Einführung in die Ökonometrie

Yannick Hoga

Universität Duisburg-Essen

Wintersemester 2022/2023

Überblick

Präliminiarien

- Kontakt:

Yannick Hoga	Martin Arnold
0201 18-34365	0201 18-33587
yannick.hoga@vwl.uni-due.de	martin.arnold@vwl.uni-due.de

- Folien, Aufgaben und Ankündigungen finden Sie auf

  https://moodle.uni-due.de/course/view.php?id=18270

  Das Kurspasswort ist "STOCKWATSON".

- Zu Illustrationszwecken nutzen wir die Programmiersprache R in der Vorlesung und den Übungen.

Ressourcen

- **Literatur**
  - Stock and Watson (2019) *Introduction to Econometrics*. Pearson Education, 4th ed.
    - Eine R basierte, interaktive Version des Buches finden Sie hier: Introduction to Econometrics with `R`
  - Wooldridge (2012) *Introductory Econometrics: A Modern Approach*. Cengage Learning, 5th ed.
  - Murray (2005) *Econometrics: A Modern Introduction*. Addison Wesley, 1st ed.
  - Baltagi (2007) *Econometrics*. Springer, 4th ed.
  - Greene (2019) *Econometric Analysis*. Pearson Education, 8th edition.
- R **packages**: `AER`, `forecast`, `tseries` und viele mehr.
- Laden Sie die obigen Pakete, können Sie die folgenden Grafiken in R replizieren.

F & A

Spezielle Fragen?

Was ist Ökonometrie?

"Broadly speaking, econometrics aims to give empirical content to economic relations for testing economic theories, forecasting, decision making, and for ex-post decision/policy evaluation."

—*The New Palgrave Dictionary of Economics, Second Edition, 2008*

Was ist Ökonometrie?

- Ökonometrie analysiert Daten mit Hilfe statistischer Methoden, um
  - ökonomische Beziehungen zu schätzen,
  - staatliche und unternehmerische Eingriffe zu evaluieren,
  - ökonomische Theorie zu testen und
  - Vorhersagen zu treffen.

Empirische ökonometrische Analyse

1. Formulieren der Forschungsfrage
2. Datenbeschaffung
3. Deskriptive Analyse der Daten
4. Statistische Analyse: Schätzung, Testen und Vorhersage

In dieser Veranstaltung nehmen wir 1. und 2. als gesetzt und beschränken unsere Diskussion auf 3. und 4.

Formulieren der Forschungsfrage

Arten von Fragen

1. **Deskriptiv:** Beschreibe Variable oder Zusammenhang
   - Erzielen Schüler aus kleineren Klasse bessere Abiturnoten?
   - (Einfach zu beantworten und nützlich, aber selten Endziel einer Forschungsarbeit.)

2. **Kausal:** Messen kausaler Effekte
   - Was ist—alles andere gleichhaltend—der Einfluss einer um X Schüler verkleinerten Klasse auf den Abiturschnitt?
   - (Schwer zu beantworten, außer wir können ein perfektes Experiment durchführen.)

3. **Prognose:** Vorhersagen für die Zukunft
   - Was ist der zu erwartende Abiturschnitt einer Klasse mit X Schülern?
   - (Schwer: Wissen die Antwort nie vorher, aber nachher mit Sicherheit.)

Datenbeschaffung I

R Beispiel

Beispiel 1.1: Daten kalifornischer Schulbezirke.

Wir haben Daten von 1999 aus 420 kalifornischen Schulbezirken erhoben. Der *TestScore* misst den Distrikt-weiten Durchschnitt von 5 Klässlern im Stanford 9 Test und die *Student-teacher ratio* (*STR*) gibt die jeweilige durchschnittliche Klassengröße an.

```
data(CASchools)
attach(CASchools)
CASchools$STR   <- students/teachers    # compute STR and append to CASchools
CASchools$score <- (read + math)/2      # compute TestScore and append to CASchools
plot(score ~ STR, data = CASchools,
     xlab = "STR", ylab = "Test Score", ylim=c(590, 720), xlim=c(12, 28))
```

Datenbeschaffung II

Beispiel

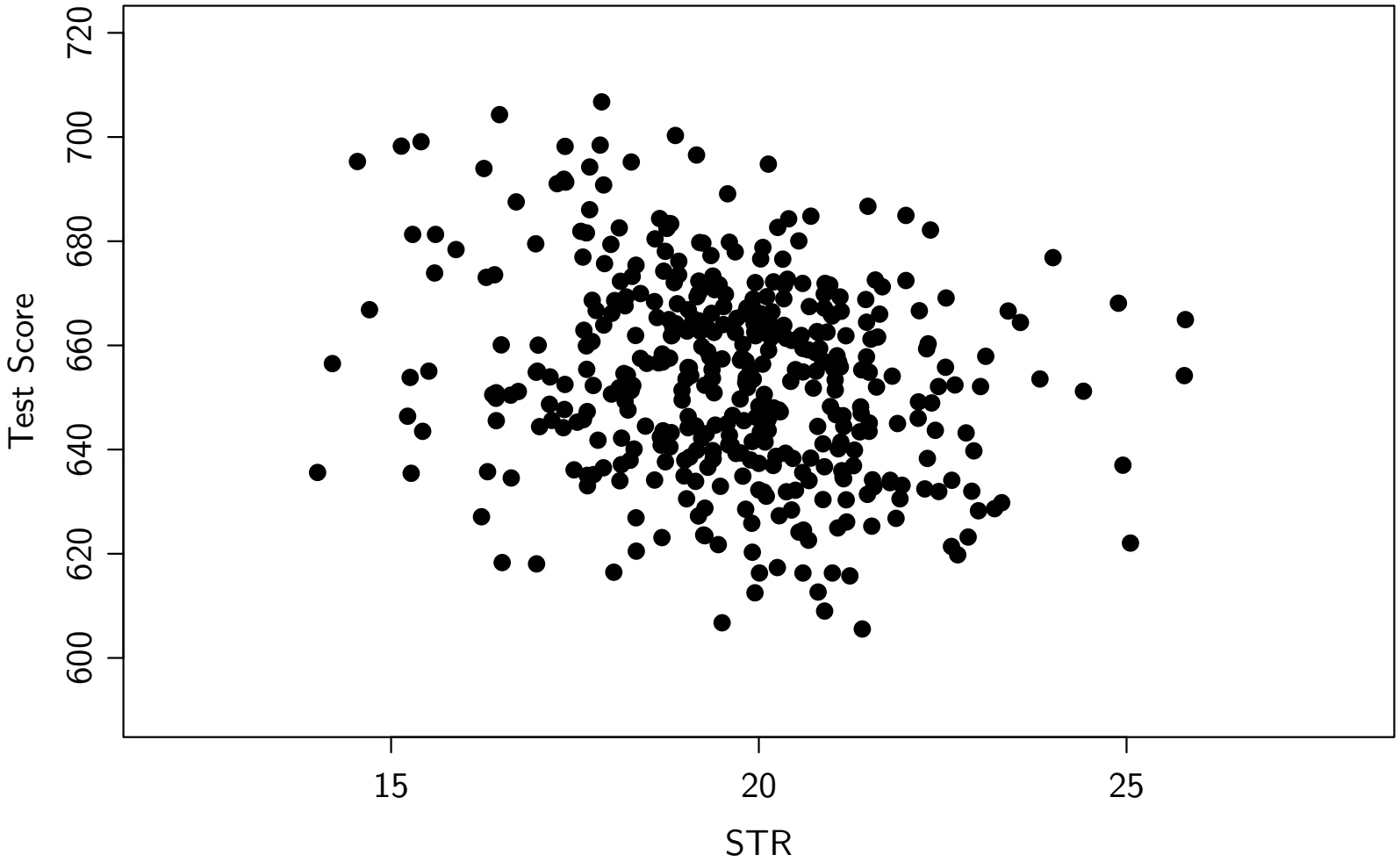


Abbildung 1: Punktediagramm von TestScore und STR.

Datenbeschaffung III

R Beispiel

- Interessante Fragen:
  - Erzielen Schüler aus kleineren Klasse bessere Leistungen? (Deskriptiv)
  - Was ist—alles andere gleichhaltend—der Einfluss einer um X Schüler verkleinerten Klasse auf die Leistung? (Kausal)
- Daten beinhalten nur deskriptive Informationen.
- Schüler aus kleineren Klassen sind in der Tat besser, aber vielleicht nur deshalb, weil sie Kinder reicher Eltern sind (ideale Lernbedingungen zu Hause, Nachhilfe etc.) oder Englisch-Muttersprachler sind?
- Was wäre, wenn wir Distrikte mit ähnlichem Durchschnittseinkommen / Anzahl Muttersprachler vergleichen könnten?
- Eine kausale Interpretation wäre dann viel glaubwürdiger!
- Das Messen kausaler Effekte ist der Hauptfokus dieser Veranstaltung.

Arten von Daten
Experimentelle vs. beobachtete Daten

- **Experimentelle Daten** stammen aus Experimenten, die Interventionen evaluieren oder kausale Effekte messen sollen.
  - Tennessee finanzierte in den 1980ern eine randomisierte Kontrollstudie, die den Einfluss der Klassengröße auf Schülerleistungen messen sollte.
  - Schüler wurden zufällig verschiedenen Klassengrößen zugeteilt und mussten jedes Jahr standardisierten Test absolvieren.
  - Randomisierte Kontrollstudien extrem teuer (manchmal unethisch), also selten in der Ökonomie.

Arten von Daten
Experimentelle vs. beobachtete Daten

- **Beobachtete Daten** sind aus Beobachtungen von tatsächlichem Verhalten außerhalb von Experimenten generiert.
  - Z.B. aus Telefonumfragen, historischen Aufzeichnungen, etc.
  - Häufig in der Ökonomie.
- Kausale Effekte aufgrund von beobachteten Daten zu schätzen ist schwierig!
  - Die "Behandlung" (z.B. Klassengröße im obigen Beispiel) ist nicht zufällig zugeteilt.
  - Also ist es schwierig den Effekt der "Behandlung" von anderen relevanten Faktoren (z.B. reiches Distrikt) zu unterscheiden.

Arten von Daten

Querschnittsdaten, Zeitreihendaten & Paneldaten

- (Experimentelle oder beobachtete) Daten kommen in drei Geschmacksrichtungen:
  1. Querschnittsdaten
  2. Zeitreihendaten
  3. Paneldaten

Arten von Daten

Querschnittsdaten, Zeitreihendaten & Paneldaten

- **Querschnittsdaten** bestehen aus einer Stichprobe von Individuen, Haushalten, Firmen, Städten, Länder oder anderer Einheiten, die zu einem bestimmten Zeitpunkt gezogen wurde.
- Z.B. Stichprobe von Studierenden im Wintersemester 2021 an der UDE:

Beob.	Alter	Größe	Studienfach
1	20	176 cm	VWL
2	24	191 cm	BWL
⋮	⋮	⋮	⋮
100	18	163 cm	Germanistik

Arten von Daten

Querschnittsdaten, Zeitreihendaten & Paneldaten

- **Zeitreihendaten** bestehen aus Beobachtungen einer oder mehrerer Variablen über die Zeit.
- Z.B. Einschreibungsdaten und Notendurchschnitt an der UDE:

Studienjahr	Anz. Studierende	Notendurchschnitt
96-97	20.021	2,05
97-98	21.259	2,03
⋮	⋮	⋮
18-19	40.935	1,94

Arten von Daten

Querschnittsdaten, Zeitreihendaten & Paneldaten

- **Paneldaten** bestehen aus Zeitreihen für jedes Querschnittselement im Datensatz.
- Z.B. Daten mehrerer Studierender über die Zeit:

ID	Studienjahr	Notendurchschnitt
1	96-97	2,84
1	97-98	2,26
2	96-97	1,05
2	97-98	1,28
⋮	⋮	⋮
500	96-97	3,05
500	97-98	2,73

Ziel der Veranstaltung

- Sie werden ...
  - ... die grundlegenden ökonometrischen Werkzeuge kennen lernen,
  - ... verstehen, was durch die (unreflektierte) Anwendung dieser "schief gehen" könnte.
- So
  - ... können Sie die Validität Ihnen präsentierter ökonometrischer Studien einschätzen
  - ... und eigene empirische Studien durchführen.
- Hauptsächlich beschäftigen wir uns dabei mit der Schätzung **kausaler** Effekte aufgrund von **beobachteten Querschnittsdaten**.

Überblick

Wahrscheinlichkeitstheorie und Statistik
(SW Kapitel 2 & 3)

Überblick

- **Wahrscheinlichkeitstheorie**
- Statistik

Zufallsvariable

- Die meisten Untersuchungsgegenstände in der Ökonomie können als **Zufallsvariablen (ZVen)** aufgefasst werden.
  - Aktienkurse, Löhne, Nachfrage nach einem neuen Produkt zu gegebenem Preis etc.
- Informell: Eine Zufallsvariable ist eine numerische Variable mit einem Zufallselement.
- Also könnte die Zufallsvariable auch einen anderen Wert angenommen haben.
- Zwei Arten: **Diskrete** und **stetige** ZVen
  - Diskret: Nimmt nur endlich viele Werte an
  - Stetig: Nimmt Werte auf Kontinuum an. (Hier der häufigere Fall.)

Verteilung und Momente der Verteilung

- Die **Verteilung** einer ZV Y ist eindeutig festgelegt durch ihre **Verteilungsfunktion**

$$F(y) = \mathrm{P}(Y \leq y), \qquad y \in \mathbb{R}.$$

- $\mathrm{E}(Y)$ bezeichnet den **Erwartungswert** von Y.
  - Der Erwartungswert ist das zentrale Lagemaß.
  - Y diskret (mit k verschiedenen Werten): $\mathrm{E}(Y) = \sum_{i=1}^{k} y_i \mathrm{P}(Y = y_i)$.
- $\mathrm{Var}(Y) = \mathrm{E}(Y - \mathrm{E}(Y))^2$ bezeichnet die **Varianz** von Y.
  - Die Varianz ist das zentrale Streuungsmaß.
- $\mathrm{sd}(Y) = \sqrt{\mathrm{Var}(Y)}$ bezeichnet die **Standardabweichung** von Y.

Verteilung und ihre Momente

Rechenregeln

Seien X und Y zwei Zufallsvariablen und $a, b \in \mathbb{R}$. Dann gelten:

- $\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y)$.
- $\mathrm{E}(aX + b) = a\,\mathrm{E}(X) + b$ ("Linearität des Erwartungswertes").
- $\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X)$.
- $\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$ ("Verschiebungssatz").
- Sind X und Y unabhängig, so gilt $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Gemeinsame Verteilung und ihre Momente

- Die **gemeinsame Verteilung** zweier ZVen X und Y ist eindeutig festgelegt durch ihre **gemeinsame Verteilungsfunktion**

$$F(x, y) = \mathrm{P}(X \leq x, Y \leq y), \qquad x, y \in \mathbb{R}.$$

- Die **Kovarianz** zwischen X und Y ist

$$\mathrm{Cov}(X, Y) = \mathrm{E}[\{X - \mathrm{E}(X)\}\{Y - \mathrm{E}(Y)\}].$$

- Der **Korrelationskoeffizient** von X und Y ist

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}.$$

- Der Korrelationskoeffizient ist ein Maß des *linearen* Zusammenhangs.
  - $-1 \leq \mathrm{Corr}(X, Y) \leq 1$.
  - $\mathrm{Corr}(X, Y) = -1/0/1$ bedeutet perfekter negativer/kein/perfekter positiver linearer Zusammenhang; `StockReturnCorrelation.R`

Gemeinsame Verteilung und ihre Momente

Rechenregeln

Seien X und Y zwei Zufallsvariablen und $a, b, c, d \in \mathbb{R}$. Dann gelten:

- $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.
- $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$.
- $\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y)$.
- $\mathrm{Cov}(aX + b, cY + d) = ac\,\mathrm{Cov}(X, Y)$, also ändert sich die Kovarianz mit den Einheiten.
- Sind X und Y unabhängig, so gilt $\mathrm{Cov}(X, Y) = \mathrm{Corr}(X, Y) = 0$. (Aber nicht anders herum!)

Bedingte Verteilung und ihre Momente

- Die **bedingte Verteilung** von Y gegeben $X = x$ ist eindeutig festgelegt durch die **bedingte Verteilungsfunktion**

$$\mathrm{P}(Y \leq y \mid X = x) \overset{\mathrm{P}(X=x)>0}{=} \frac{\mathrm{P}(Y \leq y, X = x)}{\mathrm{P}(X = x)}.$$

- ZV $\mathrm{E}(Y \mid X)$ bezeichnet den **bedingten Erwartungswert** von Y gegeben X.
- Für Realisation $X = x$ nimmt sie deterministischen Wert $\mathrm{E}(Y \mid X = x)$ an.
  - Y diskret (mit k verschiedenen Werten y_i):
    $\mathrm{E}(Y \mid X = x) = \sum_{i=1}^{k} y_i \mathrm{P}(Y = y_i \mid X = x)$.
- ZV $\mathrm{Var}(Y \mid X) = \mathrm{E}\left[\{Y - \mathrm{E}(Y \mid X)\}^2 \mid X\right]$ bezeichnet die **bedingte Varianz** von Y gegeben X.
- Für Realisation $X = x$ nimmt sie deterministischen Wert $\mathrm{Var}(Y \mid X = x)$ an.

Bedingte Verteilung und ihre Momente
Rechenregeln

Seien X, Y, Z ZVen und $a, b \in \mathbb{R}$. Dann gelten:

- $\mathrm{E}(aY + bZ \mid X) = a\,\mathrm{E}(Y \mid X) + b\,\mathrm{E}(Z \mid X)$.
- $\mathrm{E}(g(X) \mid X) = g(X)$ für jede Funktion $g(\cdot)$.
- Y und X unabhängig $\implies \mathrm{E}(Y \mid X) = \mathrm{E}(Y) \implies \mathrm{Cov}(X, Y) = 0$.
- $\mathrm{E}(Y) = \mathrm{E}[\mathrm{E}(Y \mid X)]$, weil (für diskretes X/Y mit Werten $x_1, \ldots, x_I / y_1, \ldots, y_k$)

$$
\begin{aligned}
\mathrm{E}[\mathrm{E}(Y \mid X)] &= \sum_{i=1}^{I} \mathrm{E}(Y \mid X = x_i)\mathrm{P}(X = x_i) \\
&= \sum_{i=1}^{I} \sum_{j=1}^{k} y_j \mathrm{P}(Y = y_j \mid X = x_i)\mathrm{P}(X = x_i) \\
&= \sum_{j=1}^{k} y_j \mathrm{P}(Y = y_j) = \mathrm{E}(Y).
\end{aligned}
$$

Bedingte Verteilung und ihre Momente
Beispiel

- Sei $Y = \textit{wage}$ und $X = \textit{gender} \in \{\textit{male}, \textit{female}\}$.
- Dann ist $\mathrm{E}(\textit{wage} \mid \textit{gender})$ eine ZV, die zwei Werte annehmen kann:

$$\mathrm{E}(\textit{wage} \mid \textit{gender} = \textit{male}) \quad \text{und} \quad \mathrm{E}(\textit{wage} \mid \textit{gender} = \textit{female}).$$

  - $\mathrm{E}(\textit{wage} \mid \textit{gender} = \textit{male})$ = Durchschnittslohn aller Männer.
  - $\mathrm{E}(\textit{wage} \mid \textit{gender} = \textit{female})$ = Durchschnittslohn aller Frauen.

- Der Erwartungswert von $\mathrm{E}(\textit{wage} \mid \textit{gender})$ ist daher

$$\begin{aligned}\mathrm{E}(\textit{wage}) &= \mathrm{E}[\mathrm{E}(\textit{wage} \mid \textit{gender})] \\ &= \mathrm{E}(\textit{wage} \mid \textit{gender} = \textit{male})\mathrm{P}(\textit{gender} = \textit{male}) \\ &\quad + \mathrm{E}(\textit{wage} \mid \textit{gender} = \textit{female})\mathrm{P}(\textit{gender} = \textit{female})\end{aligned}$$

Schätzung des Erwartungswertes

- Viel Schätztheorie im Folgenden gleicht der Schätztheorie für den Erwartungswert.
- Der Erwartungswert von Y ist als Parameter der Verteilung in der Grundgesamtheit i.A. unbekannt.
- Wir müssen ihn aufgrund von Beobachtungen schätzen.
- Dazu nehmen wir eine **einfache Zufallsstichprobe** $Y_1, \ldots, Y_n$ aus Y an, d.h.
  - Y_i sind **unabhängig, identisch verteilt (u.i.v.)** mit derselben Verteilung wie Y.
- Ein intuitiver Schätzer für den "durchschnittlichen" Wert $\mathrm{E}(Y)$ ist der **Stichprobendurchschnitt** $\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$.
- Da $\overline{Y}$ wieder eine ZV ist, hat auch sie eine Verteilung – die sogenannte **Stichprobenverteilung**.

Schätzung des Erwartungswertes

- $\overline{Y}$ ist der natürliche Schätzer des Erwartungswertes, aber:
  1. Was sind Eigenschaften von $\overline{Y}$?
  2. Warum sollten wir $\overline{Y}$ benutzen und nicht einen anderen Schätzer?
     - Y_1 (die erste Beobachtung)
     - ungleiche Gewichte – keinen einfachen Durchschnitt
     - Median von $Y_1, \ldots, Y_n$
- Nehme im Folgenden immer eine einfache Zufallsstichprobe $Y_1, \ldots, Y_n$ an!

1. Eigenschaften von $\overline{Y}$

Erwartungswert und Varianz

- Erwartungswert:

$$\mathrm{E}(\overline{Y}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(Y_i) = \mathrm{E}(Y).$$

- Varianz:

$$\begin{aligned}\mathrm{Var}(\overline{Y}) &= \frac{1}{n^2}\,\mathrm{Var}\left(\sum_{i=1}^{n} Y_i\right)\\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(Y_i)\\ &= \mathrm{Var}(Y)/n.\end{aligned}$$

1. Eigenschaften von $\overline{Y}$

Erwartungswert und Varianz

- Implikationen von

$$\mathrm{E}(\overline{Y}) = \mathrm{E}(Y),$$
$$\mathrm{Var}(\overline{Y}) = \mathrm{Var}(Y)/n :$$

1. $\overline{Y}$ ist ein **unverzerrter** Schätzer für $\mathrm{E}(Y)$.
2. $\sqrt{\mathrm{Var}(\overline{Y})}$ is invers proportional zu $\sqrt{n}$.
   - Die Streuung der Stichprobenverteilung von $\overline{Y}$ ist proportional zu $\frac{1}{\sqrt{n}}$.
   - Größere Stichprobe heißt weniger Unsicherheit, aber skaliert durch Quadratwurzel.

1. Eigenschaften von $\overline{Y}$

Stichprobenverteilung

- Für endliche n ist die Stichprobenverteilung von $\overline{Y}$ – über $\mathrm{E}(\overline{Y})$ und $\mathrm{Var}(\overline{Y})$ hinaus – sehr kompliziert.
- Für $n \to \infty$ gilt aber:

1. **Gesetz der großen Zahlen**: $\overline{Y}$ ist **konsistent** für $\mathrm{E}(Y)$, d.h. $\overline{Y} \xrightarrow{p} \mathrm{E}(Y)$, d.h. für alle $\varepsilon > 0$ gilt

$$\mathrm{P}\big(|\overline{Y} - \mathrm{E}(Y)| < \varepsilon\big) \to 1, \qquad n \to \infty.$$

   - Intuition: $\overline{Y}$ zentriert sich immer enger um $\mathrm{E}(Y)$.

1. Eigenschaften von $\overline{Y}$

Stichprobenverteilung

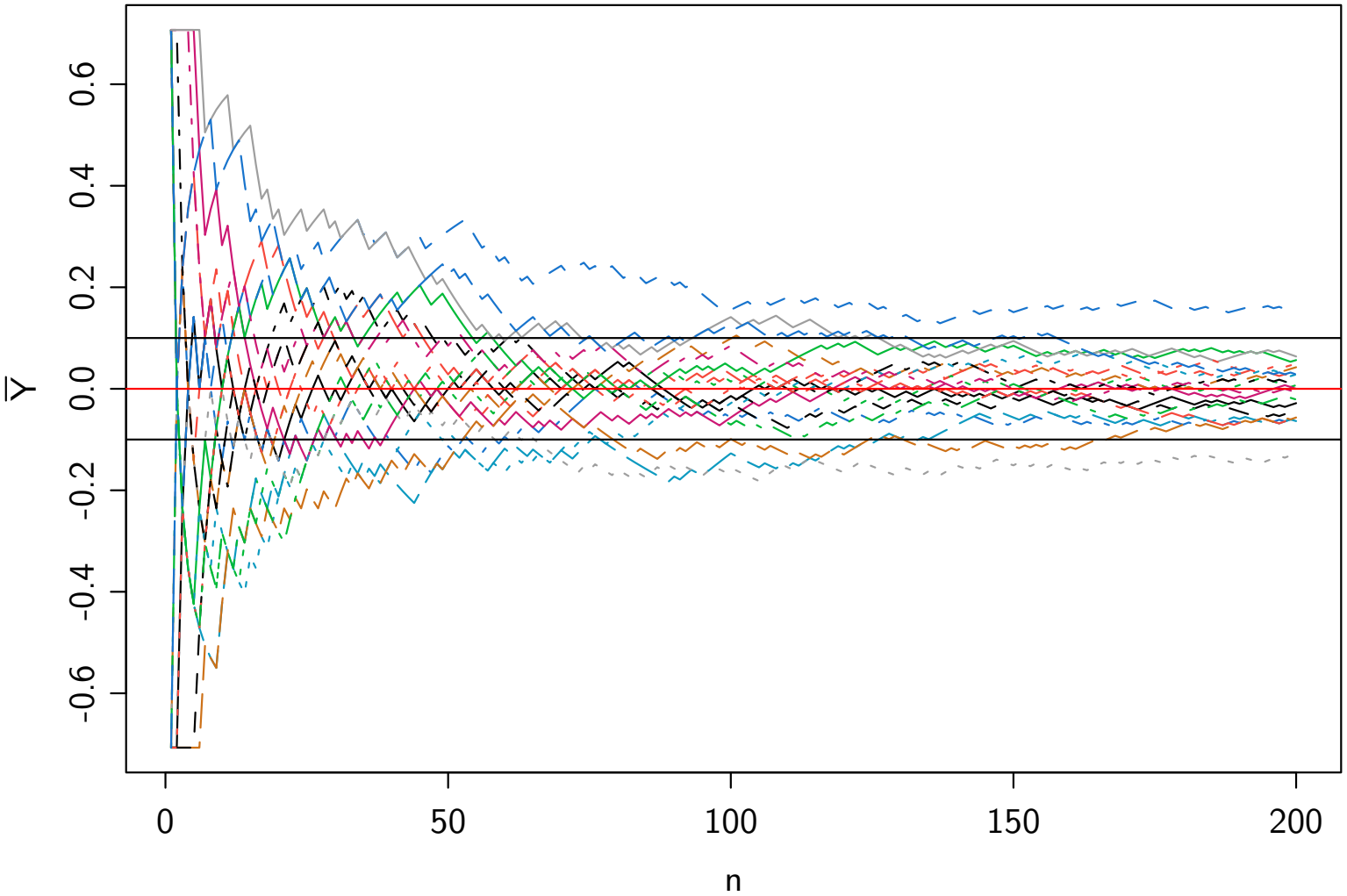


Abbildung 2: Gesetz der grossen Zahlen

1. Eigenschaften von $\overline{Y}$

Stichprobenverteilung

2. **Zentraler Grenzwertsatz (ZGWS)**: Wenn $\text{Var}(Y) < \infty$, dann $\sqrt{n}[\overline{Y} - \text{E}(Y)]/\,\text{sd}(Y) \xrightarrow{d} N(0,1)$ für $n \to \infty$, d.h.

$$\text{P}\left(\sqrt{n}[\overline{Y} - \text{E}(Y)]/\,\text{sd}(Y) \leq x\right) \to \Phi(x), \qquad x \in \mathbb{R},$$

wobei $\Phi(\cdot)$ die Verteilungsfunktion einer $N(0,1)$-ZV.

