

Modul 2: Einfache Lineare Regression

PI 6250 – Ökonometrie I

Max Heinze (mheinze@wu.ac.at)

Department für Volkswirtschaftslehre, WU
Wien

Basierend auf einem Foliensatz von [Simon Heß](#)

6. März 2025

Motivation

Das bivariate lineare Modell

Ein Schätzer

Eigenschaften des OLS-Schätzers

Was haben diese Schlagzeilen gemeinsam?

Vox POLICY & POLITICS

Want to live longer, even if you're poor? Then move to a big city in California.

Updated by Ezra Klein on April 13, 2016, 1:30 p.m. ET [@ezraklein](#)

TWEET SHARE (8)

REUTERS
World ▾
Asian Markets

World Bank boosts Sri Lanka economic forecasts after inflation

By Uditha Jayasinghe
October 3, 2023 10:17 AM GMT+2 · Updated a day ago

NEWS
MEDICAL LIFE SCIENCES

MEDICAL HOME LIFE SCIENCES HOME

About COVID-19 News Health A-Z Drugs Medical Devices

Conditional cash transfer programs linked to reduction in child mortality in Latin America

on Sport Culture Lifestyle More ▾

ve & sex Home & garden Health & fitness Family Travel Money

Diet of fish 'can prevent' teen violence

New study reveals that the root cause of crime may be biological, not social

FOOD
Study Shows That Drinking 3-4 Cups of Coffee a Day Makes You Live Longer



The New York Times
A Half-Tablespoon of Olive Oil a Day May Promote Heart Health

Americans who ate at least one-and-a-half teaspoons of olive oil a day were at lower risk of heart disease than those who ate none.

Bedingte Erwartung von y

Die Aussagen auf der vorherigen Folie betreffen alle die **bedingte Erwartung** einer **abhängigen Variable y** , gegeben eine **erklärende Variable x** .

- Manche Aussagen sind trotzdem **Unsinn**.
- Wir werden lernen, zu zeigen, wieso.

Bedingte Erwartungen sind ein wichtiges Maß, das eine **abhängige Variable y** mit einer **erklärenden Variable x** in Relation setzt, zum Beispiel so:

$$E(y | x) = 0.4 + 0.5x$$

Auf diese Weise können wir Variation in der **abhängigen Variable y** in zwei Komponenten unterteilen:

- Variation, die von der **erklärenden Variable x** ausgeht, und
- Variation, die zufällig entsteht oder von unbeobachteten Faktoren ausgeht.

Evaluierung von Politikmaßnahmen

Wenn wir bestimmte **Maßnahmen evaluieren**, sind wir oft daran interessiert, **Unterschiede** zwischen verschiedenen Gruppen zu verstehen.

Zwei Beispiele:

- Effekte eines Medikaments auf die Gesundheit der Patient:innen in einer randomisierten Doppelblindstudie

$$E(\text{Gesundheit} \mid \text{Medikament} = 1) - E(\text{Gesundheit} \mid \text{Medikament} = 0)$$

- Gender Pay Gap für ein bestimmtes Bildungsniveau

$$E(\log(\text{Lohn}) \mid \text{Männlich} = 1, \dots) - E(\log(\text{Lohn}) \mid \text{Männlich} = 0, \dots)$$

In beiden Fällen untersuchen wir den **durchschnittlichen Behandlungseffekt** (engl. **average treatment effect, ATE**): der durchschnittliche Effekt einer „Behandlung“ relativ zu keiner „Behandlung“.

Vorhersagen

Wir können auch daran interessiert sein, ein **Ergebnis** für eine bestimmte Ausgangssituation **vorherzusagen**.

Angenommen, wir kennen die Verteilung von **Schulklassengröße** und **Prüfungsergebnissen**. Für einen neuen Bezirk können wir nur die Klassengröße. **Was ist die beste Vorhersage für die Prüfungsergebnisse im neuen Bezirk?**

- Der bedingte Mittelwert?
- Der bedingte Median?
- Der bedingte Modalwert?
- Etwas anderes?

Wenn wir eine **quadratische Verlustfunktion** minimieren, wird unsere beste Vorhersage der **bedingte Mittelwert** sein.

Verlustfunktion (engl. **loss function**): Eine Funktion, die Abweichungen vom wahren Wert nach einem bestimmten System „bestraft“.

Motivation

Das bivariate lineare Modell

Ein Schätzer

Eigenschaften des OLS-Schätzers

Logarithmische Transformationen

Bedingte Erwartungsfunktion

Wir wollen jetzt die **Bedingte Erwartungsfunktion** einer bestimmten **Zufallsvariable y** in Abhängigkeit von einer anderen **Zufallsvariable x** modellieren.

Der einfachste Weg, das zu tun: wir unterstellen eine **lineare Funktion**.

$$E(y_i \mid x_i) = \beta_0 + \beta_1 x_i,$$

wobei

- β_0 und β_1 **Parameter** der Funktion sind
- i ein Index für Beobachtungen ist
- y_i die **abhängige Variable**, **erklärte Variable**, **Outcome-Variable**, der **Regressand** ... ist, und
- x_i die **erklärende Variable**, **unabhängige Variable**, der **Regressor**, ... ist.

• Mehr

Bedingte Erwartungsfunktion

$$E(y_i \mid x_i) = \beta_0 + \beta_1 x_i,$$

Diese Funktion gibt uns eine Information über den **Erwartungswert** von y_i für einen bestimmten Wert x_i , **und nur das.**

- Wir können nicht herauslesen, welchen **Wert von** y_i wir für ein bestimmtes x_i bekommen.
- Wir bekommen auch keine Informationen über die Verteilung von y_i und x_i abseits des bedingten Erwartungswerts.

Angenommen, die bedingte Erwartungsfunktion für **Prüfungsergebnisse** gegeben eine bestimmte Klassengröße ist

$$E(\text{Prüfungsergebnisse}_i \mid \text{Klassengröße}_i) = 720 - 0.6 \times \text{Klassengröße}_i,$$

Bedingte Erwartungsfunktion

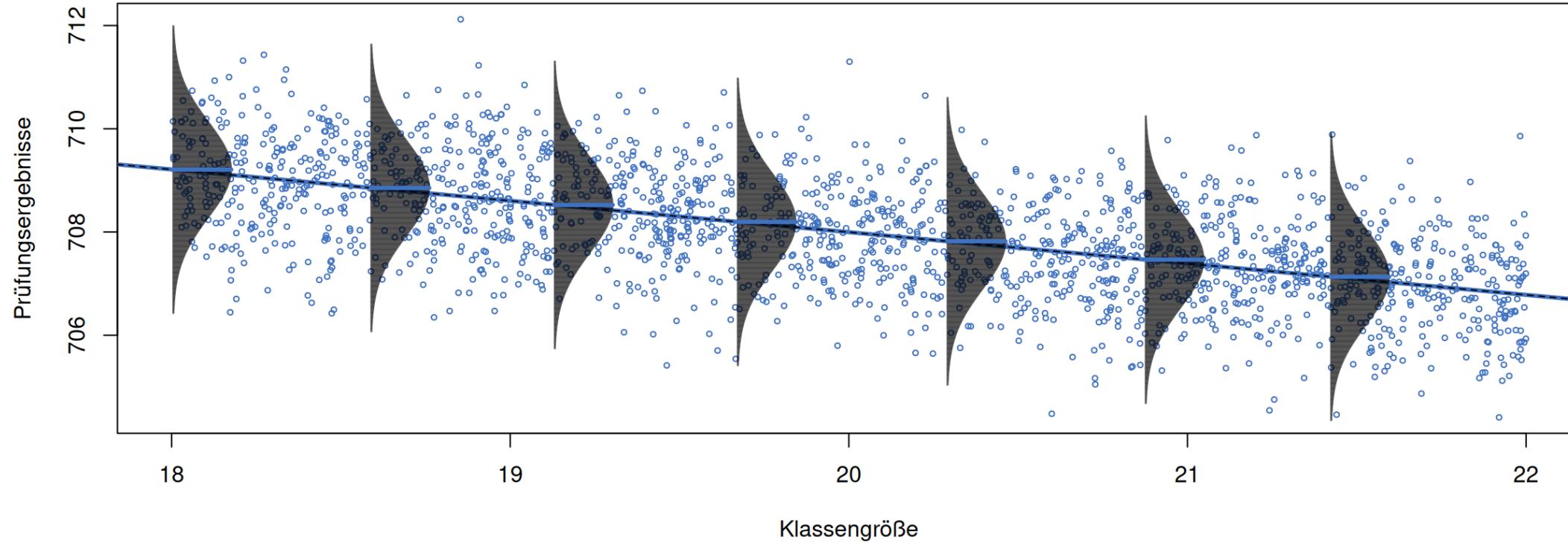
Angenommen, die bedingte Erwartungsfunktion für **Prüfungsergebnisse** gegeben eine bestimmte Klassengröße ist

$$E(\text{Prüfungsergebnisse}_i \mid \text{Klassengröße}_i) = 720 - 0.6 \times \text{Klassengröße}_i,$$

was können wir dann über die Prüfungsergebnisse in einem **neuen Bezirk** mit einer Klassengröße von 20 sagen?

- Der Erwartungswert für die Prüfungsergebnisse ist 708 Punkte.
- Die tatsächlichen Prüfungsergebnisse können darüber oder darunter liegen:
- Es gibt einen gewissen Fehler, bzw. eine **unbeobachtete Komponente**.
- Wir erwarten im Mittel einen Wert von 0 für diesen **Fehlerterm** (engl. **error term**):
 $u_i := y_i - E(y_i \mid x_i) = y_i - \beta_0 + \beta_1 x_i, \quad E(u_i \mid x_i) = 0.$
- Außerdem nehmen wir an, dass sein Erwartungswert unabhängig von x_i ist:
 $E(u_i \mid x_i) = E(u_i) = 0$ (engl. **zero conditional mean assumption**).

Visualisierung der bedingten Erwartungsfunktion



In blau sehen wir unsere **bedingte Erwartungsfunktion**. Für eine Klassengröße von 18 erwarten wir einen bestimmten Wert. Die **tatsächlichen Werte** sind um diesen Wert herum **verteilt**. Das trifft auf jeden Punkt entlang der Funktion zu.

Regressionsmodell in der Grundgesamtheit

Wir können unsere Überlegungen zur **bedingen Erwartungsfunktion** und zum **Vorhersagefehler** zusammenführen und erhalten ein **lineares Regressionsmodell**:

$$y_i = \beta_0 + \beta_1 x_i + u_i,$$

wobei

- $\beta_0 + \beta_1 x_i$ die **Regressionsfunktion** der **Grundgesamtheit** (engl. **population regression function, PRF**) ist,
- u_i der **Vorhersagefehler** bzw. **Fehlerterm** der **Grundgesamtheit** (engl. **population prediction error** bzw. **error term**) ist,
- β_0 der **konstante** Parameter (engl. **intercept**) ist, der den vorhergesagten Wert bei $x_i = 0$ abbildet, und
- β_1 der **Steigungsparameter** (engl. **slope**) ist, der den erwarteten Unterschied der vorhergesagten Werte für y_i bei einer Änderung von x_i um eine Einheit darstellt.

Regressionsmodell in der Grundgesamtheit

$$y_i = \beta_0 + \beta_1 x_i + u_i,$$

In unserem Beispiel von vorher:

$$\text{Prüfungsergebnisse}_i = \beta_0 - \beta_1 \times \text{Klassengröße}_i + u_i.$$

In diesem Fall ist:

$$\beta_1 = \frac{d E(\text{Prüfungserg.}_i | \text{Klassengr.}_i)}{d \text{Klassengröße}_i}$$

der **erwartete Unterschied** in den Prüfungsergebnissen, wenn wir die durchschnittliche Klassengröße um eine Einheit variieren.

$$\beta_0 = E(\text{Prüfungserg.}_i | \text{Klassengr.}_i = 0)$$

der **erwartete Wert** für das Prüfungsergebnis, wenn in einem Bezirk durchschnittlich 0 Schüler:innen in einer Klasse sind.

