aligned}$$

Die Annahme, dass $E(\boldsymbol{X}_t^\top u_t) = \mathbf{0}$ ist jedoch schwächer als $E(u_t|\boldsymbol{X}_t) = 0$!

Dann liefert Annahme 4.4 über das Gesetz der großen Zahlen (4.32), dass

$$\frac{\boldsymbol{X}^\top \boldsymbol{u}}{n} \xrightarrow{p} E(\boldsymbol{X}_t^\top u_t) = \boldsymbol{0}. \tag{4.45}$$

Mittels (4.37) und (4.38) erhalten wir

$$\begin{aligned} \text{plim}_{n\to\infty} \hat{\beta}_{\text{OLS}} &= \beta + \text{plim}_{n\to\infty} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u} \\ &= \beta + \text{plim}_{n\to\infty} \left(\frac{\boldsymbol{X}^\top \boldsymbol{X}}{n} \right)^{-1} \text{plim}_{n\to\infty} \left(\frac{\boldsymbol{X}^\top \boldsymbol{u}}{n} \right) \\ &= \beta + \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{-1} \boldsymbol{0} \\ &= \beta. \end{aligned} \tag{4.46}$$

Siehe `ConsistencyOLS.R`.

Um die asymptotische Verteilung herzuleiten, treffen wir zusätzlich

Annahme 4.6

$$\boldsymbol{\Omega} = Var\left(\boldsymbol{X}_t^\top u_t\right) \tag{4.47}$$

existiert und ist endlich.

Unter (4.44) ist

$$\boldsymbol{\Omega} = Var\left(\boldsymbol{X}_t^\top u_t\right) = E\left(u_t^2 \boldsymbol{X}_t^\top \boldsymbol{X}_t\right) \tag{4.48}$$

Der ZGWS (4.39) impliziert mit Annahme 4.4 und (4.44), dass

$$n^{1/2}\left(\frac{\boldsymbol{X}^\top \boldsymbol{u}}{n}\right) = n^{1/2}\left(\frac{\sum_{t=1}^n \boldsymbol{X}_t^\top u_t}{n}\right) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Omega}) \tag{4.49}$$

Wir müssen nicht mehr annehmen, dass $\boldsymbol{u}$ selbst normalverteilt ist!

Wir haben bisher keine zusätzliche Annahme getroffen, die besagen würde, dass

$$\boldsymbol{\Omega} = \sigma^2 \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}. \tag{4.50}$$

(4.50) ergibt sich aus Annahmen 3.2, 4.4 und 4.5, da

$$Var(\boldsymbol{u}|\boldsymbol{X}) = \sigma^2 \boldsymbol{I}_n \overset{\text{u.i.v.}}{\Rightarrow} E(u_t^2|\boldsymbol{X}_t) = \sigma^2$$

und

$$Var\left(\boldsymbol{X}_t^\top u_t\right) = E\left(u_t^2 \boldsymbol{X}_t^\top \boldsymbol{X}_t\right) = E\left(E(u_t^2|\boldsymbol{X}_t)\boldsymbol{X}_t^\top \boldsymbol{X}_t\right) = \sigma^2 E\left(\boldsymbol{X}_t^\top \boldsymbol{X}_t\right) \tag{4.51}$$

Aus (4.37), (4.42) und (4.49) folgt

$$\begin{aligned} n^{1/2}(\hat{\beta}_{\text{OLS}} - \beta) &= \left(\frac{\boldsymbol{X}^\top \boldsymbol{X}}{n}\right)^{-1} \left(\frac{\boldsymbol{X}^\top \boldsymbol{u}}{n^{1/2}}\right) \\ &\xrightarrow{d} \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{-1} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Omega}) \\ &= \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{-1} \boldsymbol{\Omega} \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{-1}\right), \end{aligned} \tag{4.52}$$

Wenn wir die Annahme bedingter Homoskedastie (4.50) treffen, erhalten wir

$$\begin{aligned} n^{1/2}(\hat{\beta}_{\text{OLS}} - \beta) &\xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{-1} \sigma^2 \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}} \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{-1}\right) \\ &= \mathcal{N}\left(\boldsymbol{0}, \sigma^2 \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{-1}\right) \end{aligned} \tag{4.53}$$

Wegen (4.42) wird $\boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}$ konsistent geschätzt durch $\boldsymbol{X}^\top \boldsymbol{X}/n$,

$$\frac{\boldsymbol{X}^\top \boldsymbol{X}}{n} \xrightarrow{p} \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}.$$

Daher ist

$$n^{1/2}(\hat{\beta}_{\text{OLS}} - \beta) \overset{appr}{\sim} \mathcal{N}\left(\mathbf{0}, \sigma^2 \left(\frac{\boldsymbol{X}^\top \boldsymbol{X}}{n}\right)^{-1}\right), \tag{4.54}$$

oder

$$\hat{\beta}_{\text{OLS}} \overset{appr}{\sim} \mathcal{N}\left(\beta, \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\right) \tag{4.55}$$

Die Verwendung einer konsistenten Schätzung von σ^2 komplettiert das Vorgehen: Betrachten wir das asymptotische Verhalten von (3.8),

$$\hat{\sigma}^2 = \frac{\boldsymbol{y}^\top (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}}{n-k} = \frac{\boldsymbol{u}^\top (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{u}}{n-k}. \tag{4.56}$$

Daher gilt

$$\hat{\sigma}^2 = \frac{\frac{\boldsymbol{u}^\top \boldsymbol{u}}{n} - \frac{\boldsymbol{u}^\top \boldsymbol{X}}{n} \left(\frac{\boldsymbol{X}^\top \boldsymbol{X}}{n} \right)^{-1} \frac{\boldsymbol{X}^\top \boldsymbol{u}}{n}}{\frac{n-k}{n}}. \tag{4.57}$$

Dies ist eine stetige Funktion in drei Argumenten

$$\begin{aligned}
\frac{\boldsymbol{X}^\top \boldsymbol{X}}{n} &\xrightarrow{p} \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}} \\
\frac{\boldsymbol{u}^\top \boldsymbol{u}}{n} &= \frac{1}{n} \sum_{t=1}^{n} u_t^2 \xrightarrow{p} \sigma^2 \qquad \text{(warum?)} \\
\frac{\boldsymbol{X}^\top \boldsymbol{u}}{n} &\xrightarrow{p} \boldsymbol{0} \qquad \text{(vgl. (4.45)).}
\end{aligned}$$

Daher gilt

$$\begin{aligned}
\operatorname{plim}_{n\to\infty} \hat{\sigma}^2 &= \frac{\operatorname{plim}\left(\frac{\boldsymbol{u}^\top \boldsymbol{u}}{n}\right) - \operatorname{plim}\left(\frac{\boldsymbol{u}^\top \boldsymbol{X}}{n}\right) \operatorname{plim}\left(\frac{\boldsymbol{X}^\top \boldsymbol{X}}{n}\right)^{-1} \operatorname{plim}\left(\frac{\boldsymbol{X}^\top \boldsymbol{u}}{n}\right)}{\lim \frac{n-k}{n}} \\
&= \sigma^2. \qquad (4.58)
\end{aligned}$$

Damit kann die asymptotische Verteilung (4.53) konsistent geschätzt und gemäß (4.55) als Approximation für die Verteilung von $\hat{\beta}_{\text{OLS}}$ in endlichen Stichproben genutzt werden.

Ferner ist

$$\frac{(\boldsymbol{X}^\top \boldsymbol{X})^{1/2}(\hat{\beta}_{\text{OLS}} - \beta)}{\hat{\sigma}} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_k), \tag{4.59}$$

da aus (4.37), (4.42), (4.53) folgt, dass

$$\begin{aligned} \frac{(\boldsymbol{X}^\top \boldsymbol{X})^{1/2}(\hat{\beta}_{\text{OLS}} - \beta)}{\hat{\sigma}} &= \frac{(\boldsymbol{X}^\top \boldsymbol{X}/n)^{1/2}}{\hat{\sigma}} n^{1/2}(\hat{\beta}_{\text{OLS}} - \beta) \\ &\xrightarrow{d} \frac{\boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{1/2}}{\sigma} \mathcal{N}\left(\mathbf{0}, \sigma^2 \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{-1}\right) =: \mathcal{B} \end{aligned}$$

und, wie in (4.19),

$$Var(\mathcal{B}) = \frac{\boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{1/2}}{\sigma} \sigma^2 \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{-1} \frac{\boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{1/2}}{\sigma} = \boldsymbol{I}_k.$$

Damit sind t-Statistiken wie in (4.17) unter H_0 asymptotisch standardnormalverteilt.

Sie entsprechen Elementen des Vektors (4.59), wenn wir den unbekannten wahren Parameter β mit dem unter H_0 angenommenen Wert ersetzen und im Nenner das entsprechende Diagonalelement wie in (4.17) betrachten.

`BechdelTest.R`

Asymptotische Tests haben unter $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ die Eigenschaft

$$P\left(\tau(\boldsymbol{y}, \boldsymbol{\theta}_0) \in \mathcal{A}(\boldsymbol{y}, \boldsymbol{\theta}_0)\right) \to \alpha$$

wenn $n \to \infty$.

Asymptotische Konfidenzregionen und Tests

Mit (4.53) $(n^{1/2}(\hat{\beta}_{\text{OLS}} - \beta) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \sigma^2 \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{-1}\right))$ gilt, dass

$$n^{1/2} \boldsymbol{R}^\top (\hat{\beta}_{\text{OLS}} - \beta) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \sigma^2 \boldsymbol{R}^\top \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{-1} \boldsymbol{R}\right),$$

sowie, mit einer ähnlichen Überlegung zu (4.18) und (4.5),

$$\frac{n(\boldsymbol{R}^\top (\hat{\beta}_{\text{OLS}} - \beta))^\top \left(\boldsymbol{R}^\top \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{-1} \boldsymbol{R}\right)^{-1} \boldsymbol{R}^\top (\hat{\beta}_{\text{OLS}} - \beta)}{\sigma^2} \xrightarrow{d} \chi_q^2, \tag{4.60}$$

wobei $\boldsymbol{R} : k \times q$ vollen Spaltenrang hat.

Da

$$\hat{\sigma}^2 \left(\frac{\boldsymbol{X}^\top \boldsymbol{X}}{n} \right)^{-1} \xrightarrow{p} \sigma^2 \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{-1}, \tag{4.61}$$

erhalten wir ebenfalls (analog zu (4.59)) und mit (4.37) und (4.39).

$$\frac{(\boldsymbol{R}^\top(\hat{\beta}_{\text{OLS}} - \beta))^\top \left(\boldsymbol{R}^\top(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{R}\right)^{-1} \boldsymbol{R}^\top(\hat{\beta}_{\text{OLS}} - \beta)}{\hat{\sigma}^2} \xrightarrow{d} \chi^2_q.$$

Die Waldstatistik, um $H_0 : \boldsymbol{R}^\top \beta = \boldsymbol{r}$ zu testen, ist gegeben durch:

$$W = \frac{(\boldsymbol{R}^\top \hat{\beta}_{\text{OLS}} - \boldsymbol{r})^\top (\boldsymbol{R}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{R})^{-1} (\boldsymbol{R}^\top \hat{\beta}_{\text{OLS}} - \boldsymbol{r})}{\hat{\sigma}^2}, \tag{4.62}$$

und unter H_0 gilt

$$W \xrightarrow{d} \chi^2_q. \tag{4.63}$$

Beachte, dass

$$W = q \cdot F \tag{4.64}$$

mit F aus (4.20).

`BechdelTest.R`.

Sei

$$F_{\chi^2_q}(c^*_\alpha) = 1 - \alpha,$$

also ist c^*_α das $1-\alpha$-Quantil der χ^2_q-Verteilung.