- Beachte, dass $\sqrt{n}[\overline{Y} - \text{E}(Y)]/\,\text{sd}(Y)$ **standardisiert** ist (also Erwartungswert=0 und Varianz=1 hat), da

$$\sqrt{n}\frac{\overline{Y} - \text{E}(Y)}{\sqrt{\text{Var}(Y)}} = \frac{\overline{Y} - \text{E}(\overline{Y})}{\sqrt{\text{Var}(\overline{Y})}}.$$

- Der ZGWS impliziert, dass $\overline{Y}$ für große n ungefähr normalverteilt ist: $\overline{Y} \overset{d}{\approx} N\left(\text{E}(Y), \frac{\text{Var}(Y)}{n}\right)$.

1. Eigenschaften von $\overline{Y}$

Stichprobenverteilung von $\overline{Y}$ für $\mathrm{P}(Y = 1) = 1 - \mathrm{P}(Y = -1) = 1/2$

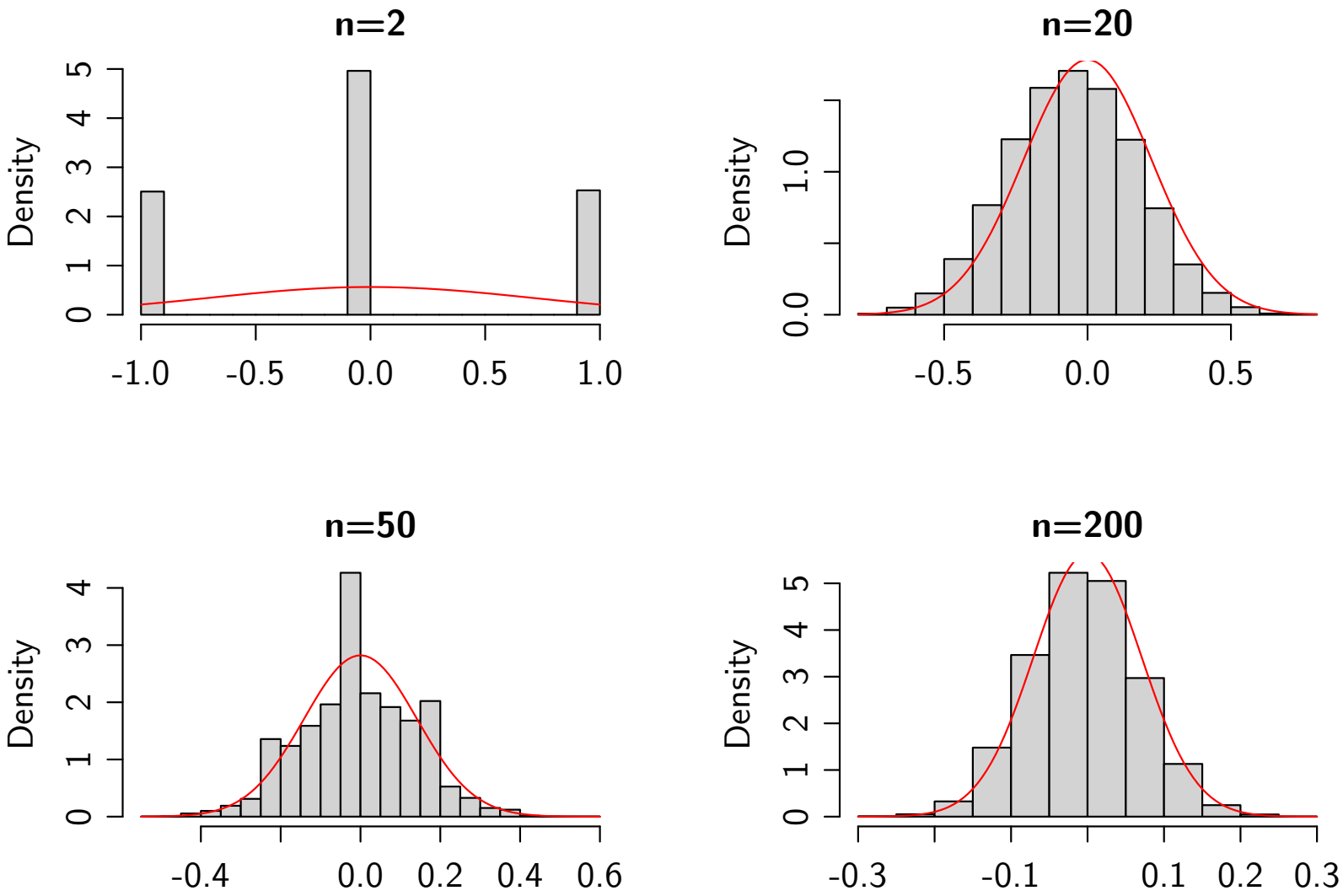


Abbildung 3: Zentraler Grenzwertsatz

1. Eigenschaften von $\overline{Y}$

Zusammenfassung: Stichprobenverteilung von $\overline{Y}$

Für u.i.v. $Y_1, \ldots, Y_n$ mit $0 < \text{Var}(Y) < \infty$ gilt:

- Die exakte Stichprobenverteilung von $\overline{Y}$ hat Erwartungswert $\text{E}(Y)$ ("$\overline{Y}$ ist ein unverzerrter Schätzer für $\text{E}(Y)$") und Varianz $\text{Var}(Y)/n$
- Bis auf Erwartungswert und Varianz von $\overline{Y}$ ist die exakte Verteilung kompliziert und hängt von der Verteilung von Y ab (der Populationsverteilung)
- Wenn $n \to \infty$, vereinfacht sich die Stichprobenverteilung:
  - Konsistenz: $\boxed{\overline{Y} \xrightarrow{p} \text{E}(Y)}$
  - Asymptotische Normalität: $\boxed{\sqrt{n}\frac{\overline{Y}-\text{E}(Y)}{\sqrt{\text{Var}(Y)}} \xrightarrow{d} N(0,1)}$

2. Warum $\overline{Y}$?

- $\overline{Y}$ ist unverzerrt: $\mathrm{E}(\overline{Y}) = \mathrm{E}(Y)$.
- $\overline{Y}$ ist konsistent: $\overline{Y} \xrightarrow{p} \mathrm{E}(Y)$.
- $\overline{Y}$ ist der "kleinste Quadrate" Schätzer für $\mathrm{E}(Y)$: $\overline{Y}$ löst

$$\min_m \sum_{i=1}^n (Y_i - m)^2,$$

  also minimiert $\overline{Y}$ die Summe der quadrierten "Residuen".
- Optionale Herleitung:

$$\frac{\partial}{\partial m} \sum_{i=1}^n (Y_i - m)^2 = \sum_{i=1}^n \frac{\partial}{\partial m} (Y_i - m)^2 = -2 \sum_{i=1}^n (Y_i - m).$$

  Setzte die Ableitung gleich Null und bezeichne die Lösung mit $\widehat{m}$:

$$\sum_{i=1}^n Y_i = \sum_{i=1}^n \widehat{m} = n\widehat{m} \qquad \text{oder} \qquad \widehat{m} = \frac{1}{n} \sum_{i=1}^n Y_i = \overline{Y}.$$

2. Warum $\overline{Y}$?

- **Hauptgrund**: $\overline{Y}$ hat eine kleinere Varianz als alle anderen linearen unverzerrten Schätzer:
  - Betrachte den Schätzer $\widehat{\mu}_Y = \frac{1}{n}\sum_{i=1}^{n} a_i Y_i$, wobei die $\{a_i\}$ so konstruiert sind, dass $\widehat{\mu}_Y$ unverzerrt ist; dann gilt $\text{Var}(\overline{Y}) \leq \text{Var}(\widehat{\mu}_Y)$.
- $\overline{Y}$ ist der **Maximum Likelihood Schätzer** für $\text{E}(Y)$ unter der Annahme von normalverteilten $Y_1, \ldots, Y_n$.
- Aber $\overline{Y}$ ist nicht der einzige Schätzer für $\text{E}(Y)$ – können Sie sich eine Situation vorstellen, in der sie stattdessen den Median verwenden sollten?

Überblick

- Wahrscheinlichkeitstheorie
- **Statistik**

Hypothesentests

- Was sind Hypothesentests?
- Sei H_0 eine Hypothese (die **Nullhypothese**), die Sie testen wollen.
- Wir nehmen an, diese sei wahr, und fragen uns, ob die Welt mit ihr konsistent ist.
- Dazu brauchen wir drei Zutaten:

  1. Die konkrete Nullhypothese H_0
  2. Daten (entweder beobachtete oder experimentelle) zum Testen von H_0
  3. Eine **Ablehnungsregel**, die sagt, wann H_0 aufgrund von Daten zu verwerfen ist

Beispielhypothesen

- H_0: Es gibt keine Erdanziehung
  - Daten: Schmeißen Sie ihren Stift vom Tisch.
  - Ablehnungsregel: Lehne ab, wenn der Stift herunterfällt.
  - Problem: Keine Probleme – können Sicherheit über H_0 gewinnen.
- H_0: Alle Schwäne sind weiß
  - Experiment: Finden Sie 100 Schwäne und schauen, ob alle weiß sind.
  - Ablehnungsregel: Lehne ab, wenn ein Schwan nicht weiß ist.
  - Problem: Was, wenn alle Schwäne weiß sind?
- H_0: Der Erwartungswert der ZV Y ist 1
  - Daten: Ziehe eine einfache Stichprobe und gucke, ob $\overline{Y} \approx 1$.
  - Ablehnungsregel: Lehne ab, wenn $|\overline{Y} - 1| \gg 0$.
  - Problem: Was, wenn $\overline{Y} = 0.9$ oder $\overline{Y} = 2$?

Mögliche Alternativen

- Im Folgenden betrachten wir letzteres Problem, also H_0: $\mathrm{E}[Y] = \mu_{Y,0}$.
- Für unser Beispiel mit dem Erwartungswert sind verschiedene **Alternativen** denkbar:

$$
\begin{aligned}
H_0 &: \ \mathrm{E}(Y) = \mu_{Y,0} \quad \text{vs.} \quad H_1 : \ \mathrm{E}(Y) > \mu_{Y,0} \quad (\text{1-seitig}, >) \\
H_0 &: \ \mathrm{E}(Y) = \mu_{Y,0} \quad \text{vs.} \quad H_1 : \ \mathrm{E}(Y) < \mu_{Y,0} \quad (\text{1-seitig}, <) \\
H_0 &: \ \mathrm{E}(Y) = \mu_{Y,0} \quad \text{vs.} \quad H_1 : \ \mathrm{E}(Y) \neq \mu_{Y,0} \quad (\text{2-seitig})
\end{aligned}
$$

- Je nach Alternative ändert sich der Ablehnbereich, aber die Teststatistik bleibt immer dieselbe.

Entscheidungen und Fehler

- Egal wie H_0, die Alternative oder die Ablehnungsregel aussehen, ...
- ... es gibt immer nur vier mögliche Szenarien:

Wirklichkeit / Entscheidung	H_0 wahr	H_0 falsch
lehne H_0 nicht ab	korrekt	falsch (**Fehler 2. Art**)
lehne H_0 ab	falsch (**Fehler 1. Art**)	korrekt

Wann verwerfen?

- Wir können also zwei Arten von Fehlern machen:
  - **Fehler 1. Art**: $\mathrm{P}(\text{lehne } H_0 \text{ ab} \mid H_0 \text{ ist wahr})$.
  - **Fehler 2. Art**: $\mathrm{P}(\text{lehne } H_0 \text{ nicht ab} \mid H_1 \text{ ist wahr})$.
- Idealerweise sollten beide Fehler klein sein.
- Es gibt einen Tradeoff zwischen Fehler 1. Art und Fehler 2. Art.
  - Teste $H_0: \ \mathrm{E}(Y) = 1$ vs. $H_1: \ \mathrm{E}(Y) \neq 1$
  - Lehne ab, wenn $|\overline{Y} - 1|$ größer als ein x.
  - Intuitiv: Großes $x \ \rightarrow$ kleiner Fehler 1. Art, aber großer Fehler 2. Art.

Beispiel

Beispiel 2.1: Schwangerschaft.

- H_0: Person nicht schwanger
  - Daten: Miss Bauchumfang
  - Ablehnungsregel: Verwirf H_0, wenn Bauchumfang $> x$ cm.
  - Intuitiv: Großes $x \rightarrow$ kleiner Fehler 1. Art, aber großer Fehler 2. Art.

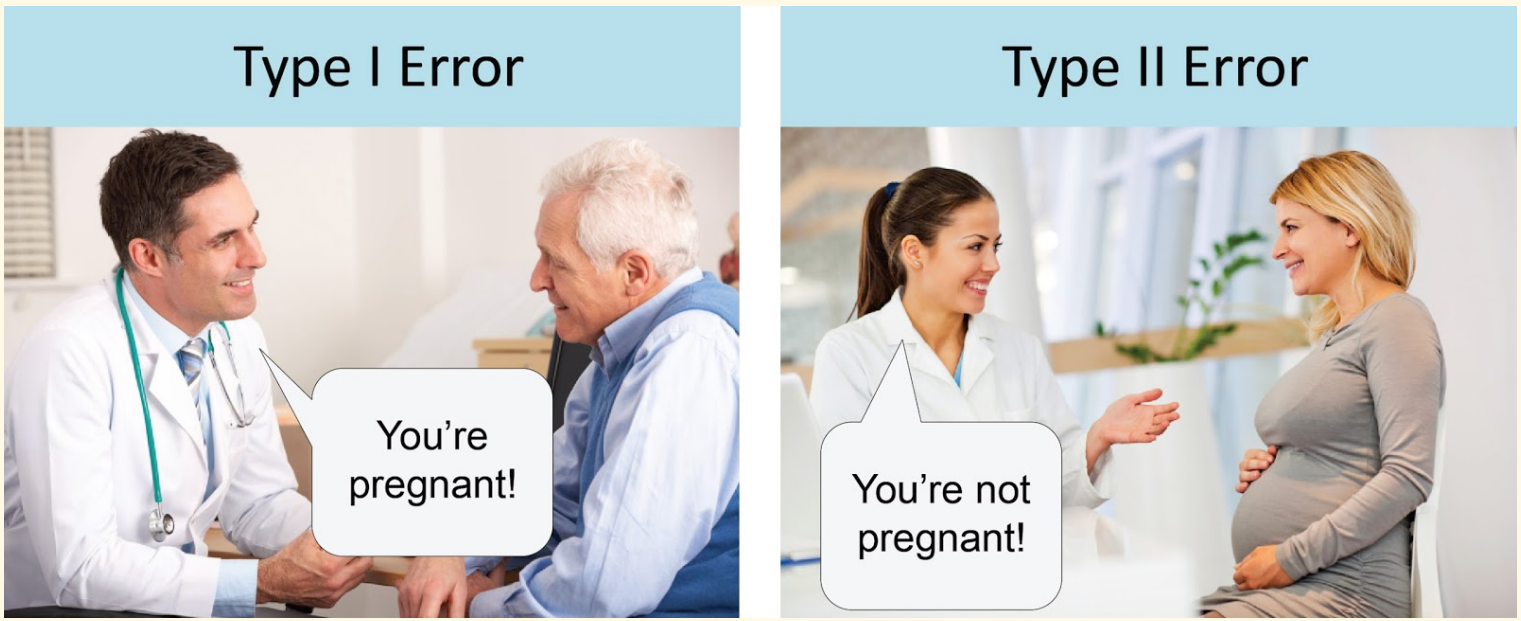


Abbildung 4: unbiasedresearch.blogspot.com

Wann verwerfen?

- Lösung für Tradeoff zwischen Fehler 1. Art und Fehler 2. Art:
  - Fixiere das **Signifikanzniveau**, also die Wahrscheinlichkeit eines Fehlers 1. Art, bei $\alpha \in (0, 1)$.
  - Oft: $\alpha \in \{0.01,\ 0.05,\ 0.10\}$.
- Der Fehler 2. Art wird **nicht** kontrolliert – man hofft meist, dass er klein ist!
- Der ***p*-Wert** (auch genannt das **marginale Signifikanzniveau**) ist die Wahrscheinlichkeit unter H_0 eine Statistik zu ziehen (z.B. $\overline{Y}$), die mindestens genau so stark gegen H_0 spricht, wie die tatsächlich gezogene.
  - Also: Wenn p-Wert kleiner als 0.05, würde man H_0 zum Signifikanzniveau $\alpha = 0.05$ verwerfen.
- Oft: Besser p-Wert angeben, als einfach Testentscheidung "ja/nein" – p-Wert ist informativer.

Berechnung des p-Wertes

- Sei $\overline{Y}^{act}$ der tatsächlich beobachtete Wert von $\overline{Y}$ und

$$H_0: \ \mathrm{E}(Y) = \mu_{Y,0} \qquad \text{vs.} \qquad H_1: \ \mathrm{E}(Y) \neq \mu_{Y,0} \quad \text{(2-seitig)}.$$

- Um asymptotische Normalität von $\overline{Y}$ zu nutzen, schätzen wir $\mathrm{Var}(Y)$.
- Benutze dazu

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \textbf{Stichprobenvarianz} \text{ von } Y.$$

- **Fakt**: Wenn $Y_1 \ldots, Y_n$ u.i.v. und $\mathrm{E}(Y^4) < \infty$, dann $s_Y^2 \xrightarrow{p} \mathrm{Var}(Y)$. (Stichwort: Gesetz der großen Zahlen.)
- Vorausschauend definieren wir

$$t = \sqrt{n}\frac{\overline{Y} - \mu_{Y,0}}{s_Y} (\approx N(0,1)) \qquad \text{und} \qquad t^{act} = \sqrt{n}\frac{\overline{Y}^{act} - \mu_{Y,0}}{s_Y}.$$

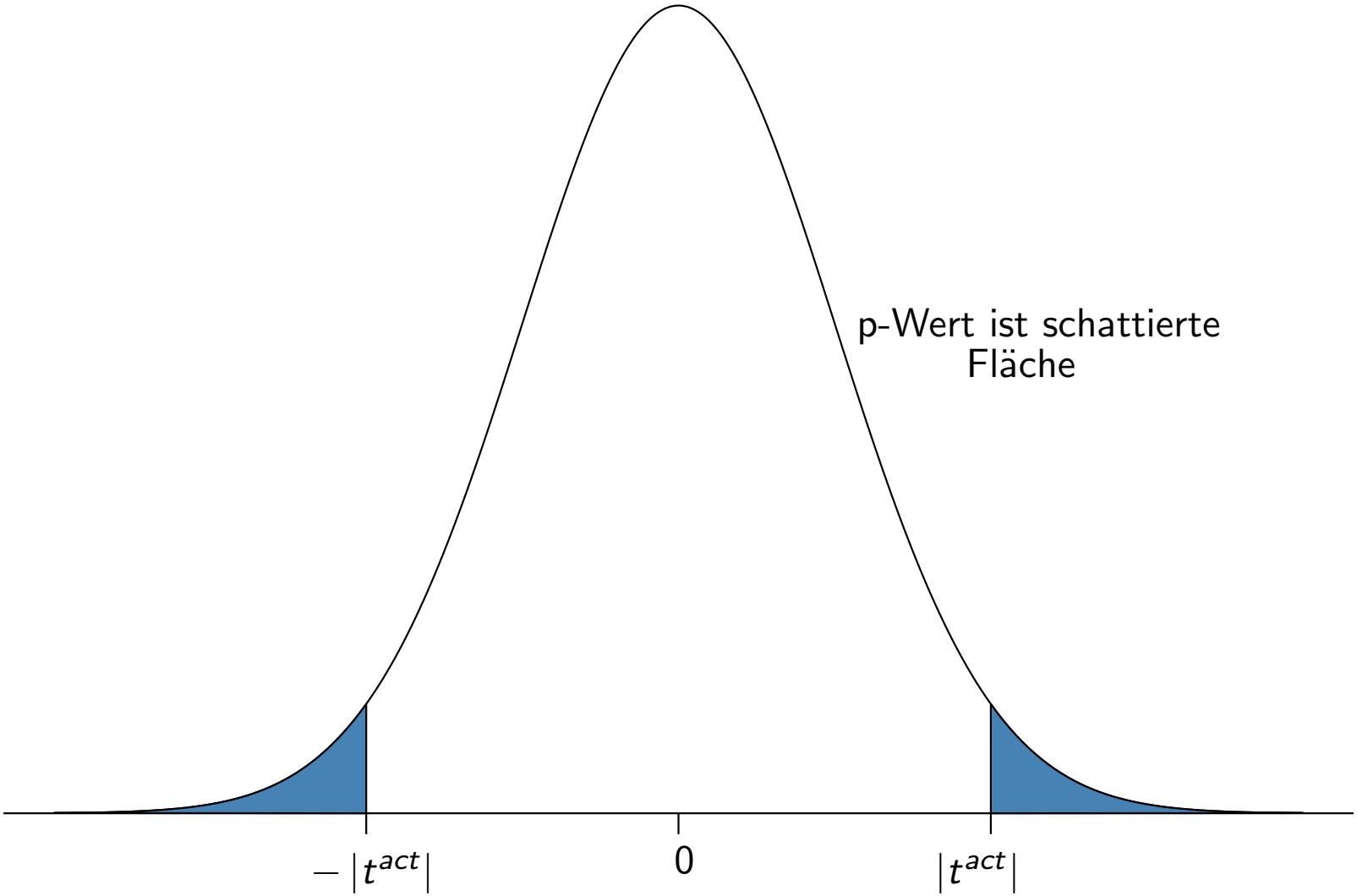
Berechnung des p-Wertes

- Dann:

$$
\begin{aligned}
p\text{-Wert} &= \mathrm{P}_{H_0}\left(|\overline{Y} - \mu_{Y,0}| > |\overline{Y}^{act} - \mu_{Y,0}|\right) \\
&= \mathrm{P}_{H_0}\left(\left|\sqrt{n}\frac{\overline{Y} - \mu_{Y,0}}{\sigma_Y}\right| > \left|\sqrt{n}\frac{\overline{Y}^{act} - \mu_{Y,0}}{\sigma_Y}\right|\right) \\
&= \mathrm{P}_{H_0}\left(|t| > \left|t^{act}\right|\right) \\
&\cong \text{Wahrscheinlichkeit, dass } N(0,1)\text{-ZV ausserhalb von } \left|t^{act}\right|.
\end{aligned}
$$

- Wir lehnen $H_0: \ \mathrm{E}(Y) = \mu_{Y,0}$ zugunsten von $H_1: \ \mathrm{E}(Y) \neq \mu_{Y,0}$ zum Niveau $\alpha = 5\%$ ab, wenn
  - p-Wert $< 0.05 = \alpha$, oder (äquivalent)
  - $|t| > \Phi^{-1}(1 - \alpha/2) = 1.96$.

Berechnung des p-Wertes



Konfidenzintervalle

- **Konfidenzintervalle (KIs)** sind informativer als *Punkt*schätzungen.
- Konfidenzintervalle sind *Intervall*schätzungen, die wir mit gewisser Konfidenz machen.
- Beispiel: Ich glaube, in 80% der Fälle liegt die Größe des Kindes zwischen Mutter und Vater.
  - [Größe der Mutter, Größe des Vaters] ist mein subjektives 80%-Konfidenzintervall für die Größe des Kindes als Erwachsener.
- In der Ökonometrie brauchen wir formalere Konfidenzintervalle, die auf Daten basieren.
- Für ein Konfidenzintervall für $\mathrm{E}(Y)$ (berechnet aus $Y_1, \ldots, Y_n$) nutzen wir, dass, wenn $\mathrm{E}(Y) = \mu_{Y,0}$,

$$t = \sqrt{n}\frac{\overline{Y} - \mu_{Y,0}}{s_Y} \approx N(0,1).$$

Konfidenzintervalle

- Ein $(1-\alpha)$-Konfidenzintervall $[K_{\text{min}}, K_{\text{max}}]$ beinhaltet den wahren Wert $\mathrm{E}(Y)$ mit Wahrscheinlichkeit $(1-\alpha)$, d.h.

$$\mathrm{P}(K_{\text{min}} \leq \mathrm{E}(Y) \leq K_{\text{max}}) = 1-\alpha.$$

- Was ist hier zufällig?
  - Die Grenzen $K_{\text{min}} = K_{\text{min}}(Y_1, \ldots, Y_n)$ und $K_{\text{max}} = K_{\text{max}}(Y_1, \ldots, Y_n)$.
- Was ist hier nicht zufällig?
  - Der Erwartungswert $\mathrm{E}(Y)$ ist nicht zufällig – wir kennen ihn nur nicht.
- Im Folgenden bekommen wir immer nur **asymptotische** $(1-\alpha)$-KIs:

$$\mathrm{P}(K_{\text{min}} \leq \mathrm{E}(Y) \leq K_{\text{max}}) \to 1-\alpha, \qquad n \to \infty.$$

Konfidenzintervalle

- Da $t \xrightarrow{d} N(0,1)$ für $n \to \infty$, gilt

$$
\begin{aligned}
1-\alpha &\leftarrow \mathrm{P}\big(-\Phi^{-1}(1-\alpha/2) \leq t \leq \Phi^{-1}(1-\alpha/2)\big) \\
&= \mathrm{P}\Big(-\Phi^{-1}(1-\alpha/2) \leq \sqrt{n}\frac{\overline{Y}-\mu_{Y,0}}{s_Y} \leq \Phi^{-1}(1-\alpha/2)\Big) \\
&= \mathrm{P}\Big(\overline{Y}-\Phi^{-1}(1-\alpha/2)\frac{s_Y}{\sqrt{n}} \leq \mu_{Y,0} \leq \overline{Y}+\Phi^{-1}(1-\alpha/2)\frac{s_Y}{\sqrt{n}}\Big).
\end{aligned}
$$

- Diese Rechnung zeigt auch, dass das $(1-\alpha)$-KI

$$
\Big[\overline{Y}-\Phi^{-1}(1-\alpha/2)\frac{s_Y}{\sqrt{n}},\ \overline{Y}+\Phi^{-1}(1-\alpha/2)\frac{s_Y}{\sqrt{n}}\Big]
$$

genau aus den Werten μ_Y besteht, für die die Hypothese $H_0: \mathrm{E}(Y) = \mu_Y$ z.N. α nicht abgelehnt werden kann.