Skalierungseffekte

$$\beta_1 = \frac{d E(\text{Prüfungserg.}_i \mid \text{Klassengr.}_i)}{d \text{ Klassengröße}_i} \quad \beta_0 = E(\text{Prüfungserg.}_i \mid \text{Klassengr.}_i = 0)$$

Wie ändern sich diese beiden Parameter, wenn wir die **Skalierung** der Variablen ändern?
Messen wir beispielsweise die Klassengröße in Zehnern:

$$\text{Prüfungsergebnisse}_i = \beta_0^\bullet - \beta_1^\bullet \times \frac{\text{Klassengröße}_i}{10} + u_i.$$

Wir sehen:

$$\beta_0^\bullet = \beta_0 \quad \text{und} \quad \beta_1^\bullet = 10 \times \beta_1.$$

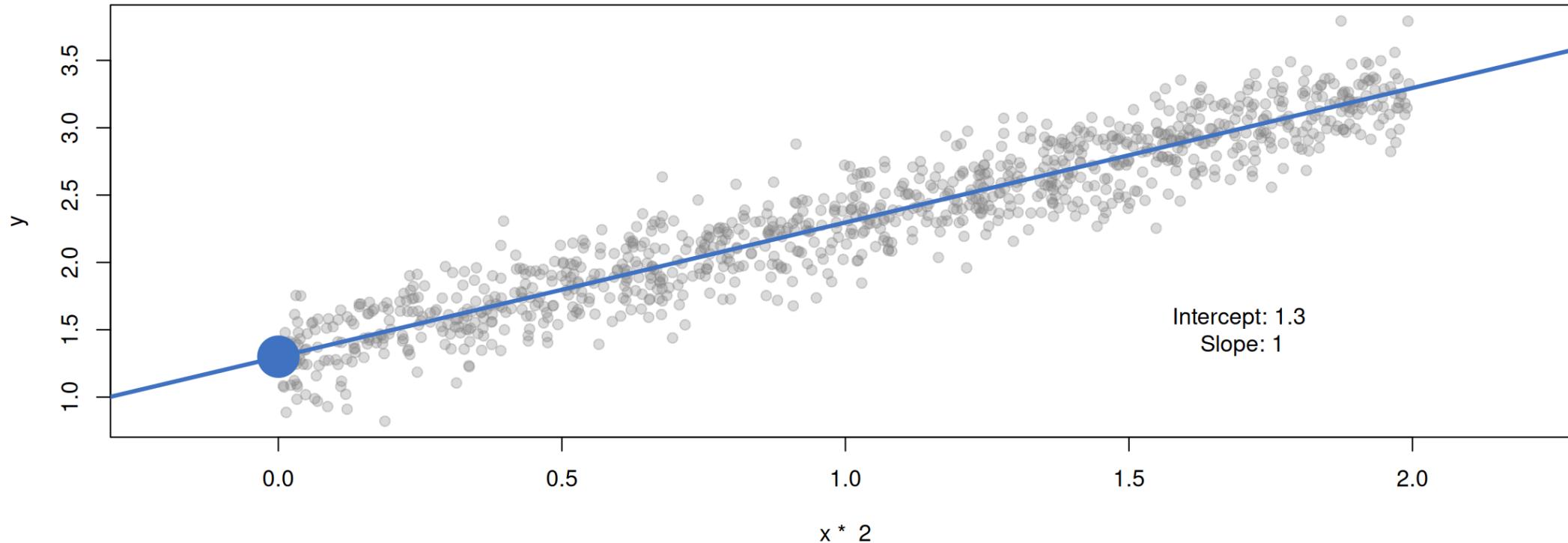
Die Regressionskonstante verändert sich nicht, der Steigungsparameter aber wird skaliert.



Übungsaufgabe

Was passiert, wenn wir die **abhängige Variable** (statt der unabhängigen Variable) skalieren?

Visualisierung der Skalierungseffekte



Auf dieser Folie **skalieren** wir die x_i -Werte in mehreren Schritten von Faktor 1 bis 2. Wir sehen, dass die **Konstante unverändert** bleibt, die **Steigung** sich aber **ändert**.

Motivation

Das bivariate lineare Modell

Ein Schätzer

Eigenschaften des OLS-Schätzers

Logarithmische Transformationen

Der Satz von Gauß-Markow

Grundgesamtheit vs. Stichprobe

Nichts, was wir bisher besprochen haben, hatte mit tatsächlichen **Daten** zu tun.

- Wir haben bisher Zusammenhänge in der **Grundgesamtheit** (engl. **population**) besprochen.
- Das **Regressionsmodell** der **Grundgesamtheit** beschreibt einen **hypothetischen Zusammenhang** zwischen mehreren Variablen. Wir können uns vorstellen, dass die Daten von PRF und Fehlerterm generiert werden.
- Wir kennen die Parameter β_0 und β_1 aus der PRF nicht.
- Daher müssen wir die Parameter **schätzen**. Wir benötigen dafür **Daten**, also eine **Stichprobe** (engl. **sample**).
- Wir werden im Folgenden Konzepte diskutieren, die sehr ähnlich zu denen aussehen, die wir vorher besprochen haben (z.B. eine Regressionsfunktion).
- Daher in Erinnerung behalten: Es gibt eine **Grundgesamtheit** und einen Zusammenhang zwischen mehreren Variablen darin. Wir können diesen Zusammenhang aber nur im Rahmen einer **Stichprobe** schätzen.

Zufallsstichprobe

Wir haben vorher diskutiert, wie Schulklassengröße und Prüfungsergebnisse in der **Grundgesamtheit** miteinander verbunden sind. Wir können β_0 und β_1 aber in der Praxis **nicht beobachten**. Daher benötigen wir eine **Stichprobe**, um sie schätzen zu können.

Wir sammeln also **Daten**:

$$\left. \begin{array}{l} \{y_1, x_1\} \\ \{y_2, x_2\} \\ \{y_3, x_3\} \\ \vdots \\ \{y_N, x_N\} \end{array} \right\} \quad \{y_i, x_i\}_{i=1}^N \quad \text{zufällig gezogen aus einer Grundgesamtheit} \quad F_{y,x}(\cdot, \cdot),$$

für die wir $E(y | x)$ mithilfe einer linearen bedingten Erwartungsfunktion approximieren wollen.

Zufallsstichprobe

Wie sieht eine **Zufallsstichprobe** in unserem Beispiel von vorher aus?

Wir bereiten zuerst den Datensatz wieder auf.

R Code [⟳ Start Over](#)

[▷ Run Code](#)

```
1 # Pakete laden
2 library(AER) # Enthält unseren Datensatz
3 library(dplyr) # Enthält mutate()
4
5 # Daten laden
6 data("CASchools")
7
8 # Variablen berechnen mit mutate()
9 CASchools <- CASchools |>
10   mutate(student_teacher_ratio = students / teachers,
11         test_score = (read + math)/2)
12
13 # Nur die Variablen behalten, die uns interessieren
14 CASchools <- select(CASchools, student_teacher_ratio, test_score)
```

Zufallsstichprobe

Wie sieht eine **Zufallsstichprobe** in unserem Beispiel von vorher aus?

R Code

⟳ Start Over

▷ Run Code

```
1 # head() gibt uns die ersten 6 Beobachtungen eines Datensatzes  
2  
3 head(CASchools)
```

Wir sehen hier fixe Zahlen. Allerdings sind diese Zahlen **Realisierungen von Zufallsvariablen**, und jedes Mal, wenn wir eine neue **Zufallsstichprobe** ziehen, werden wir andere Werte erhalten.

Zufallsstichprobe

Ziehen wir zur Veranschaulichung eine **Stichprobe** aus einer Standard-**Normalverteilung** und berechnen den Mittelwert.

R Code

[⟳ Start Over](#)

[▷ Run Code](#)

```
1 this_sample <- rnorm(n = 1000, mean = 0, sd = 1)
2
3 mean(this_sample)
```

Wenn wir diese Berechnung mehrmals durchführen, bekommen wir immer einen Mittelwert, der in der Nähe von 0 liegt, aber wir bekommen **jedes Mal** einen **anderen Wert**. Je mehr Beobachtungen wir sammeln (z.B. `n=10^6`), desto näher werden die meisten dieser Werte an 0 liegen.

Wir suchen einen Schätzer

Wir wollen eine **Regressionslinie** mit Konstanter $\tilde{\beta}_0$ und Steigung $\tilde{\beta}_1$ anpassen:

$$y_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_i,$$

die die folgenden **Vorhersagefehler** minimiert:

$$\hat{u}_i = y_i - \tilde{\beta}_0 + \tilde{\beta}_1 x_i.$$

- \hat{u}_i ist das **Residuum** (engl. **residual**), und ist nicht dasselbe wie der **Fehlerterm**.
 - Das **Residuum** ist der Unterschied zwischen unserer angepassten Regressionslinie und dem tatsächlich beobachteten Wert y_i .
 - Der **Fehlerterm** ist die zufällige oder unbeobachtete Komponente aus dem datengenerierenden Prozess der Grundgesamtheit.
- $\tilde{\beta}_0$ und $\tilde{\beta}_1$ sind unsere **angepassten Koeffizienten** für Konstante und Steigung, und sind nicht dasselbe wie die **Parameter** β_0 und β_1 **aus der Grundgesamtheit**.

OLS-Schätzer

Wie finden wir unter allen $\tilde{\beta}_0$ und $\tilde{\beta}_1$ diejenigen Parameter $\hat{\beta}_0$ und $\hat{\beta}_1$, die den Vorhersagefehler minimieren?

Vorschlag: Wir nehmen die Summe aller Residuen.

- Macht das Sinn? **Nein.**
- Positive und negative Residuen würden einander aufheben.

Besserer Vorschlag: Wir nehmen die Summe aller **Quadrat**e der Residuen. So bestrafen wir positive und negative Residuen gleichermaßen. Wir suchen also das Minimum von:

$$S(\tilde{\beta}_0, \tilde{\beta}_1) = \sum_{i=1}^N \left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i \right)^2.$$

Wir nennen den resultierenden Schätzer **Kleinste-Quadrat**-Schätzer (engl. **least squares estimator**) bzw. Gewöhnlicher Kleinste-Quadrat-Schätzer (engl. **ordinary least squares**, OLS).

OLS-Schätzer (Quadrate minimieren)

$$S(\tilde{\beta}_0, \tilde{\beta}_1) = \sum_{i=1}^N \left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i \right)^2.$$

Wir beginnen damit, die Funktion nach $\tilde{\beta}_0$ abzuleiten und die Ableitung gleich Null zu setzen:

$$\frac{\partial S}{\partial \tilde{\beta}_0} = -2 \sum_{i=1}^N \left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i \right) = 0,$$

Das gibt uns

$$\sum_{i=1}^N y_i = n\tilde{\beta}_0 + \tilde{\beta}_1 \sum_{i=1}^N x_i.$$

OLS-Schätzer (Quadrate minimieren)

Als nächstes leiten wir nach $\tilde{\beta}_1$ ab:

$$\frac{\partial S}{\partial \tilde{\beta}_1} = -2 \sum_{i=1}^N x_i (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i) = 0,$$

Wir erhalten

$$\sum_{i=1}^N x_i y_i = \tilde{\beta}_0 \sum_{i=1}^N x_i + \tilde{\beta}_1 \sum_{i=1}^N x_i^2.$$

OLS-Schätzer (Quadrate minimieren)

Wir notieren ab jetzt $\bar{x} = \frac{1}{n} \sum_{i=1}^N x_i$ und $\bar{y} = \frac{1}{n} \sum_{i=1}^N y_i$. Dann erhalten wir aus der ersten Bedingung erster Ordnung:

$$\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1 \bar{x}.$$

Wenn wir das in die zweite Bedingung erster Ordnung einsetzen, erhalten wir:

$$\sum_{i=1}^N x_i (y_i - \bar{y}) = \tilde{\beta}_1 \sum_{i=1}^N x_i (x_i - \bar{x}).$$

OLS-Schätzer (Quadrate minimieren)

Weil $\sum_{i=1}^N x_i (x_i - \bar{x}) = \sum_{i=1}^N (x_i - \bar{x})^2$ und

$\sum_{i=1}^N x_i (y_i - \bar{y}) = \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})$ (Siehe Appendix A-1 in Wooldridge):

$$\hat{\beta}_1 = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2}, = \frac{\widehat{\text{Cov}}(x_i, y_i)}{\widehat{\text{Var}}(x_i)}$$

solange $\sum_{i=1}^N (x_i - \bar{x})^2 > 0$.