Dann ist eine asymptotische $(1-\alpha)100\%$ Konfidenzregion für $\boldsymbol{r}$ gegeben durch

$$\left\{ \boldsymbol{r} \;\middle|\; \frac{(\boldsymbol{R}^\top \hat{\beta}_{\text{OLS}} - \boldsymbol{r})^\top \left\{ \boldsymbol{R}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{R} \right\}^{-1} (\boldsymbol{R}^\top \hat{\beta}_{\text{OLS}} - \boldsymbol{r})}{\hat{\sigma}^2} \leq c^*_\alpha \right\}.$$

Heteroskedastie

Was sollten wir tun, wenn die Annahme bedingter Homoskedastie (4.50),

$$\frac{\boldsymbol{X}^\top \boldsymbol{u}}{n^{1/2}} \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}),$$

zweifelhaft ist? Standardfehler gemäß (4.53) sind dann „falsch", d.h. (4.59) trifft nicht mehr zu.

Wir benötigen dann für (4.52) noch einen Schätzer für $\boldsymbol{\Omega}$ aus (4.48).

Wir können $\boldsymbol{\Omega}$ schätzen mit

$$\widehat{\boldsymbol{\Omega}} := \frac{1}{n} \sum_{t=1}^{n} \hat{u}_t^2 \boldsymbol{X}_t^\top \boldsymbol{X}_t = \frac{\boldsymbol{X}^\top \widehat{\boldsymbol{U}} \boldsymbol{X}}{n}, \tag{4.65}$$

wobei $\hat{u}_t$, $t = 1, \ldots, n$, OLS-Residuen sind.

Hierbei ist

$$\widehat{\boldsymbol{U}} = \begin{pmatrix} \hat{u}_1^2 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \hat{u}_n^2 \end{pmatrix}.$$

Unter geeigneten Regularitätsbedingungen gilt $\widehat{\boldsymbol{\Omega}} \xrightarrow{p} \boldsymbol{\Omega}$.

Der **Heteroskedastie-konsistente-Kovarianz-Matrix-Schätzer** (HCCME) (oder Eicker-White-Schätzer) von $\boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{-1} \boldsymbol{\Omega} \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{-1}$ ist dann gegeben durch

$$n(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \widehat{\boldsymbol{U}} \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}. \tag{4.66}$$

Die hieraus resultierenden Standardfehler sind extrem nützlich. In der t-Statistik (4.17) ersetzen wir $\text{se}(\hat{\beta}_i)$ durch einen robusten Standardfehler basierend auf (4.66).

Siehe `TestingWithRobustStandardErrors.R`.

Die Waldstatistik (4.62) für $H_0 : \boldsymbol{R}^\top \beta = \boldsymbol{r}$ wird zu

$$(\boldsymbol{R}^\top \hat{\beta}_{\text{OLS}} - \boldsymbol{r})^\top \left(\boldsymbol{R}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \widehat{\boldsymbol{U}} \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{R} \right)^{-1} (\boldsymbol{R}^\top \hat{\beta}_{\text{OLS}} - \boldsymbol{r}) \xrightarrow{d} \chi^2_q$$

Verbesserungen können resultieren, wenn Alternativen für $\hat{u}_t^2$ genutzt werden:

- $\hat{u}_t^2 n/(n-k)$,
- $\hat{u}_t^2/(1-h_t)$, $h_t = \boldsymbol{e}_t^\top \boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{e}_t$,
- $\hat{u}_t^2/(1-h_t)^2$.

Diese kompensieren, dass Beobachtungen mit großen Varianzen die Schätzungen stark beeinflussen.

Überblick

Die Annahme der Linearität (vgl. (1.12)) kann abgeschwächt werden, so dass $E(y_t|X_t)$ eine nichtlineare Funktion von X_t und, was noch entscheidender ist, der Parameter sein kann.

Dieses Thema wird ausführlich im Kurs „Microeconometrics“ besprochen.

Überblick

Einleitung

In diesem Kapitel widmen wir uns Modellen, in denen die Störterme $\boldsymbol{u}$ eine Kovarianzmatrix $\boldsymbol{V}$ haben, die sich von einem skalaren Vielfachen der Einheitsmatrix unterscheiden:

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\beta + \boldsymbol{u} \\ \boldsymbol{u}|\boldsymbol{X} &\sim (\boldsymbol{0}, \boldsymbol{V}). \end{aligned} \tag{6.1}$$

Wenn

- $\boldsymbol{V} = \sigma^2 \boldsymbol{I}$ gilt, dann sind wir wieder im klassischen linearen Modell
- $\boldsymbol{V} \neq \sigma^2 \boldsymbol{I}$ und $\boldsymbol{V}$ diagonal ist, dann liegt Heteroskedastie vor.
- $\boldsymbol{V}$ nicht diagonal ist, sind die Störterme autokorreliert. (Siehe z.B. ARMA-Modelle weiter unten.)

Es gibt zwei Strategien im Umgang mit $\boldsymbol{V} \neq \sigma^2 \boldsymbol{I}$:

1. Wir benutzen weiter OLS und nutzen Standardfehler, die gegen $\boldsymbol{V} \neq \sigma^2 \boldsymbol{I}$ robust sind. Für diagonale $\boldsymbol{V}$ haben wir das im vorletzten Kapitel (HCCME) gemacht. (Es gibt auch gegen Autokorrelation robuste Kovarianzmatrix „HAC"-Schätzer, die wir in diesem Kurs jedoch nicht betrachten.)
2. Wir benutzen neue, möglicherweise effizientere, Schätzer, die die Struktur von $\boldsymbol{V}$ berücksichtigen.

Wir beschäftigen uns jetzt mit der zweiten Alternative: Als Alternative zum HCCME versuchen wir die Struktur von $\boldsymbol{V}$ miteinzubeziehen, um effizientere Schätzer zu erhalten.

Wenn $\boldsymbol{V}$ in

$$\boldsymbol{V} = \sigma^2 \boldsymbol{\Sigma} \tag{6.2}$$

bis auf ein skalares Vielfaches bekannt ist—$\boldsymbol{\Sigma}$ wird also als bekannt vorausgesetzt—können wir den **generalized least squares (GLS)**-Schätzer formulieren und zeigen, dass er BLUE ist.

GLS

Betrachten wir eine „Matrixwurzel“ $\boldsymbol{\Psi}$ von $\boldsymbol{\Sigma}^{-1}$, so dass (mehr Details wie man $\boldsymbol{\Psi}$ findet unten)

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Psi}\boldsymbol{\Psi}^\top. \tag{6.3}$$

Wir transformieren (6.1) via

$$\widetilde{\boldsymbol{y}} = \boldsymbol{\Psi}^\top \boldsymbol{y}, \qquad \widetilde{\boldsymbol{X}} = \boldsymbol{\Psi}^\top \boldsymbol{X}, \qquad \widetilde{\boldsymbol{u}} = \boldsymbol{\Psi}^\top \boldsymbol{u}. \tag{6.4}$$

Wir erhalten dann

$$\widetilde{\boldsymbol{y}} = \widetilde{\boldsymbol{X}}\beta + \widetilde{\boldsymbol{u}} \qquad \text{mit} \qquad \widetilde{\boldsymbol{u}}|\widetilde{\boldsymbol{X}} \sim (\boldsymbol{0}, \sigma^2 \boldsymbol{I}),$$

da

$$Var(\widetilde{\boldsymbol{u}}|\widetilde{\boldsymbol{X}}) = \boldsymbol{\Psi}^\top Var(\boldsymbol{u}|\widetilde{\boldsymbol{X}})\boldsymbol{\Psi} \overset{(6.2)}{=} \boldsymbol{\Psi}^\top \sigma^2 \boldsymbol{\Sigma}\boldsymbol{\Psi} = \sigma^2 \boldsymbol{\Psi}^\top (\boldsymbol{\Psi}^\top)^{-1}\boldsymbol{\Psi}^{-1}\boldsymbol{\Psi},$$

weil (6.3) impliziert, dass

$$\boldsymbol{\Sigma} = (\boldsymbol{\Psi}\boldsymbol{\Psi}^\top)^{-1} \overset{\text{M20.5}}{=} (\boldsymbol{\Psi}^\top)^{-1}\boldsymbol{\Psi}^{-1}$$

Wir befinden uns nun wieder im klassischen Modell.

Der OLS-Schätzer basierend auf den transformierten Daten $\widetilde{\boldsymbol{y}}$ und $\widetilde{\boldsymbol{X}}$,

$$\widetilde{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}})^{-1} \widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{y}},$$

ist der **GLS-Schätzer**

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{GLS}} &\overset{(6.4)}{=} (\boldsymbol{X}^\top \boldsymbol{\Psi}\boldsymbol{\Psi}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\Psi}\boldsymbol{\Psi}^\top \boldsymbol{y} \\ &\overset{(6.3)}{=} (\boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{y}. \end{aligned} \tag{6.5}$$

Er ist BLUE, da $\widetilde{\boldsymbol{\beta}}$ der KQ-Schätzer für das klassische Modell ist.

Basierend auf ähnlichen Annahmen wie denen in Kapiteln 4 erhalten wir

$$n^{1/2}(\hat{\boldsymbol{\beta}}_{\text{GLS}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \sigma^2 \operatorname{plim}_{n\to\infty} \left(\frac{\boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{X}}{n}\right)^{-1}\right),$$

was wie zuvor GLS-basierte t-Tests etc. liefert.

Für den Spezialfall, bei dem $\boldsymbol{\Sigma}$ eine Diagonalmatrix ist mit Elementen σ_{tt}, $t = 1, \ldots, n$, ist der GLS-Schätzer gegeben durch

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = \hat{\boldsymbol{\beta}}_{\text{WLS}} = \left(\sum_{t=1}^{n} \sigma_{tt}^{-1} \boldsymbol{X}_t^\top \boldsymbol{X}_t \right)^{-1} \sum_{t=1}^{n} \sigma_{tt}^{-1} \boldsymbol{X}_t^\top \boldsymbol{y}_t$$

und wird normalerweise als **weighted least squares** (WLS) bezeichnet.

Je größer die Varianz σ_{tt} ist, desto weniger stark wird die Beobachtung gewichtet.

Alternative Motivation für GLS: $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ minimiert die folgende Funktion in Bezug auf $\boldsymbol{\beta}$:

$$\widetilde{\boldsymbol{u}}^\top \widetilde{\boldsymbol{u}} = \boldsymbol{u}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{u} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Feasible GLS

In der Praxis ist $\boldsymbol{\Sigma}$ normalerweise unbekannt, so dass GLS nicht direkt implementierbar ist. Wir betrachten daher nun Fälle, in denen $\boldsymbol{V}$ eine bekannte Funktion einer fixen Anzahl an Parametern $\boldsymbol{\theta} : m \times 1$ ist, die nicht größer wird, wenn n steigt.

(Die Anzahl nicht doppelter Elemente von $\boldsymbol{\Sigma}$ entspräche ansonsten $n(n+1)/2$. Dies steigt mit n. Eine wichtige Bedingung für die Konsistenz des Schätzers ist jedoch, dass die Anzahl der Parameter fix ist. Wir haben daher ein Problem, wenn die Anzahl der Parameter von n abhängt.)

Dann können wir einen konsistenten Schätzer für $\boldsymbol{\theta}$ suchen, also einen Schätzer $\hat{\boldsymbol{\theta}}$, für den gilt

$$\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}.$$

Wir arbeiten dann mit $\boldsymbol{V}(\hat{\boldsymbol{\theta}})$, was wir berechnen können. Dies bringt uns zum **feasible GLS (FGLS)**-Schätzer.

FGLS hat typischerweise dieselbe asymptotische Verteilung wie GLS:

Nehmen wir an, dass (mit Verwendung der Notation $\widehat{\boldsymbol{V}} = \boldsymbol{V}(\hat{\boldsymbol{\theta}})$) gilt

$$\frac{\boldsymbol{X}^\top \widehat{\boldsymbol{V}}^{-1} \boldsymbol{X}}{n} - \frac{\boldsymbol{X}^\top \boldsymbol{V}^{-1} \boldsymbol{X}}{n} \xrightarrow{p} \mathbf{O} \tag{6.6}$$

Wir müssen dies explizit annehmen, da—obwohl $\boldsymbol{X}^\top \boldsymbol{V}^{-1} \boldsymbol{X}$ eine $k \times k$ Matrix ist, wobei k nicht von n abhängt und für ein einziges n eine stetige Funktion von $\boldsymbol{\theta}$ ist, da $\boldsymbol{V}^{-1}$ eine stetige Funktion von $\boldsymbol{\theta}$ ist—wir die 5. asymptotische Regel nicht verwenden können. Der Grund hierfür ist, dass die Funktion sich in Abhängigkeit von n verändert.