Zusammenfassung

- Ausgehend von den beiden Annahmen

  1. $Y_1, \ldots, Y_n$ sind u.i.v.,
  2. $0 < \mathrm{E}(Y^4) < \infty$ haben wir hergeleitet:

  - Schätztheorie (Stichprobenverteilung von $\overline{Y}$)
  - Hypothesentests (große-n Verteilung der t-Statistik und Berechnung des p-Wertes)
  - Konfidenzintervalle (konstruiert durch das Invertieren der t-Statistik)

- Sind die Annahmen 1. & 2. in der Praxis plausibel? Ja.

Überblick

Lineare Regression mit einem Regressor

(SW Kapitel 4)

Lineare Regression mit einem Regressor

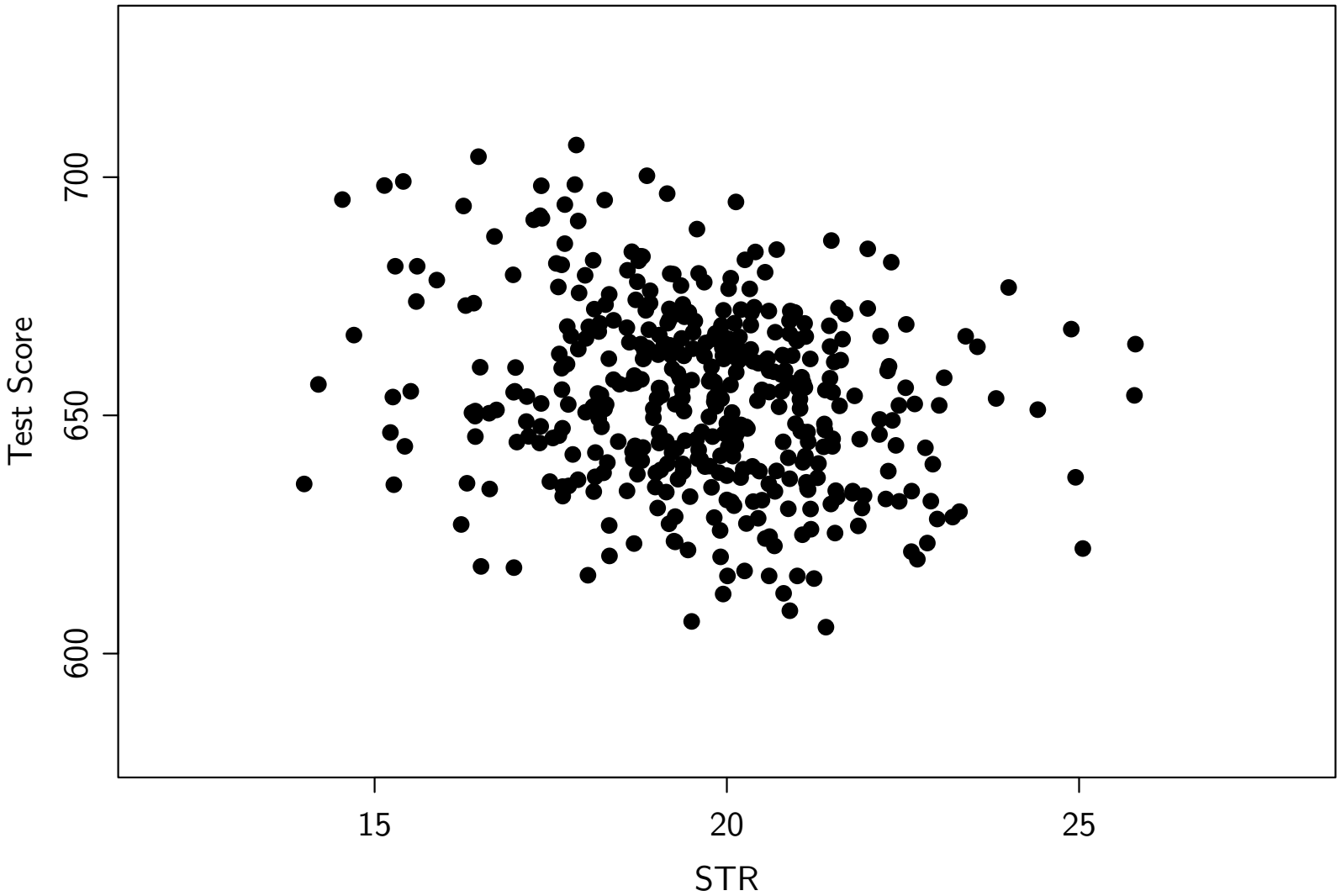


Abbildung 5: Punktediagramm von TestScore und STR.

Lineare Regression mit einem Regressor

- Das einfachste (ökonometrische) Modell ist
  - linear
  - mit zwei Variablen – und einer unverzichtbaren zufälligen Komponente:

$$Y = \beta_0 + \beta_1 X + u.$$

- **Lineare Regression** gibt uns die Möglichkeit Schätzungen und Folgerungen über den unbekannten "Einflussparameter" β_1 (mathematisch : Steigungsparameter) in einer **Grundgesamtheit** zu machen.
- Unser Ziel ist es, einen (hoffentlich kausalen) Einfluss auf Y durch eine Veränderung von X zu schätzen – d.h. β_1.
- Aber erstmal beschäftigen wir uns mit der Prognose von Y durch X.

Das (einfache) lineare Regressionsmodell

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \qquad i = 1, \ldots, n.$$

- X ist die **unabhängige Variable** (exogen, r.S.) oder **Regressor**.
- Y ist die **abhängige Variable** (endogen, l.S.) oder **Regressand**.
- $\mathrm{E}(Y \mid X) = \beta_0 + \beta_1 X$ ist die **Populationsregressionsgerade**.
- β_0 = **Achsenabschnitt**.
- β_1 = **Steigungskoeffizient** (Regressionskoeffizient).
- u_i = der Regressions**fehler** (Fehlerterm).
  - Der Fehlerterm enthält ausgelassene Faktoren oder mögliche Messfehler von Y.

Terminologie auf einen Blick

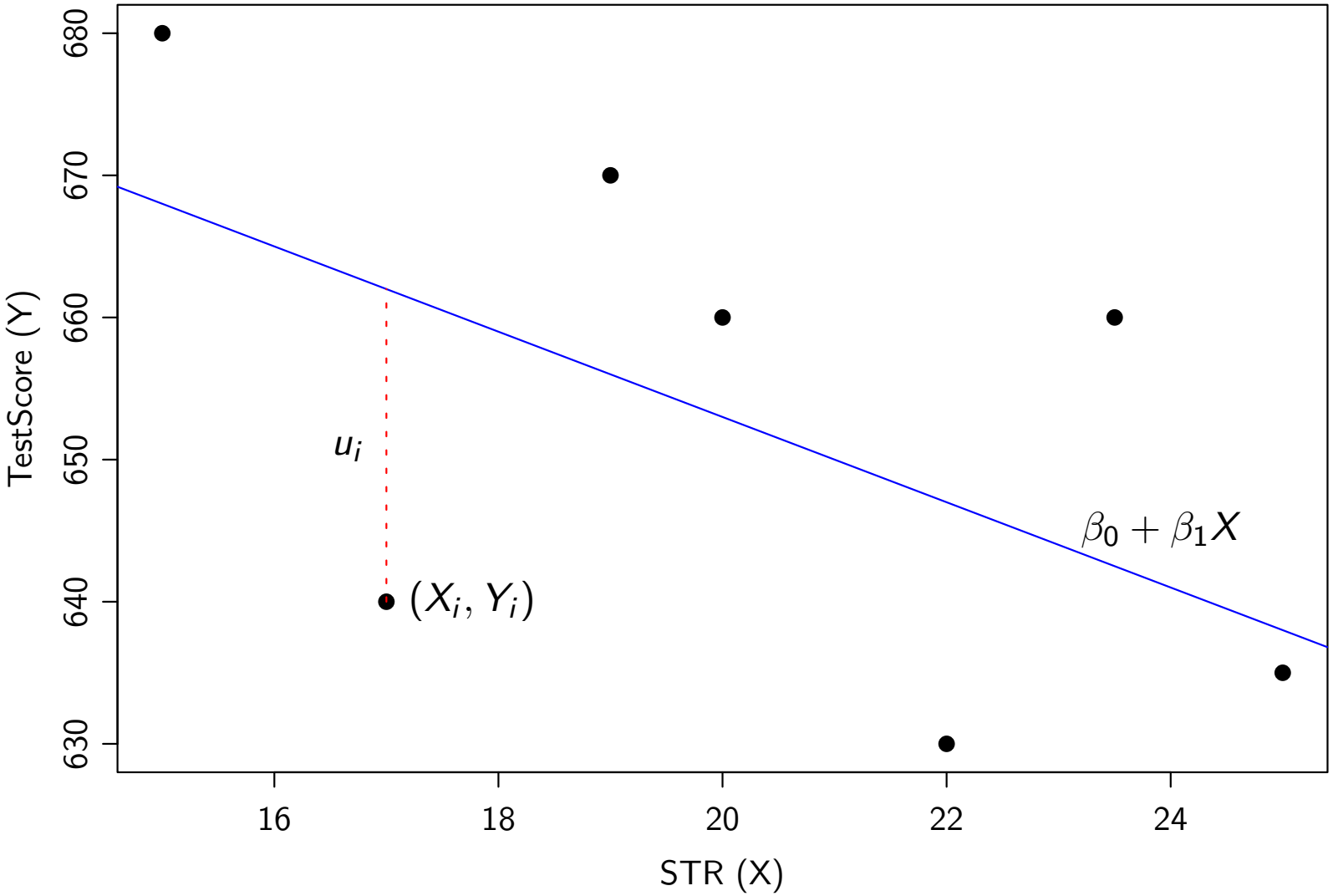


Abbildung 6: Punktediagramm von TestScore und STR.

Interpretationen des Regressionsmodells

$$Y = \beta_0 + \beta_1 X + u.$$

- Deskriptive Interpretation: "Leute mit **einer Einheit** höherem X, haben im Schnitt β_1 höheres Y."
- Kausale Interpretation: "Wenn wir – **alles andere gleichhaltend** – X um eine Einheit erhöhen, **verändert** sich Y um β_1."
- Prognoseinterpretation: "Wenn ich eine Person mit $X + 1$ sehe, **prognostiziere** ich, dass sich sein Y um β_1 von einer Person mit X unterscheidet."

Das Untersuchungsobjekt

- Die statistische Inferenz für β_0 und β_1 gleicht i.A. der des Schätzens des Erwartungswertes.
- Inferenz des Parameters umfasst:
- Schätzung (Kapitel 3):
  - Wie sollten wir eine Linie durch die Daten legen, um den Steigungskoeffizienten zu schätzen?
    (Antwort: **Kleinste Quadrate (KQ)**, oder synonym **ordinary least squares (OLS)**).
  - Was sind die Vor- und Nachteile von KQ?
- Hypothesentests und Konfidenzintervalle (Kapitel 4):
  - Wie kann man testen, ob der Steigungsparameter 0 ist?
  - Wie kann man ein Konfidenzintervall bilden?

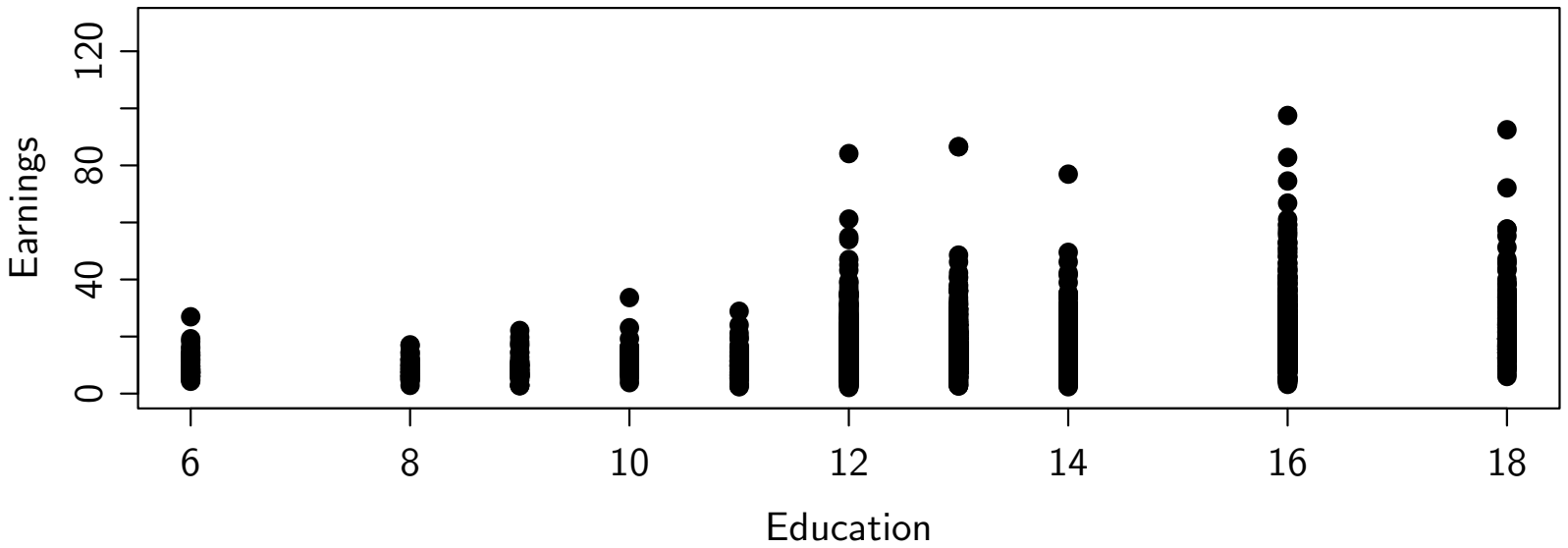
Überblick

- **Der KQ-Schätzer**
- Maße der Anpassungsgüte
- Die Annahmen der KQ-Schätzung
- Die Stichprobenverteilung des KQ-Schätzers

Schätzung
Prognose

- Wie bereits erwähnt, vergessen wir die kausale Analyse vorerst.
- Stellen Sie sich Folgendes vor:
  - Sie haben Daten $(X_i, Y_i)_{i=1,\ldots,n}$, z.B. $X = STR$, $Y = TestScore$.
  - Sie wissen X und wollen Y prognostizieren, also $\mathrm{E}(Y \mid X)$ berechnen.
  - Ihr einziges Ziel ist eine gute Prognose.
- Weitere Beispiele:
  - Inflationsrate dieses Jahr, Arbeitslose nächstes Jahr.
  - Aktienkurs heute, Aktienkurs morgen.

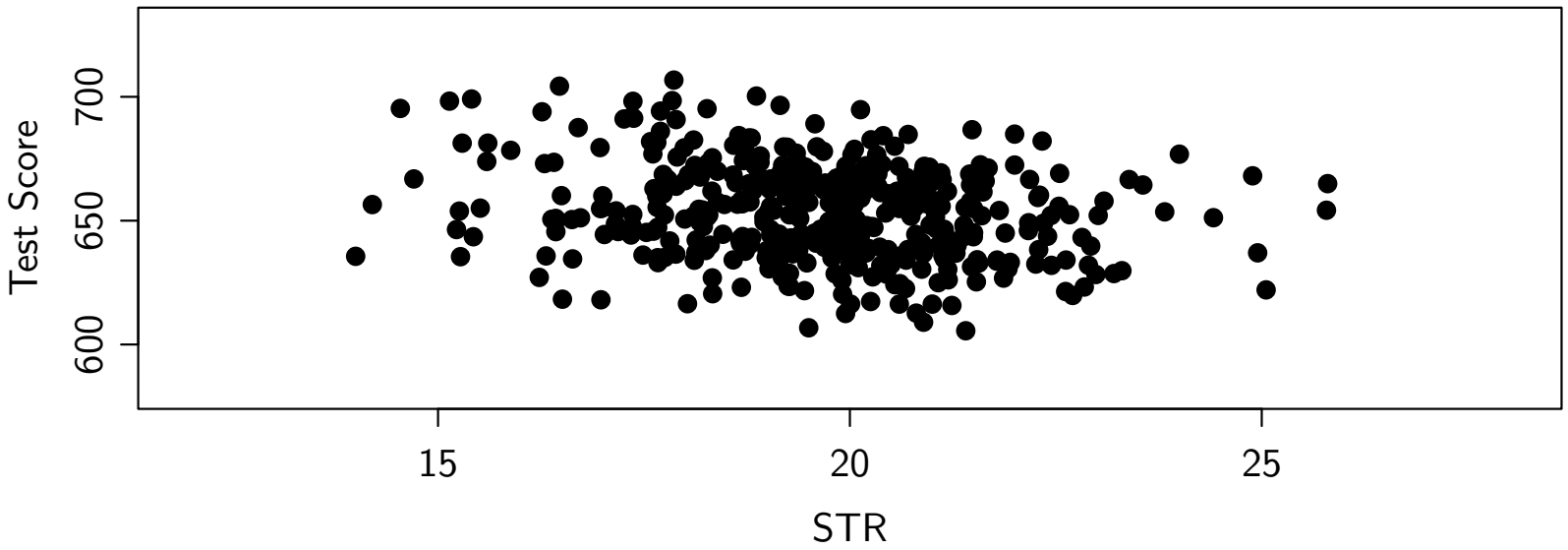
Schätzung
Prognose



- $\mathrm{E}(Y \mid X)$ zu schätzen ist einfach, wenn X diskret ist.
- Berechne einfach

$$\frac{1}{\#\text{ Menschen mit } X_i = x} \sum_{i:X_i=x} Y_i.$$

Schätzung
Prognose



- $\mathrm{E}(Y \mid X)$ zu schätzen ist schwer wenn X stetig ist.
- Deshalb brauchen wir ein Modell...

Schätzung
Prognose

- Für die Prognose brauchen wir ein Modell:

$$\mathrm{E}(Y \mid X) = \beta_0 + \beta_1 X.$$

- Nachdem wir $X = x^*$ beobachtet haben, wollen wir Y vorhersagen:

$$Y^* = \mathrm{E}(Y \mid X = x^*) = \beta_0 + \beta_1 x^*.$$

- Ambitionierter als die deskriptive Analyse:
  - Wollen die Zukunft vorhersagen, nicht nur die Vergangenheit beschreiben.
- Weniger ambitioniert als kausale Analyse:
  - X muss nicht notwendigerweise kausal für Y sein, um gute Vorhersage zu liefern.

Schätzung
Prognose

Beispiel 3.1: Kalifornische Schulbezirke.

- Denken Sie ans *TestScore*/*STR* Beispiel, um sich die unterschiedlichen Fragestellungen zu verdeutlichen:
  - Deskriptiv: Welchen TestScore haben Schüler aus einem Distrikt mit $STR = 23$ durchschnittlich?
  - Prognose: Gegeben ein Distrikt mit $STR = 23$, was ist meine beste Prognose für TestScore?
  - Kausal: Wenn ich – alles andere gleichhaltend – die *STR* auf 23 verändere, was passiert mit TestScore?
- Die Fragen werden der Reihe nach schwieriger!

Schätzung
Prognose

- Definiere den **geschätzten Wert** $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$.
- Eine gute Vorhersage minimiert den Abstand zwischen $\widehat{Y}_i$ und Y_i.
- Daher: Wähle $\widehat{\beta}_0$ und $\widehat{\beta}_1$ so, dass $(\widehat{Y}_i - Y_i)^2$ möglichst klein für alle Beobachtungen.
- **Kleinste Quadrate** ("ordinary least squares" oder "OLS") Schätzer der unbekannten Parameter β_0 und β_1 löst

$$\min_{b_0, b_1} \sum_{i=1}^{n} [Y_i - \widehat{Y}_i]^2 = \min_{b_0, b_1} \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2.$$

- Analogie: Der Kleinste Quadrate Schätzer $\overline{Y}$ von μ_Y löst

$$\min_{m} \sum_{i=1}^{n} (Y_i - m)^2.$$

Schätzung
Lösung

- Der KQ-Schätzer löst das Minimierungsproblem:

$$\min_{b_0, b_1} \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2.$$

- Dieses kann gelöst werden...
- Das Ergebnis ist der KQ-Schätzer von β_0 und β_1:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2},$$
$$\widehat{\beta}_0 = \overline{Y} - \overline{X}\widehat{\beta}_1.$$

Schätzung

R Beispiel

```
linear_model <- lm(score ~ STR, data = CASchools)
summary(linear_model)
```

```
##
## Call:
## lm(formula = score ~ STR, data = CASchools)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -47.73 -14.25   0.48  12.82  48.54
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   698.93       9.47   73.82  < 2e-16 ***
## STR            -2.28       0.48   -4.75  2.8e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.6 on 418 degrees of freedom
## Multiple R-squared:  0.0512, Adjusted R-squared:  0.049
## F-statistic: 22.6 on 1 and 418 DF,  p-value: 2.78e-06
```

Schätzung

Beispiel

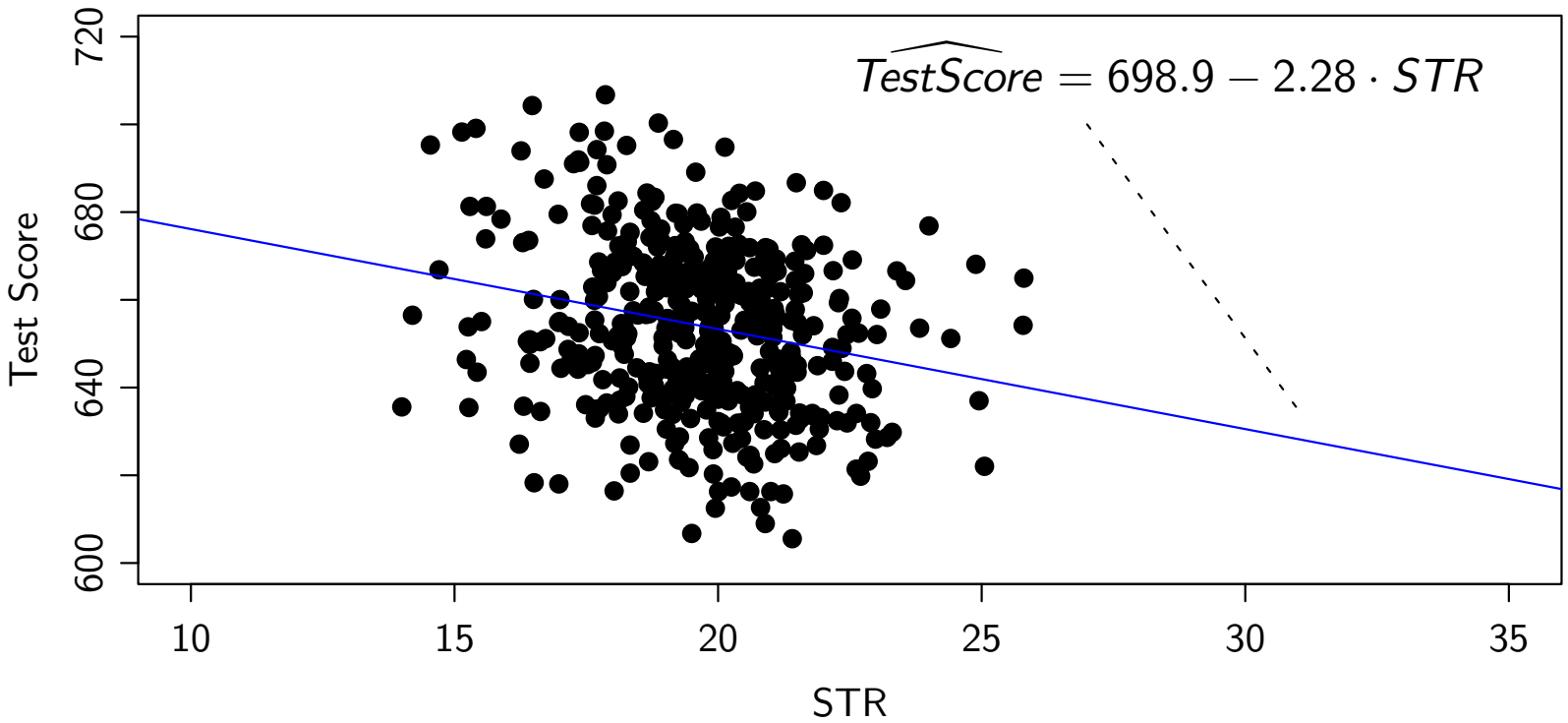


Abbildung 7: Streudiagramm von TestScore und STR

Schätzung

R Beispiel

Beispiel 3.1: Fortsetzung.

Interpretation der geschätzten Regressionsgeraden

$$\widehat{TestScore} = 698.9 - 2.28 \cdot STR.$$

Geschätzte Werte und Residuen:

- Ein Schuldistrikt in den Daten ist Antelope, CA, für den $STR = 19.33$ und $TestScore = 657.8$ gilt.
- Geschätzter Wert: $\widehat{Y}_{Antelope} = 698.9 - 2.28 \cdot 19.33 = 654.8$.
- Residuum: $\widehat{u}_{Antelope} = 657.8 - 654.8 = 3.0$.

Schätzung

R Beispiel

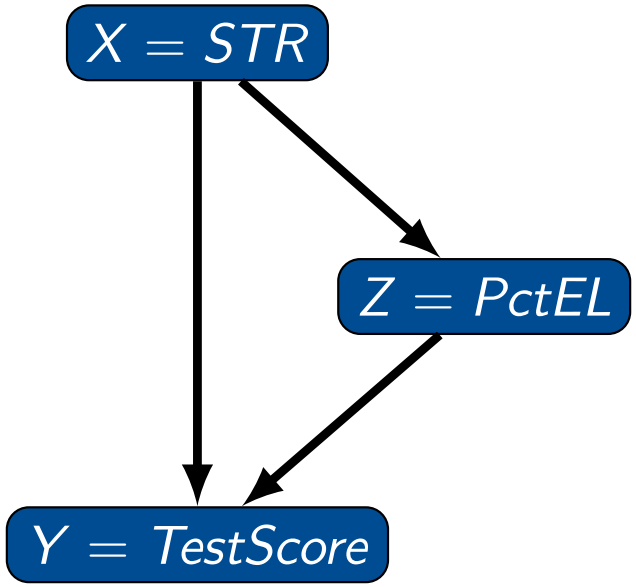
Beispiel 3.1: Fortsetzung.