Und von vorher:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Diese Schätzer **minimieren die Summe der Residuenquadrate**.

OLS-Schätzer (Momentenmethode)

Alternativ können wir die Schätzer über die **Momentenmethode** (engl. **method of moments**) herleiten. Wir können dabei die folgenden (vorher besprochenen) Annahmen als **Momentenbedingungen** (engl. **moment conditions**) verwenden:

- $E(u_i) = 0$ (sonst wäre die Linie einfach zu weit unten/oben)
- $\text{Cov}(x_i, u_i) = E(x_i u_i) = 0$ (sonst wäre die Linie schief) • **Beweis**

Als ersten Schritt ersetzen wir die **Momente der Grundgesamtheit** durch die **Stichprobenmomente**:

$$\frac{1}{n} \sum_{i=1}^n x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{1}{n} \sum_{i=1}^n y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = 0$$

Momente sind wichtige Größen zur Beschreibung von Zufallsvariablen. Der k -te Moment einer Zufallsvariable ist definiert als $m_k(X) := E(X^k)$ – der erste Moment ist also der Erwartungswert, der zweite die Varianz, ... Stichprobenmomente sind analog als $N^{-1} \sum_{i=1}^N x_i^k$ definiert.

OLS-Schätzer (Momentenmethode)

$$\frac{1}{n} \sum_{i=1}^n x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{1}{n} \sum_{i=1}^n y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = 0$$

Diese Ausdrücke sind **äquivalent** zu denen, die wir durch **Ableiten der Verlustfunktion** erhalten haben. Insofern können wir genau so fortsetzen wie vorher und erhalten:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Wir haben **denselben Schätzer** durch **zwei verschiedene Methoden** erhalten.

Momente sind wichtige Parameter, mit denen wir Zufallsvariablen beschreiben können. Der k -te Moment einer Zufallsvariable ist definiert als $m_k(X) := E(X^k)$ – der erste Moment ist also der Erwartungswert, der zweite die Varianz, ... Stichprobenmomente sind analog als $N^{-1} \sum_{i=1}^N x_i^k$ definiert.

Motivation

Das bivariate lineare Modell

Ein Schätzer

Eigenschaften des OLS-Schätzers

Logarithmische Transformationen

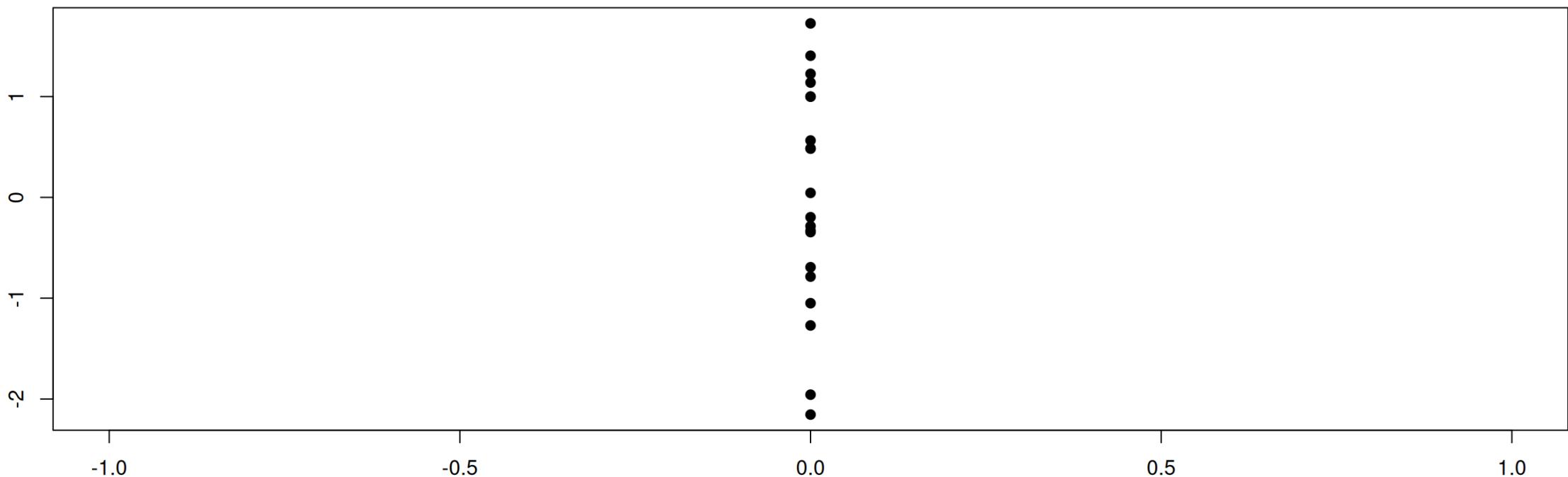
Der Satz von Gauß-Markow

Erwartungswert des OLS-Schätzers

Variation in X

Wir können unseren Schätzer für die Steigung nur berechnen, wenn die **Varianz in x_i nicht 0** ist (andernfalls würden wir durch 0 dividieren):

$$\hat{\beta}_1 = \frac{\widehat{\text{Cov}}(x_i, y_i)}{\widehat{\text{Var}}(x_i)}$$



Die Residuen sind im Mittel 0

Die Residuen sind die Differenz zwischen **tatsächlich beobachtetem Wert** und dem **angepassten Wert**:

$$\hat{u}_i = y_i - \hat{y}_i$$

Als wir vorher nach $\tilde{\beta}_0$ abgeleitet haben, hatten wir:

$$\frac{\partial S}{\partial \tilde{\beta}_0} = -2 \sum_{i=1}^N (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i) = 0,$$

was impliziert, dass die **Summe** (und somit das Mittel) der **Residuen 0 ist**.

Intuition: Wären die Residuen im Mittel positiv oder negativ, könnten wir die Linie nach unten bzw. oben verschieben und eine bessere Anpassung erreichen.

Die Residuen sind nicht mit x_i korreliert

Als wir vorher nach $\tilde{\beta}_1$ abgeleitet haben, hatten wir:

$$\frac{\partial S}{\partial \tilde{\beta}_1} = -2 \sum_{i=1}^N x_i (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i) = 0.$$

Das impliziert:

$$\sum_{i=1}^N (x_i - \bar{x}) \hat{u}_i = 0$$

Das impliziert wiederum, dass die Korrelation zwischen den x_i und den Residuen 0 ist.

Intuition: Wären die Residuen mit den x_i korreliert, könnten wir eine bessere Anpassung erreichen, indem wir unsere Linie neigen.

Dekomposition der Varianz von y

Wir können die Variation in y in einen **erklärten Teil**, also **Variation, die von Variation in x ausgeht**; und in einen **nicht erklärten Teil**, also einen Teil, der von **unbeobachteten Faktoren** ausgeht, aufteilen:

$$\sum_{i=1}^N (y_i - \bar{y})^2 = \sum_{i=1}^N (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^N \hat{u}_i^2$$

oder auch

Totale Quadratsumme = **Erklärte Quadratsumme** + **Residuenquadratsumme**

Total Sum of Squares = **Explained Sum of Squares** + **Residual Sum of Squares**

$$SST = SSE + SSR$$

• Beweis

Anpassungsgüte

Das **Bestimmtheitsmaß R^2** (engl. **coefficient of determination**) ist eine Maßzahl zur **Anpassungsgüte** (engl. **goodness of fit**) und gibt an, welcher Anteil der Variation durch unser Modell erklärt wird:

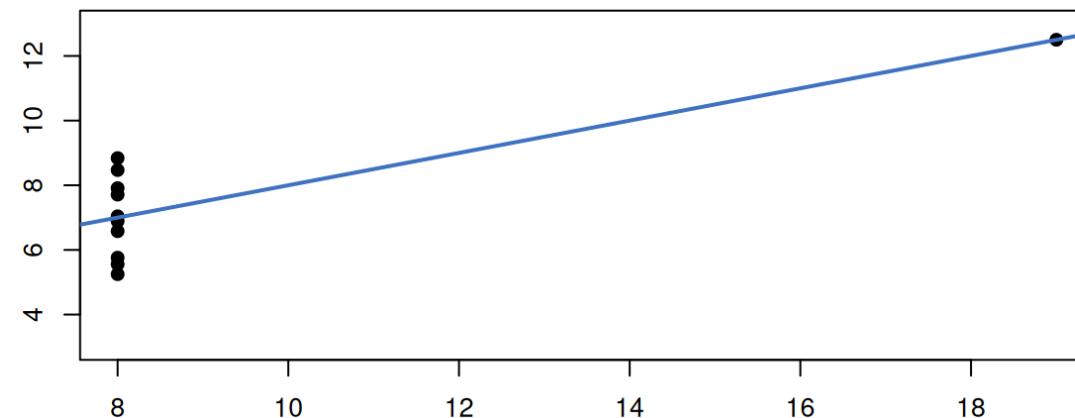
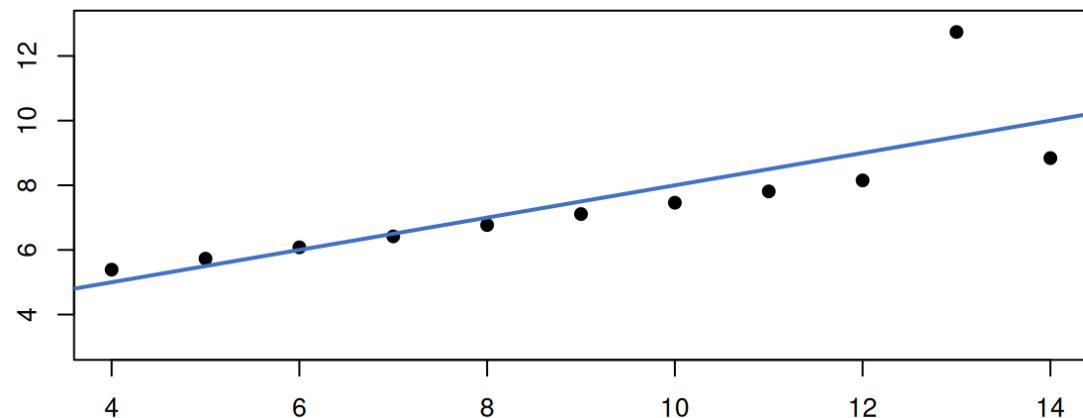
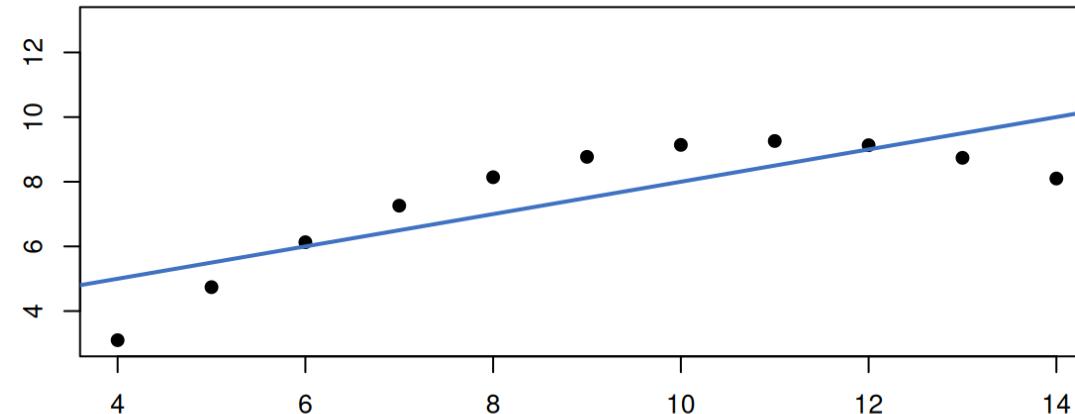
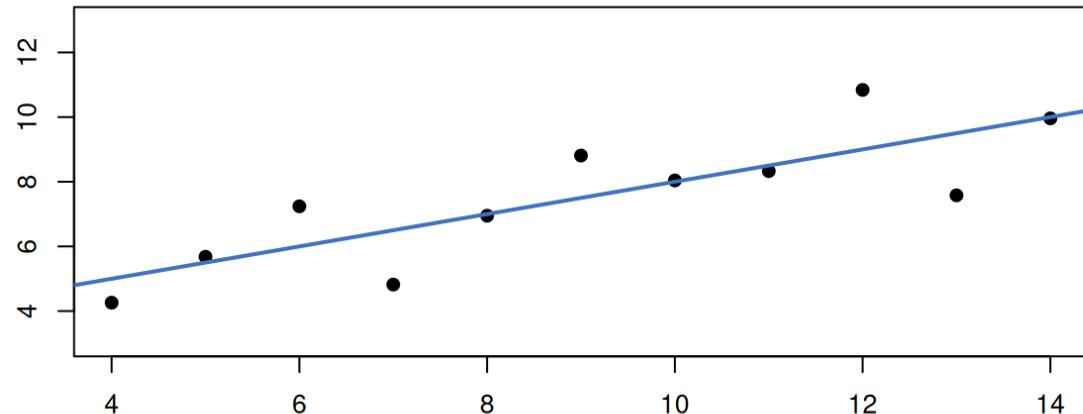
$$R^2 = \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\text{SSR}}{\text{SST}}.$$