Wenn jedoch Regularitätsbedingungen erfüllt sind (was für die folgenden Beispiele der Fall ist), gilt (6.6) sowie auch das folgende stärkere Resultat (wir überspringen technische Details)

$$\frac{\boldsymbol{X}^\top \widehat{\boldsymbol{V}}^{-1}\boldsymbol{u}}{n^{1/2}} - \frac{\boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{u}}{n^{1/2}} \xrightarrow{p} \mathbf{0}.$$

Dann hat

$$\hat{\boldsymbol{\beta}}_{\text{FGLS}} = (\boldsymbol{X}^\top \widehat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \widehat{\boldsymbol{V}}^{-1}\boldsymbol{y}$$

dieselbe asymptotische Verteilung wie $\hat{\boldsymbol{\beta}}_{\text{GLS}}$, d.h.

$$n^{1/2}(\hat{\boldsymbol{\beta}}_{\text{FGLS}} - \hat{\boldsymbol{\beta}}_{\text{GLS}}) \xrightarrow{p} \mathbf{0}.$$

ARMA Modelle

Die Kovarianzmatrix der Störterme wird oft mittels eines ARMA-Modells konstruiert.

Häufig wird ein AR(1)-Prozess verwendet:

$$u_t = u_{t-1}\rho + \epsilon_t, \qquad \epsilon_t \overset{\text{u.i.v.}}{\sim} (0, \sigma_\epsilon^2),$$

mit ϵ_t unabhängig von u_{t-j}, $j > 0$.

Stationarität gilt, wenn die unbedingte Verteilung von u_t nicht von t abhängt, oder schwächer, wenn die ersten zwei Momente (wenn sie existieren) nicht von t abhängen.

Wenn also $E(u_t) = 0$ und $Var(u_t) = \sigma^2$ für alle t, dann gilt, bei Unabhängigkeit von ϵ_t und u_{t-1}

$$Var(u_t) = \rho^2 Var(u_{t-1}) + \sigma_\epsilon^2,$$

und Stationarität impliziert

$$\sigma^2 = \rho^2 \sigma^2 + \sigma_\epsilon^2.$$

Also gilt

$$\sigma^2 = \frac{\sigma_\epsilon^2}{1 - \rho^2}.$$

Wenn $|\rho| < 1$ liegt Stationarität vor, wenn der Prozess bei $t \to -\infty$ beginnt oder wenn

$$u_0 \sim \left(0, \frac{\sigma_\epsilon^2}{1 - \rho^2}\right).$$

Die Kovarianzmatrix ist gegeben durch (siehe Übung)

$$Var(\boldsymbol{u}|\boldsymbol{X}) = \boldsymbol{V}(\rho, \sigma_\epsilon^2) = \frac{\sigma_\epsilon^2}{1-\rho^2} \begin{pmatrix} 1 & \rho & \cdots & \rho^{n-2} & \rho^{n-1} \\ \rho & 1 & \cdots & \rho^{n-3} & \rho^{n-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho^{n-2} & \rho^{n-3} & \cdots & 1 & \rho \\ \rho^{n-1} & \rho^{n-2} & \cdots & \rho & 1 \end{pmatrix}.$$

Die Matrixzerlegung gemäß (6.3) ist gegeben durch

$$\boldsymbol{\Psi}(\rho) = \begin{pmatrix} (1-\rho^2)^{1/2} & -\rho & 0 & \cdots & 0 & 0 \\ 0 & 1 & -\rho & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -\rho \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}, \tag{6.7}$$

was eine sogenannte **Bandmatrix** mit zwei Diagonalen ist.

Denn: Direkte Multiplikation zeigt, dass

$$\boldsymbol{\Psi}^\top(\rho)\boldsymbol{V}(\rho,\sigma_\epsilon^2)\boldsymbol{\Psi}(\rho) = \sigma_\epsilon^2\boldsymbol{I}_n.$$

Daraus folgt mit M20.5

$$\left(\boldsymbol{\Psi}^\top(\rho)\boldsymbol{V}(\rho,\sigma_\epsilon^2)\boldsymbol{\Psi}(\rho)\right)^{-1} = \boldsymbol{\Psi}(\rho)^{-1}\boldsymbol{V}(\rho,\sigma_\epsilon^2)^{-1}\boldsymbol{\Psi}^\top(\rho)^{-1} = \sigma_\epsilon^{-2}\boldsymbol{I}_n.$$

Wir sehen, dass

$$\boldsymbol{V}^{-1}(\rho,\sigma_\epsilon^2) = \sigma_\epsilon^{-2}\boldsymbol{\Psi}(\rho)\boldsymbol{\Psi}^\top(\rho),$$

und der GLS-Schätzer (6.5) ist gegeben durch (σ_ϵ^{-2} kürzt sich raus)

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = \left(\boldsymbol{X}^\top\boldsymbol{\Psi}(\rho)\boldsymbol{\Psi}^\top(\rho)\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}(\rho)\boldsymbol{\Psi}^\top(\rho)\boldsymbol{y}.$$

Dies wäre der BLUE-Schätzer, wenn ρ bekannt wäre.

Für den FGLS-Schätzer ersetze ρ durch eine konsistente Schätzung $\hat{\rho}$.
Z.B. können wir die Residuen $\hat{u}_t$ einer KQ-Regression von $\boldsymbol{y}$ auf $\boldsymbol{X}$ auf deren verzögerte Werte $\hat{u}_{t-1}$ regressieren.

Wir können andere ARMA-Prozesse in ähnlicher Weise verwenden. Z.B. erhalten wir für einen MA(1)-Prozess

$$u_t = \epsilon_t + \alpha\epsilon_{t-1}, \qquad \epsilon_t \overset{\text{u.i.v.}}{\sim} (0, \sigma_\epsilon^2)$$

dass

$$\boldsymbol{V}(\alpha, \sigma_\epsilon^2) = \sigma_\epsilon^2 \begin{pmatrix} 1+\alpha^2 & \alpha & 0 & \cdots & 0 & 0 \\ \alpha & 1+\alpha^2 & \alpha & \cdots & 0 & \vdots \\ \vdots & \vdots & \vdots & & \vdots & \alpha \\ 0 & 0 & 0 & \cdots & \alpha & 1+\alpha^2 \end{pmatrix}.$$

Faktorisierung

Es war ein bisschen ad hoc die Matrix $\boldsymbol{\Psi}$ zu finden, so dass $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Psi}\boldsymbol{\Psi}^\top$, wie wir es in den ARMA-Fällen vorher getan haben. Wir besprechen nun kurz Möglichkeiten für die allgemeine Wahl von $\boldsymbol{\Psi}$.

Die untere Dreiecks-Cholesky-Zerlegung ist eine Möglichkeit (siehe Resultat 30).

Eine weitere Möglichkeit für $\boldsymbol{\Psi}$ ist gegeben durch eine symmetrische Matrix $\boldsymbol{\Sigma}^{-1/2}$ über die Singulärwert- (SVD) oder die Eigenwertzerlegung von $\boldsymbol{\Sigma}$:

$$\begin{aligned} \boldsymbol{\Sigma} &= \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^\top \\ \boldsymbol{P}^\top\boldsymbol{P} &= \boldsymbol{P}\boldsymbol{P}^\top = \boldsymbol{I}, \\ \boldsymbol{D} &> \mathbf{0}, \quad \text{diagonal} \end{aligned}$$

Dann gilt

$$\boldsymbol{\Sigma}^{1/2} = \boldsymbol{P}\boldsymbol{D}^{1/2}\boldsymbol{P}^\top,$$

und $\boldsymbol{\Sigma}^{-1/2}$ ist dessen Inverse, gegeben durch $\boldsymbol{P}\boldsymbol{D}^{-1/2}\boldsymbol{P}^\top$.

Test auf Heteroskedastie

Oft wollen wir testen, ob Homoskedastie vorliegt, d.h. $H_0 : E(u_t^2) = \sigma^2$, $t = 1, \ldots, n$ gegen die Alternative $E(u_t^2) \neq E(u_s^2)$, $s \neq t$, $s, t = 1, \ldots, n$.

Betrachte die restriktivere Alternative

$$E(u_t^2) = h(\delta + \boldsymbol{Z}_t \gamma),$$

wobei h die **skedastische Funktion** ist. h sei stetig differenzierbar und $h \geq 0$, muss ansonsten jedoch nicht weiter spezifiziert werden. $\boldsymbol{Z}_t$ ist ein beobachteter $1 \times r$ Vektor (der Elemente von $\boldsymbol{X}_t$ enthalten kann).

H_0 wird zu $\gamma = \mathbf{0}$, während H_1 durch $\gamma \neq \mathbf{0}$ gegeben ist.

Um einen Test durchführen zu können, benötigen wir quadrierte Residuen, die für u_t^2 konsistent sind. Eine offensichtliche Wahl sind die quadrierten OLS-Residuen $\hat{u}_t^2$.

White-Test

Bezeichne den Vektor von r unterschiedlichen (nicht konstanten) Elementen von $\boldsymbol{X}_t^\top \boldsymbol{X}_t$ als $\boldsymbol{Z}_t$. Unter H_0 sind die u_t^2 keine Funktion von $\boldsymbol{Z}_t$.

Berechne das R^2 einer Regression von $\hat{u}_t^2$ auf eine Konstante und $\boldsymbol{Z}_t$. Wir können dann zeigen, dass (dies ist ein recht allgemeines Resultat!)

$$nR^2 \xrightarrow{d} \chi_r^2$$

Schreibe dazu (4.24) im „χ^2-Format" (4.64):

$$r \cdot F_{\beta_2} = \frac{\text{RSSR} - \text{USSR}}{\text{USSR}/(n-k)}$$

Ersetze, was asymptotisch keinen Unterschied macht, $n-k$ durch n und bemerke, dass unter H_0 $\text{USSR}/(n-k)$ und RSSR/n äquivalent sind. Nun ist mit (2.19) für diese H_0 (vgl. (4.26))

$$\frac{\text{RSSR} - \text{USSR}}{\text{RSSR}} = R^2$$

Zum Beispiel erhalten wir für $\boldsymbol{X}_t = (1,\ q_t,\ q_t^2,\ p_t)$

$$\boldsymbol{X}_t^\top \boldsymbol{X}_t = \begin{pmatrix} 1 \\ q_t \\ q_t^2 \\ p_t \end{pmatrix} (1,\ q_t,\ q_t^2,\ p_t) = \begin{pmatrix} 1 & q_t & q_t^2 & p_t \\ & q_t^2 & q_t^3 & q_t p_t \\ & & q_t^4 & q_t^2 p_t \\ \bullet & & & p_t^2 \end{pmatrix},$$

so dass

$$\boldsymbol{Z}_t = (q_t,\ q_t^2,\ p_t,\ q_t^3,\ q_t p_t,\ q_t^4,\ q_t^2 p_t,\ p_t^2)$$

und $r = 8$.

`white test.R`

Test auf Autokorrelation

Wir verwenden den unten beschriebenen t-Test. (Diese Prozedur ist asymptotisch valide, auch wenn erklärende Variablen nur „vorbestimmt“ statt exogen sind—wir verfolgen das aber nicht im Detail weiter.)

Die Prozedur ist wie folgt. Seien $\hat{\boldsymbol{u}}$ die OLS-Residuen aus

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{u},$$

Wir bezeichnen den Vektor der verzögerten Residuen mit $\hat{\boldsymbol{u}}_{-1} = (\hat{u}_1, \ldots, \hat{u}_{n-1})^\top$. Dann können wir folgende Regression verwenden

$$\boldsymbol{y} = \boldsymbol{X}\beta + b_\rho \hat{\boldsymbol{u}}_{-1} + \text{Fehler},$$

um $H_0 : b_\rho = 0$ mit einer üblichen t-Statistik (4.17) zu testen (y_1 und $\boldsymbol{X}_1$ müssen aus dem Datensatz gestrichen werden oder $\hat{\boldsymbol{u}}_{-1}$ am Anfang um eine Null ergänzt werden). `breusch godfrey.R`

Was tun wir, wenn wir Autokorrelation finden?