Geschätzte Steigung von $\widehat{\beta}_1 = -2.28$:

- **Deskriptiv**: Distrikte mit einer durchschnittlich um einen Schüler größeren Klasse haben durchschnittlich einen um 2.28 Punkte niedrigeren TestScore.
- **Prognose**: Meine Prognose für den Unterschied in *TestScore* für zwei Distrikte mit einer um 1 verschiedenen *STR* ist 2.28.
- **Kausal**: Ist Verschlechterung von $\widehat{\frac{\Delta \text{Test Score}}{\Delta STR}} = -2.28$ kausal interpretierbar?
  - Nein: Durch die um 1 erhöhte STR wurde nicht alles andere gleich gehalten – Distrikte mit um 1 erhöhter *STR* haben mglw. weniger Muttersprachler (*PctEL*).
  - Es könnten also eigentlich Sprachschwierigkeiten der Grund für die im Schnitt schlechteren Testergebnisse sein.
  - D.h. unsere Schätzung $\widehat{\beta}_1 = -2.28$ fängt auch den negativen Effekt von Sprachschwierigkeiten mit ein – nicht nur den von größeren Klassen!

Schätzung

Beispiel



Schätzung

R Beispiel

Beispiel 3.1: Fortsetzung.

Geschätzter Achsenabschnitt von $\widehat{\beta}_0 = 698.9$:

- **Deskriptiv**: Der Achsenabschnitt bedeutet (wörtlich), dass *entsprechend der Regressionsgeraden*, Distrikte mit null Schülern pro Lehrer einen geschätzten Test-Score von 698.9 haben.
- **Prognose**: Gegeben ein Distrikt mit $STR = 0$ ist meine beste Prognose $TestScore = 698.9$.
- Beide Interpretationen machen keinen Sinn – die Regressionsgerade extrapoliert außerhalb der Daten, in diesem Fall hat der Achsenabschnitt keine ökonomische Bedeutung.

Schätzung
Eigenschaften der Residuen

- Aus der Definition von $\widehat{\beta}_0$ und $\widehat{\beta}_1$ folgen Eigenschaften für die Residuen.
- Das Residuum $\widehat{u}_i = Y_i - \widehat{Y}_i$ misst die Güte der Vorhersage.
- Zwar ist das Residuum nicht immer null, aber die $\widehat{u}_i$ sind im Schnitt null:

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i = 0.$$

- Die Stichprobenkovarianz der X_i und $\widehat{u}_i$ ist auch null:

$$\frac{1}{n}\sum_{i=1}^{n}X_i\widehat{u}_i = 0.$$

- Beweisen Sie diese beiden algebraischen Eigenschaften, die ohne (!) weitere Modellannahmen gelten.

Warum „Regression“?

Beispiel 3.2: Galton (1886).

- Betrachten Sie die Regression von Galton (1886):

$$midchild = \beta_0 + \beta_1 \cdot midparent + u.$$

- *midparent* (*midchild*) ist die mittlere Körpergröße der Eltern (Kinder).
- Welche Werte erwarten Sie für die Parameter β_0 und β_1?
- Siehe auch `bundesliga.R`.

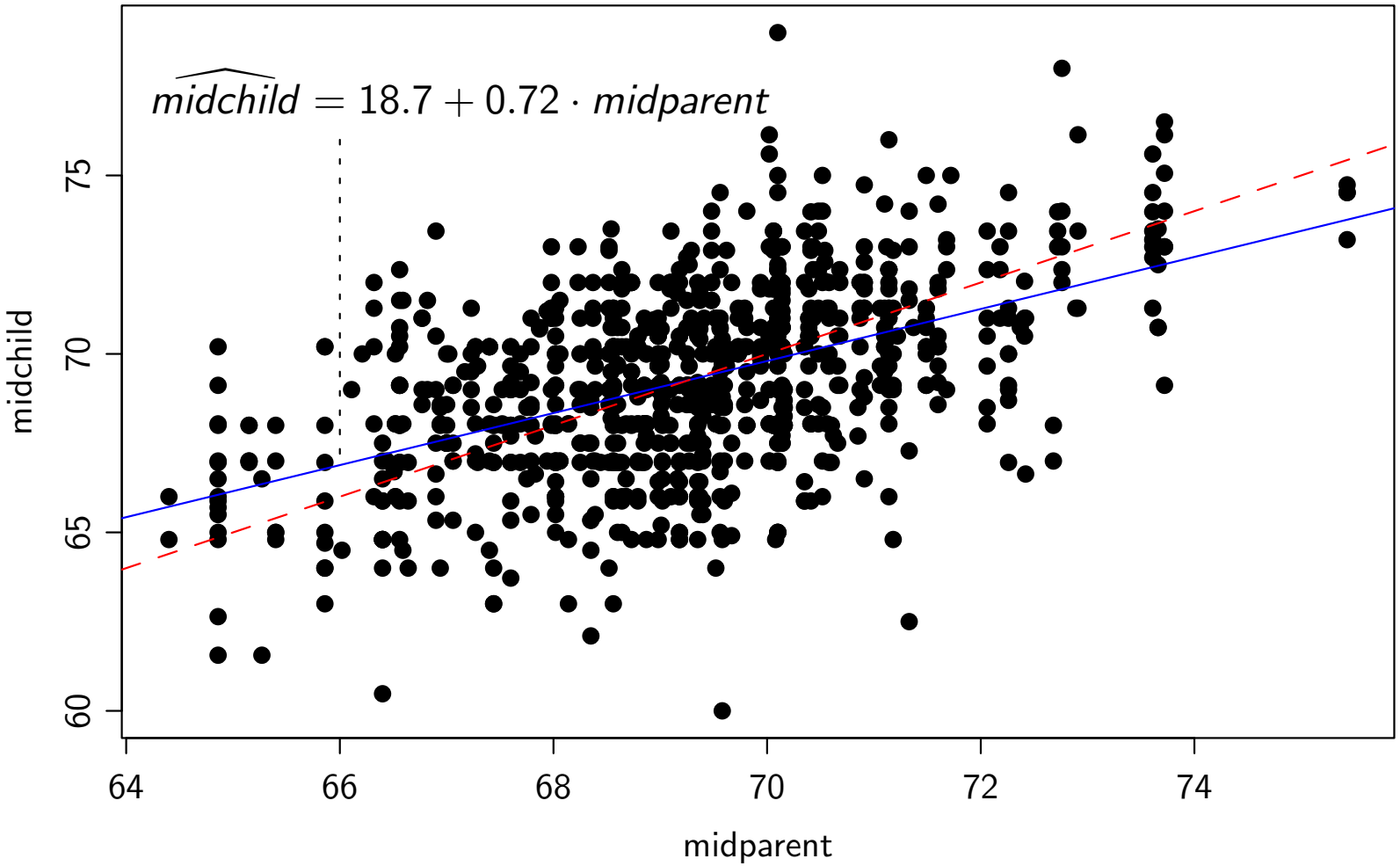
Beispiel I

Regression towards mediocrity

```r
Galton <- read.csv("Daten/Galton.csv")
# see https://github.com/vincentarelbundock/Rdatasets/raw/master/csv/mosaicData/Galt

# Create variables
Galton$midparent <- (Galton$father + Galton$mother * 1.08) / 2
Galton$midchild  <- ifelse(Galton$sex=='M', Galton$height, Galton$height*1.08)

# Run regression and
linear_model     <- lm(midchild ~ midparent, data = Galton)

# Plot results
plot(midchild ~ midparent, data = Galton,
     xlab = "midparent", ylab = "midchild")
abline(linear_model, col="blue")
text(64, 77, labels = "$\\widehat{midchild}=18.7+0.72\\cdot midparent$", pos=4)
segments(66, 76, 66, 67, lty="dotted")
abline(coef = c(0,1), col="red", lty="dashed")
```

Beispiel II

Regression towards mediocrity



Überblick

- Der KQ-Schätzer
- **Maße der Anpassungsgüte**
- Die Annahmen der KQ-Schätzung
- Die Stichprobenverteilung des KQ-Schätzers

Parameterschätzung ist erst der Anfang

- Wir würden gerne wissen, wie gut unser Model die Daten erklärt.
- Diese Frage ist relevant für deskriptive Analyse und Prognosen, ...
- ... aber nicht so sehr für kausale Analyse.
- Trotzdem wollen wir einige Maße für die **Anpassungsgüte** definieren.
  - R^2, Standardfehler der Regression, RMSE

Maße der Anpassungsgüte

Bausteine

- **Total Sum of Squares (TSS)** misst die Gesamtvariation in Y:

$$TSS = \sum_{i=1}^{n} (Y_i - \overline{Y}_n)^2.$$

- **Explained Sum of Squares (ESS)** misst den Teil der Variation in Y, der durch den Regressor erklärt wird:

$$ESS = \sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y}_n)^2.$$

- **Residual Sum of Squares (RSS)** misst den Teil der Variation in Y der nicht durch den Regressor erklärt wird:

$$RSS = \sum_{i=1}^{n} \widehat{u}_i^2.$$

Maße der Anpassungsgüte
Bausteine

- Intuitiv sollte die Gesamtvariation gleich der erklärten plus der unerklärten Variation sein, d.h. $TSS = ESS + RSS$.
- Dies lässt sich leicht nachprüfen:

$$\begin{aligned} TSS &= \sum_{i=1}^{n} (Y_i - \overline{Y}_n)^2 \\ &= \sum_{i=1}^{n} (Y_i - \widehat{Y}_i + \widehat{Y}_i - \overline{Y}_n)^2 \\ &= \dots \\ &= RSS + ESS. \end{aligned}$$

- Füllen Sie die Leerstellen aus!

Maße der Anpassungsgüte
R^2

- Das R^2 ist der Anteil erklärter Variation von Y_i durch die Regression:

$$R^2 = \frac{ESS}{TSS} = \frac{ESS}{ESS + RSS}.$$

- $0 \leq R^2 \leq 1$
  - $R^2 = 0$ bedeutet $ESS = 0$.
  - $R^2 = 1$ bedeutet $ESS = TSS$.
- Für Regressionen mir nur einem X: $R^2 = \widehat{\text{Corr}}(X, Y)^2$.
- Nur nützlich, wenn Konstante in Regression!

Maße der Anpassungsgüte
R^2

- Kleines R^2 der Regression $Y = \beta_0 + \beta_1 X + u$ deutet auf schlechte Anpassung hin.
- Mögliche Gründe:

1. Zu viel Variablen außer X, die auch Y beeinflussen.
   - Kleines R^2 nicht notwendigerweise schlimm, solange es eine plausible Theorie dafür gibt, dass die anderen Faktoren nicht mit X korrelieren.
2. Nichtlineare Beziehung zwischen X und Y (z.B. $Y = \beta_0 + \beta_1 \log(X) + u$).
   - Kleines R^2 schlimm; ziehe Variablentransformation in Betracht.

Maße der Anpassungsgüte
Standardfehler der Regression (SER)

- Der **SER** misst die Spannweite der Verteilung von u.
- Der SER ist (fast) die Stichproben-Standardabweichung der KQ-Residuen:

$$SER = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(\widehat{u}_i - \overline{\widehat{u}})^2} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\widehat{u}_i^2}.$$

- (Erinnern Sie sich, dass $\overline{\widehat{u}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i = 0$?)
- Der SER
  - besitzt die Einheit von Y,
  - misst die durchschnittliche "Größe" der KQ-Residuen (der durchschnittliche "Fehler", der durch die Regression gemacht wurde).

Maße der Anpassungsgüte
Root Mean Squared Error (RMSE)

- Der **RMSE** ist dem SER sehr ähnlich:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i^2}.$$

- Der einzige Unterscheid zum SER ist die Division durch $1/n$ anstatt $1/(n-2)$.
- F: Warum Division durch $(n-2)$ im SER?
- A: Korrektur um die "Freiheitsgrade", da für den SER zwei Parameter (β_0 und β_1) geschätzt wurden.

Maße der Anpassungsgüte

R Beispiel

Beispiel 3.2: Fortsetzung.

Die Regression von TestScore auf STR liefert:

$$\widehat{TestScore} = 698.9 - 2.28 \cdot STR, \quad R^2 = 0.05, \quad SER = 18.6.$$

- STR erklärt nur einen kleinen Anteil der Variation in TestScore.
- Bedeutet dies, dass die STR "unwichtig" ist (z.B. in einem politischen Sinne)?

Überblick

- Der KQ-Schätzer
- Maße der Anpassungsgüte
- **Die Annahmen der KQ-Schätzung**
- Die Stichprobenverteilung des KQ-Schätzers

Keine Eigenschaften ohne Annahmen

- Was sind genau die Eigenschaften des KQ-Schätzers?
- Es gibt eine Reihe von allgemeinen Gütekriterien für Schätzer:
  - **Unverzerrtheit**: $\mathrm{E}[\widehat{\beta}_1] = \beta_1$ – im Mittel schätzen wir richtig!
  - **Niedrige Varianz**: $\mathrm{Var}(\widehat{\beta}_1 - \beta_1)$ – die Streuung um den wahren Wert ist gering
  - **Konsistenz**: $\widehat{\beta}_1 \xrightarrow{p} \beta_1$ – in großen Stichproben nähern wir uns dem wahren Wert an!
- Um solche Eigenschaften zu zeigen, müssen wir einige Annahmen über die Beziehung von Y und X machen, und wie diese erhoben wurden.
- Diese drei Annahmen sind die **(klassischen) KQ-Annahmen**.

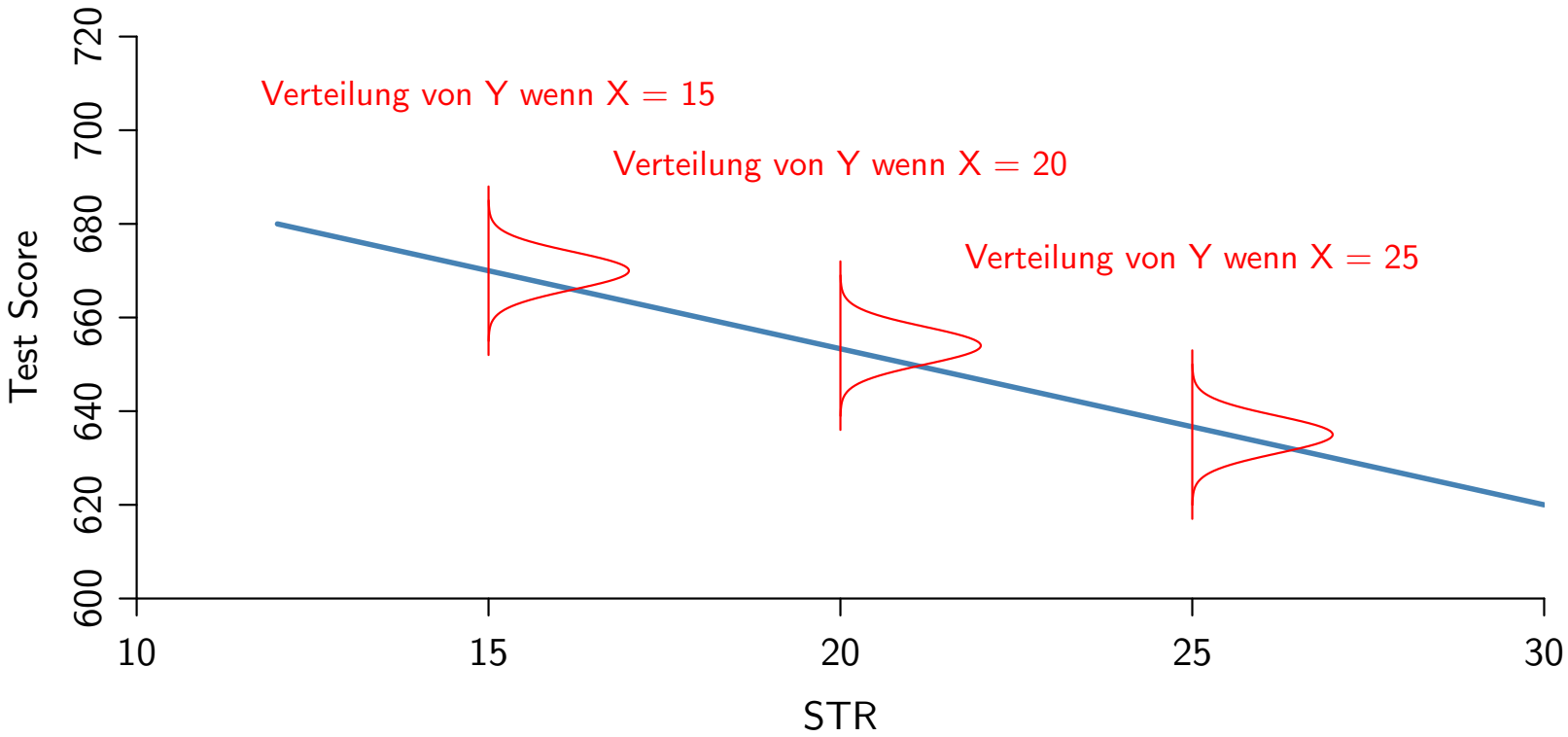
Die KQ-Annahmen

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \quad i = 1, \ldots, n. \tag{3.1}$$

1. Die bedingte Verteilung von u gegeben X hat einen Mittelwert von null, das heißt, $\mathrm{E}(u|X = x) = 0$.
   - Dies impliziert, dass $\widehat{\beta}_1$ unverzerrt ist.
   - Diese Annahme kombiniert mit (3.1) ist entscheidend!
2. (X_i, Y_i), $i = 1, \ldots, n$, sind *gemeinsam* u.i.v.
   - Dies gilt sofern X, Y zufällig gezogen wurden.
   - Dies liefert die Stichprobenverteilung von $\widehat{\beta}_0$ und $\widehat{\beta}_1$.
3. Größere Ausreißer von X oder Y sind selten.
   - Technisch, X und Y haben endliche vierte Momente.
   - Ausreißer können zu bedeutungslosen Ergebnissen von $\widehat{\beta}_1$ führen.

Die KQ-Annahme Nr. 1: Exogenität

- Für jeden Wert von X ist der Erwartungswert von u null:



Die KQ-Annahme Nr. 1: Exogenität

Die Annahme $\mathrm{E}(u \mid X) = 0$ impliziert drei wichtige Dinge:

1. $\mathrm{E}(u \mid X)$ ändert sich nicht, wenn sich X ändert.
2. $\mathrm{E}(u) = 0$, da
$$\mathrm{E}(u) = \mathrm{E}(\mathrm{E}(u \mid X)) = \mathrm{E}(0) = 0.$$
3. Die nützlichste Implikation ist, dass
$$\mathrm{E}(uX) = \mathrm{E}(\mathrm{E}(uX \mid X)) = \mathrm{E}(X\,\mathrm{E}(u \mid X)) = \mathrm{E}(X \cdot 0) = 0,$$
und somit
$$\mathrm{Cov}(u, X) = \mathrm{E}(uX) - \mathrm{E}(u)\,\mathrm{E}(X) = 0.$$

Die KQ-Annahme Nr. 1: Exogenität

Eine Möglichkeit, sich diese Annahme plausibel zu machen, ist ein **(ideales, zufälliges) kontrolliertes Experiment**:

- X ist zufällig Leuten zugeteilt (Schüler sind zufällig verschiedenen Klassengrößen zugeteilt; Patienten zufällig zu medizinischen Behandlungen).
- Da X zufällig zugeordnet wurde, sind alle anderen individuellen Charakteristika – die Dinge die u beeinflussen – unabhängig von X verteilt.
- Also gilt in einem ideal zufälligen Experiment KQA Nr. 1, $\mathrm{E}(u|X = x) = 0$.
- In echten Experimenten, oder bei beobachteten Daten, sollte genau darüber nachgedacht werden, ob $\mathrm{E}(u|X = x) = 0$ plausibel ist.
  - Was wäre, wenn Eltern leistungsfähigerer Schüler, ihre Kinder eher an Schulen in Distrikten mit niedriger STR (z.B. $STR < 20$) anmelden? Dann ist $\mathrm{E}(u \mid STR < 20) > 0$ plausibel!

Die KQ-Annahme Nr. 1: Exogenität

Drei Interpretationen von $\mathrm{E}(u \mid X = x) = 0$:

1. Deskriptive Analyse: u hat keine Bedeutung
   - $\mathrm{E}(u \mid X) = 0$ gilt automatisch, wenn die bedingte Erwartung linear ist, weil $u = Y - \mathrm{E}(Y \mid X)$.
   - Überprüfung: Punktediagramm von Y und X sollte "linear" aussehen.
2. Kausale Analyse: u hat Bedeutung als "unbeobachtete" Faktoren
   - $\mathrm{E}(u \mid X) = 0$ gilt nicht automatisch.
   - Überprüfung: Muss gerechtfertigt werden durch ökonomische Theorie oder gesunden Menschenverstand.
3. Prognose: Wie in deskriptiver Analyse
   - Aber muss auch für x^* gerechtfertigt sein!

KQ-Annahme Nr. 1: Exogenität

- Das Modell $Y = \beta_0 + \beta_1 X + u$ zusammen mit Annahme, dass X zufällig zugeteilt wurde, $\mathrm{E}(u \mid X) = 0$, implizieren:
  1. Der Zusammenhang zwischen Y und X ist linear.
  2. Der Koeffizient ist derselbe für alle.
  3. u hat keine Beziehung zu X.

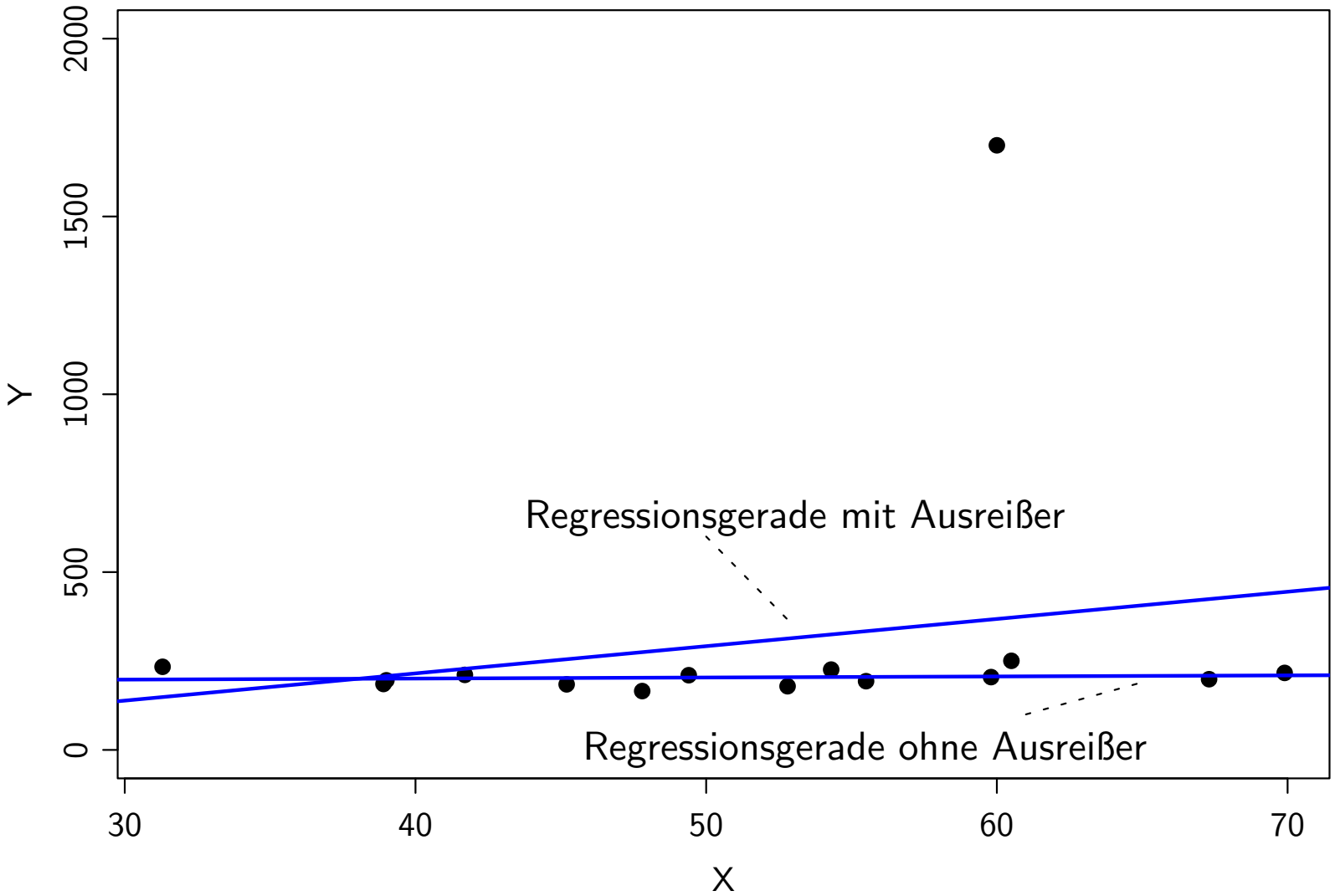
KQ-Annahme Nr. 2: u.i.v. Stichproben

- Ist automatisch erfüllt, sofern das Objekt (Individuum, Distrikt) rein zufällig gezogen wurde.
- Dann wird für jedes Objekt in der Stichprobe X und Y aufgezeichnet.
- Hauptsächlich wird man im Bereich der Zeitreihen auf nicht-u.i.v. Stichproben treffen – dies wird neue Schwierigkeiten machen.

KQ-Annahme Nr. 3: Große Ausreißer sind selten

- Technisch: $\mathrm{E}(X^4) < \infty$ und $\mathrm{E}(Y^4) < \infty$
- Ein großer Ausreißer ist ein Extremwert von X oder Y.
- Der Grundgedanke dieser Annahme ist, dass große Ausreißer das Resultat stark beeinflussen können.
- Beispiel: Wenn X und Y beschränkt sind, dann besitzen diese endliche vierte Momente. (Standardisierte Test Scores erfüllen dies automatisch; STR etc. erfüllen dies auch).