- R^2 liegt immer zwischen 0 und 1.
- Bei einem R^2 von 1 liegen alle Beobachtungen auf einer Geraden.
- R^2 wird manchmal verwendet, um Modelle zu vergleichen. **Das ist aber meistens eine schlechte Idee.**
 - Es gibt keinen Schwellenwert für ein „gutes“ R^2 .
 - Es gibt „schlechte“ Modelle, die gut an einen Datensatz angepasst sind.
 - Es gibt Modelle mit niedrigem R^2 , die uns wichtige Zusammenhänge aufzeigen.

Anpassungsgüte

Anscombe-Quartett

In allen vier Beispielen ist $R^2 = 0.67$.



Das bivariate lineare Modell

Ein Schätzer

Eigenschaften des OLS-Schätzers

Logarithmische Transformationen

Der Satz von Gauß-Markow

Erwartungswert des OLS-Schätzers

Varianz des OLS-Schätzers

Logarithmische Transformation der abhängigen Variable

Wir beginnen mit einem Beispiel. Nehmen wir an, der **Lohn**, den eine Person erhält, hängt von der **Ausbildung** der Person ab:

$$\text{Lohn}_i = f(\text{Ausbildung}_i)$$

Ist es plausibler, dass ein zusätzliches Ausbildungsjahr den Lohn immer um die gleiche **Menge** erhöht, oder um den gleichen **Faktor**?

Das **5. Jahr** Ausbildung erhöht den
Lohn um **1 Euro**

und

Das **12. Jahr** Ausbildung erhöht den
Lohn um **1 Euro**

Das **5. Jahr** Ausbildung erhöht den
Lohn um **8 Prozent**

und

Das **12. Jahr** Ausbildung erhöht den
Lohn um **8 Prozent**

Logarithmische Transformation der abhängigen Variable

Wir können eine derartige Beziehung **mit Logarithmen approximieren**:

$$\log(\text{Lohn}_i) = \beta_0 + \beta_1 \text{Ausbildung}_i + u_i.$$

Das ist äquivalent zu:

$$\text{Lohn}_i = \exp(\beta_0 + \beta_1 \text{Ausbildung}_i + u_i).$$

Die Beziehung ist **nicht-linear** in y (Lohn) und x (Ausbildung), aber sie ist **linear** in $\log(y)$ und x .

Wir können die Regression genau so **mit OLS schätzen** wie vorher, indem wir $y_i^* = \log(y_i)$ definieren und folgendes Modell schätzen:

$$y_i^* = \beta_0 + \beta_1 x_i + u_i$$

Logarithmische Transformation der unabhängigen Variable

Analog zu vorher können wir auch die **unabhängige Variable** (x) logarithmieren. Die **Interpretation** im vorherigen Beispiel wäre:

Eine Erhöhung der Ausbildung um **1 Prozent** (egal von welchem Niveau) erhöht den Lohn um eine Bestimmte Anzahl **Euro**.

Wir definieren $x_i^* = \log(x_i)$ **schätzen** das Modell:

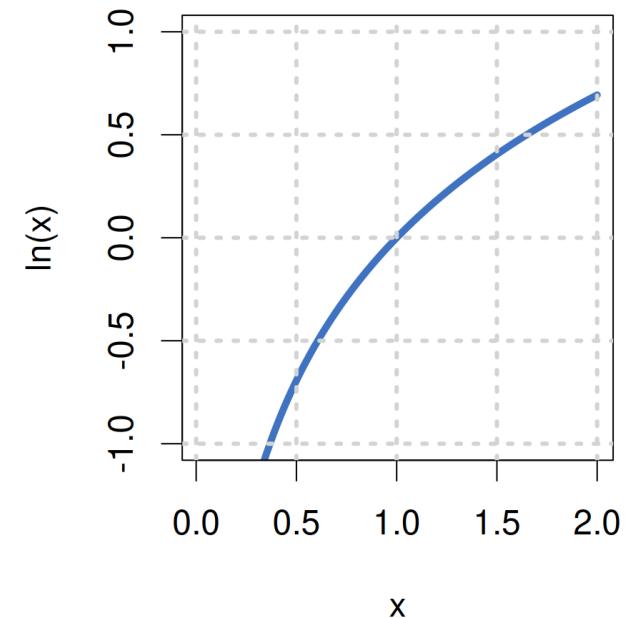
$$y_i = \beta_0 + \beta_1 x_i^* + u_i.$$

Natürlicher Logarithmus

Wenn wir den **natürlichen Logarithmus** für unsere Transformation verwenden, ist die Interpretation der Koeffizienten sehr einfach:

- **Absolute** Veränderungen in logarithmierten Variablen entsprechen **ungefähr** einer **relativen** Veränderung der nicht-logarithmierten Variable mit demselben numerischen Wert.
- Ein Anstieg von x um 1 Prozent **entspricht ungefähr** einem Anstieg von $\log(x)$ um 0.01:

$$\begin{aligned}\log(1.01x) &= \log(x) + \log(1.01) \\ &= \log(x) + 0.00995 \\ &\approx \log(x) + 0.01\end{aligned}$$



- Die Approximation funktioniert am besten für **kleinere** Prozentwerte.

Überblick über Log-Transformationen

- **Nicht transformierte** Modelle erlauben uns Aussagen über die Beziehung zwischen **absoluten Veränderungen** zweier Variablen.
- Modelle, bei denen wir **eine Seite logarithmieren**, erlauben uns Aussagen über **Semi-Elastizitäten**.
- Modelle, bei denen wir **beide Seiten logarithmieren**, erlauben uns Aussagen über **Elastizitäten**.

Modell	Abh. Variable	Unabh. Variable	Interpretation
Level-Level	y	x	+1 in $x \Leftrightarrow +\beta_1$ in y
Level-Log	y	$\log(x)$	+1% in $x \Leftrightarrow +\beta_1/100$ in y
Log-Level	$\log(y)$	x	+1 in $x \Leftrightarrow +\beta_1 \times 100\%$ in y
Log-Log	$\log(y)$	$\log(x)$	+1% in $x \Leftrightarrow +\beta_1\%$ in y

Ein Schätzer
Eigenschaften des OLS-Schätzers
Logarithmische Transformationen

Der Satz von Gauß-Markow

Erwartungswert des OLS-Schätzers
Varianz des OLS-Schätzers
Regressionen mit nur einem Parameter

BLUE

Wenn wir annehmen, dass unser lineares Modell korrekt ist, können wir einige Aussagen über **Erwartungswert** und **Varianz** des OLS-Schätzers treffen.

Der **Satz von Gauß-Markow** (engl. **Gauss-Markov Theorem**) besagt, dass der OLS-Schätzer der „beste lineare unverzerrte Schätzer“ ist, oder auch der

Best Linear Unbiased Estimator (BLUE)

- Dass der OLS-Schätzer ein **linearer Schätzer** ist, wissen wir bereits.
- **Unverzerrt** (engl. **unbiased**) bedeutet, dass der **Erwartungswert** des Schätzers dem wahren Parameter entspricht.
- Der **beste** Schätzer ist ein Schätzer dann, wenn er unter allen unverzerrten linearen Schätzern die geringste Varianz hat. Das besprechen wir im nächsten Abschnitt.

Modellannahmen

Damit wir mithilfe des **Satzes von Gauß-Markow** beweisen können, dass der OLS-Schätzer **BLUE** ist, benötigen wir **vier Annahmen** hinsichtlich unseres Modells:

(i) Satz von Gauß-Markow: Annahmen für Einfache Lineare Regression (SLR)

- (1) Linearität in Parametern
- (2) Zufallsstichprobe
- (3) Variation in x
- (4) Exogener Fehlerterm

(SLR.1) Linearität in Parametern

Die Regressionsfunktion der Grundgesamtheit (PRF) muss **linear** in ihren **Parametern** sein:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

- Transformationen (z.B. logarithmische) sind kein Problem, da die PRF trotzdem eine lineare Kombination **der Parameter** bleibt.
- Wenn wir nur von einem „linearen Modell“ sprechen, ist unklar, ob wir die Parameter oder x meinen.
- Ein Beispiel für ein Modell, das nicht linear in seinen Parametern ist, wäre:
 $y_i = 1^{\beta_0} x_i^{\beta_1} + u_i$.
- Diese Annahme dient nur dazu, die Klasse der Modelle/Schätzer (linear) zu definieren.

(SLR.2) Zufallsstichprobe

Unsere **Stichprobe** mit N Beobachtungen, $\{(y_i, x_i), i = 1, 2, \dots, N\}$ muss **zufällig** aus der Grundgesamtheit gezogen werden. Die Wahrscheinlichkeit, eine Beobachtung in die Stichprobe aufzunehmen, muss für alle gleich sein, und darf nicht davon abhängen, wen wir zuerst „gezogen“ haben.

- Es ist ziemlich leicht, diese Annahme zu verletzen:
 - Wir ziehen nur aus einem gewissen Teil der Grundgesamtheit, z.B. indem wir Studierende nur in der Mensa befragen.
 - Wir wählen einen Teil der Stichprobe abhängig von einem anderen Teil, z.B. indem wir $N/2$ Studierende zufällig befragen und dann die andere Hälfte der Stichprobe mit deren besten Freund:innen auffüllen.
- Mithilfe dieser Annahme können wir das Modell der Grundgesamtheit durch einzelne Beobachtungen beschreiben: $E(y_i | x_1, \dots, x_N) = E(y_i | x_i) = E(y | x)$
- Es gibt ökonometrische Techniken, mit denen man mit nicht-zufälligen Stichproben arbeiten kann. Damit beschäftigen wir uns in späteren Kursen.