Es ist verlockend das Problem Autokorrelatation zu „lösen“, indem ein FGLS-Schätzer angewendet wird, der Autokorrelation berücksichtigt.

Allerdings können autokorrelierte Residuen oft einfach ein Hinweis auf eine andere Art von Misspezifikation des Modells sein, z.B. das Fehlen von quadratischen Termen. Ein lineares Modell wird dann zunächst viele positive (negative) und dann viele negative (positive) Residuen liefern, wodurch Autokorrelation in den Residuen entsteht.

Einfach einen FGLS-Schätzer zu verwenden löst das Problem fälschlicherweise ausgelassener quadratischer Terme also nicht.

Fazit: Nachdenken!

Paneldaten

Wir beschäftigen uns nun mit Daten, die über zwei Dimensionen erhoben wurden, auch **Paneldaten** genannt.

Wir könnten z.B. jährliche Beobachtungen zu verschiedenen europäischen Ländern haben („Querschnittseinheiten“) oder Beobachtungen von mehreren Kindern aus einer Familie.

Wir betrachten das folgende lineare Modell

$$y_{it} = \boldsymbol{X}_{it}\beta + u_{it} \qquad i = 1, \ldots, m, \quad t = 1, \ldots, T \tag{6.8}$$

Hierbei entspricht nun m der Anzahl der Einheiten und T der Anzahl der Beobachtungen pro Einheit.

Wenn $E(u_{it}|\boldsymbol{X}_{it}) = 0$ gilt, können wir das Modell einfach mit **pooled OLS** schätzen, was die Paneldaten-Struktur ignoriert und einfach alle $n = mT$ Beobachtungen zusammenwirft.

Allerdings ist diese Annahme in Paneldaten oft unrealistisch. Häufig gibt es Faktoren, die ein Land über mehrere Zeitperioden oder alle Kinder einer Familie beeinflussen. Wenn diese Faktoren mit $\boldsymbol{X}_{it}$ in Beziehung stehen, produziert OLS inkonsistente Schätzer (vgl. (3.14)).

Wir modellieren u_{it} mit dem **Fehlerkomponentenmodell**

$$u_{it} = \eta_i + \epsilon_{it} \tag{6.9}$$

Beispiel 6.1 (Geronimus and Korenman, Quarterly Journal of Economics 1992). Wir interessieren uns für den Effekt von Teenagerschwangerschaften auf zukünftige ökonomische Ergebnisse, z.B. das Einkommen relativ zum Bedarf (wobei Letzteres von der Anzahl der Kinder abhängt). Betrachten wir

$$\begin{aligned} \log(\text{income}/\text{needs}_{fs}) &= \beta_0 + \delta_0 \text{sister2}_s + \beta_1 \text{teenbirth}_{fs} \\ &\quad + \beta_2 \text{age}_{fs} + \text{other factors} + a_f + \epsilon_{fs}, \end{aligned}$$

wobei f der Index für die Familie und s der Index für eine Schwester innerhalb der Familie ist. D.h. t wird durch s und i durch f ersetzt. Zum Beispiel ist a_f hier der familienspezifische Effekt.

Unser Fokus liegt auf β_1, d.h. der ökonomische Effekt einer Teenagergeburt ($\text{teenbirth}_{fs} = 1$). Eine Problem bei gepoolten oder Random Effects-Regressionen (siehe unten) ist, dass teenbirth_{fs} und a_f wahrscheinlich korreliert sind.

Die Bildung der ersten Differenz zwischen Schwestern bietet eine Lösung:

$$\Delta \log(\text{income}/\text{needs}_f) = \delta_0 + \beta_1 \Delta \text{teenbirth}_f + \beta_2 \Delta \text{age}_f + \ldots + \Delta \epsilon_f \quad (6.10)$$

Dafür ist es offensichtlich notwendig, dass für mindestens ein Schwesternpaar $\Delta \text{teenbirth}_f \neq 0$ ist, um β_1 zu identifizieren. D.h. es muss Familien geben, in denen genau eine der Schwestern eine Teenagergeburt hatte.

Mittels Pooled OLS (d.h. die Familienstruktur der Daten wird ignoriert) fanden Geronimus und Korenman $\hat{\beta}_1 \approx -0.3$ (in Abhängigkeit von den genauen anderen Faktoren), wohingegen die Regression mit der ersten Differenz einen Wert von $\hat{\beta}_1 = -0.08$ ergibt.

Dies suggeriert, dass der Familienhintergrund der Mutter einen stärkeren negativen Effekt auf ökonomische Outcomes hat als die Teenagerschwangerschaft selbst.

`sisters.R`

Wir beschäftigen uns jetzt mit einer allgemeineren Formulierung eines Paneldaten-Schätzers, der für $T = 2$ äquivalent zum obigen First Difference-Schätzer (d.h. eine OLS-Regression von (6.10)) ist.

Im Folgenden nehmen wir an, dass $\boldsymbol{X}_{it}$ exogen ist—verzögerte abhängige Variablen sorgen für einige zusätzliche Komplikationen mit Paneldaten.

In Matrixnotation können (6.8) und (6.9) kompakt als

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{D}\eta + \varepsilon \tag{6.11}$$

geschrieben werden, wobei $\boldsymbol{y}$ und ε n-Vektoren mit typischen Elementen y_{it} und ϵ_{it} sind und $\boldsymbol{D}$ eine $n \times m$ (eine Spalte pro Einheit) Matrix aus Dummyvariablen ist.

Wenn eine Zeile eine zu Einheit i gehörende Beobachtung enthält, dann hat $\boldsymbol{D}$ eine eins in Spalte i und ansonsten eine null, für alle $i = 1, \ldots, m$.

Die **individuenspezifischen Effekte** η müssen unabhängig von ϵ_{it} sein. Im **Fixed Effects** Paneldatenmodell dürfen sie allerdings mit den Variablen in $\boldsymbol{X}$ korrelieren.

Der **Fixed Effects Schätzer** ist der OLS-Schätzer von β in der Regression von $\boldsymbol{y}$ auf $\boldsymbol{X}$ *und* $\boldsymbol{D}$ in (6.11). Wegen seiner Berechnungsweise wird er auch **Least Squares Dummy Variablen-** oder **LSDV-Schätzer**, genannt.

Sei

$$\boldsymbol{M_D} = \boldsymbol{I} - \boldsymbol{D}(\boldsymbol{D}^\top \boldsymbol{D})^{-1}\boldsymbol{D}^\top.$$

Dann erhalten wir mittels FWL (siehe 2.4) den OLS-Schätzer durch eine Regression von $\boldsymbol{M_D y}$, die Residuen von der Regression von $\boldsymbol{y}$ auf $\boldsymbol{D}$, auf $\boldsymbol{M_D X}$, die Matrix der Residuen aus der Regression jeder Spalte von $\boldsymbol{X}$ auf $\boldsymbol{D}$.

Der Fixed Effects Schätzer ist daher

$$\hat{\beta}_{\text{FE}} = (\boldsymbol{X}^\top \boldsymbol{M_D X})^{-1}\boldsymbol{X}^\top \boldsymbol{M_D y} \tag{6.12}$$

Es ist einfach zu überprüfen (vgl. Beispiel 2.5), dass das it-te Element des Vektors $\boldsymbol{M_D z}$, für jeden Vektor $\boldsymbol{z}$, gleich $z_{it} - \bar{z}_i$ ist, der Abweichung vom Gruppenmittelwert.

Da alle Variablen in (6.12) mit $\boldsymbol{M_D}$ multipliziert werden, verwendet dieser Schätzer lediglich die Variation um den Mittelwert für jede Einheit.

Damit beseitigen wir das einheitenweise konstante η_i!

Aus diesem Grund wird der Schätzer auch **Within-Schätzer** genannt. Siehe `panel data.R` und `panel data FD and FE.R`.

Dies impliziert unglücklicherweise, dass der Schätzer nicht für erklärende Variablen, die für jede Beobachtung einer Gruppe den selben Wert annehmen, verwendet werden kann, da er nach Transformation mit $\boldsymbol{M_D}$ gleich null ist.

Können wir Paneldaten verwenden, um die Renditen von Schulbildung zu schätzen? Angeborene Fähigkeiten sind per Definition über die Zeit unveränderlich. Ein mögliches Modell ist:

$$\log(wage)_{it} = \beta_0 + \delta_0 d_{2t} + \beta_1 educ_{it} + \beta_2 gender_{it} + \eta_i + \epsilon_{it}, \qquad t = 1, 2$$

Mit den ersten Differenzen erhalten wir

$$\Delta \log(wage)_i = \delta_0 + \beta_1 \Delta educ_i + \beta_2 \Delta gender_i + \epsilon_i \tag{6.13}$$

Das Problem ist, dass wir, um Einkommen zu beobachten, arbeitende Erwachsene betrachten, bei denen sich die Bildung (wie meist auch Geschlecht) über t nicht mehr verändert. Dann können wir (6.13) jedoch nicht schätzen.

Wenn sich Bildung nur bei einigen über t verändert, ist der Schätzer für β_1 unpräzise, da der Regressor wenig Varianz hat.

Dies illustriert die generelle Lektion, dass der FE-Schätzer nur den Effekt von **zeitvariierenden Variablen** schätzen kann, da erste Differenzen/Dummies zeitunveränderliche Effekten entfernen.

Dieses Manko macht es manchmal attraktiv, einen anderen Schätzer, den **Random-Effects-Schätzer**, zu verwenden. Wir kehren zurück zu Modell (6.8) und (6.9), benötigen jedoch für diesen Schätzer die striktere Annahme, dass

$$E(\eta_i|\boldsymbol{X}) = 0 \tag{6.14}$$

Unter dieser Annahme wäre pooled OLS unverzerrt, aber wir können einen effizienteren GLS-Schätzer (6.5) finden. GLS ist effizienter, da die Fehler-Kovarianz-Matrix in diesem Model nicht homoskedastisch ist.

Nehmen wir an, die η_i sind u.i.v. mit $E(\eta_i) = 0$ und $Var(\eta_i) = \sigma_\eta^2$. Diese Annahmen motivieren die Bezeichnung „zufällige Effekte“ (random effects). Wir nehmen des Weiteren an, dass

$$E(\varepsilon\varepsilon^\top|\boldsymbol{X}) = \sigma_\epsilon^2\boldsymbol{I}$$

Wenn die zwei Fehlerkomponenten ferner unabhängig sind, gilt in (6.9)

$$\begin{aligned} Var(u_{it}) &= \sigma_\eta^2 + \sigma_\epsilon^2 \\ Cov(u_{it}, u_{is}) &= \sigma_\eta^2 \\ Cov(u_{it}, u_{js}) &= 0 \qquad \text{für alle } i \neq j. \end{aligned}$$

Wir erhalten dann die folgende Varianz-Kovarianz-Matrix $\boldsymbol{V}$:

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma} \end{pmatrix}$$

Hier gilt

$$\boldsymbol{\Sigma} = \sigma_\eta^2 \boldsymbol{\iota}\boldsymbol{\iota}^\top + \sigma_\epsilon^2 \boldsymbol{I}_T$$

mit $\boldsymbol{\iota}$ einem T-Vektor aus Einsen.

Wie gewöhnlich benötigt FGLS einen konsistenten Schätzer für $\boldsymbol{V}$, und daher hier für σ_η^2 und σ_ϵ^2. Solche existieren. Wir verfolgen dies jedoch nicht im Detail, da die Schlüsselannahme für Random Effects $E(\eta_i|\boldsymbol{X}) = 0$ (siehe (6.14)), in Anwendungen oft nicht glaubwürdig ist.

Man kann die Modelle mit einem **Hausman-Test** gegeneinander testen, siehe das folgende Kapitel und `panel data FE and RE.R`.

Überblick

Einführung

Wir betrachten nun

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{u}, \tag{7.1}$$

$$\boldsymbol{u} \sim (\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n), \tag{7.2}$$

wobei wir nun erlauben, dass (möglicherweise alle) Elemente von $\boldsymbol{X}_t$ mit u_t korreliert sind.

Die Randverteilung von u_t erfüllt somit (7.2). D.h. es kann sein, dass (4.44) nicht erfüllt ist, sondern

$$E(\boldsymbol{X}_t^\top u_t) \neq \boldsymbol{0}. \tag{7.3}$$

Unter (7.3) ist der OLS-Schätzer inkonsistent (vgl. auch (4.46)).