KQ-Annahme Nr. 3: Große Ausreißer sind selten



KQ-Annahme Nr. 3: Große Ausreißer sind selten

- Wie man Ausreißer identifiziert und behandelt, ist ein altehrwürdiges Problem in der Datenanalyse, und es gibt keine einfachen Antworten.
  - Die einfachste Methode ist ein Punktediagramm; `Anscombe.R`, `DataSaurus.R`.
- Wenn der Ausreißer ein Datenfehler (Kodierungs- bzw. Rekodierungsfehler) ist:
  - Entfernen des Datenpunktes, wenn der Datenpunkt nicht rekonstruiert werden kann.
- Aber Achtung: Selbst ein "sauberer" Datensatz hat eine "extremste" Beobachtung, die mitnichten entfernt werden sollte.
  - Im Gegenteil: Die extremsten Beobachtungen sind häufig die informativsten – präzise Schätzung braucht Variation in den Daten!

Überblick

- Der KQ-Schätzer
- Maße der Anpassungsgüte
- Die Annahmen der KQ-Schätzung
- **Die Stichprobenverteilung des KQ-Schätzers**

Zurück zur Statistik

Erinnerung:

- Der KQ-Schätzer wird basierend auf einer Stichprobe aus der Grundgesamtheit berechnet; und
- eine andere Stichprobe liefert einen anderen Wert von $\widehat{\beta}_1$.
- Dies ist die Quelle der "Stichprobenunsicherheit" von $\widehat{\beta}_1$.

Wir möchten

- die Stichprobenunsicherheit von $\widehat{\beta}_1$ quantifizieren,
- $\widehat{\beta}_1$ für Hypothesentests verwenden, z.B. $H_0 : \beta_1 = 0$,
- ein Konfidenzintervall für β_1 ausrechnen.
- Daher brauchen wir die Stichprobenverteilung des KQ-Schätzers.
- Diese erhalten wir aus den drei KQ-Annahmen...

Wahrscheinlichkeitstheoretischer Rahmen

- Der wahrscheinlichkeitstheoretische Rahmen für lineare Regression wird von den drei KQ-Annahmen zusammengefasst.
- **Population**: Die Menge der interessierenden Objekte (z.B. alle möglichen Schuldestrikte)
- **Zufallsvariablen**: (Y, X) z.B. (Test Score, STR)
- **Gemeinsame Verteilung von** (Y, X)

  1. Die Regressionsgerade ist linear
  2. $\mathrm{E}(u|X) = 0$ (1. KQ-Annahme)
  3. X, Y haben endliche vierte Momente (3. KQ-Annahme)

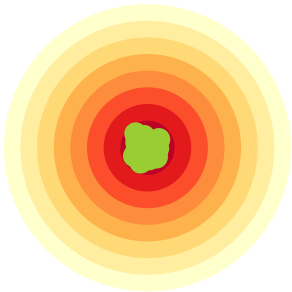
- **Datenauswahl durch reine Zufallsstichprobe:** $\{(X_i, Y_i)\}_{i=1,\ldots,n}$ sind u.i.v. (2. KQ-Annahme)

Die Stichprobenverteilung von $\widehat{\beta}_1$

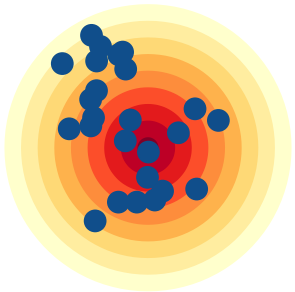
- Genau wie $\overline{Y}$ hat $\widehat{\beta}_1$ eine Stichprobenverteilung.
- Was ist $\mathrm{E}(\widehat{\beta}_1)$?
  - Falls $\mathrm{E}(\widehat{\beta}_1) = \beta_1$, so ist der KQ-Schätzer unverzerrt – eine gute Eigenschaft!
  - ... aber nicht notwendigerweise, wie die nächste Folie zeigt.
- Was ist $\mathrm{Var}(\widehat{\beta}_1)$? (Maß der Stichprobenunsicherheit)
- Wie ist die Verteilung von $\widehat{\beta}_1$ in kleinen Stichproben?
  - Im Allgemeinen kann diese sehr kompliziert sein, insbesondere nicht normal.
- Wie ist die Verteilung von $\widehat{\beta}_1$ in großen Stichproben?
  - Diese ist relativ einfach – in großen Stichproben ist $\widehat{\beta}_1$ (annähernd) normalverteilt. (So wie $\overline{Y}$.)

Unverzerrtheit

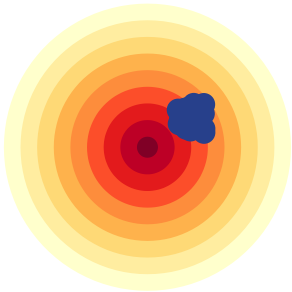
unverzerrt, niedrige Varianz



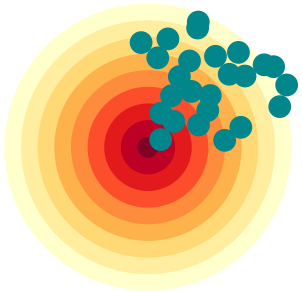
unverzerrt, hohe Varianz



verzerrt, niedrige Varianz



verzerrt, hohe Varianz



Unverzerrtheit

- In praxi
  1. sehen wir die Dartscheibe nie und ...
  2. ... dürfen auch nur einmal werfen.
- Also machen wir Annahmen darüber, wie wir werfen.
- Wenn diese erfüllt sind, haben wir Vertrauen, dass unser Wurf gut war.
- ... also ist Statistik, wie einen einzelnen Pfeil zu werfen und zu behaupten der Punkt, den wir getroffen haben, ist am ehesten die Mitte.

Erwartungswert und Varianz

Erwartungswert

- Wir können zeigen, dass

$$\widehat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^n (X_i - \overline{X}) u_i}{\sum_{i=1}^n (X_i - \overline{X})^2}.$$

- Dann folgt

$$\begin{aligned}
\mathrm{E}(\widehat{\beta}_1) - \beta_1 &= E\left[\frac{\sum_{i=1}^n (X_i - \overline{X}) u_i}{\sum_{i=1}^n (X_i - \overline{X})^2}\right] \\
&= E\left\{E\left[\frac{\sum_{i=1}^n (X_i - \overline{X}) u_i}{\sum_{i=1}^n (X_i - \overline{X})^2} | X_1, \ldots, X_n\right]\right\} \\
&= 0 \quad \text{da } \mathrm{E}(u_i | X_i = x) = 0 \text{ wegen KQA Nr. 1.}
\end{aligned}$$

- Also impliziert die KQ-Annahme Nr. 1, dass $\mathrm{E}(\widehat{\beta}_1) = \beta_1$.
- Das heißt, $\widehat{\beta}_1$ ist ein unverzerrter Schätzer für β_1; `unbiasedness.R`.

Erwartungswert und Varianz

Varianz

- Als nächstes berechnen wir $\mathrm{Var}(\widehat{\beta}_1)$:

$$\widehat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^n (X_i - \overline{X})u_i}{\sum_{i=1}^n (X_i - \overline{X})^2} = \frac{\frac{1}{n}\sum_{i=1}^n v_i}{\left(\frac{n-1}{n}\right) s_X^2},$$

  wobei

$$v_i = (X_i - \overline{X})u_i \approx (X_i - \mu_X)u_i.$$

- Wenn n groß ist, $s_X^2 \approx \sigma_X^2$ und $\frac{n-1}{n} \approx 1$.
- Also

$$\widehat{\beta}_1 - \beta_1 \approx \frac{\frac{1}{n}\sum_{i=1}^n v_i}{\sigma_X^2}.$$

Erwartungswert und Varianz

Varianz

- Daher, und weil $v_i = (X_i - \overline{X})u_i \approx (X_i - \mu_X)u_i$ fast u.i.v. (wegen KQ-Annahme Nr. 2),

$$\begin{aligned} \text{Var}(\widehat{\beta}_1) &= \text{Var}(\widehat{\beta}_1 - \beta_1) \\ &\approx \frac{\text{Var}(v_i)/n}{(\sigma_X^2)^2}, \end{aligned}$$

- Also,

$$\boxed{\text{Var}(\widehat{\beta}_1 - \beta_1) \approx \frac{1}{n} \cdot \frac{Var[(X_i - \mu_X)u_i]}{\sigma_X^4}}$$

Zusammenfassung bis jetzt

- $\widehat{\beta}_1$ ist unverzerrt: $\text{E}(\widehat{\beta}_1) = \beta_1$ – so wie $\overline{Y}$!
- $\text{Var}(\widehat{\beta}_1)$ ist invers proportional zu n – so wie $\overline{Y}$!

Erwartungswert und Varianz

Varianz

Die Mathematik

- Mit $\sigma_X^4 = [\text{Var}(X_i)]^2$ gilt

$$\text{Var}(\widehat{\beta}_1 - \beta_1) = \frac{1}{n} \cdot \frac{Var[(X_i - \mu_X)u_i]}{\sigma_X^4}.$$

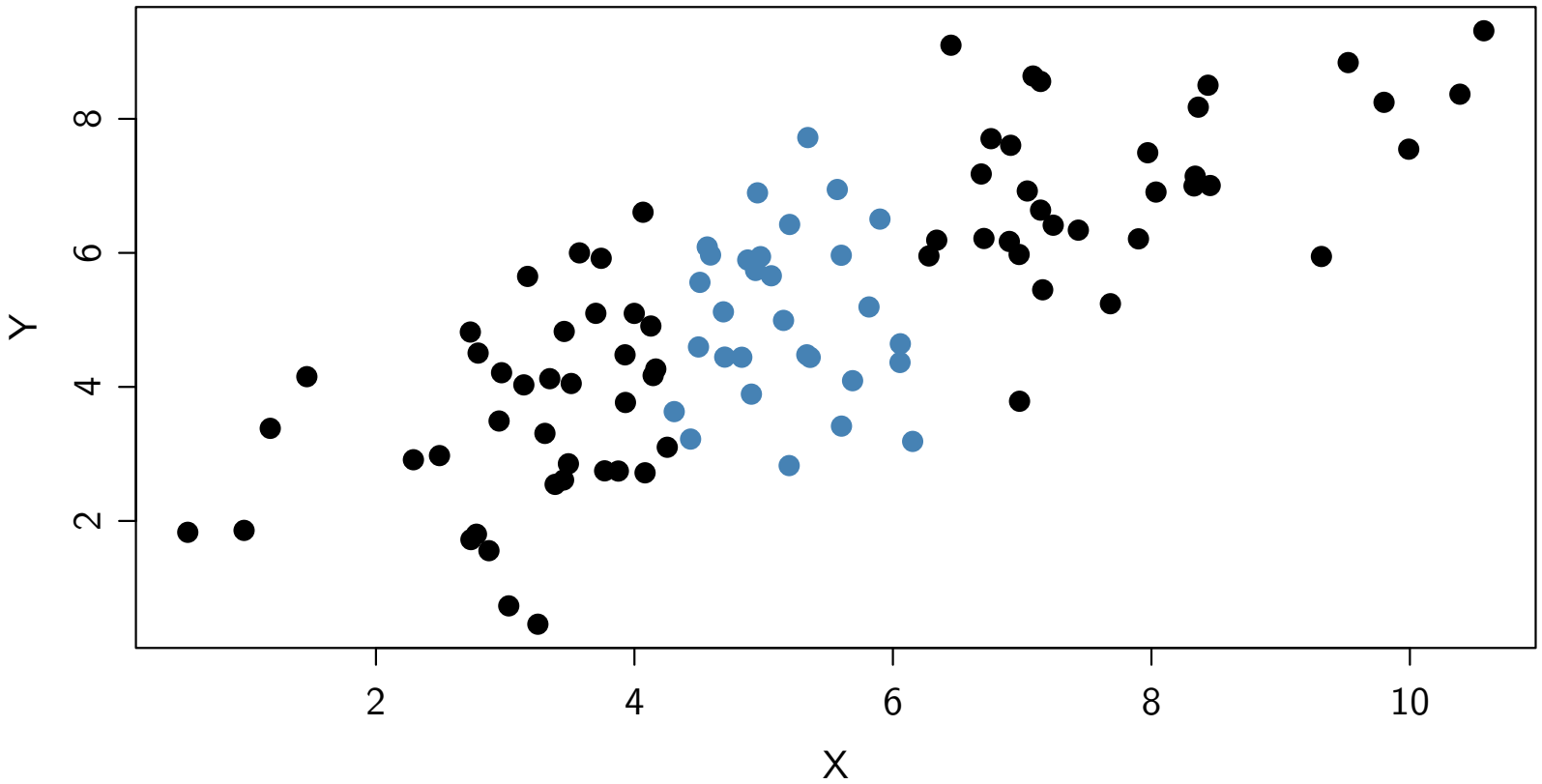
- Die Varianz von X tritt quadratisch im Nenner auf – also verursacht eine *breitere* Streuung von X eine *geringere* Varianz von β_1.

Die Intuition

- Sofern mehr Variation in X vorliegt, gibt es mehr Informationen in den Daten, die man verwenden kann, um die Regressionsgerade anzupassen.
  - Dies ist einfach zu sehen in einer Abbildung ...

Erwartungswert und Varianz

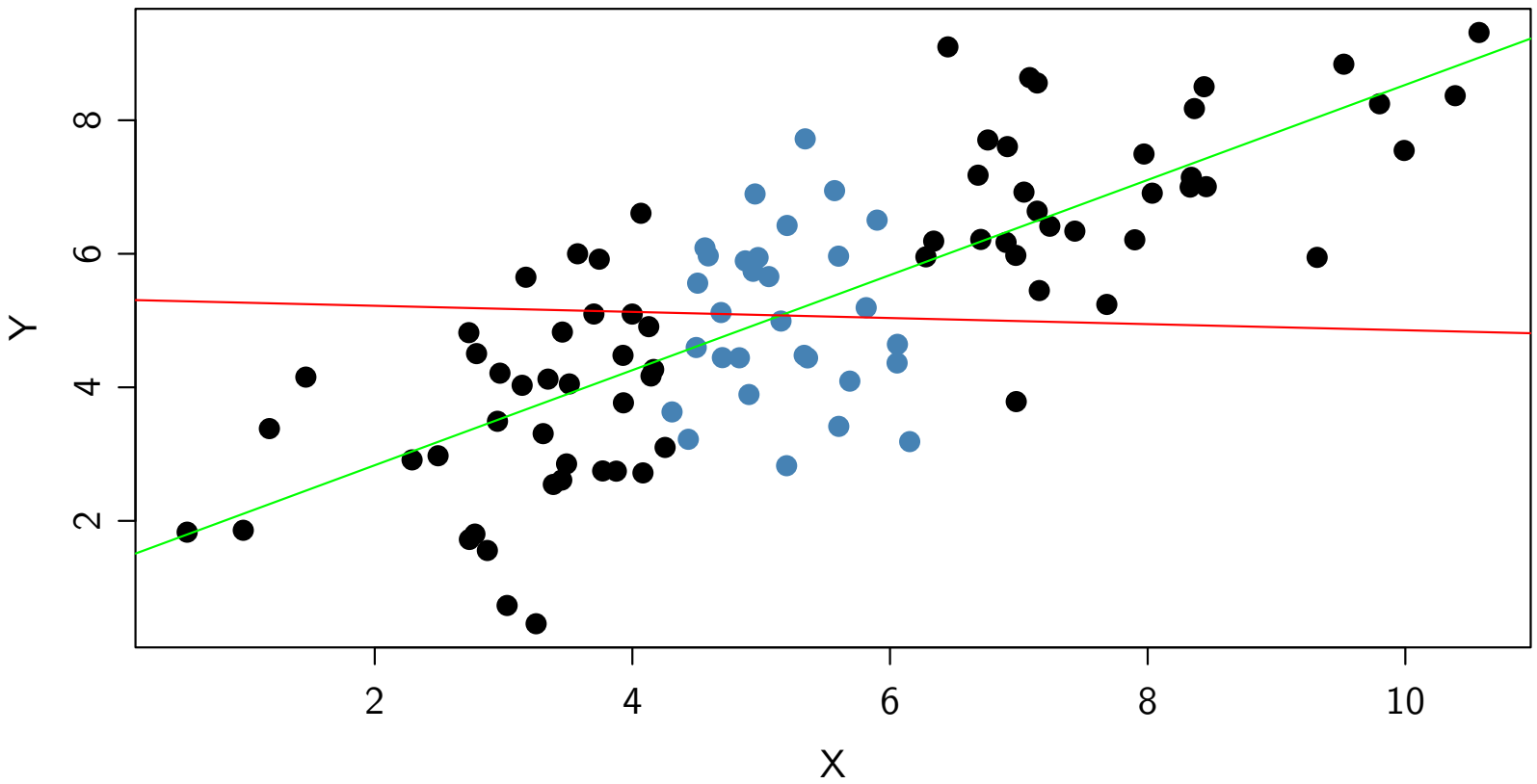
Varianz



- Es liegen die selbe Anzahl an schwarzen und blauen Beobachtungen vor.
- Welche Auswahl ergibt präzisere Schätzung der Regressionsgerade?

Erwartungswert und Varianz

Varianz



Stichprobenverteilung von $\widehat{\beta}_1$

- Die genaue Stichprobenverteilung ist kompliziert – diese hängt von der Populationsverteilung von (Y, X) ab.
- Aber wenn n sehr groß ist, gibt es zwei einfache (und gute) Approximationen:
  1. Da $\text{Var}(\widehat{\beta}_1) \propto 1/n$ und $\text{E}(\widehat{\beta}_1) = \beta_1$, $\widehat{\beta}_1 \xrightarrow{p} \beta_1$.
  2. Wenn n groß ist, ist die Stichprobenverteilung von $\widehat{\beta}_1$ gut durch eine Normalverteilung approximiert (ZGWS).
- Erinnern Sie sich an den **ZGWS**: Seien $\{v_i\}$, $i = 1, \ldots, n$ u.i.v. mit $\text{E}(v) = 0$ und $\text{Var}(v) = \sigma_v^2$. Dann gilt $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} v_i \xrightarrow{d} N(0, \sigma_v^2)$ für $n \to \infty$.

Stichprobenverteilung von $\widehat{\beta}_1$

Approximation der Verteilung von $\widehat{\beta}_1$ für großes n

- Wie gerade gesehen, ist (mit $v_i = (X_i - \overline{X})u_i$)

$$\widehat{\beta}_1 - \beta_1 = \frac{\frac{1}{n}\sum_{i=1}^n v_i}{\left(\frac{n-1}{n}\right) s_X^2} \approx \frac{\frac{1}{n}\sum_{i=1}^n v_i}{\sigma_X^2}.$$

- Wenn n groß ist, ist
  - $v_i = (X_i - \overline{X})u_i \approx (X_i - \mu_X)u_i$, was u.i.v. ist (warum?)
  - und $\text{Var}(v_i) < \infty$ (*warum*?).
- Also ist $\frac{1}{\sqrt{n}}\sum_{i=1}^n v_i$ nach dem ZGWS approximativ $N(0, \sigma_v^2)$ verteilt.
- Also gilt für große n, dass approximativ

$$\boxed{\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_v^2}{n\sigma_X^4}\right) \quad \text{wobei} \quad v_i \approx (X_i - \mu_X)u_i.}$$

Zusammenfassung: Stichprobenverteilung von $\widehat{\beta}_1$

Sofern die drei KQ-Annahmen erfüllt sind, gilt:

- Die genaue Stichprobenverteilung (in endlichen Stichproben) von $\widehat{\beta}_1$ hat:
  - $\mathrm{E}(\widehat{\beta}_1) = \beta_1$ (d.h., $\widehat{\beta}_1$ ist unverzerrt)
  - $\mathrm{Var}(\widehat{\beta}_1) = \frac{1}{n} \cdot \frac{Var[(X_i - \mu_X)u_i]}{\sigma_X^4}$
- Die (exakte) Verteilung von $\widehat{\beta}_1$ ist, anders als sein Erwartungswert und die Varianz, schwierig zu bestimmen und hängt von der Verteilung von (X, u) ab.
- $\widehat{\beta}_1 \xrightarrow{p} \beta_1$ (d.h. $\widehat{\beta}_1$ ist konsistent)
- Sofern n groß ist, $\frac{\widehat{\beta}_1 - \mathrm{E}(\widehat{\beta}_1)}{\sqrt{\mathrm{Var}(\widehat{\beta}_1)}} \approx N(0,1)$ (ZGWS)
- Dies ist genau wie bei der Stichprobenverteilung von $\overline{Y}$.

Zusammenfassung

Wir haben gesehen ...

- wie (zwei) Regressionsparameter zu schätzen sind,
- wie die Aussagekraft einer Regression zu bewerten ist, und
- wie die Stichprobenverteilung des Schätzers (für große n) ist.

Überblick

Regression mit einem einzelnen Regressor: Hypothesentests und Konfidenzintervalle

(SW Kapitel 5)

Überblick

- Jetzt, wo wir die Stichprobenverteilung des KQ-Schätzers kennen, können wir
  - Hypothesentests für β_1 durchführen und
  - Konfidenzintervalle für β_1 konstruieren.
- Auch werden wir das ein oder andere bisher Offengebliebene im Hinblick auf Regressionen diskutieren:
  - Regression falls X binär ist (0/1)
  - Heteroskedastizität und Homoskedastizität
  - Effizienz der KQ-Schätzung

Aber zunächst . . . ein Überblick (und Rückblick)

- Wir möchten β_1 mit Hilfe von Daten aus einer Stichprobe schätzen, also gibt es Stichprobenunsicherheit.
- Vier Schritte zu diesem Ziel:

1. Lege die Population des Untersuchungsobjekts genau fest.
2. Leite die Stichprobenverteilung eines Schätzers her (dazu benötigt z.B. die klassischen KQ-Annahmen).
3. Schätze die Varianz der Stichprobenverteilung (welche nach dem ZGWS alles ist, was wir kennen müssen, sofern *n* groß ist), d.h. bestimme den Standardfehler des Schätzers – *nur mit Hilfe der zur Verfügung stehenden Informationen aus der Stichprobe!*
4. Benutze den Schätzer ($\widehat{\beta}_1$) um eine Punktschätzung zu erhalten und dessen Standardfehler, um Hypothesen zu testen und Konfidenzintervalle zu konstruieren.

Das interessierende Objekt: β_1

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \quad i = 1, \ldots, n.$$

- Koeffizient $\beta_1 = \Delta Y / \Delta X$ für eine autonome Änderung von X ist idealerweise interpretierbar als kausaler Effekt.
- Unsere **KQ-Annahmen**:
  1. $\mathrm{E}(u|X = x) = 0$.
  2. (X_i, Y_i), $i = 1, \ldots, n$, sind u.i.v.
  3. Große Ausreißer sind selten: $\mathrm{E}(X^4) < \infty$, $\mathrm{E}(Y^4) < \infty$.
- Die **Stichprobenverteilung** von $\widehat{\beta}_1$: Unter den KQ-Annahmen ist $\widehat{\beta}_1$ für große n approximativ normalverteilt:

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_v^2}{n\sigma_X^4}\right), \qquad \text{wobei} \quad v_i = (X_i - \mu_X)u_i.$$

Überblick

- **Hypothesentests und der Standardfehler von $\widehat{\beta}_1$**
- Konfidenzintervalle für β_1
- Regression wenn X binär ist
- Homoskedastizität und Heteroskedastizität
- Einige weitere Details über KQ

Statistische Tests

- Das Ziel ist es, eine Hypothese wie $\beta_1 = 0$ mit Hilfe der Daten zu testen.
- Sind die Daten kompatibel mit der Hypothese oder nicht?
- **Aufbau**: Null-Hypothese und **zweiseitige** Alternative:

$$H_0 : \beta_1 = \beta_{1,0} \qquad \textit{vs.} \qquad H_1 : \beta_1 \neq \beta_{1,0}$$

  wobei $\beta_{1,0}$ der hypothetische Wert unter der Null ist.

  - Oftmals ist $\beta_{1,0} = 0$.

- Null-Hypothese und **einseitige** Alternative:

$$H_0 : \beta_1 = \beta_{1,0} \qquad \textit{vs.} \qquad H_1 : \beta_1 < \beta_{1,0} \quad \text{oder} \quad >$$

Statistische Tests

- **Generelle Herangehensweise**: Konstruiere die ***t*-Statistik**, und vergleiche mit den kritischen Werten der $N(0,1)$-Verteilung.
- Allgemein:

$$t = \frac{\text{Schätzwert} - \text{Wert unter Nullhypothese}}{\text{Standardfehler des Schätzers}}$$

  wobei der Standardfehler des Schätzers die Wurzel der Varianz ist.
- Konkret: Um $\beta_1 = \beta_{1,0}$ zu testen,

$$t = \frac{\widehat{\beta}_1 - \beta_{1,0}}{\text{SE}(\widehat{\beta}_1)}$$

  wobei $\text{SE}(\widehat{\beta}_1)$ = die Wurzel der (geschätzten) Varianz des Schätzers von β_1.

Formel für $\mathrm{SE}(\widehat{\beta}_1)$

- Erinnern Sie sich an den Ausdruck der Varianz von $\widehat{\beta}_1$ (großes n):

$$\mathrm{Var}(\widehat{\beta}_1) = \frac{Var[(X_i - \mu_X)u_i]}{n(\sigma_X^2)^2} = \frac{\sigma_v^2}{n\sigma_X^4},$$

  wobei $v_i \approx (X_i - \mu_X)u_i$.

- Die unbekannten Werte von σ_v^2 und σ_X^4 schätzen wir aus den Daten:

$$\widehat{\mathrm{Var}}(\widehat{\beta}_1) = \frac{1}{n} \cdot \frac{\text{Schätzer von } \sigma_v^2}{(\text{Schätzer von } \sigma_X^2)^2} = \frac{1}{n} \cdot \frac{\frac{1}{n-2}\sum_{i=1}^n \widehat{v}_i^2}{\left[\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2\right]^2},$$

  wobei $\widehat{v}_i = (X_i - \overline{X})\widehat{u}_i$.

Formel für $\mathrm{SE}(\widehat{\beta}_1)$

- Die Formeln sehen fies aus:

$$\widehat{\mathrm{Var}}(\widehat{\beta}_1) = \frac{1}{n} \cdot \frac{\frac{1}{n-2}\sum_{i=1}^{n} \widehat{v}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2} \quad \text{wobei } \widehat{v}_i = (X_i - \overline{X})\widehat{u}_i,$$

$$\mathrm{SE}(\widehat{\beta}_1) = \sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_1)} = \text{Standardfehler von } \widehat{\beta}_1.$$

- Aber es ist weniger kompliziert als es aussieht: Der Zähler schätzt $\mathrm{Var}(v)$, der Nenner schätzt $\mathrm{Var}(X)$.
- Warum die Korrektur um die Freiheitsgrade $n-2$? Weil zwei Koeffizienten geschätzt wurden (β_0 und β_1).
- $\mathrm{SE}(\widehat{\beta}_1)$ wird von jedem ökonometrischen Softwarepaket berechnet.
- Also besteht kein Grund die Formel auswendig zu lernen.