(SLR.3) Variation in x

Damit wir unser Modell schätzen können, benötigen wir **Variation in x** . Die x -Werte dürfen nicht alle vollständig gleich sein.

- Wir brauchen diese Annahme, weil wir sonst keinen Parameter identifizieren können.
- Wenn wir unsere Daten aus einer Grundgesamtheit ziehen, wird diese Annahme typischerweise erfüllt sein; es sei denn, die Stichprobe ist sehr klein und die Variation in der Grundgesamtheit ist minimal.
- Ein Beispiel für eine Verletzung der Annahme: Wir versuchen, einen Effekt von Klassengröße auf Prüfungsergebnisse zu schätzen. Alle Beobachtungen in der Stichprobe haben eine Klassengröße von 20.

Übrigens: Wenn wir keine Variation in y haben, werden unsere Ergebnisse nicht wahnsinnig interessant sein (unsere Regressionsgerade ist dann horizontal), aber berechnen können wir sie ohne Probleme.

(SLR.4) Exogene Fehler

Der Erwartungswert des Fehlerterms u ist für jeden x -Wert 0:

$$E(u_i | x_i) = 0$$

Diese Annahme impliziert auch die beiden **Momentenbedingungen** $E(u_i) = 0$ und $E(u_i x_i) = 0$. • **Beweis**

- In vielen Herleitungen arbeiten wir mit Erwartungswerten der Form $E(\cdot | x_i)$.
- Anders gesagt: Wir fixieren die x -Werte und suchen dann mehrere Zufallsstichproben, die diese Werte erfüllen (sich aber in u_i und daher y_i unterscheiden) (engl. x **fixed in repeated samples**).
- Das ist, besonders bei Beobachtungsdaten, nicht wahnsinnig realistisch.
- Die Annahme erlaubt uns, **dieselben Herleitungen** auch **mit nicht-fixierten x_i** anzuwenden.

Wann sind Fehler nicht exogen?

Dass $E(u_i) = 0$, ist keine besonders restriktive Annahme (notfalls verschieben wir einfach die Linie). Dass $E(u_i | x_i) = 0$, ist weit weniger trivial.

Experiment

Wir wählen zufällig eine Anzahl von Feldern aus. Dann wählen wir wiederum zufällig die Hälfte der Stichprobe aus und wenden auf diesen Feldern Dünger an. Wir notieren dann die Erträge.

Beobachtungsstudie

Wir wählen zufällig eine Anzahl von Feldern aus. Dann fragen wir die Landwirt:innen, ob sie diese Felder bedüngt haben. Wir notieren dann Düngergebrauch und die Erträge der Felder.

Im **Experiment** ist die Intervention, der Düngergebrauch, (x_i) , garantiert unabhängig von unbeobachteten Faktoren. Die Annahme, dass $E(u_i | x_i) = 0$, ist also plausibel.

In der **Beobachtungsstudie** ist die Intervention möglicherweise nicht unabhängig von unbeobachteten Faktoren. Vielleicht wird Dünger auf weniger fruchtbaren Feldern angewandt, um einen Nachteil auszugleichen? Oder auf „besseren“ Feldern, um den Ertrag noch mehr zu verbessern? Wenn wir $E(u_i | x_i) = 0$ für plausibel halten, müssen wir dafür **argumentieren**.

Eigenschaften des OLS-Schätzers

Logarithmische Transformationen

Der Satz von Gauß-Markow

Erwartungswert des OLS-Schätzers

Varianz des OLS-Schätzers

Regressions mit nur einem Parameter

Binäre erklärende Variablen

OLS ist unverzerrt

Wenn die vier Annahmen SLR.1 bis SLR.4 erfüllt sind, können wir **beweisen**, dass der OLS-Schätzer **unverzerrt** (engl. **unbiased**) ist.

Ein Schätzer ist dann unverzerrt (oder auch: erwartungstreu), wenn **sein Erwartungswert dem wahren Wert des Parameters im Modell der Grundgesamtheit entspricht**. Wir wollen also beweisen:

$$E(\hat{\beta}_j) = \beta_j \quad j = 0, 1$$

Beweis: OLS ist unverzerrt

Wir starten mit dem Ausdruck für den OLS-Schätzer:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})x_i} = \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})x_i}$$

Als ersten Schritt schreiben wir y_i als Summe seiner Bestandteile an:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i)}{\sum_{i=1}^n (x_i - \bar{x})x_i}$$

Wir teilen auf:

$$\hat{\beta}_1 = \frac{\beta_0 \sum_{i=1}^n (x_i - \bar{x}) + \beta_1 \sum_{i=1}^n (x_i - \bar{x})x_i + \sum_{i=1}^n (x_i - \bar{x})u_i}{\sum_{i=1}^n (x_i - \bar{x})x_i}$$

Beweis: OLS ist unverzerrt

$$\hat{\beta}_1 = \frac{\beta_0 \sum_{i=1}^n (x_i - \bar{x}) + \beta_1 \sum_{i=1}^n (x_i - \bar{x})x_i + \sum_{i=1}^n (x_i - \bar{x})u_i}{\sum_{i=1}^n (x_i - \bar{x})x_i}$$

Weil $\sum_{i=1}^n (x_i - \bar{x}) = 0$ und $\frac{\beta_1 \sum_{i=1}^n (x_i - \bar{x})x_i}{\beta_1 \sum_{i=1}^n (x_i - \bar{x})x_i} = 1$:

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x})u_i}{\sum_{i=1}^n (x_i - \bar{x})x_i}$$

Wir nennen $\frac{\sum_{i=1}^n (x_i - \bar{x})u_i}{\sum_{i=1}^n (x_i - \bar{x})x_i}$ **Stichprobenfehler** (engl. **sampling error**). Die Gleichung zeigt uns, dass $\hat{\beta}_1$ in einer endlichen Stichprobe der Summe aus dem wahren Parameter β_1 und einer bestimmten Linearkombination der Fehlerterme, dem Stichprobenfehler, entspricht.

Wenn wir zeigen können, dass dieser Stichprobenfehler im **Mittel 0** ist, haben wir die Unverzerrtheit des OLS-Schätzers bewiesen.

Beweis: OLS ist unverzerrt

Was ist also der Erwartungswert von $\hat{\beta}_1$?

$$E(\hat{\beta}_1 | x_1, \dots, x_N) = E\left(\beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x}) u_i}{\sum_{i=1}^n (x_i - \bar{x}) x_i} \middle| x_1, \dots, x_N\right)$$

Da der wahre Parameter β_1 keine Zufallsvariable ist, können wir ihn herausnehmen:

$$E(\hat{\beta}_1 | x_1, \dots, x_N) = \beta_1 + E\left(\frac{\sum_{i=1}^n (x_i - \bar{x}) u_i}{\sum_{i=1}^n (x_i - \bar{x}) x_i} \middle| x_1, \dots, x_N\right)$$

Weil $E(x_i | x_i) = x_i$:

$$E(\hat{\beta}_1 | x_1, \dots, x_N) = \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x}) E(u_i | x_1, \dots, x_N)}{\sum_{i=1}^n (x_i - \bar{x}) x_i}$$

Beweis: OLS ist unverzerrt

$$E(\hat{\beta}_1|x_1, \dots, x_N) = \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x}) E(u_i|x_1, \dots, x_N)}{\sum_{i=1}^n (x_i - \bar{x}) x_i}$$

Die Annahme SLR.2 erlaubt uns folgende Vereinfachung:

$$E(\hat{\beta}_1|x_1, \dots, x_N) = \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x}) E(u_i|x_i)}{\sum_{i=1}^n (x_i - \bar{x}) x_i}$$

Annahme SLR.4 besagt, dass $E(u_i|x_i) = 0$, also

$$E(\hat{\beta}_1|x_1, \dots, x_N) = \beta_1$$

Beweis: OLS ist unverzerrt

$$E(\hat{\beta}_1 | x_1, \dots, x_N) = \beta_1$$

Aufgrund des Satzes der iterierten Erwartungen ist $E(\hat{\beta}_1) = E(E(\hat{\beta}_1 | x_1, \dots, x_N))$ und somit folgt

$$E(\hat{\beta}_1) = \beta_1,$$

Der Erwartungswert des Schätzers entspricht dem wahren Parameter aus dem Modell der Grundgesamtheit, er ist also unverzerrt.

□

Beweis: OLS ist unverzerrt

Der Beweis dafür, dass auch $\hat{\beta}_0$ unverzerrt ist, ist sehr einfach. Zuerst schreiben wir $\hat{\beta}_0$ als

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Weil $E(\hat{\beta}_1 | x_1, \dots, x_N) = \beta_1$:

$$\begin{aligned} E(\hat{\beta}_0 | x_i, \dots, x_N) &= E(\bar{y} | x_1, \dots, x_N) - E(\hat{\beta}_1 \bar{x} | x_1, \dots, x_N) \\ &= E(\bar{y} | x_1, \dots, x_N) - E(\hat{\beta}_1 | x_1, \dots, x_N) \bar{x} \\ &= \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} \\ &= \beta_0. \end{aligned}$$

Auch der Schätzer $\hat{\beta}_0$ ist unverzerrt.

□

Logarithmische Transformationen

Der Satz von Gauß-Markow

Erwartungswert des OLS-Schätzers

Varianz des OLS-Schätzers

Regressionen mit nur einem Parameter

Binäre erklärende Variablen

Kausale Inferenz

(SLR.5) Homoskedastizität

Die Varianz des Fehlerterms u_i ist für alle x_i -Werte gleich:

$$\text{Var}(u_i \mid x_i) = \text{Var}(u_i) = \sigma^2$$

- Die Varianz des Fehlerterms ist eine Maßzahl für die Variation, die von unbeobachteten Faktoren ausgeht.
- Unter dieser Annahme ist diese Varianz für alle x_i -Werte gleich σ^2 .
- Wir brauchen diese Annahme **nicht**, um zu zeigen, dass der OLS-Schätzer unverzerrt ist. Aber wir brauchen sie, um zu zeigen, dass er die geringstmögliche Varianz hat.
- In echten Querschnittsdaten ist diese Annahme oft verletzt.
 - Leute mit mehr Ausbildung haben vielleicht eine größere Varianz in ihren Löhnen.
 - Später lernen wir Wege kennen, um mit einer Verletzung dieser Annahme umzugehen.

Effizienz des OLS-Schätzers

Wenn die fünf Annahmen SLR.1 bis SLR.5 erfüllt sind, können wir **beweisen**, dass der OLS-Schätzer die **niedrigstmögliche Varianz** aller unverzerrten linearen Schätzer hat.

Wir sagen dann, er ist der **beste** lineare unverzerrte Schätzer (BLUE). Diese Eigenschaft nennen wir auch **Effizienz** (engl. **efficiency**).

Wir können das **beweisen**. Dafür zeigen wir erst, dass die Varianz des OLS-Schätzers

$$\text{Var}(\hat{\beta}_1 \mid x_i) = \frac{\sigma^2}{\sum_{i=1}^N (x_i - \bar{x})^2}, \quad \text{Var}(\hat{\beta}_0 \mid x_i) = \frac{\sigma^2 N^{-1} \sum_{i=1}^N x_i^2}{\sum_{i=1}^N (x_i - \bar{x})^2}$$

ist, und dann zeigen wir, dass es keinen linearen unverzerrten Schätzer geben kann, dessen Varianz geringer ist.