Zunächst betrachten wir Fälle, in denen $E(u_t|\boldsymbol{X}_t) \neq 0$ sein kann und untersuchen die Folgen für die Schätzer aus den vorherigen Kapiteln.

Danach führen wir neue Informationen in Form von **Instrumentvariablen** ein und beschreiben die klassische **Instrumentvariablen-Schätzung** (IV) ähnlich zur MM-Schätzung.

Danach befassen wir uns mit verschiedenen Tests:

- Heteroskedastie-robuste Tests,
- Tests auf überidentifizierende Restriktionen,
- „Tests auf Exogenität“,
- Hausmantests.

Korrelation zwischen Regressoren und Störtermen

Wir betrachten die folgenden Fälle, bei denen $\boldsymbol{X}_t$ mit u_t korreliert sein kann.

Erster Fall: Omitted Variable Bias (OVB)

Es gibt Variablen, die $\boldsymbol{y}$ erklären und die mit den bereits berücksichtigten $\boldsymbol{X}$ korreliert sind, die wir jedoch nicht beobachten (können). Vgl. (3.14).

Betrachte z.B. Bildungsrenditen: Sei y_t das Einkommen, b_t eine geeignete Funktion des Geburtsjahres und s_t die in der Schule verbrachten Jahre.

Durchführen einer OLS-Regression für

$$y_t = \alpha + s_t\beta + b_t\gamma + u_t,$$

wobei $t = 1, \ldots, n$, kann zur inkonsistenten Schätzung des „wahren“ Effekts von s_t führen, da sowohl s_t als auch u_t durch dieselbe unbeobachtbare Variable („Fähigkeiten“) beeinflusst sein kann. Also ist

$$E(u_t|s_t) \neq 0.$$

Fehler in Variablen

Zweiter Fall: Sei

$$\boldsymbol{y} = \boldsymbol{x}^* \beta + \boldsymbol{u},$$

wobei $\boldsymbol{x}^*$ der Einfachheit halber $n \times 1$ sei.

Wir beobachten jedoch nicht $\boldsymbol{x}^*$, sondern $\boldsymbol{x} = \boldsymbol{x}^* + \boldsymbol{v}$, wobei die $\boldsymbol{v}$ **Messfehler** darstellen.

Es seien $\{u_i, v_i\}$ u.i.v., $E(u_i) = E(v_i) = 0$ und

$$Var(u_i, v_i) = \begin{bmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{bmatrix}$$

sowie

$$\frac{\boldsymbol{x}^{*\top} \boldsymbol{x}^*}{n} \xrightarrow{p} M > 0$$

und

$$Cov(x_i^*, v_i) = E(x_i^* v_i) = 0 \tag{7.4}$$

Das Gesetz der großen Zahlen ($\{u_i v_i\}$ und $\{v_i^2\}$ sind u.i.v.) liefert

$$\begin{aligned} \frac{\boldsymbol{u}^\top \boldsymbol{v}}{n} &= \frac{1}{n}\sum_{i=1}^{n} u_i v_i \xrightarrow{p} E(u_i v_i) = \sigma_{uv}, \\ \frac{\boldsymbol{v}^\top \boldsymbol{v}}{n} &= \frac{1}{n}\sum_{i=1}^{n} v_i^2 \xrightarrow{p} E(v_i^2) = \sigma_v^2. \end{aligned}$$

Des Weiteren gilt mit dem GGZ (4.32) und (7.4), dass

$$\frac{\boldsymbol{x}^{*\top} \boldsymbol{v}}{n} \xrightarrow{p} E(x_i^* v_i) = 0$$

Analog zeigt man, dass unter geeigneten Exogenitätsannahmen wie üblich

$$\frac{\boldsymbol{x}^{*\top} \boldsymbol{u}}{n} \xrightarrow{p} 0. \tag{7.5}$$

Also ist

$$
\begin{aligned}
\operatorname{plim}_{n\to\infty} \hat{\beta} &= \operatorname{plim} \frac{\boldsymbol{x}^\top \boldsymbol{y}}{\boldsymbol{x}^\top \boldsymbol{x}} \\
&= \frac{\operatorname{plim} \left(\frac{(\boldsymbol{x}^* + \boldsymbol{v})^\top (\boldsymbol{x}^* \beta + \boldsymbol{u})}{n} \right)}{\operatorname{plim} \left(\frac{(\boldsymbol{x}^* + \boldsymbol{v})^\top (\boldsymbol{x}^* + \boldsymbol{v})}{n} \right)} \\
&= \frac{\operatorname{plim} \left(\frac{\boldsymbol{x}^{*\top} \boldsymbol{x}^*}{n} \right) \beta + \operatorname{plim} \left(\frac{\boldsymbol{v}^\top \boldsymbol{x}^*}{n} \right) + \operatorname{plim} \left(\frac{\boldsymbol{x}^{*\top} \boldsymbol{u}}{n} \right) \beta + \operatorname{plim} \left(\frac{\boldsymbol{v}^\top \boldsymbol{u}}{n} \right)}{\operatorname{plim} \left(\frac{\boldsymbol{x}^{*\top} \boldsymbol{x}^*}{n} \right) + 2 \operatorname{plim} \left(\frac{\boldsymbol{x}^{*\top} \boldsymbol{v}}{n} \right) + \operatorname{plim} \left(\frac{\boldsymbol{v}^\top \boldsymbol{v}}{n} \right)} \\
&= \frac{M\beta + \sigma_{uv}}{M + \sigma_v^2} \\
&= \beta + \frac{\sigma_{uv} - \sigma_v^2 \beta}{M + \sigma_v^2}.
\end{aligned}
$$

Demzufolge ist der OLS-Schätzer nicht konsistent, wenn die erklärenden Variablen fehlerhaft gemessen werden.

Beachte zudem, dass der $\operatorname{plim} \hat{\beta}$ näher an null liegt als β, wenn $\sigma_{uv} = 0$: Messfehler führen zum so genannten **Attenuation Bias**.

Simultane Gleichungssysteme

Ein weiterer Fall, bei dem die Annahme $E(u_t|\boldsymbol{X}_t) = 0$ verletzt sein könnte, tritt bei **simultanen Gleichungssystemen** ein:

$$\begin{aligned} y_{1t} &= \alpha_1 + x_t\beta_1 + y_{2t}\gamma + u_{1t}, \\ y_{2t} &= \alpha_2 + x_t\beta_2 + u_{2t}, \end{aligned}$$

wobei $t = 1, \ldots, n$. Beide y_{1t} und y_{2t} sind endogen, jedoch ist y_{2t} ebenfalls ein Regressor für y_{1t}.

Wenn $Cov(u_{1t}, u_{2t}) \neq 0$, dann $Cov(u_{1t}, y_{2t}) \neq 0$ und eine Regression von $\boldsymbol{y}_1$ auf $\boldsymbol{\iota}$, $\boldsymbol{x}$ und $\boldsymbol{y}_2$ würde zu inkonsistenten OLS-Schätzern führen.

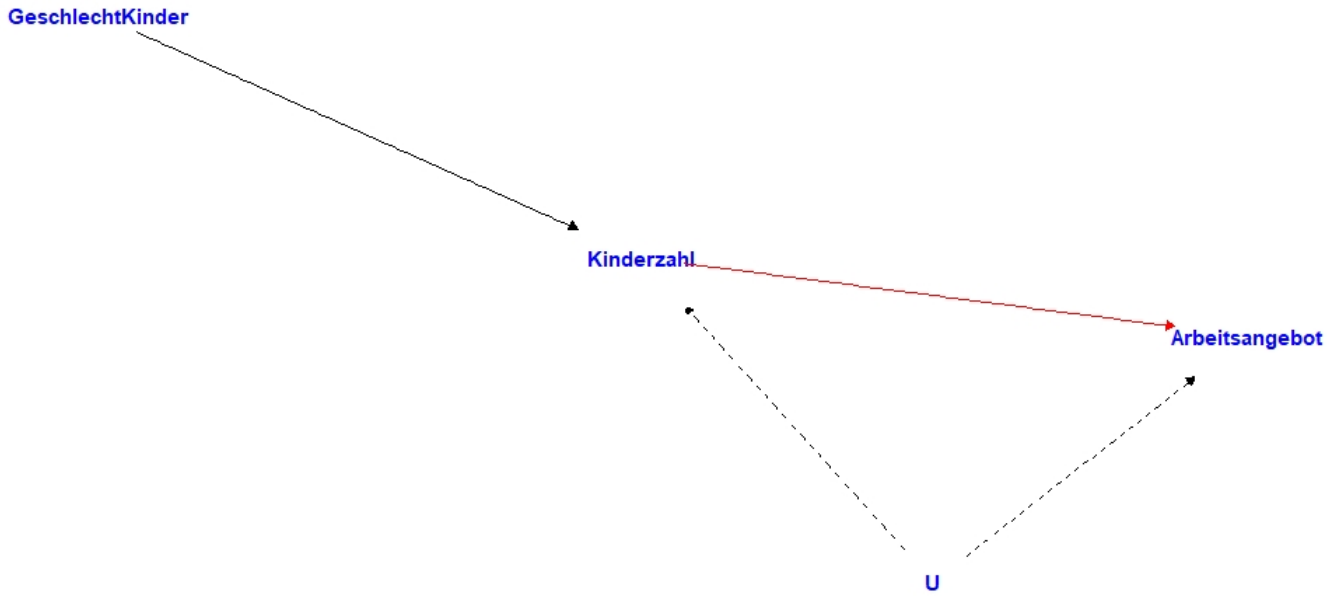
Wir befassen uns hier nicht weiter mit simultanen Gleichungssystemen.

Stichprobenverzerrungen führen ebenfalls zu Inkonsistenz.

Wenn die Stichprobe zufällig in eine **Behandlungs-** und eine **Kontrollgruppe** geteilt wird, kann der Effekt einer **Behandlung** valide geschätzt werden. Wird diese Zuteilung nicht randomisiert vorgenommen, sondern z.B. durch die Wahl der Probanden, kommt es zu Stichprobenverzerrungen.

Manchmal machen jedoch **natürliche Experimente** Beobachtungen verfügbar, die Beobachtungen aus einem Experiment mit zufällig zugewiesenen Behandlungen ähneln.

In unserem Beispiel der Bildungsrenditen könnte man möglicherweise gesetzliche Unterschiede zwischen verschiedenen Bundesstaaten ausnutzen, wenn Schüler in einigen Staaten länger in der Schule bleiben müssen als in anderen. Diese Information kann zur Konstruktion eines Instrumentes verwendet werden, um β konsistent zu schätzen.



Angrist, J. and W. Evans, "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size," American Economic Review 88(3), 1998, 450-477

Instrumentvariablenschätzung

Es gebe zusätzliche Informationen in Form von Instrumenten $\boldsymbol{W}_t$, die mit $\boldsymbol{X}_t$, jedoch nicht mit u_t korreliert sind.

Sei $\boldsymbol{W}$ eine $n \times l$ Matrix, welche die Beobachtungen der Instrumente enthält.

Annahme 7.1

$$E(\boldsymbol{W}_t^\top \boldsymbol{u}_t) = \mathbf{0}. \tag{7.6}$$

Erweitere Annahme 4.4 auf

Annahme 7.2 Die Folge $\{y_t, \boldsymbol{X}_t, \boldsymbol{W}_t\}$ ist gemeinsam u.i.v.

Dann gilt mit dem GGZ (4.32), dass

$$\frac{\boldsymbol{W}^\top \boldsymbol{X}}{n} \xrightarrow{p} E(\boldsymbol{W}_t^\top \boldsymbol{X}_t) =: \boldsymbol{S}_{\boldsymbol{W}^\top \boldsymbol{X}}$$

Annahme 7.3 $\boldsymbol{S}_{\boldsymbol{W}^\top \boldsymbol{X}}$ hat vollen Spaltenrang.

Die Annahme impliziert $l \geq k$ (vgl. Matrixreader, Definition 13).