Zusammenfassung: H_0: $\beta_1 = \beta_{1,0}$ vs. $H_1 : \beta_1 \neq \beta_{1,0}$

- Konstruiere die t-Statistik

$$t^{act} = \frac{\widehat{\beta}_1^{act} - \beta_{1,0}}{\text{SE}(\beta_1^{act})}.$$

- Lehne zum 5% Signifikanzniveau ab, sofern $|t^{act}| > 1.96$.
- Der p-Wert ist

$p = Pr[|t| > |t^{act}|] \approx$ Wahrscheinlichkeit $N(0,1)$-Verteilung außerhalb $|t^{act}|$.

  - Lehne zum 5% Signifikanzniveau ab, falls $p < 5\%$ ist.
- Dieses Verfahren beruht auf der Approximation für große n.
  - Typischerweise ist die Approximation für $n \geq 50$ gut.

Beispiel
Kalifornische Schulbezirke

Beispiel 4.0: Fortsetzung.

Erinnern Sie sich an die geschätzte Regressionsgerade

$$\widehat{TestScore} = 698.9 - 2.28 \cdot STR.$$

liefert die Standardfehler

$$\text{SE}(\widehat{\beta}_0) = 10.4, \qquad \text{SE}(\widehat{\beta}_1) = 0.52.$$

Hinweis: Unten stehend sehen Sie einen anderen SE als im Rechneroutput in Kapitel 3, da dort der Standardfehler nicht optimal berechnet wurde – dazu später mehr in diesem Kapitel!

```
vcov      <- vcovHC(linear_model, type = "HC1")
robust_se <- sqrt(diag(vcov))
robust_se
```

```
## (Intercept)         STR
##     10.3644      0.5195
```

Beispiel
Kalifornische Schulbezirke

- t-Statistik für $H_0 : \beta_1 = 0$: $t^{act} = \frac{\widehat{\beta}_1^{act} - \beta_{1,0}}{\mathrm{SE}(\widehat{\beta}_1^{act})} = \frac{-2.28-0}{0.52} = -4.38.$
- (Warum ist $H_0 : \beta_1 = 0$ besonders interessant?)
- Der kritische Wert des 2-seitigen Tests zum 1% Signifikanzniveau ist 2.58, also lehnen wir die Nullhypothese zum 1% Signifikanzniveau ab.
- Alternativ können wir den p-Wert berechnen...

Beispiel
Kalifornische Schulbezirke

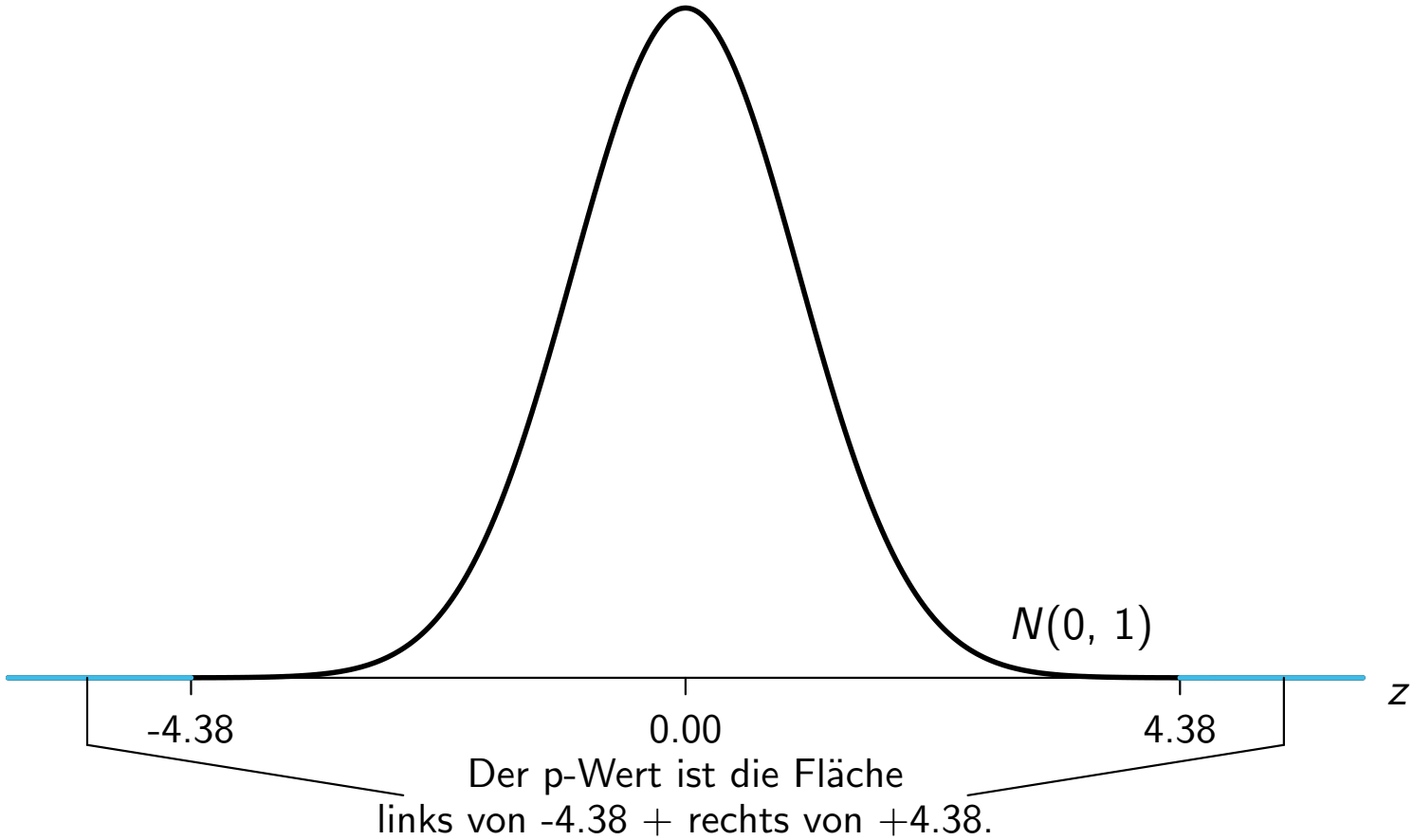


Abbildung 8: Der p-Wert, basierend auf der für große n mit der Standardnormalverteilung approximierten t-Statistik, ist $0.00001 = 10^{-5}$.

Kompaktdarstellung

- Schreibe die Standardfehler in Klammern unter die geschätzten Koeffizienten, zu welchen diese gehören:

$$\widehat{TestScore} = \underset{(10.4)}{698.9} - \underset{(0.52)}{2.28} \cdot STR, \qquad R^2 = 0.05, \qquad SER = 18.6.$$

Dieser Ausdruck enthält viele Informationen:

- Die geschätzte Regressionsgerade ist $\widehat{TestScore} = 698.9 - 2.28 \cdot STR$.
- Der Standardfehler von $\widehat{\beta}_0$ ist 10.4.
- Der Standardfehler von $\widehat{\beta}_1$ ist 0.52.
- Das R^2 ist 0.05; der Standardfehler der Regression ist 18.6.

Überblick

- Hypothesentests und der Standardfehler von $\widehat{\beta}_1$
- **Konfidenzintervalle für** β_1
- Regression wenn X binär ist
- Homoskedastizität und Heteroskedastizität
- Einige weitere Details über KQ

Plus/minus ...

- Erinnern Sie sich, dass ein 95%-Konfidenzintervall äquivalent definiert werden kann als:
  - eine mengenwertige Funktion der Daten (ein Intervall, welches eine Funktion der Daten ist), welche den wahren Wert des Parameters in 95% von wiederholt gezogenen Stichproben enthält; oder
  - die Menge der nullhypothetischen Werte, für die die Nullhypothese zum 5% Signifikanzniveau nicht abgelehnt werden kann.
- Da die t-Statistik für β_1 in großen Stichproben $N(0,1)$-verteilt ist, verläuft die Konstruktion eines 95%-Konfidenzintervalls für β_1 wie beim Erwartungswert:

$$\boxed{\text{95\%-Konfidenzintervall für } \beta_1 = \{\widehat{\beta}_1 \pm 1.96 \cdot \text{SE}(\widehat{\beta}_1)\}}$$

Beispiel
Kalfornische Schulbezirke

Beispiel 4.1:

Wir berechnen ein Konfidenzintervall für β_1 basierend auf

$$\widehat{TestScore} = \underset{(10.4)}{698.9} - \underset{(0.52)}{2.28} \cdot STR, \qquad R^2 = 0.05, \qquad SER = 18.6.$$

Das 95%-Konfidenzintervall ist

$$\begin{aligned} \{\widehat{\beta}_1 \pm 1.96 \cdot \mathrm{SE}(\widehat{\beta}_1)\} &= \{-2.28 \pm 1.96 \cdot 0.52\} \\ &= (-3.30, -1.26) \end{aligned}$$

Die folgenden beiden Aussagen sind äquivalent (warum?)

- Das 95%-Konfidenzintervall beinhaltet nicht die Null.
- Die Hypothese $\beta_1 = 0$ wird zum Niveau 5% verworfen.

Siehe hierzu `ConfidenceIntervals.R`.

Beispiel
Kalfornische Schulbezirke

Beispiel 4.1: Fortsetzung.

F: Kann das 95%-Konfidenzintervalls $(-3.30, -1.26)$ für β_1 wie folgt interpretiert werden:

Der wahre Wert β_1 liegt mit 95%-Wahrscheinlichkeit zwischen -3.30 und -1.26?

A: Diese Interpretation zeugt von falschem statistischem Denken. Sowohl β_1 als auch das Intervall $(-3.30, -1.26)$ sind nicht zufällig. β_1 liegt also entweder im Intervall oder es tut es nicht. Daher: β_1 liegt entweder mit Wahrscheinlichkeit 100% im Intervall oder mit Wahrscheinlichkeit 0% – aber sicher nicht mit Wahrscheinlichkeit 95%.

Korrekte Interpretation: Aus 100 wiederholt gezogenen Stichproben aus allen kalifornischen Schulbezirken erwarten wir, dass in 95 Fällen das Konfidenzintervall den wahren Wert β_1 enthält.

Zusammenfassung: Inferenz für β_0 und β_1

Schätzung:

- KQ-Schätzer: $\widehat{\beta}_0$ und $\widehat{\beta}_1$.
- $\widehat{\beta}_0$ und $\widehat{\beta}_1$ approximativ normalverteilt in großen Stichproben.

Testen:

- $H_0 : \beta_1 = \beta_{1,0}$ vs. $\beta_1 \neq \beta_{1,0}$.
- $t = (\widehat{\beta}_1 - \beta_{1,0}) / \mathrm{SE}(\widehat{\beta}_1)$.
- p-Wert = Fläche unter Standardnormalverteilung außerhalb von $|t^{act}|$.

Konfidenzintervalle:

- 95%-Konfidenzintervall für β_1 ist $\{\widehat{\beta}_1 \pm 1.96 \cdot \mathrm{SE}(\widehat{\beta}_1)\}$.
- Dies ist die Menge aller $\beta_{1,0}$, für die die Nullhypothese $\beta_1 = \beta_{1,0}$ zum Niveau 5% nicht verworfen werden kann.

Überblick

- Hypothesentests und der Standardfehler von $\widehat{\beta}_1$
- Konfidenzintervalle für β_1
- **Regression wenn X binär ist**
- Homoskedastizität und Heteroskedastizität
- Einige weitere Details über KQ

Unterscheidung

- Gelegentlich ist ein Regressor **binär**:
  - $X = 1$ falls weiblich, $= 0$ falls männlich.
  - $X = 1$ falls behandelt (experimentelles Medikament), $= 0$ falls nicht.
- Binäre Regressoren werden gelegentlich auch **Dummyvariablen** genannt.
- Bis jetzt wurde β_1 als "Steigungsparameter" bezeichnet, aber das macht wenig Sinn, wenn X binär ist.
- Wie interpretieren wir Regressionen mit einem binären Regressor?

Interpretation von binären Regressionen

- $Y_i = \beta_0 + \beta_1 X_i + u_i$, wobei X binär ($X_i = 0$ oder 1):
- Einerseits ist der Erwartungswert von Y_i gegeben $X_i = 0$
  - $\mathrm{E}(Y_i|X_i = 0) = \beta_0$.
- Andererseits ist der Erwartungswert von Y_i gegeben $X_i = 1$
  - $\mathrm{E}(Y_i|X_i = 1) = \beta_0 + \beta_1$.
- Also:

$$\begin{aligned}\beta_1 &= \mathrm{E}(Y_i|X_i = 1) - \mathrm{E}(Y_i|X_i = 0) \\ &= \text{Populationsunterschied der Gruppenerwartungswerte}\end{aligned}$$

Beispiel
Kalifornische Schulbezirke

Beispiel 4.2: Kalifornische Schulbezirke.

- Sei
$$D_i = \begin{cases} 1, & \text{wenn } STR_i < 20, \\ 0, & \text{wenn } STR_i \geq 20. \end{cases}$$
- KQ-Regression: $\widehat{TestScore} = \underset{(1.3)}{650.0} + \underset{(1.8)}{7.4} \cdot D$
- **Abbildung der Gruppenmittelwerte**

Klassengröße	Durchschnittlicher Score ($\overline{Y}$)	Std.abw (s_Y)	N
Klein ($STR < 20$)	657.4	19.4	238
Groß ($STR \geq 20$)	650.0	17.9	182

- Differenzen der Mittelwerte: $\overline{Y}_{klein} - \overline{Y}_{gross} = 657.4 - 650.0 = 7.4$
- Standardfehler: $\text{SE} = \sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}} = \sqrt{\frac{19.4^2}{238} + \frac{17.9^2}{182}} = 1.8$

Beispiel I

Kalifornische Schulbezirke

```
CASchools$D <- CASchools$STR < 20    # Create the dummy variable as defined above
plot(CASchools$D, CASchools$score,   # provide the data to be plotted
     pch = 20,                       # use filled circles as plot symbols
     cex = 0.5,                      # set size of plot symbols to 0.5
     col = "Steelblue",              # set the symbols' color to "Steelblue"
     xlab = expression(D[i]),        # Set title and axis names
     ylab = "Test Score")            # Plot the data
dummy_model <- lm(score ~ D, data = CASchools)
points(x = CASchools$D,             # add group specific predictions to the plot
       y = predict(dummy_model),
       col = "red",
       pch = 20)
```

Beispiel II

Kalifornische Schulbezirke

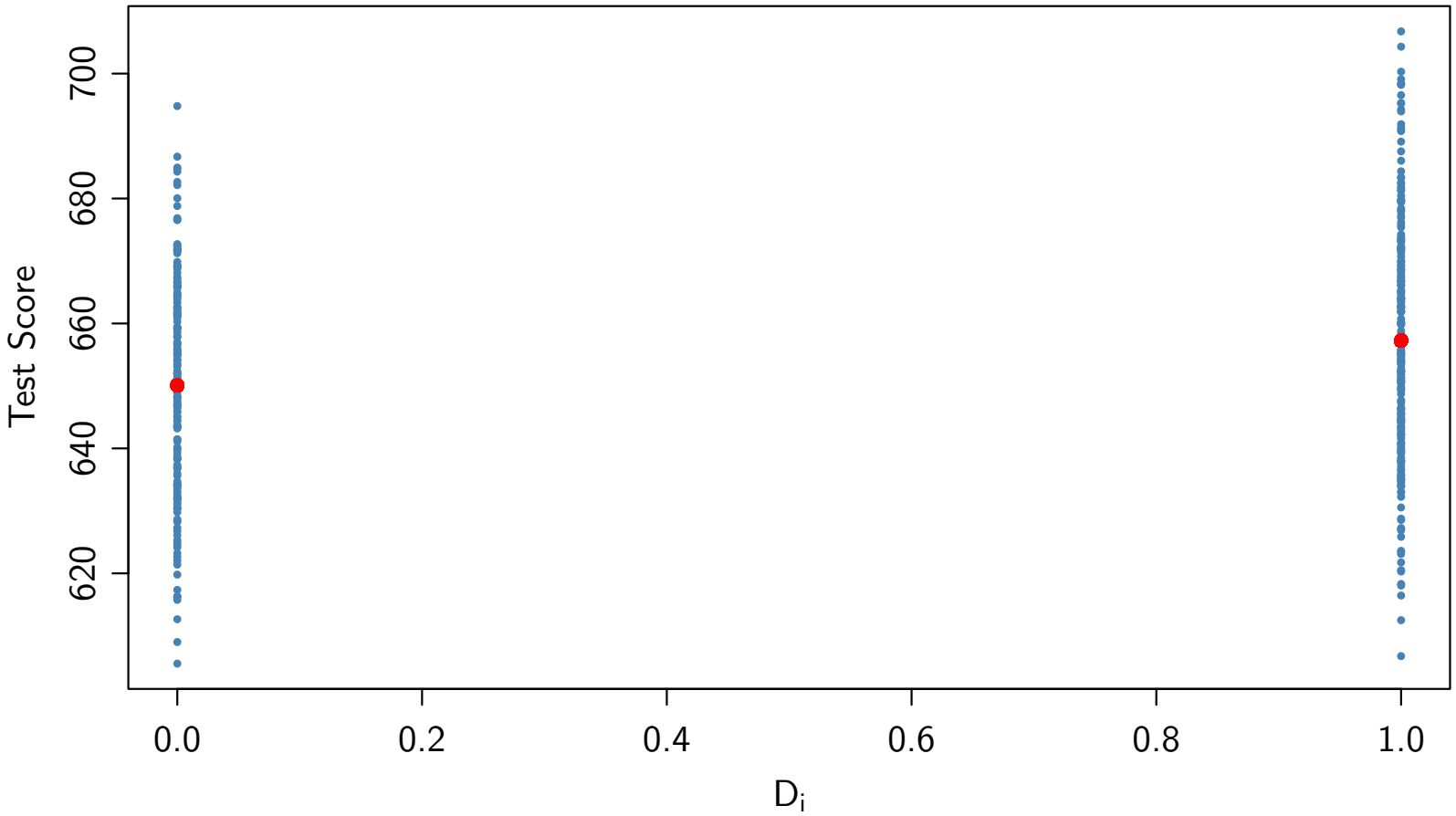


Abbildung 9: Dummy Regression

Zusammenfassung: Regression für binäres X_i

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- β_0 = Erwartungswert von Y falls $X = 0$.
- $\beta_0 + \beta_1$ = Erwartungswert von Y falls $X = 1$.
- β_1 = Unterschiede in Gruppenmittelwerten, $X = 1$ minus $X = 0$.
- SE($\widehat{\beta}_1$) hat die übliche Interpretation.
- t-Statistiken, Konfidenzintervalle werden wie üblich konstruiert.
- Dies ist eine weitere (einfache) Möglichkeit um die sogenannte "difference-in-means" Analyse durchzuführen.
- Diese Formulierung einer Regression ist besonders nützlich, wenn wir zusätzliche Regressoren haben (die wir bald haben werden).

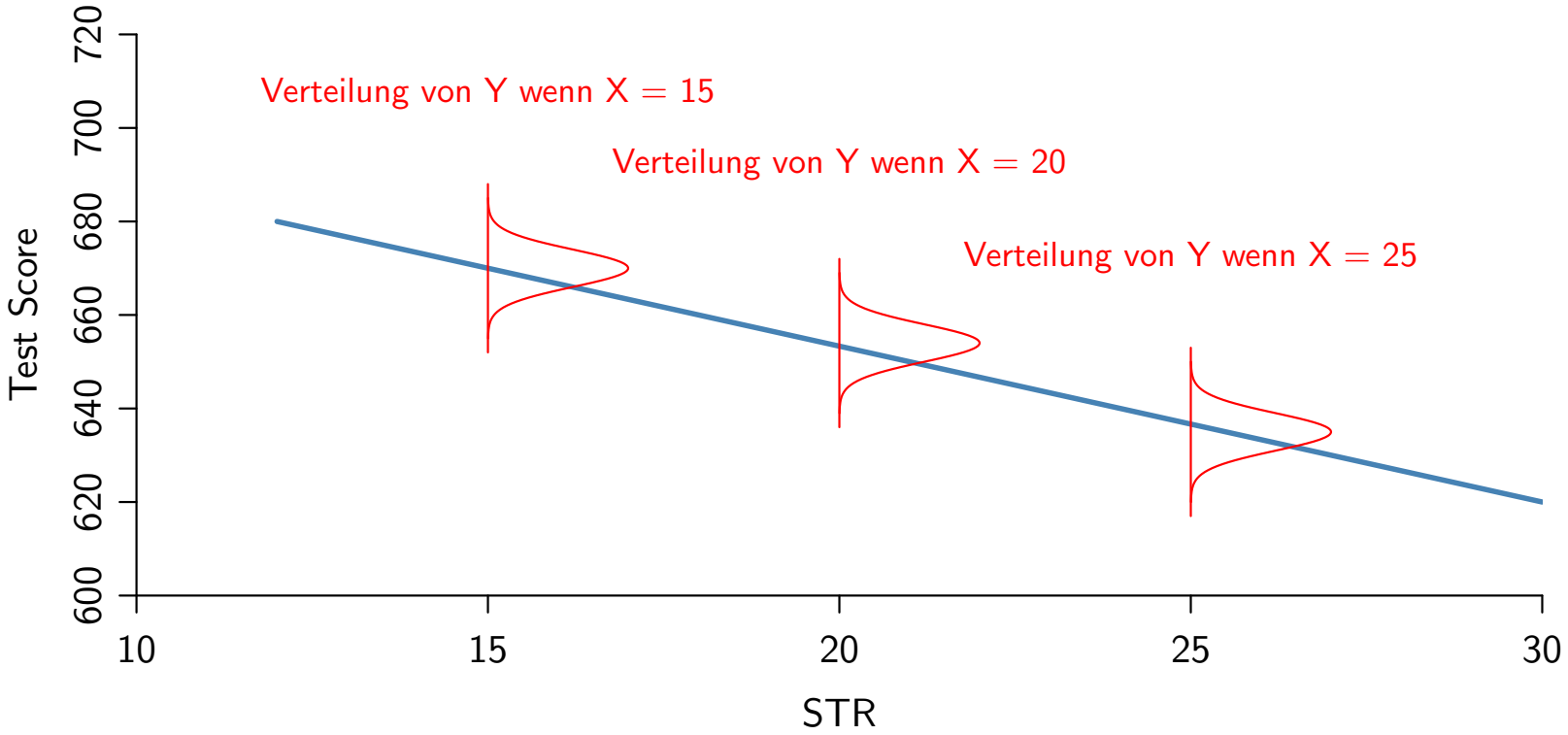
Überblick

- Hypothesentests und der Standardfehler von $\widehat{\beta}_1$
- Konfidenzintervalle für β_1
- Regression wenn X binär ist
- **Homoskedastizität und Heteroskedastizität**
- Einige weitere Details über KQ

Definition

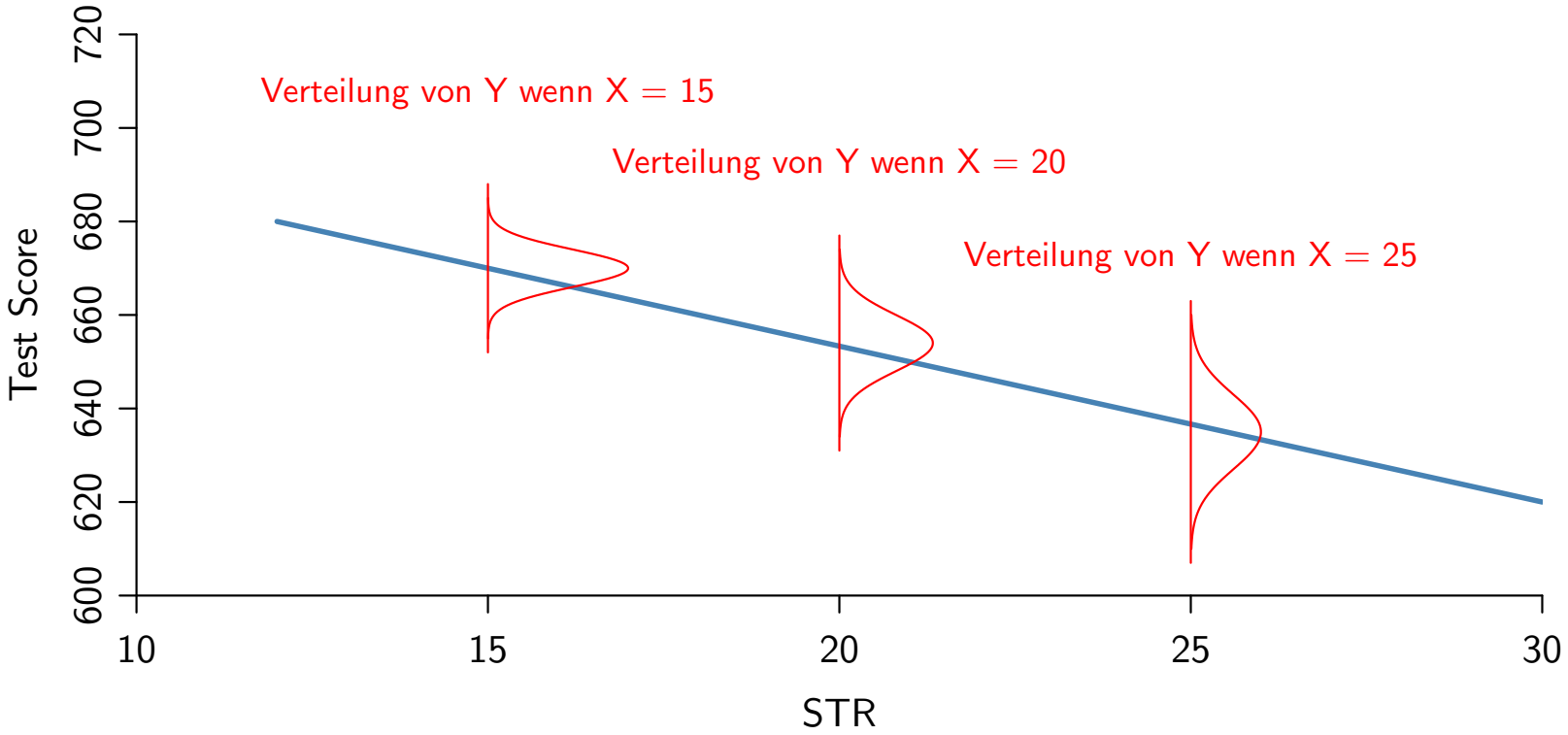
- Wenn $\text{Var}(u|X = x)$ konstant ist – d.h. falls die Varianz der bedingten Verteilung von u nicht von X abhängt –, dann ist u **homoskedastisch**.
- Andernfalls ist u **heteroskedastisch**.
- Heteroskedastizität ist der übliche Fall.
- Im Folgenden: Konsequenzen von Homo-/Heteroskedastizität für Standardfehler und Effizienz von KQ.