Beweis: Effizienz des OLS-Schätzers

Wir zeigen den Beweis für β_1 . Wir beginnen mit der **Aufteilung** des Schätzers von vorher:

$$\text{Var}(\hat{\beta}_1 \mid x_i) = \text{Var} \left(\beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x}) u_i}{\sum_{i=1}^n (x_i - \bar{x}) x_i} \middle| x_i \right)$$

Zur besseren Übersichtlichkeit schreiben wir jetzt $w_i := \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x}) x_i}$:

$$\text{Var}(\hat{\beta}_1 \mid x_i) = \text{Var} \left(\beta_1 + \sum_{i=1}^n w_i u_i \middle| x_i \right)$$

Jetzt können wir SLR.5 anwenden. Außerdem hängen die Gewichte w_i nur von x_i ab und sind somit fix:

$$\text{Var}(\hat{\beta}_1 \mid x_i) = \sigma^2 \sum_{i=1}^n w_i^2$$

Beweis: Effizienz des OLS-Schätzers

$$\text{Var}(\hat{\beta}_1 \mid x_i) = \sigma^2 \sum_{i=1}^n w_i^2$$

Jetzt können wir w_i wieder ausschreiben: Wenn $w_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})x_i}$, dann gilt auch:

$$\sum_{i=1}^n w_i^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\left(\sum_{i=1}^n (x_i - \bar{x})x_i\right)^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)^2} = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}. \text{ Somit:}$$

$$\text{Var}(\hat{\beta}_1 \mid x_i) = \frac{\sigma^2}{\sum_{i=1}^N (x_i - \bar{x})^2}$$



Übungsaufgabe

Wie können wir $\text{Var}(\hat{\beta}_0 \mid x_i) = \frac{\sigma^2 N^{-1} \sum_{i=1}^N x_i^2}{\sum_{i=1}^N (x_i - \bar{x})^2}$ herleiten?

Beweis: Effizienz des OLS-Schätzers

Jetzt widmen wir uns dem zweiten Teil: Ist diese Varianz die geringstmögliche für einen linearen unverzerrten Schätzer? Sei $\tilde{\beta}_1$ irgendein anderer linearer Schätzer, der beliebige Gewichte a_i (statt den OLS-Gewichten w_i) hat:

$$\tilde{\beta}_1 = \sum_{i=1}^N a_i y_i = \sum_{i=1}^N a_i (\beta_0 + \beta_1 x_i + u_i)$$

Da diese Gewichte a_i sich aus den x -Werten ergeben, können wir SLR.4 anwenden, um den Erwartungswert so anzuschreiben:

$$E(\tilde{\beta}_1 | x_i) = \beta_0 \sum_{i=1}^N a_i + \beta_1 \sum_{i=1}^N a_i x_i$$

Da wir voraussetzen, dass auch dieser Schätzer unverzerrt ist, können wir daraus zwei Bedingungen ableiten: $\sum_{i=1}^N a_i = 0$ und $\sum_{i=1}^N a_i x_i = 1$.

Beweis: Effizienz des OLS-Schätzers

Wir können die Gewichte von $\tilde{\beta}_1$ als die OLS-Gewichte plus eine Differenz darstellen:

$$a_i = w_i + d_i$$

Das erlaubt uns, den Schätzer wie folgt anzuschreiben (wir benutzen dieselbe Aufteilung wie vorher beim OLS-Schätzer):

$$\tilde{\beta}_1 = \beta_1 + \sum_{i=1}^N (w_i + d_i) u_i.$$

Die Varianz von $\tilde{\beta}_1$ ist somit:

$$\text{Var}(\tilde{\beta}_1 | x_i) = \sigma^2 \sum_{i=1}^N (w_i + d_i)^2 = \sigma^2 \sum_{i=1}^N (w_i^2 + 2w_i d_i + d_i^2)$$

Beweis: Effizienz des OLS-Schätzers

$$\text{Var}(\tilde{\beta}_1 | x_i) = \sigma^2 \sum_{i=1}^N (w_i + d_i)^2 = \sigma^2 \sum_{i=1}^N w_i^2 + 2w_i d_i + d_i^2$$

Weil $\sum_{i=1}^N a_i = \sum_{i=1}^N (w_i + d_i) = 0$ und $\sum_{i=1}^N w_i = 0$, muss auch

$$\sum_{i=1}^N d_i = 0$$

Außerdem:

$$\sum_{i=1}^N (w_i + d_i)x_i = \sum_{i=1}^N w_i x_i + \sum_{i=1}^N d_i x_i = 1 \Rightarrow \sum_{i=1}^N d_i x_i = 0$$

Beweis: Effizienz des OLS-Schätzers

$$\text{Var}(\tilde{\beta}_1 | x_i) = \sigma^2 \sum_{i=1}^N (w_i + d_i)^2 = \sigma^2 \sum_{i=1}^N w_i^2 + 2w_i d_i + d_i^2$$

Weil $\sum_{i=1}^N d_i = 0$ und $\sum_{i=1}^N d_i x_i = 0$, gilt für den mittleren Term:

$$\sum_{i=1}^N w_i d_i = \frac{\sum_{i=1}^N (x_i - \bar{x})}{\sum_{i=1}^N (x_i - \bar{x})^2} d_i = \frac{1}{\sum_{i=1}^N (x_i - \bar{x})^2} \sum_{i=1}^N x_i d_i - \frac{\bar{x}}{\sum_{i=1}^N (x_i - \bar{x})^2} \sum_{i=1}^N d_i = 0$$

Also reduziert sich der Ausdruck für die Varianz auf

$$\text{Var}(\tilde{\beta}_1 | x_i) = \sigma^2 \sum_{i=1}^N w_i^2 + \sigma^2 \sum_{i=1}^N d_i^2$$

Der Unterschied zur Varianz des OLS-Schätzers ist der **rechte Term**. Da dieser Term nie negativ sein kann, muss die Varianz von $\tilde{\beta}_1$ immer gleich oder größer sein als die von $\hat{\beta}_1$.

Schätzer für σ^2

Zurück zur Varianz des OLS-Schätzers:

$$\text{Var}(\hat{\beta}_1 \mid x_i) = \frac{\sigma^2}{\sum_{i=1}^N (x_i - \bar{x})^2}$$

Wenn wir diese Varianz aus den Daten berechnen wollen, haben wir ein Problem: **Wir kennen σ^2 nicht.**

Unter SLR.1 bis SLR.5 können wir allerdings einen unverzerrten Schätzer für die Varianz finden, und zwar:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^N \hat{u}_i^2}{n - 2},$$

also die **Residuenquadratsumme geteilt durch $n - 2$.**

Standardfehler der Regression

Wenn wir die **Wurzel** aus dem Schätzer für die Varianz des Fehlerterms ziehen, erhalten wir

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2}.$$

Wir nennen diese Größe den **Standardfehler der Regression**. Er ist zwar kein unverzerrter, aber ein **konsistenter** Schätzer für σ . Wir können damit den Standardfehler von $\hat{\beta}_1$, ein Schätzer für die Standardabweichung von $\hat{\beta}_1$ bestimmen:

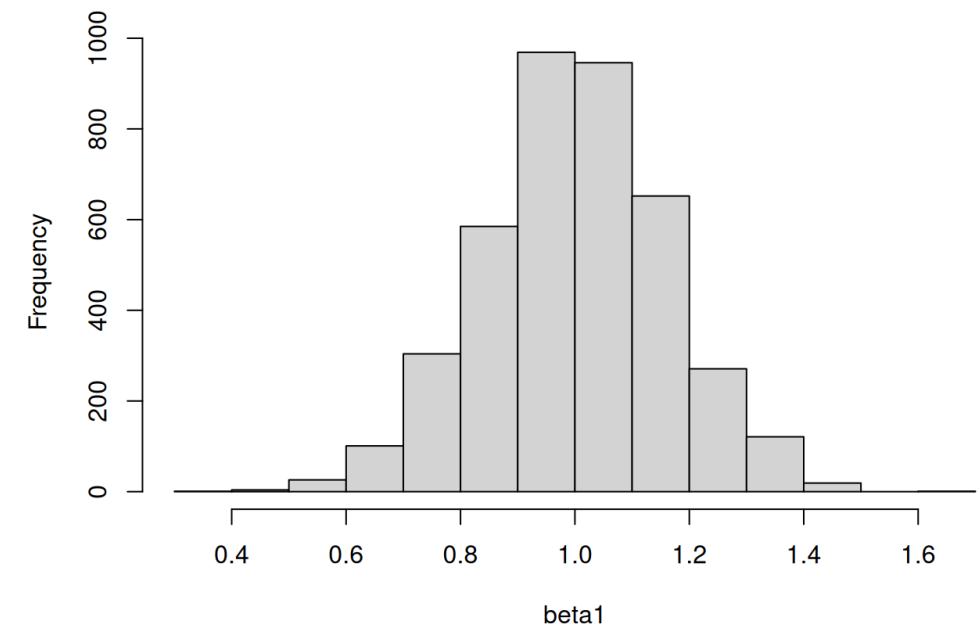
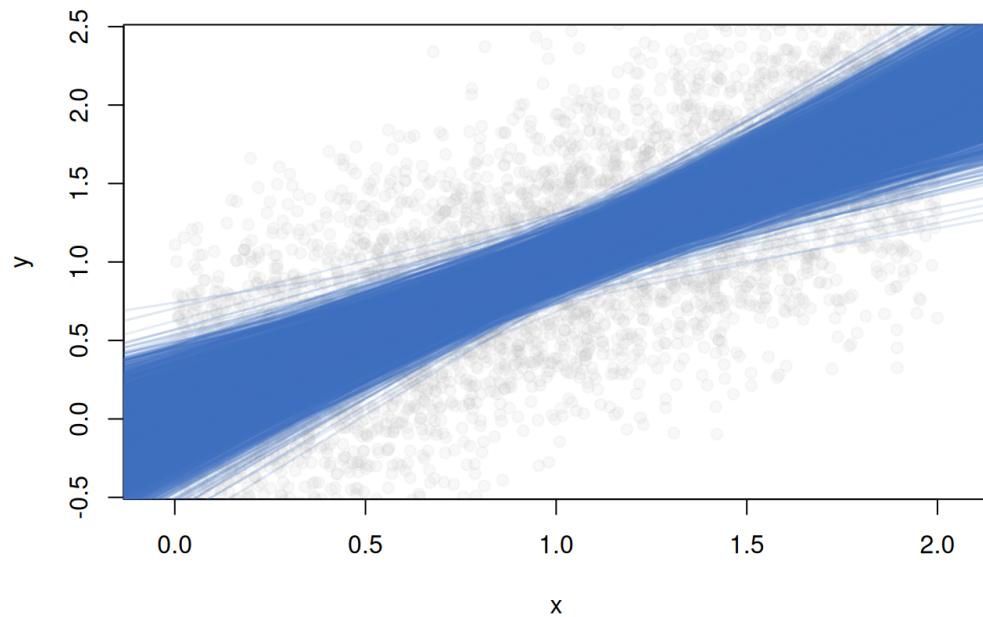
$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2}}$$

Analog können wir den Standardfehler von $\hat{\beta}_0$ bestimmen. Wir können somit messen, wie „genau“ die Koeffizienten geschätzt sind.

Unverzerrtheit bedeutet, dass der Erwartungswert eines Schätzers dem wahren Parameter entspricht. **Konsistenz** bedeutet, dass der Schätzer bei größer werdender Stichprobe zum wahren Parameter konvergiert.

Visualisierung

Wir simulieren 4000 Stichproben aus einer Grundgesamtheit und schätzen 4000 Mal den β_1 -Koeffizienten.



In diesem Beispiel ist die Standardabweichung der β_1 -Koeffizienten 0.161. Der Standardfehler ist 0.1637897.

Der Satz von Gauß-Markow
Erwartungswert des OLS-Schätzers
Varianz des OLS-Schätzers

Regressionen mit nur einem Parameter

Binäre erklärende Variablen
Kausale Inferenz
Appendix

Regressionen ohne Konstante

Was passiert, wenn wir statt dem Modell $y = \beta_0 + \beta_1 x + u$ folgendes Modell schätzen?

$$y = \beta_1 x + u$$

Das bedeutet nichts anderes, als dass wir eine Restriktion $\beta_0 = 0$ auferlegen und somit die Regressionsgerade durch den Ursprung geht.