Analog gilt

$$\frac{\boldsymbol{W}^\top \boldsymbol{W}}{n} \xrightarrow{p} E(\boldsymbol{W}_t^\top \boldsymbol{W}_t) =: \boldsymbol{S}_{\boldsymbol{W}^\top \boldsymbol{W}} > 0. \tag{7.7}$$

Natürlich können (und sollten) vorbestimmte Regressoren (also solche Regressoren $\boldsymbol{X}_1 \subset \boldsymbol{X}$, für die wie in (4.44) $E(\boldsymbol{X}_{1t}^\top u_t) = \mathbf{0}$ gilt) als Instrumente verwendet werden.

Wenn $l = k$, ist das Modell exakt identifiziert (**just identified**). Wenn $l > k$ gilt, ist das Modell überidentifiziert (**overidentified**).

Der **IV-Schätzer** ist gegeben durch

$$\begin{aligned}\hat{\beta}_{\text{IV}} &= \arg\min_{\beta} \left\{ (\boldsymbol{y} - \boldsymbol{X}\beta)^\top \boldsymbol{P}_{\boldsymbol{W}} (\boldsymbol{y} - \boldsymbol{X}\beta) \right\} \\ &= \left(\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X} \right)^{-1} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{y}. \end{aligned} \tag{7.8}$$

(7.8) zeigt man analog zu (1.21).

Wenn entweder $\boldsymbol{W} = \boldsymbol{X}$ oder $l = n$ gilt, ist $\hat{\beta}_{\text{IV}} = \hat{\beta}_{\text{OLS}}$.

Wie in (3.3) gilt $\hat{\beta}_{\text{IV}} = \beta + \left(\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{u}$.

Wir sehen

$$\begin{aligned}\operatorname{plim}_{n\to\infty} \left(\frac{\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X}}{n}\right) &= \operatorname{plim}_{n\to\infty} \left\{ \left(\frac{\boldsymbol{X}^\top \boldsymbol{W}}{n}\right) \left(\frac{\boldsymbol{W}^\top \boldsymbol{W}}{n}\right)^{-1} \left(\frac{\boldsymbol{W}^\top \boldsymbol{X}}{n}\right) \right\} \\ &= \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{W}} \boldsymbol{S}^{-1}_{\boldsymbol{W}^\top \boldsymbol{W}} \boldsymbol{S}_{\boldsymbol{W}^\top \boldsymbol{X}} =: \tilde{\boldsymbol{S}}. \end{aligned} \tag{7.9}$$

Annahme 7.2 und (7.6) liefern mit dem GGZ (4.32), dass

$$\operatorname{plim}_{n\to\infty} \frac{\boldsymbol{W}^\top \boldsymbol{u}}{n} = \boldsymbol{0}.$$

Daraus folgt

$$\begin{aligned}\operatorname{plim}_{n\to\infty} \left(\frac{\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{u}}{n}\right) &= \operatorname{plim}_{n\to\infty} \left(\frac{\boldsymbol{X}^\top \boldsymbol{W} (\boldsymbol{W}^\top \boldsymbol{W})^{-1} \boldsymbol{W}^\top \boldsymbol{u}}{n}\right) \\ &= \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{W}} \boldsymbol{S}^{-1}_{\boldsymbol{W}^\top \boldsymbol{W}} \boldsymbol{0} = \boldsymbol{0}. \end{aligned} \tag{7.10}$$

Damit gilt

$$\begin{aligned} \text{plim}_{n\to\infty} \hat{\beta}_{IV} &= \beta + \text{plim}_{n\to\infty} \left(\frac{\boldsymbol{X}^\top \boldsymbol{P}_W \boldsymbol{X}}{n} \right) \text{plim}_{n\to\infty} \left(\frac{\boldsymbol{X}^\top \boldsymbol{P}_W \boldsymbol{u}}{n} \right) \\ &= \beta, \end{aligned}$$

d.h. der IV-Schätzer ist unter Annahme (7.6) konsistent.

Die folgende Annahme 7.4 für den IV-Fall entspricht z.B. Annahme 4.6 mit (4.50).

Annahme 7.4

$$Var(\boldsymbol{W}_t^\top u_t) = \sigma^2 \boldsymbol{S}_{\boldsymbol{W}^\top \boldsymbol{W}}$$

Dann liefert der ZGWS (4.39) mit Annahme 7.2, dass

$$n^{1/2} \left(\frac{\boldsymbol{W}^\top \boldsymbol{u}}{n} \right) = n^{1/2} \left(\frac{\sum_{t=1}^n \boldsymbol{W}_t^\top u_t}{n} \right) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \sigma^2 \boldsymbol{S}_{\boldsymbol{W}^\top \boldsymbol{W}}\right) \qquad (7.11)$$

Die asymptotische Verteilung folgt dann wie in (4.53) mit (7.11) und (7.9) als

$$\begin{aligned} n^{1/2}(\hat{\beta}_{\text{IV}} - \beta) &= \left(\frac{\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X}}{n}\right)^{-1} \frac{\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{u}}{n^{1/2}} \\ &\xrightarrow{d} \tilde{\boldsymbol{S}}^{-1} \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{W}} \boldsymbol{S}_{\boldsymbol{W}^\top \boldsymbol{W}}^{-1} \mathcal{N}\left(\boldsymbol{0}, \sigma^2 \boldsymbol{S}_{\boldsymbol{W}^\top \boldsymbol{W}}\right) \\ &\sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{V}_{\text{IV}}\right), \end{aligned} \tag{7.12}$$

wobei mit (3.6)

$$\begin{aligned} \boldsymbol{V}_{\text{IV}} &= \tilde{\boldsymbol{S}}^{-1} \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{W}} \boldsymbol{S}_{\boldsymbol{W}^\top \boldsymbol{W}}^{-1} \sigma^2 \boldsymbol{S}_{\boldsymbol{W}^\top \boldsymbol{W}} \left(\tilde{\boldsymbol{S}}^{-1} \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{W}} \boldsymbol{S}_{\boldsymbol{W}^\top \boldsymbol{W}}^{-1}\right)^\top \\ &= \sigma^2 \tilde{\boldsymbol{S}}^{-1} \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{W}} \boldsymbol{S}_{\boldsymbol{W}^\top \boldsymbol{W}}^{-1} \boldsymbol{S}_{\boldsymbol{W}^\top \boldsymbol{X}} \tilde{\boldsymbol{S}}^{-1} = \sigma^2 \tilde{\boldsymbol{S}}^{-1} \tilde{\boldsymbol{S}} \tilde{\boldsymbol{S}}^{-1} \\ &= \sigma^2 \tilde{\boldsymbol{S}}^{-1}, \end{aligned}$$

was man auch schreiben kann als $\sigma^2 \operatorname{plim}_{n\to\infty} \left(\frac{\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X}}{n}\right)^{-1}$.

2SLS

In der Praxis nennt man (7.8) vor allem IV, wenn $l = k$ und sonst **two-stage least squares**: 2SLS-Schätzer.

Dieser Name kann wie folgt erklärt werden. Betrachte

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\beta + \boldsymbol{u}, \\ \boldsymbol{X} &= \boldsymbol{W}\boldsymbol{\Pi} + \boldsymbol{E}. \end{aligned}$$

In der **ersten Stufe** (first-stage) regressieren wir jede Spalte von $\boldsymbol{X}$ auf die Spalten von $\boldsymbol{W}$:

$$\widehat{\boldsymbol{\Pi}} = (\boldsymbol{W}^\top \boldsymbol{W})^{-1} \boldsymbol{W}^\top \boldsymbol{X}.$$

In der **zweiten Stufe** (second-stage) regressieren wir $\boldsymbol{y}$ auf

$$\widehat{\boldsymbol{X}} = \boldsymbol{W}\widehat{\boldsymbol{\Pi}}.$$

Damit erhalten wir

$$\begin{aligned}
\hat{\beta}_{\text{2SLS}} &= \left(\widehat{\boldsymbol{X}}^\top \widehat{\boldsymbol{X}}\right)^{-1} \widehat{\boldsymbol{X}}^\top \boldsymbol{y} \\
&= \left(\widehat{\boldsymbol{\Pi}}^\top \boldsymbol{W}^\top \boldsymbol{W} \widehat{\boldsymbol{\Pi}}\right)^{-1} \widehat{\boldsymbol{\Pi}}^\top \boldsymbol{W}^\top \boldsymbol{y} \\
&= \left(\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{y} \\
&= \hat{\beta}_{\text{IV}}
\end{aligned}$$

Hypothesentests

Tests können anhand der asymptotischen Verteilung (7.12) konstruiert werden:

$$n^{1/2}(\hat{\beta}_{\text{IV}} - \beta) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{V}_{\text{IV}})$$

Eine Teststatistik für $H_0 : \boldsymbol{R}^\top \beta = \boldsymbol{r}$, wobei die $k \times q$ Matrix $\boldsymbol{R}$ vollen Spaltenrang hat, ist analog zu (4.62) gegeben durch

$$\begin{aligned} \tau &= n(\boldsymbol{R}^\top \hat{\beta}_{\text{IV}} - \boldsymbol{r})^\top (\boldsymbol{R}^\top \widehat{\boldsymbol{V}}_{\text{IV}} \boldsymbol{R})^{-1} (\boldsymbol{R}^\top \hat{\beta}_{\text{IV}} - \boldsymbol{r}) \\ &= \frac{(\boldsymbol{R}^\top \hat{\beta}_{\text{IV}} - \boldsymbol{r})^\top \left\{\boldsymbol{R}^\top (\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1} \boldsymbol{R}\right\}^{-1} (\boldsymbol{R}^\top \hat{\beta}_{\text{IV}} - \boldsymbol{r})}{\hat{\sigma}^2} \end{aligned} \tag{7.13}$$

wobei

$$\hat{\sigma}^2 = \frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}_{\text{IV}})^\top (\boldsymbol{y} - \boldsymbol{X}\hat{\beta}_{\text{IV}})}{n}$$

Unter H_0 ist $\tau \xrightarrow{d} \chi^2_q$. Wie gewöhnlich verwerfen wir H_0, wenn τ den kritischen Wert, das $1 - \alpha$-Quantil der χ^2_q-Verteilung, übersteigt.

Testen von überidentifizierenden Restriktionen

Nehmen wir an, dass $\boldsymbol{W} = (\boldsymbol{W}_1, \boldsymbol{W}_2)$ wobei $\boldsymbol{W}_1$ k Spalten und $\boldsymbol{W}_2$ $l - k$ Spalten hat. Der Grad der Überidentifizierung ist gegeben durch $l - k$.

Um β konsistent zu schätzen, benötigen wir lediglich $\boldsymbol{W}_1$. Wir können das Modell formulieren als

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\beta + \boldsymbol{W}_2\gamma + \boldsymbol{u}, && (7.14) \\ \boldsymbol{X} &= \boldsymbol{W}\boldsymbol{\Pi} + \boldsymbol{E}, && (7.15) \end{aligned}$$

und die überidentifizierenden Restriktionen H_0: $\gamma = \mathbf{0}$ testen.