Homoskedastizität in einem Bild



- $\mathrm{E}(u|X = x) = 0$ (u erfüllt die KQ-Annahme Nr. 1).
- Die Varianz von u hängt **nicht** von x ab.

Heteroskedastizität in einem Bild



- $\mathrm{E}(u|X = x) = 0$ (u erfüllt die KQ-Annahme Nr. 1).
- Die Varianz von u hängt **in der Tat** von x ab: u ist heteroskedastisch.

Heteroskedastisch oder homoskedastisch?

Beispiel aus der Arbeitsmarktökonomie

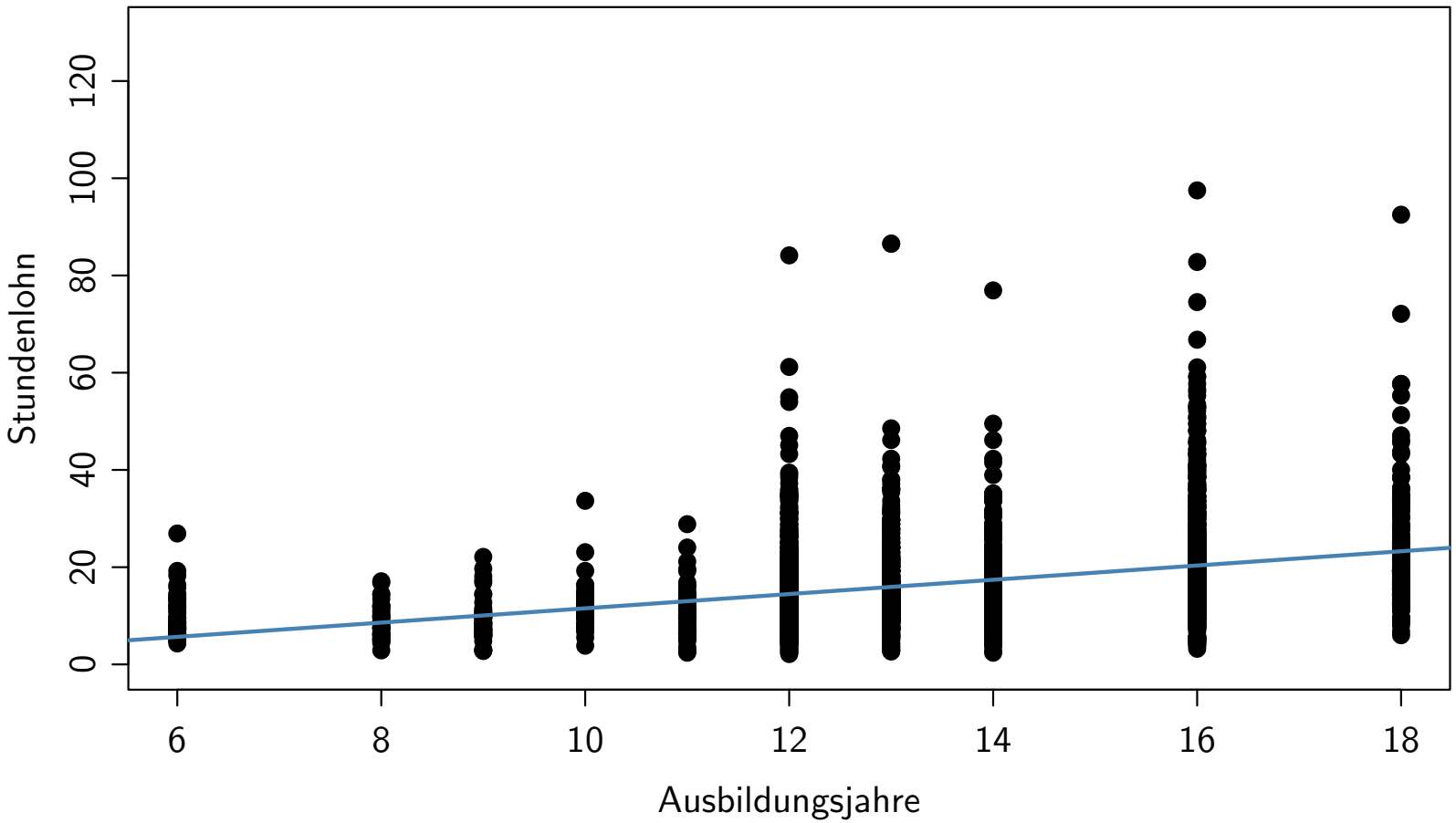


Abbildung 10: Durchschnittlicher Stundenlohn vs. Ausbildungsjahre (Datenquelle: CPS)

Heteroskedastisch oder homoskedastisch?
Kalifornische Schulbezirke

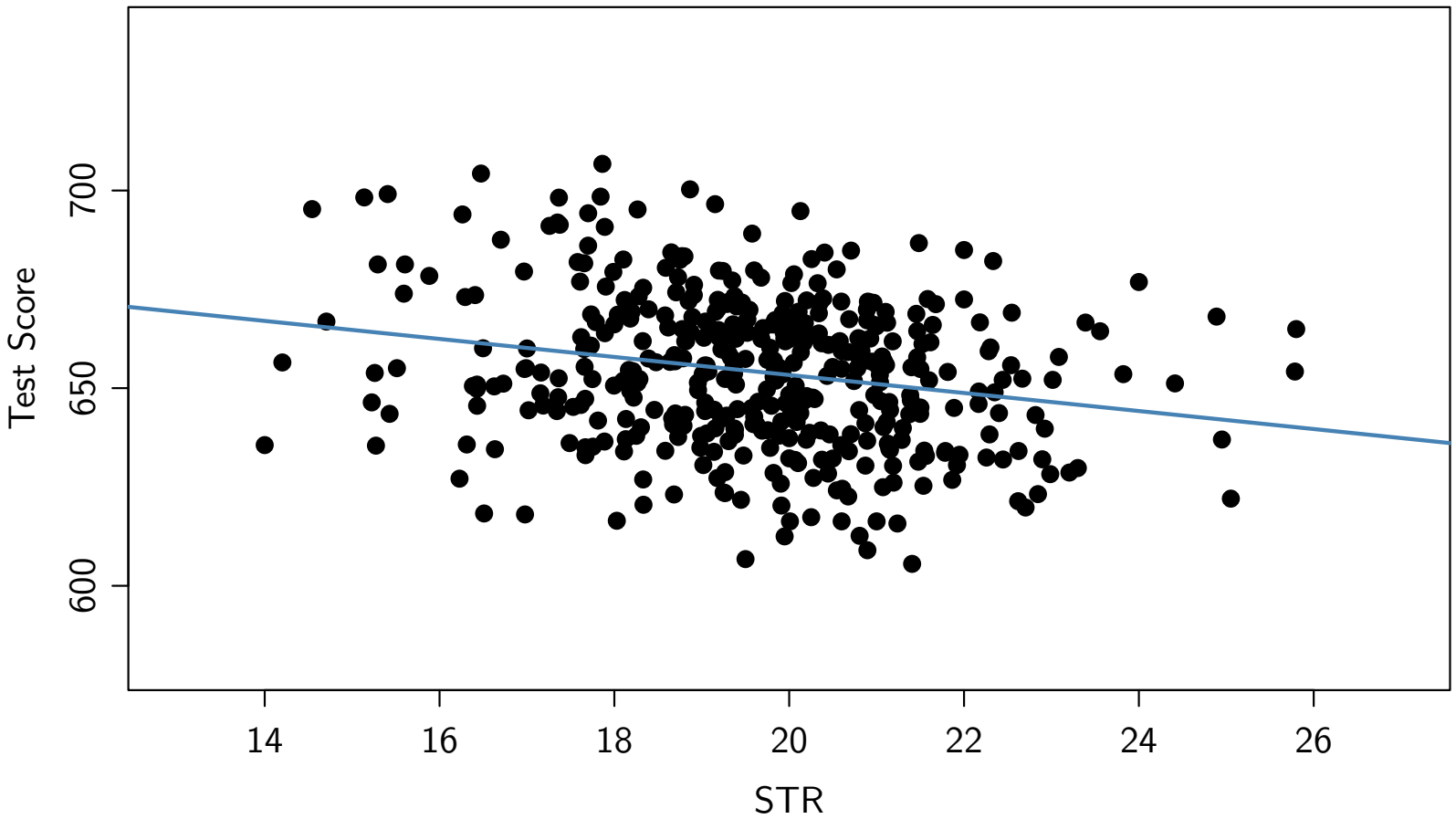


Abbildung 11: TestScore vs. STR.

Was bisher geschah . . .

- Bis jetzt haben wir (stillschweigend) angenommen, dass u heteroskedastisch sein könnte.
- Erinnern Sie sich an die drei KQ-Annahmen:
  1. $\mathrm{E}(u|X = x) = 0$
  2. (X_i, Y_i), $i = 1, \ldots, n$, sind u.i.v.
  3. Große Ausreißer sind selten
- Heteroskedastizität und Homoskedastizität betreffen $\mathrm{Var}(u|X = x)$.
- Da wir nicht explizit homoskedastische Störterme angenommen haben, haben wir implizit heteroskedastische in Kauf genommen.

Was, wenn Störterme homoskedastisch sind?

- Man kann zeigen, dass KQ die kleinste Varianz unter den in Y linearen unverzerrten Schätzern hat ... bekannt als das Gauss-Markov Theorem.
- Die Formel für die Varianz von $\widehat{\beta}_1$ und der KQ-Standardfehler vereinfachen sich: Falls $\text{Var}(u_i|X_i = x) = \sigma_u^2$, dann

$$\text{Var}(\widehat{\beta}_1) = \frac{Var[(X_i - \mu_X)u_i]}{n(\sigma_X^2)^2} = \frac{E[(X_i - \mu_X)^2 u_i^2]}{n(\sigma_X^2)^2}$$
$$= \frac{\sigma_u^2}{n\sigma_X^2}.$$

- Anmerkung: $\text{Var}(\widehat{\beta}_1)$ ist invers proportional zu $\text{Var}(X)$: mehr Ausweitung von X bedeutet mehr Informationen über $\widehat{\beta}_1$.
- Wir haben das bereits diskutiert, aber diese Formel macht es deutlicher.

Was, wenn Störterme homoskedastisch sind?

- Zusammen mit dieser nur bei Homoskedastie gültigen Formel für die Varianz von $\widehat{\beta}_1$, erhalten wir nur bei Homoskedastie konsistente Standardfehler:
- **Formel für nur bei Homoskedastie gültige Standardfehler:**

$$\boxed{\mathrm{SE}(\widehat{\beta}_1) = \sqrt{\frac{1}{n} \cdot \frac{\frac{1}{n-2}\sum_{i=1}^{n} \widehat{u}_i^2}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}}.}$$

- Vorteil: Formel ist einfacher.
- Nachteil: Nur gültig bei Homoskedastie.
- Zur Unterscheidung bezeichnen wir die bisherigen "normalen" Standardfehler als **Heteroskedastie-robuste Standardfehler**, weil sie unabhängig davon gültig sind, ob die Störterme heteroskedastisch sind oder nicht.

Praktische Implikationen

- **Warnung:** Die nur bei Homoskedastie konsistenten Standardfehler werden in fast allen Softwarepaketen standardmäßig ausgegeben!
- Um die "Heteroskedastie-robusten" Standardfehler zu erhalten, muss die Standardeinstellung geändert werden.
- Sollte das nicht getan werden und die Störterme sind in der Tat heteroskedastisch, werden falsche Standardfehler (falsche t-Statistiken und falsche Konfidenzintervalle) berechnet.
  - Üblicherweise sind die nur Homoskedastie-konsistenten SE zu klein!
  - Siehe hierzu `ConfidenceIntervalsHeteroskedasticity.R`.
- Also immer Heteroskedastie-robuste Standardfehler benutzen, um auf der sicheren Seite zu sein ...
  - ... oder im Zweifel beide berechnen und größeren SE benutzen!

Überblick

- Hypothesentests und der Standardfehler von $\widehat{\beta}_1$
- Konfidenzintervalle für β_1
- Regression wenn X binär ist
- Homoskedastizität und Heteroskedastizität
- **Einige weitere Details über KQ**

KQ-Eigenschaften

- Wir haben schon eine Menge über KQ gelernt:
  - KQ ist unverzerrt und konsistent;
  - wir haben eine Formel für Heteroskedastie-robuste Standardfehler;
  - und wir können Konfidenzintervalle konstruieren und Hypothesen testen.
- Ein weiterer guter Grund um KQ zu verwenden ist, dass es jeder andere auch tut – und daher jeder versteht was man macht.
- Im Endeffekt ist KQ die Sprache der Regressionsanalyse und wenn man einen anderen Schätzer verwendet, spricht man eine andere Sprache.

KQ-Eigenschaften

- Trotzdem werden vielleicht noch Fragen offen sein:
  - Ist das wirklich ein guter Grund um KQ anzuwenden?
  - Gibt es keine anderen Schätzer, die besser sein könnten – insbesondere welche mit kleinerer Varianz?
  - Wie robust ist KQ?
- Wir werden nun diese Fragen beantworten – allerdings werden wir einige stärkere Annahmen als die letzten drei KQ-Annahmen treffen müssen.

Die “klassischen” KQ-Annahmen

Klassische KQ-Annahmen

1. $\mathrm{E}(u|X = x) = 0$.
2. (X_i, Y_i), $i = 1, \ldots, n$, sind u.i.v.
3. Große Ausreißer sind selten ($\mathrm{E}(Y^4) < \infty$, $\mathrm{E}(X^4) < \infty$).
4. u ist homoskedastisch.
5. u ist $N(0, \sigma^2)$ verteilt.

- Die Annahmen 4. und 5. sind restriktiver – also in der Praxis seltener erfüllt.
- Wenn man jedoch diese Annahmen trifft, macht dies einige mathematische Berechnungen einfacher und man kann starke Ergebnisse beweisen – Ergebnisse mit Gültigkeit, sollten die Annahmen Bestand haben.
- Wir beginnen mit einer Diskussion der Effizienz von KQ.

Der KQ-Schätzer ist der Beste

Theorem 4.3: Gauss–Markov Theorem.

Unter den erweiterten KQ-Annahmen 1–4 hat $\widehat{\beta}_1$ die kleinste Varianz unter allen linearen und unverzerrten Schätzern (Schätzer, die lineare Funktionen von $Y_1, \ldots, Y_n$ sind).

- $\widehat{\beta}_1$ ist ein linearer Schätzer, d.h. dieser kann als lineare Funktion von $Y_1, \ldots, Y_n$ geschrieben werden:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} = \frac{\sum_{i=1}^n (X_i - \overline{X}) Y_i}{\sum_{i=1}^n (X_i - \overline{X})^2} = \sum_{i=1}^n w_i Y_i,$$

  wobei $\sum_{i=1}^n (X_i - \overline{X})\overline{Y} = \overline{Y} \sum_{i=1}^n (X_i - \overline{X}) = 0$ und $w_i = \frac{(X_i - \overline{X})}{\sum_{i=1}^n (X_i - \overline{X})^2}$.

- Das GM Theorem sagt, dass **unter allen möglichen** $\{w_i\}$ die KQ-Gewichte die kleinste $\text{Var}(\widehat{\beta}_1)$ ergeben.

KQ ist noch besser

- Unter **allen fünf** erweiterten KQ-Annahmen – inklusive normalverteilter Störterme – hat $\widehat{\beta}_1$ die kleinste Varianz von **allen** konsistenten Schätzern (lineare oder nichtlineare Funktionen von $Y_1, \ldots, Y_n$), wenn $n \to \infty$.
  - Siehe `efficiencyOLSvsLAD.R`.
- Dies ist ein starkes Ergebnis – es besagt, dass wenn (zusätzlich zu den KQ-Annahmen 1–3) die Störterme homoskedastisch und normalverteilt sind, dann ist KQ eine bessere Wahl als alle anderen konsistenten Schätzer.
- Und da ein inkonsistenter Schätzer eine schlechte Wahl ist, besagt dies, dass KQ tatsächlich die beste zu verwendende Methode ist – **sofern alle fünf erweiterten KQ-Annahmen erfüllt sind.**

Einige nicht so gute Dinge an KQ

Die bisherigen Ergebnisse sind beeindruckend, aber diese Ergebnisse – und der KQ-Schätzer – haben wichtige Grenzen.

1. Das GM Theorem ist vielleicht gar nicht so zwingend:
   - Die Bedingung der Homoskedastie ist oft nicht erfüllt. (Homoskedastie ist besonders.)
   - Das Ergebnis gilt nur für lineare Schätzer – nur eine kleine Gruppe von Schätzern.
2. Das stärkste Optimalitätsergebnis setzt homoskedastische normalverteilte Störterme voraus – nicht plausibel in Anwendungen.

Einige nicht so gute Dinge an KQ

3. KQ reagiert stärker auf Ausreißer als andere Schätzer.
   - Bei großen Ausreißern können andere Schätzer effizienter sein (eine kleinere Varianz haben).
   - Einer dieser Schätzer ist der Schätzer der kleinsten absoluten Abweichungen (LAD):

$$\min_{b_0, b_1} \sum_{i=1}^{n} |Y_i - (b_0 + b_1 X_i)|.$$

   - Siehe nochmal `efficiencyOLSvsLAD.R`.

In fast allen angewandten Regressionsanalysen wird KQ verwendet, und daher werden wir diese Methode auch im Kurs weiter behandeln.

Zusammenfassung

Wir haben gesehen ...

- wie Hypothesentests für Parameter durchgeführt werden
- und wie Konfidenzintervalle zu konstruieren sind, wenn wir mit KQ regressieren.

Überblick

Einführung in die multiple Regression

(SW Kapitel 6)

Überblick

- **Omitted variable bias**
- Multiples Regressionsmodell
- Anpassungsmaße für multiple Regressionen
- KQ-Annahmen für die multiple Regression
- Stichprobenverteilung des KQ-Schätzers

Omitted variable bias

- Wir möchten folgendes Modell zur kausalen Analyse benutzen:

$$Y = \beta_0 + \beta_1 X + u.$$

- Bei der kausalen Analyse ist die KQ-Annahme 1, $\mathrm{E}(u \mid X) = 0$, häufig anrüchig.
- Insbesondere gibt es häufig eine Variable Z, die Einfluss auf Y hat und mit X korreliert.
- Der Fehlerterm u enthält dann alle diese Variablen.
  - Z.B. $u = \beta_2 Z + \varepsilon$ mit $Z = \delta X + \xi$ und $\mathrm{E}(\varepsilon \mid X) = \mathrm{E}(\xi \mid X) = 0$ führt zu
  - $\mathrm{E}(u \mid X) = \beta_2 \delta X \neq 0$.
- **Problem**: $\mathrm{E}(u \mid X) \neq 0$ führt zu verzerrten Schätzungen von β_1.

Mechanismus

Definition 5.1:

Omitted variable bias entsteht, wenn ein ausgelassener Faktor Z

1. eine Determinante von Y ist (d.h. Z ist ein Teil von u);
2. mit dem Regressor X korreliert (d.h. $\text{Corr}(X, Z) \neq 0$).

Beispiel
Kalifornische Schulbezirke

Beispiel 5.2:

Erinnern Sie sich:

$$Y = \beta_0 + \beta_1 X + u$$

mit $Y = TestScore$, $X = STR$.

Betrachte $Z =$ Anteil an Nicht-Muttersprachlern im Distrikt. Z ist eine omitted variable, denn:

1. Z ist eine Determinante von Y: Fähigkeit Englisch zu sprechen hat Einfluss auf TestScores.
2. Z ist mit X korreliert: Distrikte mit vielen Immigranten sind oftmals weniger wohlhabend und haben daher weniger Budget für Schulen – und daher höhere STR.

Somit ist $\widehat{\beta}_1 = -2.28$ verzerrt. Was ist die Richtung der Verzerrung?

Beispiel
Kalifornische Schulbezirke

	$STR < 20$		$STR \geq 20$		Differenzen	
	TestScore	n	TestScore	n	Diff.	t-Stat.
Alle Distrikte	657.4	238	650.0	182	7.4	4.04
Anteil Nicht-Muttersprachler						
$< 1.9\%$	664.5	76	665.4	27	-0.9	-0.30
$1.9 - 8.8\%$	665.2	64	661.8	44	3.3	1.13
$8.8 - 23.0\%$	654.9	54	649.7	50	5.2	1.72
$> 23\%$	636.7	44	634.8	61	1.9	0.68

Tabelle 1: Differenzen in TestScores für Distrikte mir niedriger und hoher STR, sortiert nach Z

1. Distrikte mit weniger englischlernenden Personen haben höhere Test-Scores.
2. Distrikte mit weniger Prozent *EL (PctEL)* haben kleinere Klassen.
3. Unter den Distrikten mit vergleichbaren *PctEL* ist der Effekt der Klassengröße klein.

Beispiel
Kalifornische Schulbezirke

Beispiel 5.2: Fortsetzung.

- Distrikte mit um 1 erhöhter STR haben durchschnittlich -2.28 niedrigeren TestScore.
- Distrikte mit um 1 erhöhter STR haben aber auch höheres Z, also mehr Nicht-Muttersprachler.
- Also: *TestScore* ist um durchschnittlich 2.28 niedriger, weil (kausal!)
  1. STR um 1 erhöht und
  2. Z erhöht.
- Also: $\widehat{\beta}_1 = -2.28$ fängt nicht nur partiellen Effekt von $(STR \to STR + 1)$ ein, sondern auch Erhöhung von Z.
- Also: $\widehat{\beta}_1 = -2.28$ nach unten verzerrt.
- Es gibt aber auch eine Formel...

Eine Formel für den omitted variable bias

- Erinnern Sie sich an die Gleichung

$$\widehat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^n (X_i - \overline{X})u_i}{\sum_{i=1}^n (X_i - \overline{X})^2} \approx \frac{\frac{1}{n}\sum_{i=1}^n v_i}{\frac{n-1}{n}s_X^2},$$

  wobei $v_i = (X_i - \mu_X)u_i \approx (X_i - \overline{X})u_i$.

- Unter der KQ-Annahme 1 gilt,

$$E[(X_i - \mu_X)u_i] = \text{Cov}(X_i, u_i) = 0$$

- Aber was wenn $E[(X_i - \mu_X)u_i] = \text{Cov}(X_i, u_i) = \sigma_{Xu} \neq 0$?

Eine Formel für den omitted variable bias

- Im Allgemeinen (also, selbst wenn KQ-Annahme 1 nicht erfüllt ist),

$$\begin{aligned}
\widehat{\beta}_1 - \beta_1 &= \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2} \\
&\xrightarrow{p} \frac{\sigma_{Xu}}{\sigma_X^2} \\
&= \left(\frac{\sigma_u}{\sigma_X}\right) \cdot \left(\frac{\sigma_{Xu}}{\sigma_X \sigma_u}\right) = \left(\frac{\sigma_u}{\sigma_X}\right) \rho_{Xu},
\end{aligned}$$

  wobei $\rho_{Xu} = \text{Corr}(X, u)$.

- Wenn KQ-Annahme 1 erfüllt ist, dann $\rho_{Xu} = 0$, aber wenn dies nicht gilt ...

Eine Formel für den omitted variable bias

- Die omitted variable bias Formel:

$$\boxed{\widehat{\beta}_1 \xrightarrow{p} \beta_1 + \left(\frac{\sigma_u}{\sigma_X}\right) \rho_{Xu}}$$

- Wenn ein ausgelassener Faktor Z beides ist:
  1. Eine Determinante von Y (d.h. enthalten in u) und
  2. korreliert mit X,
- dann ist $\rho_{Xu} \neq 0$ und der KQ-Schätzer $\widehat{\beta}_1$ inkonsistent (und verzerrt).

Beispiel
Kalifornische Schulbezirke

Beispiel 5.2: Fortsetzung.

Erinnern Sie sich:

$$Y = \beta_0 + \beta_1 X + u$$

mit $Y = \mathit{TestScore}$, $X = \mathit{STR}$, $\rho_{Xu} = \text{Corr}(X, u)$.

- Wie beeinflusst $Z =$ Anteil Nicht-Muttersprachler $Y = \mathit{TestScore}$? Negativ.
- Wie korreliert $Z =$ Anteil Nicht-Muttersprachler $X = \mathit{STR}$? Positiv.
- In welche Richtung geht die Verzerrung? Positiv mal negativ ist negativ: $\rho_{Xu} < 0$.
- Was ist die Intuition? Wenn es keinen kausalen Effekt $\mathit{STR} \to \mathit{TestScore}$ gibt, haben Distrikte mit höherer STR mehr Nicht-Muttersprachler und deshalb niedrigeren $\mathit{TestScore}$. Wenn es einen (negativen) kausalen Effekt gibt, fängt KQ beide Effekte ein.

Exkurs: Kausalitäts- und Regressionsanalyse

- Was ist genau genommen ein kausaler Effekt?
- Die allgemeine Definition von Kausalität ist nicht präzise genug für unsere Zwecke.
- In diesem Kurs definieren wir einen kausalen Effekt als den Effekt, der in einem **idealen randomisierten kontrollierten Experiment** gemessen wurde.

Exkurs: Kausalitäts- und Regressionsanalyse

Ideal randomisiert kontrolliertes Experiment

- **Ideal**: Alle Subjekte folgen genau der vorgesehenen Behandlung – perfekte Regelbefolgung, keine Fehler im Bericht, etc.!
- **randomisiert**: Subjekte der zu untersuchenden Population werden zufällig zu einer Behandlungs- oder Kontrollgruppe zugeordnet (also gibt es keine Störfaktoren).
- **kontrolliert**: Eine Kontrollgruppe zu haben, erlaubt das Messen von Differenzeffekten einer Behandlung.
- **Experiment**: Die Behandlung ist zugeteilt als Teil des Experiments: Die Subjekte haben keine Entscheidung, also gibt es keine "umgekehrte Kausalität" in welcher Subjekte die Behandlung nach eigener Einschätzung der spezifischen Erfolgsaussichten auswählen.

Exkurs: Kausalitäts- und Regressionsanalyse
Beispiel

Beispiel 5.2: Fortsetzung.

- Entwerfen Sie ein ideales randomisiert kontrolliertes Experiment zur Messung des Effektes von *STR* auf *TestScore* ...
- Wie unterscheiden sich unsere Beobachtungen von diesem Ideal?
  - Die Behandlung ist nicht zufällig zugeteilt.
  - Betrachte *PctEL* – Prozent Englischlernende – im Distrikt.
  - Es erfüllt plausibel die zwei Bedingungen für den omitted variable bias: $Z = PctEL$ ist:
    1. eine Determinante von $Y = TestScore$ und
    2. korreliert mit dem Regressor $X = STR$.
  - Kontroll- und Behandlungsgruppen (d.h. große und kleine Klassen) unterscheiden sich systematisch – $\text{Corr}(STR, PctEL) \neq 0$.