Der OLS-Schätzer in diesem Fall ist

$$\hat{\beta}_1 = \frac{\sum_{i=1}^N x_i y_i}{\sum_{i=1}^N x_i^2}.$$



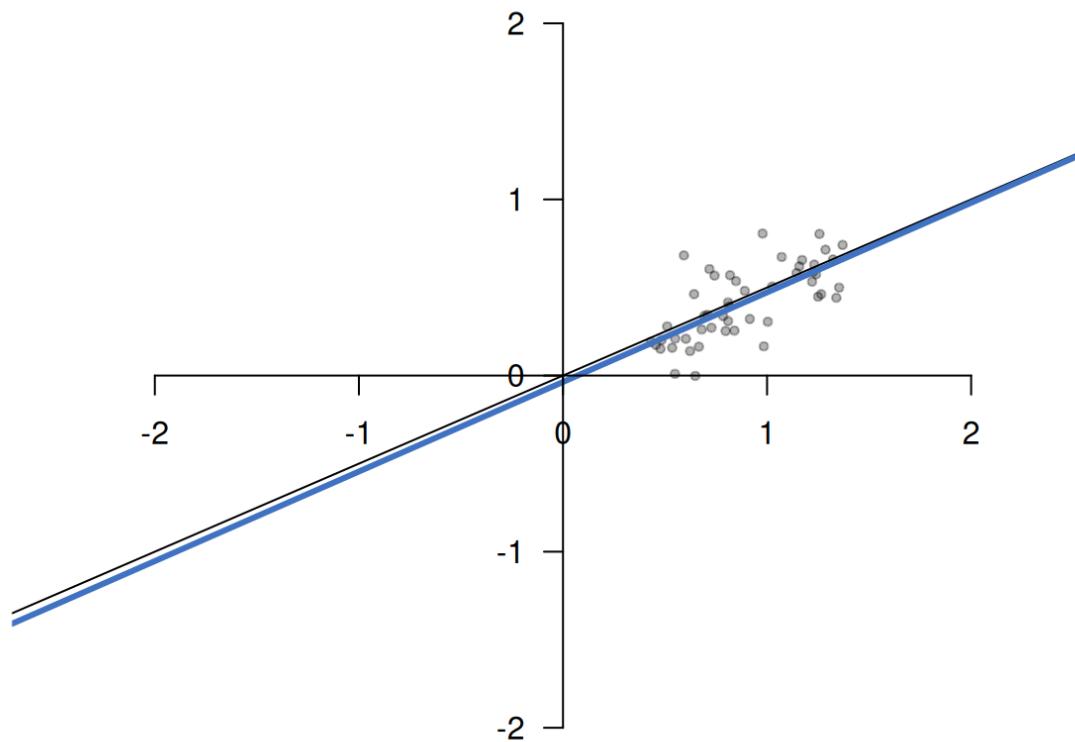
Übungsaufgabe

Wie können wir diesen Schätzer herleiten?

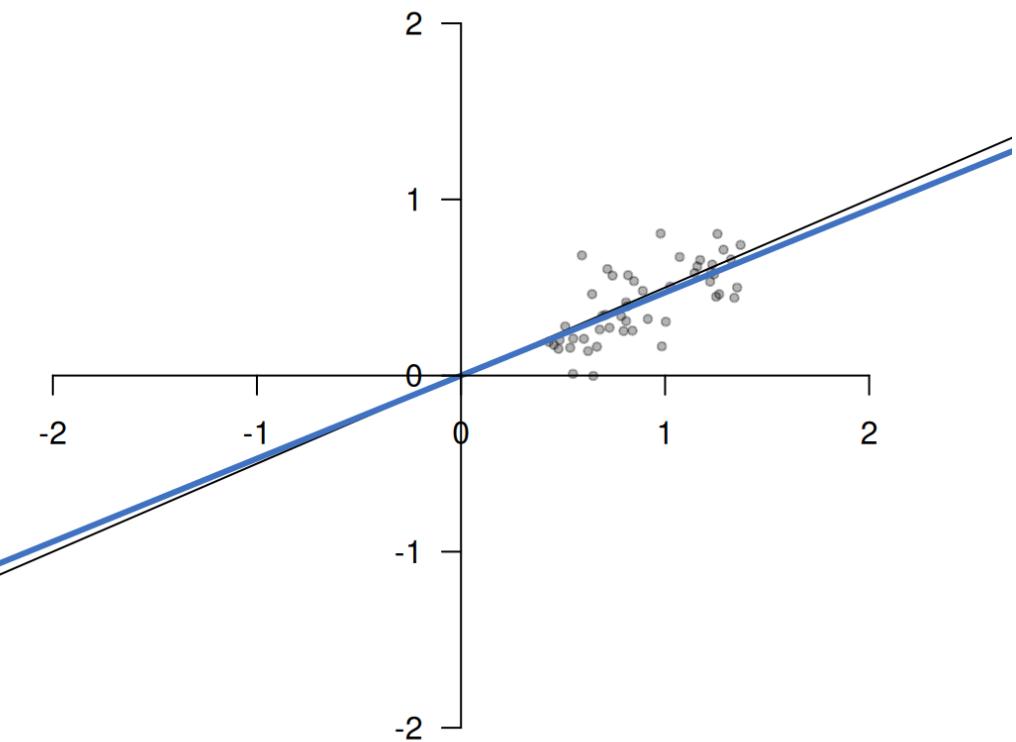
Verzerrung von Regressionen ohne Konstante

Wenn unser **wahres Modell** der Grundgesamtheit **keine Konstante** hat, dann ist dieser Schätzer **unverzerrt**:

OLS ohne Konstante

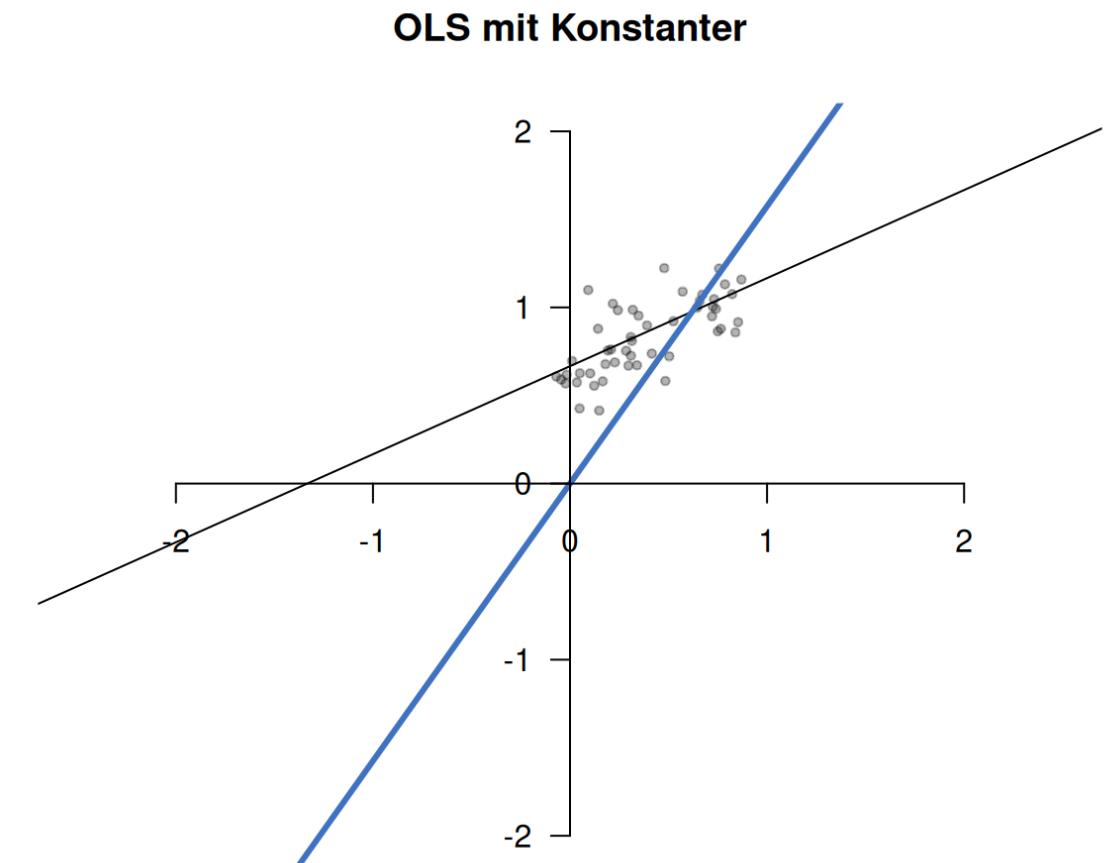
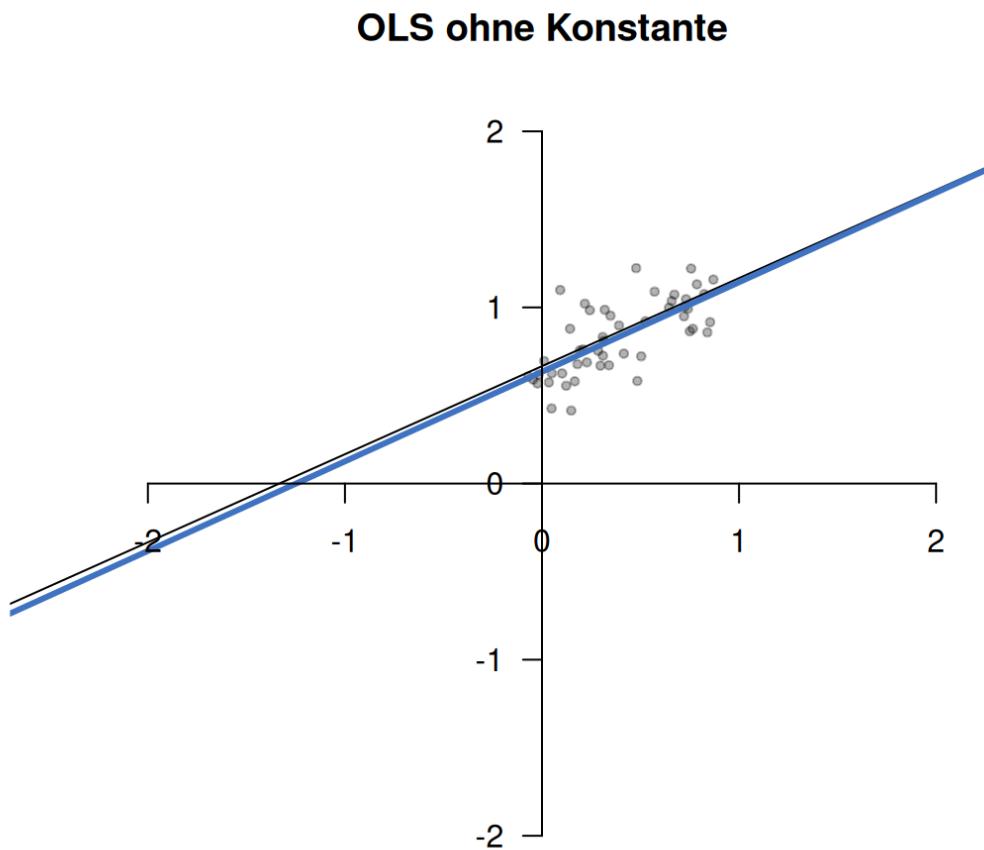


OLS mit Konstanter



Verzerrung von Regressionen ohne Konstante

Wenn unser **wahres Modell** der Grundgesamtheit **keine Konstante** hat, dann ist dieser Schätzer **verzerrt**:



Verzerrung von Regressionen ohne Konstante

Der OLS-Schätzer in einer Regression ohne Konstante ist nur dann unverzerrt, wenn die Konstante im wahren Modell auch 0 ist.