Definiere $\hat{\boldsymbol{u}}_{\text{IV}} = \boldsymbol{y} - \boldsymbol{X}\hat{\beta}_{\text{IV}}$. Eine Teststatistik (von der man zeigen kann, dass sie ein Spezialfall von (7.13) ist) für H_0 ist

$$\tau_{\text{IV}} = \frac{n\hat{\boldsymbol{u}}_{\text{IV}}^\top \boldsymbol{P}_{\{\boldsymbol{P}_{\boldsymbol{W}}(\boldsymbol{X},\boldsymbol{W}_2)\}}\hat{\boldsymbol{u}}_{\text{IV}}}{\hat{\boldsymbol{u}}_{\text{IV}}^\top \hat{\boldsymbol{u}}_{\text{IV}}} = \frac{n\hat{\boldsymbol{u}}_{\text{IV}}^\top \boldsymbol{P}_{\boldsymbol{W}}\hat{\boldsymbol{u}}_{\text{IV}}}{\hat{\boldsymbol{u}}_{\text{IV}}^\top \hat{\boldsymbol{u}}_{\text{IV}}}.$$

Die zweite Gleichung folgt aus

$$\boldsymbol{P}_{\{\boldsymbol{P}_{\boldsymbol{W}}(\boldsymbol{X},\boldsymbol{W}_2)\}} = \boldsymbol{P}_{\boldsymbol{W}},$$

da

$$\boldsymbol{P}_{\boldsymbol{W}}(\boldsymbol{X},\boldsymbol{W}_2) = \boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top(\boldsymbol{X},\boldsymbol{W}_2) = \boldsymbol{W}\boldsymbol{C},$$

wobei $\boldsymbol{C}$, welches implizit definiert ist, nicht singulär ist, ähnlich zu Annahme 7.3. Daher gilt wegen, wie in (2.13),

$$\boldsymbol{P}_{\boldsymbol{W}\boldsymbol{C}} = \boldsymbol{W}\boldsymbol{C}((\boldsymbol{W}\boldsymbol{C})^\top\boldsymbol{W}\boldsymbol{C})^{-1}(\boldsymbol{W}\boldsymbol{C})^\top = \boldsymbol{W}\boldsymbol{C}\boldsymbol{C}^{-1}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}(\boldsymbol{C}^\top)^{-1}\boldsymbol{C}^\top\boldsymbol{W}^\top,$$

dass

$$\boldsymbol{P}_{\{\boldsymbol{P}_{\boldsymbol{W}}(\boldsymbol{X},\boldsymbol{W}_2)\}} = \boldsymbol{P}_{\boldsymbol{W}\boldsymbol{C}} = \boldsymbol{P}_{\boldsymbol{W}}.$$

Unter $H_0 : \boldsymbol{\gamma} = \boldsymbol{0}$ hat die **Teststatistik für überidentifizierende Restriktionen** eine asymptotische Chi-Quadrat-Verteilung:

$$\tau_{\text{IV}} = \frac{n\hat{\boldsymbol{u}}_{\text{IV}}^\top\boldsymbol{P}_{\boldsymbol{W}}\hat{\boldsymbol{u}}_{\text{IV}}}{\hat{\boldsymbol{u}}_{\text{IV}}^\top\hat{\boldsymbol{u}}_{\text{IV}}} \xrightarrow{d} \chi^2_{l-k}. \tag{7.16}$$

Der Test ist auch bekannt als **Sargan-Test**.

`dwh Hansen IV.R`

Diese Tests auf Überidentifikation werden ebenfalls **Tests für die Validität der Instrumente** genannt. Wie sollten wir eine Ablehnung interpretieren?

- Sind Instrumente mit dem Fehlerterm korreliert und daher unzulässig?
- Wurden einige Variablen fälschlicherweise ausgelassen?

Im Folgenden betrachten wir einen Spezialfall genauer.

Ein Spezialfall: „Test auf Exogenität“

Betrachten wir ein Modell, bei dem wir nicht sicher sind, ob einige erklärende Variablen exogen sind. Sei

$$\boldsymbol{y} = \boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2 + \boldsymbol{u}, \tag{7.17}$$

wobei wir annehmen, dass $E(\boldsymbol{X}_{1t}^\top u_t) = \boldsymbol{0}$, aber wir uns nicht sicher sind, ob $E(\boldsymbol{X}_{2t}^\top u_t) = \boldsymbol{0}$. Daher sind die Variablen in $\boldsymbol{X}_2$ potentiell endogen.

Sei $\boldsymbol{X}_1$ $n \times k_1$ und $\boldsymbol{X}_2$ $n \times k_2$. Daher ist $k = k_1 + k_2$. Wir haben l Instrumente $\boldsymbol{W}$, bestehend aus den k_1 Variablen in $\boldsymbol{W}_1 = \boldsymbol{X}_1$ und k_2 Instrumente $\boldsymbol{W}_2$, die verwendet werden könnten, wenn $\boldsymbol{X}_2$ tatsächlich endogen ist. Außerdem haben wir noch $l - k$ überschüssige Instrumente $\boldsymbol{W}_3$.

Sei

$$\boldsymbol{X}_2 = \boldsymbol{W}\boldsymbol{\Pi} + \boldsymbol{E}. \tag{7.18}$$

Wenn $\boldsymbol{X}_2$ endogen ist, dann sind die Elemente von $\boldsymbol{E}$ mit $\boldsymbol{u}$ korreliert. Modelliere $\boldsymbol{u}$ durch $\boldsymbol{E}$,

$$\boldsymbol{u} = \boldsymbol{E}\boldsymbol{\lambda} + \varepsilon,$$

wobei wir annehmen können (lineare Projektion), dass $\boldsymbol{E}$ und ε unkorreliert sind. Schreibe (7.17) dann als

$$\boldsymbol{y} = \boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2 + \boldsymbol{E}\boldsymbol{\lambda} + \varepsilon, \tag{7.19}$$

wobei somit nun $\boldsymbol{X}_2$ und ε unkorreliert sind.

Um

$$E(u_t|\boldsymbol{X}_{2t}) = E(u_t|\boldsymbol{E}_t) = \boldsymbol{E}_t\boldsymbol{\lambda} = 0,$$

$t = 1, \ldots, n$ zu testen, testen wir also $\boldsymbol{\lambda} = \boldsymbol{0}$.

Allerdings ist $\boldsymbol{E}$ nicht beobachtbar. Wir können dennoch weiter machen.

Betrachte zunächst den exakt identifizierten Fall $l = k$. Für jede Spalte in $\boldsymbol{X}_2$ haben wir ein Instrument in $\boldsymbol{W}_2$ und es gibt keine weiteren Instrumente in $\boldsymbol{W}_3$.

Schreibe das Modell (7.19) mit $\boldsymbol{E} = \boldsymbol{X}_2 - \boldsymbol{W}\boldsymbol{\Pi}$ als

$$\boldsymbol{y} = \boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2(\beta_2 + \boldsymbol{\lambda}) - \boldsymbol{W}\boldsymbol{\Pi}\boldsymbol{\lambda} + \varepsilon, \tag{7.20}$$

wobei

$$\boldsymbol{W}\boldsymbol{\Pi}\boldsymbol{\lambda} = (\boldsymbol{X}_1, \boldsymbol{W}_2)\begin{pmatrix}\boldsymbol{\Pi}_1 \\ \boldsymbol{\Pi}_2\end{pmatrix}\boldsymbol{\lambda} = \boldsymbol{X}_1\boldsymbol{\Pi}_1\boldsymbol{\lambda} + \boldsymbol{W}_2\boldsymbol{\Pi}_2\boldsymbol{\lambda}. \tag{7.21}$$

Also kann (7.20) geschrieben werden als

$$\begin{aligned}\boldsymbol{y} &= \boldsymbol{X}_1(\beta_1 - \boldsymbol{\Pi}_1\boldsymbol{\lambda}) + \boldsymbol{X}_2(\beta_2 + \boldsymbol{\lambda}) - \boldsymbol{W}_2\boldsymbol{\Pi}_2\boldsymbol{\lambda} + \varepsilon \\ &= \boldsymbol{X}_1\beta_1^* + \boldsymbol{X}_2\beta_2^* + \boldsymbol{W}_2\gamma + \varepsilon,\end{aligned} \tag{7.22}$$

wobei γ nicht restringiert ist, da $\boldsymbol{\Pi}_2$ für $l = k$ quadratisch und nichtsingulär ist.

Testen von $\gamma = \mathbf{0}$ in (7.22) ist lediglich ein Test für die Validität der Instrumente, welche nun gegeben sind durch $(\boldsymbol{W}_1, \boldsymbol{X}_2, \boldsymbol{W}_2)$.

Die Teststatistik kann ähnlich zu (7.16) formuliert werden, wobei nun der restringierte IV-Schätzer der OLS-Schätzer einer Regression auf $\boldsymbol{X}_1$ und $\boldsymbol{X}_2$ ist.

Die Teststatistik und deren asymptotische Verteilung unter $H_0 : \boldsymbol{\lambda} = \mathbf{0}$ oder äquivalent $\gamma = \mathbf{0}$ sind gegeben durch

$$\tau = \frac{\hat{\boldsymbol{u}}_{\text{OLS}}^\top \boldsymbol{P}_{(\boldsymbol{X}, \boldsymbol{W}_2)} \hat{\boldsymbol{u}}_{\text{OLS}}}{\hat{\sigma}^2} = \frac{\hat{\boldsymbol{u}}_{\text{OLS}}^\top \boldsymbol{P}_{\{(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{W}_2\}} \hat{\boldsymbol{u}}_{\text{OLS}}}{\hat{\sigma}^2} \xrightarrow{d} \chi^2_{k_2}. \tag{7.23}$$

Die zweite Gleichung folgt mit dem FWL-Theorem 2.4 und (2.9), da

$$\boldsymbol{P}_{(\boldsymbol{X}, \boldsymbol{W}_2)} \hat{\boldsymbol{u}}_{\text{OLS}} = \boldsymbol{P}_{\{(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{W}_2\}} (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}}) \hat{\boldsymbol{u}}_{\text{OLS}} = \boldsymbol{P}_{\{(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{W}_2\}} \hat{\boldsymbol{u}}_{\text{OLS}}$$

Die Anzahl der Freiheitsgrade entspricht der Anzahl der zusätzlichen Instrumente in $(\boldsymbol{W}_1, \boldsymbol{X}_2, \boldsymbol{W}_2)$, also k_2.

`dwh Hansen IV.R`

Durbin-Wu-Hausman-Test

Im Fall von Überidentifikation mit $l > k$ enthält $\boldsymbol{W}_3$ zusätzliche Instrumente.

Wir können $\boldsymbol{E}$ in (7.19) durch $\widehat{\boldsymbol{E}}$, die Residuen der Regression der ersten Stufe aus (7.18), ersetzen, da diese in Wahrscheinlichkeit gegen $\boldsymbol{E}$ konvergieren.

Wir können $\gamma = \mathbf{0}$ in der Regression

$$\boldsymbol{y} = \boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2 + \widehat{\boldsymbol{E}}\gamma + \text{residuals}. \tag{7.24}$$

testen. Die Teststatistik und die asymptotische Verteilung sind gegeben durch

$$\tau = \frac{\hat{\boldsymbol{u}}_{\text{OLS}}^\top \boldsymbol{P}_{(\boldsymbol{X},\widehat{\boldsymbol{E}})} \hat{\boldsymbol{u}}_{\text{OLS}}}{\hat{\sigma}^2} = \frac{\hat{\boldsymbol{u}}_{\text{OLS}}^\top \boldsymbol{P}_{(\boldsymbol{X},\boldsymbol{W}\widehat{\boldsymbol{\Pi}})} \hat{\boldsymbol{u}}_{\text{OLS}}}{\hat{\sigma}^2} = \frac{\hat{\boldsymbol{u}}_{\text{OLS}}^\top \boldsymbol{P}_{\{(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{W}\widehat{\boldsymbol{\Pi}}\}} \hat{\boldsymbol{u}}_{\text{OLS}}}{\hat{\sigma}^2} \xrightarrow{d} \chi^2_{k_2}. \tag{7.25}$$

Im exakt identifizierten Fall entspricht diese Teststatistik der in (7.23), da $\widehat{\boldsymbol{\Pi}}$ dann quadratisch und invertierbar ist, vgl. (2.13). `dwh Hansen IV.R`

Der Kontrastvektor

Betrachten wir den Fall, ähnlich zum vorherigen Abschnitt, bei dem alle Variablen in $\boldsymbol{X}$ potenziell endogen sind und wir für jede ein Instrument haben, das in der $n \times k$ Matrix $\boldsymbol{W}$ enthalten ist. Da $l = k$, ist das Modell exakt identifiziert, wenn nicht davon ausgegangen werden kann, dass die erklärenden Variablen exogen sind.

Die Teststatistik auf Exogenität kann wie in (7.23) formuliert werden

$$\tau = \frac{\hat{\boldsymbol{u}}_{\text{OLS}}^{\top} \boldsymbol{P}_{(\boldsymbol{X},\boldsymbol{W})} \hat{\boldsymbol{u}}_{\text{OLS}}}{\hat{\sigma}^2} \xrightarrow{d} \chi_k^2.$$

Diese Teststatistik kann ebenfalls auf Grundlage des **Kontrastvektors**, der Differenz zwischen $\hat{\beta}_{\text{OLS}}$ und $\hat{\beta}_{\text{IV}}$, konstruiert werden.

Da

$$\begin{aligned}\hat{\beta}_{\text{IV}} &= (\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{y}, \\ \hat{\beta}_{\text{OLS}} &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y},\end{aligned}$$

erhalten wir

$$\begin{aligned}\hat{\beta}_{\text{IV}} - \hat{\beta}_{\text{OLS}} &= (\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}}) \boldsymbol{y} \\ &= (\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \hat{\boldsymbol{u}}_{\text{OLS}}. \qquad (7.26)\end{aligned}$$

Wir sehen

$$\boldsymbol{P}_{(\boldsymbol{X},\boldsymbol{W})} = \boldsymbol{P}_{(\boldsymbol{X},\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X})}$$

Dies gilt, da

$$(\boldsymbol{X}, \boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}) = (\boldsymbol{X}, \boldsymbol{W})\boldsymbol{A},$$
$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{I}_k & \mathbf{O} \\ \mathbf{O} & (\boldsymbol{W}^\top \boldsymbol{W})^{-1}\boldsymbol{W}^\top \boldsymbol{X} \end{pmatrix},$$

wobei $\boldsymbol{A}$ eine nichtsinguläre $2k \times 2k$ Matrix ist.

Daher kann $\hat{\sigma}^2\tau$ umgeschrieben werden in

$$\begin{aligned}\hat{\sigma}^2\tau &= \hat{\boldsymbol{u}}_{\text{OLS}}^\top \boldsymbol{P}_{(\boldsymbol{X}, \boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X})}\hat{\boldsymbol{u}}_{\text{OLS}} \\ &= \hat{\boldsymbol{u}}_{\text{OLS}}^\top (\boldsymbol{X}, \boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X}) \begin{pmatrix} \boldsymbol{X}^\top \boldsymbol{X} & \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X} \\ \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X} & \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{X} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{X}^\top \\ \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \end{pmatrix} \hat{\boldsymbol{u}}_{\text{OLS}},\end{aligned}$$

Und da $\boldsymbol{X}^\top \hat{\boldsymbol{u}}_{\text{OLS}} = \boldsymbol{0}$,

$$\hat{\sigma}^2\tau = \hat{\boldsymbol{u}}_{\text{OLS}}^\top \boldsymbol{P_W X}(\boldsymbol{O}, \boldsymbol{I}_k) \begin{pmatrix} \boldsymbol{X}^\top \boldsymbol{X} & \boldsymbol{X}^\top \boldsymbol{P_W X} \\ \boldsymbol{X}^\top \boldsymbol{P_W X} & \boldsymbol{X}^\top \boldsymbol{P_W X} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{O} \\ \boldsymbol{I}_k \end{pmatrix} \boldsymbol{X}^\top \boldsymbol{P_W} \hat{\boldsymbol{u}}_{\text{OLS}} \quad (7.27)$$

Hierbei ist

$$(\boldsymbol{O}, \boldsymbol{I}_k) \begin{pmatrix} \boldsymbol{X}^\top \boldsymbol{X} & \boldsymbol{X}^\top \boldsymbol{P_W X} \\ \boldsymbol{X}^\top \boldsymbol{P_W X} & \boldsymbol{X}^\top \boldsymbol{P_W X} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{O} \\ \boldsymbol{I}_k \end{pmatrix}$$

das südöstliche Element der Inverse, mit Matrixreader, Resultat 23 gleich

$$\begin{aligned} & [\boldsymbol{X}^\top \boldsymbol{P_W X} - \boldsymbol{X}^\top \boldsymbol{P_W X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{P_W X}]^{-1} \\ = \quad & (\boldsymbol{X}^\top \boldsymbol{P_W X})^{-1} [(\boldsymbol{X}^\top \boldsymbol{P_W X})^{-1} - (\boldsymbol{X}^\top \boldsymbol{X})^{-1}]^{-1} (\boldsymbol{X}^\top \boldsymbol{P_W X})^{-1} \end{aligned}$$

Gleichzeitig ist, wiederum mit Resultat 23, $[(\boldsymbol{X}^\top \boldsymbol{P_W X})^{-1} - (\boldsymbol{X}^\top \boldsymbol{X})^{-1}]^{-1}$ das südöstliche Element von

$$\begin{pmatrix} \boldsymbol{X}^\top \boldsymbol{X} & \boldsymbol{I}_k \\ \boldsymbol{I}_k & (\boldsymbol{X}^\top \boldsymbol{P_W X})^{-1} \end{pmatrix}^{-1}$$

Demnach können wir (7.27) schreiben als

$$
\begin{aligned}
& \hat{\boldsymbol{u}}_{\text{OLS}}^{\top} \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X} \left(\mathbf{O}, (\boldsymbol{X}^{\top} \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1}\right) \times \\
& \qquad \begin{pmatrix} \boldsymbol{X}^{\top}\boldsymbol{X} & \boldsymbol{I}_k \\ \boldsymbol{I}_k & (\boldsymbol{X}^{\top} \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{O} \\ (\boldsymbol{X}^{\top} \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1} \end{pmatrix} \boldsymbol{X}^{\top} \boldsymbol{P}_{\boldsymbol{W}} \hat{\boldsymbol{u}}_{\text{OLS}} \\
\overset{(7.26)}{=} \quad & (\hat{\beta}_{\text{IV}} - \hat{\beta}_{\text{OLS}})^{\top} (\mathbf{O}, \boldsymbol{I}_k) \begin{pmatrix} \boldsymbol{X}^{\top}\boldsymbol{X} & \boldsymbol{I}_k \\ \boldsymbol{I}_k & (\boldsymbol{X}^{\top} \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{O} \\ \boldsymbol{I}_k \end{pmatrix} (\hat{\beta}_{\text{IV}} - \hat{\beta}_{\text{OLS}}) \\
= \quad & (\hat{\beta}_{\text{IV}} - \hat{\beta}_{\text{OLS}})^{\top} \left\{ (\boldsymbol{X}^{\top} \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1} - (\boldsymbol{X}^{\top}\boldsymbol{X})^{-1} \right\}^{-1} (\hat{\beta}_{\text{IV}} - \hat{\beta}_{\text{OLS}})
\end{aligned}
$$

Wir können τ kann also mithilfe des Kontrastvektors, (4.61) und (7.12) schreiben als

$$\tau = n(\hat{\beta}_{\text{IV}} - \hat{\beta}_{\text{OLS}})^\top \left\{ \widehat{\boldsymbol{V}}_{\text{IV}} - \widehat{\boldsymbol{V}}_{\text{OLS}} \right\}^{-1} (\hat{\beta}_{\text{IV}} - \hat{\beta}_{\text{OLS}}),$$

(plus einem Term, der aufgrund der Wahl von $\hat{\sigma}^2$ in Wahrscheinlichkeit gegen null konvergiert , welche sich von der gewöhnlichen, für $\widehat{\boldsymbol{V}}_{\text{OLS}}$ verwendeten, unterscheidet—wegen (4.38) und unter der Nullhypothese würde dieser Unterschied die asymptotische Verteilung nicht beeinflussen.)

Wenn die Variablen in $\boldsymbol{X}$ tatsächlich exogen sind, dann ist OLS asymptotisch effizient und IV konsistent. Wenn die erklärenden Variablen endogen sind, dann ist OLS inkonsistent, jedoch ist IV immer noch konsistent.

Dies ist ein Beispiel für einen sogenannten **Hausman-Test**.

Hausman-Tests

Ein Hausman-Test kann in einem generellen Kontext mit zwei Schätzern $\hat{\boldsymbol{\theta}}$ und $\tilde{\boldsymbol{\theta}}$ formuliert werden, welche die folgenden Bedingungen erfüllen:

(i) $\hat{\boldsymbol{\theta}}$ ist asymptotisch effizient unter H_0 und inkonsistent unter H_1,

(ii) $\tilde{\boldsymbol{\theta}}$ ist sowohl unter H_0 als auch unter H_1 konsistent.

Da $\hat{\boldsymbol{\theta}}$ unter H_0 asymptotisch effizient ist, ist es asymptotisch unkorreliert mit allen konsistenten Schätzer von $\mathbf{0}$.

Insbesondere ist $\hat{\boldsymbol{\theta}}$ unter der Nullhypothese asymptotisch unkorreliert mit dem Kontrastvektor $\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}$, das bedeutet

$$n^{1/2} \begin{pmatrix} \hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \\ \hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}} \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \text{AVar}(\hat{\boldsymbol{\theta}}) & \mathbf{O} \\ \mathbf{O} & \text{AVar}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) \end{pmatrix}\right).$$

Daher gilt mit (3.6)

$$n^{1/2}\begin{pmatrix}\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\\ \tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}\end{pmatrix}=\begin{pmatrix}\boldsymbol{I} & \mathbf{O}\\ \boldsymbol{I} & -\boldsymbol{I}\end{pmatrix}n^{1/2}\begin{pmatrix}\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\\ \hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}}\end{pmatrix}$$
$$\xrightarrow{d}\mathcal{N}\left(\mathbf{0},\begin{pmatrix}\mathrm{AVar}(\hat{\boldsymbol{\theta}}) & \mathrm{AVar}(\hat{\boldsymbol{\theta}})\\ \mathrm{AVar}(\hat{\boldsymbol{\theta}}) & \mathrm{AVar}(\hat{\boldsymbol{\theta}})+\mathrm{AVar}(\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}})\end{pmatrix}\right).$$

Daraus folgt

$$\mathrm{AVar}(\tilde{\boldsymbol{\theta}})=\mathrm{AVar}(\hat{\boldsymbol{\theta}})+\mathrm{AVar}(\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}}),$$

so dass $\mathrm{AVar}(\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}})=\mathrm{AVar}(\tilde{\boldsymbol{\theta}})-\mathrm{AVar}(\hat{\boldsymbol{\theta}})$ bzw.

$$n^{1/2}(\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}})\xrightarrow{d}\mathcal{N}\left(\mathbf{0},\mathrm{AVar}(\tilde{\boldsymbol{\theta}})-\mathrm{AVar}(\hat{\boldsymbol{\theta}})\right).$$

Wir erhalten also unter der Nullhypothese

$$n(\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}})^{\top}\left\{\mathrm{AVar}(\tilde{\boldsymbol{\theta}})-\mathrm{AVar}(\hat{\boldsymbol{\theta}})\right\}^{-1}(\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}})\xrightarrow{d}\chi^2_k,$$

wenn $\boldsymbol{\theta}$ ein k Vektor und $\mathrm{AVar}(\hat{\boldsymbol{\theta}}-\tilde{\boldsymbol{\theta}})$ nichtsingulär ist.

Heteroskedastie-robuste Tests

Heteroskedastie-konsistente Kovarianzmatrix-Schätzung in Gegenwart von Instrumenten erfolgt ähnlich wie für KQ-Schätzung.

D.h., wenn $\boldsymbol{u}|\boldsymbol{W} \sim (\boldsymbol{0}, \boldsymbol{D})$ und $\boldsymbol{D}$ diagonal, jedoch nicht notwendigerweise gleich $\sigma^2\boldsymbol{I}$ ist, dann gilt unter den üblichen anderen Annahmen

$$n^{1/2}(\hat{\beta}_{\text{IV}} - \beta) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{V}_{\text{IV,rob}}),$$

wobei, analog zu (4.52),

$$\boldsymbol{V}_{\text{IV,rob}} = \text{plim}_{n\to\infty}\, n(\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{D} \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1}.$$

Um $\boldsymbol{W}^\top \boldsymbol{D} \boldsymbol{W}$ in $\boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{D} \boldsymbol{P}_{\boldsymbol{W}}$ zu schätzen, können wir, analog zu (4.65),

$$\boldsymbol{W}^\top \widehat{\boldsymbol{D}} \boldsymbol{W} = \sum_{t=1}^{n} \boldsymbol{W}_t^\top \boldsymbol{W}_t \hat{u}_{\text{IV},t}^2$$

verwenden, wobei $\boldsymbol{W}_t$ eine Reihe von $\boldsymbol{W}$ ist. Dann können Teststatistiken mit der robusten Schätzung $\widehat{\boldsymbol{V}}_{\text{IV,rob}}$ formuliert werden. `IV robust.R`