- Wenn das der Fall ist, ist es eigentlich sogar vorzuziehen, ein Modell ohne Konstante zu schätzen (weil wir sonst unnötige Struktur vorgeben).
- Es ist aber fast nie der Fall.
 - Und außerdem wissen wir nie, ob es wirklich der Fall ist.
 - Wir sollten also **nie** eine Regression ohne Konstante rechnen, wenn wir nicht durchschlagende theoretische Gründe dafür haben (wie gesagt: haben wir selten).



Übungsaufgabe

Wie können wir beweisen, dass der Schätzer im oben genannten Fall verzerrt ist?

Regressionen ohne erklärende Variablen

Was passiert, wenn wir statt dem Modell $y = \beta_0 + \beta_1 x + u$ folgendes Modell schätzen?

$$y = \beta_0 + u$$

Das bedeutet nichts anderes, als dass wir eine Restriktion $\beta_1 = 0$ auferlegen und somit die Regressionsgerade horizontal ist.

Der OLS-Schätzer in diesem Fall ist

$$\hat{\beta}_0 = \bar{y},$$

der **Mittelwert** der y -Werte.



Übungsaufgabe

Wie können wir diesen Schätzer herleiten?

Erwartungswert des OLS-Schätzers

Varianz des OLS-Schätzers

Regressions mit nur einem Parameter

Binäre erklärende Variablen

Kausale Inferenz

Appendix

Qualitative und quantitative Information

Bisher sind wir immer von **erklärenden Variablen** mit einer **quantitativen Interpretation** ausgegangen (Ausbildungsjahre, Klassengröße, ...). Wie können wir **qualitative Information** ins Modell einbeziehen?

Angenommen, wir wollen den **Gender Pay Gap** analysieren und sind daher daran interessiert, ob ein Individuum eine Frau ist oder nicht. Wir können eine Variable wie folgt definieren:

$$\text{Frau}_i = \begin{cases} 1 & \text{wenn } i \text{ eine Frau ist,} \\ 0 & \text{andernfalls} \end{cases}$$

Wir nennen eine solche Variable eine **binäre Variable** oder **Dummy-Variable**.

Ein anderes Beispiel wäre ein **Arbeitstrainingsprogramm**. Die Variable $\text{Programmteilnahme}_i$ ist dann 1 für alle Personen, die an dem Programm teilgenommen haben, und 0 für alle anderen.

Interpretation

Wir haben also ein Modell der Form

$$y = \beta_0 + \beta_1 x + u,$$

wo x eine Dummy-Variable ist. Unsere Annahmen SLR.1 bis SLR.5 gelten nach wie vor. Das bedeutet:

$$\begin{aligned} E(y | x = 1) &= \beta_0 + \beta_1, \\ E(y | x = 0) &= \beta_0. \end{aligned}$$

Wir können also β_1 als den **erwarteten Unterschied in y zwischen den beiden Gruppen** interpretieren, und β_1 als den **mittleren Wert in der Gruppe $x = 0$** . Daraus folgt, dass der mittlere Wert in der Gruppe $x = 1$ dann $\beta_0 + \beta_1$ entspricht.

Wir können auch komplexere qualitative Information als nur „ja/nein“ mit Dummy-Variablen kodieren. Dazu benötigen wir aber die Techniken multipler linearer Regression aus dem nächsten Modul.

Varianz des OLS-Schätzers
Regressionen mit nur einem Parameter
Binäre erklärende Variablen

Kausale Inferenz

Appendix

Kontrafaktisches Ergebnis

Wir haben an mehreren Stellen davon gesprochen, dass wir **Behandlungen** oder **Interventionen** (engl. **treatment**) evaluiieren wollen.

- Jetzt, wo wir Dummy-Variablen kennen, wissen wir, wie wir Behandlungsteilnahme modellieren können.
- Wir können also unsere Stichprobe in eine **Behandlungsgruppe** und eine **Kontrollgruppe** aufteilen.
- Grundsätzlich gibt es für jedes Individuum zwei mögliche **Ergebniszustände**:
 - $y_i(1)$ ist das Ergebnis, wenn i an der Intervention teilgenommen hat.
 - $y_i(0)$ ist das Ergebnis, wenn i nicht teilgenommen hat.
- Wir können aber immer nur **einen Zustand beobachten**, da wir keine alternative Realität besuchen können.
- Den nicht beobachteten Zustand bezeichnen wir als **kontrafaktisches Ergebnis** (engl. **counterfactual outcome**).

Kausale Effekte

Es gibt also für jedes Individuum zwei mögliche **Zustände**, von denen wir nur einen beobachten können.

- Könnten wir beide Zustände beobachten, könnten wir sehr leicht einen **kausalen Effekt** isolieren. Wir müssten einfach rechnen

$$\text{Kausaler Effekt}_i = y_i(1) - y_i(0)$$

- Dieser Effekt hat ein Subskript i , er ist also möglicherweise für verschiedene Individuen unterschiedlich.
- Wir werden diesen Effekt **nie** beobachten können, da wir nur eine Realität beobachten. Dieses Problem bezeichnen wir als **fundamentales Problem kausaler Inferenz** (engl. **fundamental problem of causal inference**).
- Wir benötigen also **alternative Strategien**, uns diesem Effekt anzunähern.

Durchschnittlicher Behandlungseffekt (ATE)

Ein Effekt, den wir schätzen können, ist der **durchschnittliche Behandlungseffekt** (engl. **average treatment effect, ATE**):

$$\text{ATE} = E(\text{Kausaler Effekt}_i) = E(y_i(1) - y_i(0)) = E(y_i(1)) - E(y_i(0)).$$

Wenn die Annahmen SLR.1 bis SLR.4 halten, ist der OLS-Schätzer β_1 ein **unverzerrter Schätzer** für den **durchschnittlichen Behandlungseffekt**.

Wir kommen zurück zu dem, was wir vorher schon einmal besprochen haben: Die Annahme SLR.4 (also in diesem Kontext: Die Fehler sind unabhängig von der Zugehörigkeit zur Behandlungsgruppe x) hält nur dann garantiert, wenn die **Zuweisung zur Behandlungsgruppe zufällig** ist, zum Beispiel in einer **randomisierten kontrollierten Studie**.

In Kontexten, in denen eine zufällige Zuweisung zu Behandlungsgruppen nicht möglich ist, können wir mit den bisherigen Methoden keine validen Aussagen über Behandlungseffekte treffen. In Modul 3 diskutieren wir, wie wir dieses Problem mit Methoden **multipler linearer Regression** angehen können.

Regressionen mit nur einem Parameter

Binäre erklärende Variablen

Kausale Inferenz

Appendix

Beste lineare Vorhersagefunktion

Warum verwenden wir die **lineare bedingte Erwartungsfunktion** zur Vorhersage?

Mit einer **quadratischen Verlustfunktion**:

- Wenn $y_i = \beta_0 + \beta_1 x_i + u_i$ das wahre Modell ist, und
- wenn $E(y_i^2) < \infty$, $E(x_i^2) < \infty$, und $\text{Var}(x_i) > 0$,
- können wir zeigen, dass die lineare bedingte Erwartungsfunktion $E(y_i|x_i) = \beta_0 + \beta_1 x_i$ die **beste lineare Vorhersagefunktion** von y_i ist,
- also die eindeutige Lösung von

$$(\beta_0, \beta_1) = \arg \min_{b_0 \in \mathbb{R}, b_1 \in \mathbb{R}} E((y_i - b_0 - b_1 x_i)^2).$$

Wenn wir also die gemeinsame Verteilung von x und y kennen, y mit einem linearen Modell vorhersagen wollen, und den die erwarteten Fehlerquadrate minimieren wollen, ist die lineare bedingte Erwartungsfunktion die beste Funktion, die wir dazu verwenden können.

Beste lineare Vorhersagefunktion

Zwei Bemerkungen:

- (1) **Explizite Lösungen** für Steigung und Konstante in Abhängigkeit der (unbeobachteten) Momente der Grundgesamtheit sind:

$$\beta_1 = \frac{\text{Cov}(x, y)}{\text{Var}(x)} \quad \text{und} \quad \beta_0 = \mathbb{E}(y) - \beta_1 \mathbb{E}(x).$$

- (2) Ein ähnliches Ergebnis wie auf der vorherigen Folie hält auch in allgemeinerer Form: Wenn wir eine quadratische Verlustfunktion verwenden, ist die beste Vorhersagefunktion unbekannter y immer eine bedingte Erwartungsfunktion; auch dann, wenn wir mit nicht-linearen Funktionen arbeiten.

Warum eine quadratische Verlustfunktion?

- Vorwiegend, weil die analytischen Eigenschaften quadratischer Verlustfunktionen bekannt und bequem sind.
- Man kann auch andere Verlustfunktionen verwenden.
 - z.B. führt eine **Absolutbetrag-Verlustfunktion** der Form $|\cdot|$ zum **bedingten median** als Lösung.

• Zurück

Beweis: $\text{Cov}(u, x) = \text{E}(ux)$

$$\text{Cov}(u_i, x_i) = \text{E}(u_i x_i) - \text{E}(u_i)\text{E}(x_i)$$

Weil wir annehmen, dass $\text{E}(u_i) = 0$,

$$\text{Cov}(u_i, x_i) = \text{E}(x_i u_i)$$



• Zurück

Beweis: SST = SSE + SSR

$$\begin{aligned} \text{SST} &= \sum (y_i - \bar{y})^2 \\ &= \sum (y_i - \bar{y} + \underbrace{\hat{y}_i - \hat{y}_i}_{=0})^2 \\ &= \sum ((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}))^2 \\ &= \sum (\hat{u}_i + (\hat{y}_i - \bar{y}))^2 \\ &= \sum (\hat{u}_i^2 + 2\hat{u}_i(\hat{y}_i - \bar{y}) + (\hat{y}_i - \bar{y})^2) \\ &= \sum \hat{u}_i^2 + 2 \sum \hat{u}_i(\hat{y}_i - \bar{y}) + \sum (\hat{y}_i - \bar{y})^2 \\ &= \text{SSR} + 2 \underbrace{\sum \hat{u}_i(\hat{y}_i - \bar{y})}_{=0, \text{ siehe rechts}} + \text{SSE} \\ &= \text{SSR} + \text{SSE} \end{aligned}$$

□

$$\begin{aligned} \sum \hat{u}_i(\hat{y}_i - \bar{y}) &= \sum \hat{u}_i \hat{y}_i - \bar{y} \sum \hat{u}_i \\ &= \sum \hat{u}_i (\hat{\beta}_0 + \hat{\beta}_1 x_i) - \bar{y} \sum \hat{u}_i \\ &= \hat{\beta}_0 \underbrace{\sum \hat{u}_i}_{=0} + \hat{\beta}_1 \underbrace{\sum \hat{u}_i x_i}_{=0} - \bar{y} \underbrace{\sum \hat{u}_i}_{=0} \\ &= 0 \end{aligned}$$

• Zurück

Beweis: $E(u_i \mid x_i) = 0 \Rightarrow E(u_i x_i) = 0$ und $E(u_i) = 0$

Teil 1: $E(u_i \mid x_i) = 0 \Rightarrow E(u_i) = 0$

Zuerst wenden wir den Satz der iterierten Erwartungen an: $E(u_i) = E(E(u_i \mid x_i))$.
Dann nutzen wir die Annahme, dass $E(u_i \mid x_i) = 0$: $E(E(u_i \mid x_i)) = E(0) = 0$. \square

Teil 2: $E(u_i \mid x_i) = 0 \Rightarrow E(u_i x_i) = 0$

Wir wenden wieder den Satz der iterierten Erwartungen an: $E(u_i x_i) = E(E(u_i x_i \mid x_i))$
Da $E(x_i \mid x_i) = x_i$, ist $E(E(u_i x_i \mid x_i)) = E(E(u_i \mid x_i)x_i)$ Dann nutzen wir die
Annahme, dass $E(u_i \mid x_i) = 0$: $E(E(u_i \mid x_i)x_i) = E(0x_i) = 0$. \square

• Zurück