Exkurs: Kausalitäts- und Regressionsanalyse
Beispiel

Beispiel 5.2: Fortsetzung.

- Randomisierte kontrollierte Experimente:
  - Randomisierung + Kontrollgruppen bedeutet, dass alle Unterschiede zwischen der Behandlungs- und der Kontrollgruppe zufällig sind, also keinen systematischen Einfluss auf die Behandlung haben.
- **Problem:** Distrikte mit großen (Kontrolle) und kleinen (Behandlung) Klassen unterscheiden sich systematisch in *PctEL*.
- **Mögliche Lösung:** Untersuchung des Effekts der Klassengröße von Distrikten mit *gleicher PctEL*.
  - Sofern der einzige systematische Unterschied von Distrikten mit großen und kleinen Klassengrößen der Anteil *PctEL* ist, dann sind wir zurück im randomisierten kontrollierten Experiment – innerhalb von jeder *PctEL* Gruppe.
  - Dies ist ein Weg für den Effekt von *PctEL* zu "kontrollieren" um den Effekt von *STR* zu schätzen.

Zurück zum omitted variable bias

Drei Wege um den omitted variable bias zu umgehen:

1. Führe randomisiertes kontrolliertes Experiment durch, in dem Behandlung (STR) zufällig zugeordnet: Dann ist $PctEL$ noch immer eine Determinante von $TestScore$, aber $PctEL$ ist unkorreliert mit STR.
   - In der Praxis unrealistisch.
2. Wende die obige "Aufgliederung" an, mit feinen Abstufungen der STR und $PctEL$ – innerhalb jeder Gruppe haben alle Klassen die gleiche $PctEL$, also kontrollieren wir für $PctEL$.
   - Allerdings werden wir schnell nicht genug Daten haben; und was ist mit anderen Daten, wie Einkommen und Ausbildung der Eltern?
3. Führe $PctEL$ als weiteren Regressor in eine multiple Regression ein.
   - Dann ist $PctEL$ nicht mehr in u, und u dann (hoffentlich) unkorreliert mit Regressoren.

Überblick

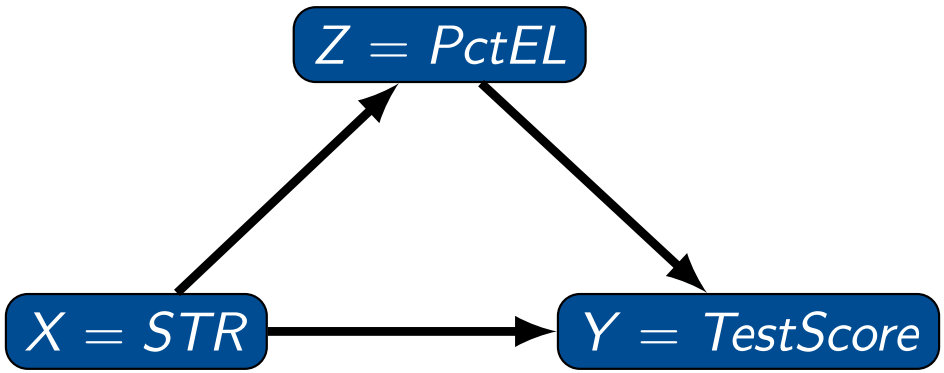
- Omitted variable bias
- **Multiples Regressionsmodell**
- Anpassungsmaße für multiple Regressionen
- KQ-Annahmen für die multiple Regression
- Stichprobenverteilung des KQ-Schätzers

Modell und Notation

- Um Verzerrungen durch eine ausgelassene Variable zu umgehen, fügen wir die ausgelassene Variable der Regression hinzu.
- Im Schuldaten-Beispiel betrachten wir also

$$TestScore = \beta_0 + \beta_1 STR + \beta_2 PctEL + u.$$

- Bildlich: Wir schütten den Transmissionskanal ($STR \to PctEL$) zu.



Modell und Notation

- I.A. ist es unrealistisch anzunehmen, dass es nur **eine** ausgelassene Variable gibt.
- Daher behandeln wir jetzt das **multiple Regressionsmodell** mit k Regressoren:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i, \qquad i = 1, \ldots, n.$$

Vorteile

Das multiple Regressionsmodell

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + u$$

erlaubt uns

1. OVB zu eliminieren/mildern,
2. die Effekte mehrerer Variablen gleichzeitig zu untersuchen,
3. Trade-offs zwischen den Variable zu untersuchen,
4. die Vorhersagekraft mehrerer Variablen gleichzeitig zu benutzen.

Interpretation von β_1

Multiples Regressionsmodell:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + u.$$

Interpretationen von β_1:

- **Kausal:** Wenn wir X_1 um eine Einheit erhöhen und alles andere gleichhalten, ändert sich Y um β_1.
- **Deskriptiv:** Untersuchungsobjekte mit demselben $X_2, \ldots, X_k$, aber eine Einheit höherem X_1, haben im Schnitt ein um β_1 höheres Y.
- **Vorhersage:** Wenn ich ein Untersuchungsobjekt mit $(X_1 + 1, X_2, \ldots, X_k)$ sehe, prognostiziere ich ein um β_1 verschiedenes Y verglichen mit einem Objekt mit $(X_1, X_2, \ldots, X_k)$.

KQ-Schätzung

- Bei k Regressoren löst der KQ-Schätzer

$$\min_{b_0,\ldots,b_k} \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_{1i} + \ldots + b_k X_{ki})]^2.$$

- Der KQ-Schätzer minimiert die durchschnittlichen quadrierten Differenzen zwischen den tatsächlichen Werten von Y_i und der Prognose (angepasste Werte) $\widehat{Y}_i = b_0 + b_1 X_{1i} + \ldots + b_k X_{ki}$ gemäß der geschätzten Gerade.
- Dieses Minimierungsproblem kann man analytisch lösen.
- Das liefert den KQ-Schätzer von $\beta_0, \ldots, \beta_k$.
- Hierfür ist Matrixalgebra nötig, wobei für uns die genauen Formeln nicht so wichtig sind.

Beispiel

Kalifornische Schulbezirke

```
cor(CASchools$STR, CASchools$english)
```

```
## [1] 0.1876
```

```
mult.mod <- lm(score ~ STR + english, data = CASchools)
mult.mod
```

```
##
## Call:
## lm(formula = score ~ STR + english, data = CASchools)
##
## Coefficients:
## (Intercept)          STR      english
##      686.03        -1.10        -0.65
```

Beispiel
Kalifornische Schulbezirke

Beispiel 5.3:

Betrachte die beiden geschätzten Regressionsgeraden:

$$(1) \qquad \widehat{TestScore} = 698.9 - 2.28 \cdot STR,$$

$$(2) \qquad \widehat{TestScore} = 686.0 - 1.10 \cdot STR - 0.65 \cdot PctEL.$$

- Der Koeffizient von *STR* in (2) ist der Effekt auf *TestScore* einer Erhöhung um 1 von *STR*, wenn *PctEL* konstant gehalten wird.
- Wie vermutet hat die positive Korrelation von *PctEL* mit *STR* ($\text{Corr}(STR, PctEL) = 0.19$) den KQ-Schätzer für β_1 in (1) nach unten verzerrt – der Koeffizient von *STR* fällt um die Hälfte!
- Warum? $\widehat{\beta}_1$ hat zusätzlich zu dem negativen Effekt von ($STR \rightarrow TestScore$) auch noch den negativen Effekt von ($PctEL \rightarrow TestScore$) eingefangen.

Überblick

- Omitted variable bias
- Multiples Regressionsmodell
- **Anpassungsmaße für multiple Regressionen**
- KQ-Annahmen für die multiple Regression
- Stichprobenverteilung des KQ-Schätzers

Parameterschätzung ist erst der Anfang

- Wieder wollen wir wissen, wie gut unser Modell die Daten erklärt.
  - Relevant für deskriptive Analyse und Prognosen
  - Weniger relevant für kausale Analyse
- Dafür benutzen wir (alt) R^2, *SER*, *RMSE* und (neu) das **adjustierte** R^2, $\overline{R}^2$.

Maße der Anpassungsgüte
Bausteine

- Definiere die **geschätzten Werte** und **Residuen** als

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_{1i} + \ldots + \widehat{\beta}_k X_{ki},$$
$$\widehat{u}_i = Y_i - \widehat{Y}_i.$$

- Damit haben die folgenden Größen dieselbe Bedeutung wie im einfachen linearen Regressionsmodell:

$$TSS = \sum_{i=1}^{n} (Y_i - \overline{Y}_n)^2,$$
$$ESS = \sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y}_n)^2,$$
$$RSS = \sum_{i=1}^{n} \widehat{u}_i^2.$$

Maße der Anpassungsgüte
SER und *RMSE*

- Der SER und der RMSE messen die Spannweite der Verteilung von u:

$$SER = \sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n}\widehat{u}_i^2},$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i^2}.$$

Maße der Anpassungsgüte

R^2 und $\overline{R}^2$

- Der Anteil an Variation in Y_i, der durch die $X_1, \ldots, X_k$ erklärt wird, ist

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y}_n)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2}.$$

- R^2 wächst immer, wenn man einen Regressor hinzufügt (Warum?), ...
- ... auch wenn der Regressor nicht relevant ist!
- Also: Das Hinzufügen von "Unsinns"-Regressoren führt zu besserer Anpassung, gemessen in R^2.
- Das ist schlecht für ein "Gütemaß"... (`R2noise.R`)

Maße der Anpassungsgüte

R^2 und $\overline{R}^2$

- Das **adjustierte** R^2, $\overline{R}^2$, korrigiert dieses Problem, indem es das Hinzufügen weiterer Regressoren bestraft:

$$\overline{R}^2 = 1 - \left(\frac{n-1}{n-k-1}\right)\frac{SSR}{TSS}.$$

- Beachte:

  1. $\overline{R}^2$ wächst nicht notwendigerweise, wenn weiterer Regressor hinzugefügt wird.
  2. $\overline{R}^2 < R^2 = 1 - SSR/TSS$.

Maße der Anpassungsgüte

R^2 und $\overline{R}^2$

Wenn wir einfach R^2 maximieren, verlieren wir unser eigentliches Ziel aus dem Auge: ein guter Schätzer des Effekts von (z.B.) der Klassengröße.

- Ein hohes R^2 (oder $\overline{R}^2$) **bedeutet**, dass die Regressoren die Variation in Y erklären.
- Ein hohes R^2 (oder $\overline{R}^2$) **bedeutet nicht**, dass keine Verzerrung durch ausgelassene Variablen vorliegt.
- Ein hohes R^2 (oder $\overline{R}^2$) **bedeutet nicht**, dass der kausale Effekt β_1 unverzerrt geschätzt wird.
- Ein hohes R^2 (oder $\overline{R}^2$) **bedeutet nicht**, dass die benutzten erklärenden Variablen statistisch signifikant sind – prüfe dies mit Hypothesentests.

Beispiel
Kalifornische Schulbezirke

Beispiel 5.4:

$$\widehat{TestScore} = 698.9 - 2.28 \cdot STR,$$
$$R^2 = .05, \quad SER = 18.6.$$
$$\widehat{TestScore} = 686.0 - 1.10 \cdot STR - 0.65 \cdot PctEL,$$
$$R^2 = 0.426, \quad \overline{R}^2 = 0.424, \quad SER = 14.5.$$

1. Bessere Anpassung: Die "lange" Regression erklärt mehr als 2/5 der Variation in *TestScore*!
2. Bessere Vorhersagen: *SER* ist niedriger in der langen Regression!
3. Da n groß und $k = 2$, unterscheiden sich R^2 und $\overline{R}^2$ kaum.

Überblick

- Omitted variable bias
- Multiples Regressionsmodell
- Anpassungsmaße für multiple Regressionen
- **KQ-Annahmen für die multiple Regression**
- Stichprobenverteilung des KQ-Schätzers

Etwas präziser

Das allgemeine Modell:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i, \quad i = 1, \ldots, n.$$

KQ-Annahmen im multiplen Regressionsmodell

1. Die bedingte Verteilung von u gegeben die X hat einen Erwartungswert von Null, d.h. $\mathrm{E}(u|X_1 = x_1, \ldots, X_k = x_k) = 0$.
2. $(X_{1i}, \ldots, X_{ki}, Y_i)$, $i = 1, \ldots, n$, sind u.i.v.
3. Ausreißer sind selten: $X_1, \ldots, X_k$, und Y besitzen vierte Momente: $\mathrm{E}(X_{1i}^4) < \infty, \ldots, \mathrm{E}(X_{ki}^4) < \infty, \mathrm{E}(Y_i^4) < \infty$.
4. Keine **perfekte Multikollinearität**.

Annahme Nr. 1

Der bedingte Erwartungswert von u gegeben die X der Regression ist Null

$$\mathrm{E}(u|X_1 = x_1, \ldots, X_k = x_k) = 0$$

- Gleiche Interpretation wie bei einem einzelnen Regressor.
- Wenn eine ausgelassene Variable (1) in die Regression gehört und (2) korreliert ist mit einem X der Regression, ist die Annahme nicht erfüllt.
- Wenn sie nicht erfüllt ist, resultiert Verzerrung durch ausgelassene Variablen.
- Die Lösung – *wenn möglich* – ist es, die ausgelassene Variable zur Regression hinzuzufügen.

Annahmen Nr. 2 und 3

$(X_{1i}, \ldots, X_{ki}, Y_i)$**,** $i = 1, \ldots, n$**, sind u.i.v.**

- Diese ist automatisch erfüllt, wenn die Daten als einfache Zufallsstichprobe erhoben werden.

Ausreißer sind selten (endliche vierte Momente)

- Dieselbe Annahme wie bei einem einzelnen Regressor.
- Wie vorher kann KQ stark auf Ausreißer reagieren, also ...
- ... müssen Sie Ihre Daten (Punktediagramme!) sorgfältig darauf untersuchen, dass es keine komischen Werte gibt (z.B. Eingabefehler).

Annahme Nr. 4

Keine perfekte Multikollinearität

- **Perfekte Multikollinearität** liegt vor, wenn einer der Regressoren X_j eine exakte lineare Funktion der anderen Regressoren X_k $(k \neq j)$ ist, d.h.

$$X_j = \alpha_0 + \sum_{k \neq j} \alpha_k X_k.$$

- Das ist neu!

Multikollinearität
Beispiele

Beispiel 5.5: Kalifornische Schulbezirke.

1. *STR* zweimal: Intuitiv, wenn $X_1 = STR = X_2$, was ist der Effekt von X_1, gegeben konstantes X_2?
2. Regressiere *TestScore* auf eine Konstante, auf D, und auf B, wobei:
   - $D_i = 1$ wenn $STR < 20$, $= 0$ sonst;
   - $B_i = 1$ wenn $STR \geq 20$, $= 0$ sonst, also $B_i = 1 - D_i$ – perfekte Multikollinearität.
   - Beachte, dass die Konstante – die Variable mit Koeffizient β_0 – auch ein Regressor ist. Der Wert ist 1 für jede Beobachtung!
   - Dieses Beispiel ist ein Spezialfall von...

Die Dummyvariablen-Falle

- Es liegen mehrere binäre (Dummy-)Variablen vor, die sich gegenseitig ausschließen und ausschöpfend sind, d.h. es gibt multiple Kategorien und jede Beobachtung fällt in genau eine Kategorie.
  - z.B. Männer und Frauen.
- Wenn man all diese Dummyvariablen und eine Konstante einfügt, liegt perfekte Multikollinearität vor – auch bekannt als **Dummyvariablen-Falle**.
- Frage: Warum liegt hier perfekte Multikollinearität vor?
- Auswege aus der Dummyvariablen-Falle:
  1. Lasse eine der Gruppen weg (z.B. Frauen), oder
  2. lasse die Konstante weg.
- Was implizieren 1. oder 2. für die Interpretation der Koeffizienten?

Imperfekte Multikollinearität
Definition

- Imperfekte und perfekte Multikollinearität sind trotz der ähnlichen Bezeichnungen recht verschiedenartige Probleme.
- Warum diese Bezeichnung?
- **Imperfekte Multikollinearität** tritt auf, wenn zwei oder mehr Regressoren sehr hoch korreliert sind.
- Wenn zwei Regressoren hoch korreliert sind, sieht ihr Punktediagramm einer geraden Line ähnlich, jedoch ist diese Kollinearität imperfekt, außer wenn die Korrelation genau ± 1 ist.
  - Siehe hierzu `NearMulticollinearity.R`.

Imperfekte Multikollinearität
Konsequenzen

- Imperfekte Multikollinearität führt dazu, dass ein oder mehrere Regressionskoeffizienten ungenau geschätzt werden.
- Der Koeffizient von X_1 ist der Effekt von X_1, wenn man X_2 konstant hält.
- Wenn nun X_1 und X_2 aber stark korrelieren, gibt es sehr wenig Variation in X_1, wenn X_2 konstant gehalten wird.
- Daher wird die Varianz des KQ-Schätzers des Koeffizienten von X_1 groß sein!
- Imperfekte Multikollinearität führt also typischerweise zu großen Standardfehlern für einen oder mehrere der Koeffizienten.

Überblick

- Omitted variable bias
- Multiples Regressionsmodell
- Anpassungsmaße für multiple Regressionen
- KQ-Annahmen für die multiple Regression
- **Stichprobenverteilung des KQ-Schätzers**

Asymptotik...

- Unter den vier KQ-Annahmen gilt:
- $\mathsf{E}(\widehat{\beta}_1) = \beta_1$, $\mathsf{Var}(\widehat{\beta}_1) \propto 1/n$.
- Bis auf Erwartungswert und Varianz ist die exakte (endliches-n) Verteilung von $\widehat{\beta}_1$ sehr kompliziert; aber für großes n ...

  1. ... ist $\widehat{\beta}_1$ konsistent: $\widehat{\beta}_1 \xrightarrow{p} \beta_1$ (Gesetz der großen Zahl).
  2. $\frac{\widehat{\beta}_1 - \mathsf{E}(\widehat{\beta}_1)}{\sqrt{\mathsf{Var}(\widehat{\beta}_1)}}$ ist approximativ $N(0,1)$-verteilt (Zentraler Grenzwertsatz).

- Ebenso für $\widehat{\beta}_2, \ldots, \widehat{\beta}_k$.
- Konzeptionell nichts neues...
- ...abgesehen davon, dass $(\widehat{\beta}_1, \ldots, \widehat{\beta}_k)'$ eine gemeinsame Verteilung, und daher u.a. eine Kovarianzmatrix haben...

Kompakte Darstellung

- Schreibe das multiple lineare Regressionsmodell

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_k X_{k,i} + u_i \\ &= \boldsymbol{x}_i' \boldsymbol{\beta} + u_i, \end{aligned}$$

  wobei $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)'$ und $\boldsymbol{x}_i' = (1, X_{1,i}, \ldots, X_{k,i})$.

- Stapeln der $i = 1, \ldots, n$ Gleichungen liefert

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

  wobei $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$, $\boldsymbol{u} = (u_1, \ldots, u_n)'$ und

$$\boldsymbol{X} = \begin{pmatrix} 1 & X_{11} & \ldots & X_{k1} \\ \vdots & & & \vdots \\ 1 & X_{1n} & \ldots & X_{kn} \end{pmatrix}.$$

Kompakte Darstellung

- Ausgehend von $\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{u}$ ist die KQ-Prozedur dann

$$\widehat{\beta} = \arg\min_{\boldsymbol{b}} \left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\right)' \left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\right).$$

- Die Lösung ist

$$\begin{aligned}
\widehat{\beta} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{Y} \\
&= \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{x}_i Y_i\right) \\
&= \beta + \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{x}_i u_i\right),
\end{aligned}$$

wobei wir im letzten Schritt $Y_i = \boldsymbol{x}_i'\beta + u_i$ eingesetzt haben.

Kompakte Darstellung

- Wie gesehen:

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{x}_i u_i \right).$$

- Daher implizieren das **multivariate** Gesetz der großen Zahlen und ZGWS unter den KQ-Annahmen 1–4, dass

$$\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta},$$
$$\widehat{\boldsymbol{\beta}} \overset{approx}{\sim} N_{k+1}\left(\boldsymbol{\beta}, \text{Cov}\left(\widehat{\boldsymbol{\beta}}\right)\right).$$

Überblick

- Omitted variable bias
- Multiples Regressionsmodell
- Anpassungsmaße für multiple Regressionen
- KQ-Annahmen für die multiple Regression
- Stichprobenverteilung des KQ-Schätzers
- **Variablen hinzunehmen oder nicht?**

Irrelevanter Regressor

- Unter den KQ-Annahmen 1–4, betrachte die "lange" Regression

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + u.$$

- Nehme zusätzlich (wie beim GM Theorem) Homoskedastie an: $\text{Var}(u \mid X_1, \ldots, X_k) = \sigma^2$.
- Was, wenn X_k irrelevant ist, d.h. $\beta_k = 0$?
- Dann sollten wir die "kurze" Regression betrachten:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{k-1} X_{k-1} + u.$$

- Für die kurze Regression sind die Annahmen des (multiplen) GM Theorem's erfüllt, und KQ damit BLU.
- Auch unter $\beta_k = 0$ ist KQ der langen Regression unverzerrt (wegen KQA 1–4).
- Daher: KQ in der langen Regression hat größere Varianz.

Ausgelassener Regressor

- Wenn wir nicht wissen, dass $\beta_k = 0$, riskieren wir eine relevante Variable in der kurzen Regression auszulassen:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{k-1} X_{k-1} + u.$$

- Wenn $\beta_k \neq 0$, führt Auslassen von X_k zu Verzerrungen **aller** Schätzer $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_{k-1}$.

Variablen hinzunehmen oder nicht?

- Schwierige Frage: Keine allgemeingültige Antwort.
- Daumenregel:
  - Für große n: Im Zweifel hinzunehmen; es sei denn, es gibt sehr gute Gründe es nicht zu tun.
  - Für kleine n: Wähle die Variablen sehr sorgfältig.

Ein Grund nicht hinzuzunehmen
Beispiel

- Die Variable resultiert direkt aus einer vorhandenen erklärenden Variablen im Modell.

Beispiel 5.6: Lipitor.

- Beispiel: Wir wollen den Effekt von Lipitor (X) auf die Herzinfarktrate (Y) untersuchen.
- Lipitor wirkt hauptsächlich durch Senkung des Cholesterinspiegels (Z), welche das Risiko eines Herzinfarktes reduziert.

Ein Grund nicht hinzuzunehmen
Beispiel

Beispiel 5.6: Fortsetzung.

- Betrachte die Regression:

$$Herzinfarkt = \beta_0 + \beta_1 \cdot Lipitor + \beta_2 \cdot Cholesterol + u.$$

- Was sollte β_1 hier sein?
- β_1 sollte sehr klein sein, selbst wenn Lipitor effektiv ist, weil Lipitor hauptsächlich durch Senkung des Cholesterinspiegels wirkt.
- Daher: Wenn wir die Gesamteffektivität von Lipitor messen wollen, sollten wir nicht gleichzeitig den Effekt des Cholesterinspiegels herausrechnen.
- Ein weiteres Beispiel ist:

$$Gehalt = \beta_0 + \beta_1 \cdot Sport + \beta_2 \cdot Gesundheit + u.$$

Überblick

Hypothesentests und Konfidenzintervalle in multiplen Regressionen

(SW Kapitel 7)

Überblick

- **Tests und Konfidenzintervalle für einzelne Koeffizienten**
- Testen gemeinsamer Hypothesen
- Präsentation von Regressionsergebnissen

Implikationen der multivariaten Normalverteilung

- Die gemeinsame Verteilung der KQ-Schätzer ist asymptotisch multivariat normal, so dass näherungsweise

$$\frac{\widehat{\beta}_1 - \mathsf{E}(\widehat{\beta}_1)}{\sqrt{\mathsf{Var}(\widehat{\beta}_1)}} \sim N(0,1).$$

- Also können Hypothesen über β_1 mit der üblichen t-Statistik getestet werden, und 95%-Konfidenzintervalle werden konstruiert als $\{\widehat{\beta}_1 \pm 1.96 \cdot \mathsf{SE}(\widehat{\beta}_1)\}$.
- Ebenso für $\beta_2, \ldots, \beta_k$ (und β_0, wenn es interessant wäre).

Beispiel

Beispiel 6.1: Kalifornische Schulbezirke.

$$(1) \quad \widehat{TestScore} = \underset{(10.4)}{698.9} - \underset{(0.52)}{2.28} \cdot STR$$

$$(2) \quad \widehat{TestScore} = \underset{(8.7)}{686.0} - \underset{(0.43)}{1.10} \cdot STR - \underset{(0.031)}{0.650} \cdot PctEL$$

- Das 95%-Konfidenzintervall für den Koeffizienten von STR in (2) ist $\{-1.10 \pm 1.96 \cdot 0.43\} = (-1.95, -0.26)$.
- Die t-Statistik für $\beta_{STR} = 0$ ist $t^{act} = -1.10/0.43 = -2.54$, so dass wir die Hypothese zum 5%-Signifikanzniveau verwerfen.

Homoskedastie vs. Heteroskedastie

Benutze Heteroskedastie-robuste Standardfehler

(aus genau den gleichen Gründen wie bei einem einzelnen Regressor.)

Beispiel

Kalifornische Schulbezirke

```
coeftest(mult.mod, vcov. = vcovHC, type = "HC1")
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  686.032      8.728   78.60   <2e-16 ***
## STR           -1.101      0.433   -2.54    0.011 *
## english       -0.650      0.031  -20.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Überblick

- Tests und Konfidenzintervalle für einzelne Koeffizienten
- **Testen gemeinsamer Hypothesen**
- Präsentation von Regressionsergebnissen

Untersuchung mehrerer Parameter

Beispiel 6.2: Kalifornische Schulbezirke.

- Prüfen wir, ob die Ressourcen von Schulen (wovon die Klassengröße nur ein Aspekt ist) wichtig sind oder nicht.
- Sei $Expn$ = Ausgaben pro Schüler, und betrachte das Regressionsmodell:

$$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i.$$

- Die Nullhypothese, dass "Ressourcen unwichtig sind", und die Alternative, dass sie es sind, entspricht:

$$H_0 : \beta_1 = 0 \text{ und } \beta_2 = 0$$

vs.

$$H_1 : \beta_1 \neq 0 \text{ oder } \beta_2 \neq 0.$$

Untersuchung mehrerer Parameter

- Eine **gemeinsame Hypothese** spezifiziert einen Wert für zwei oder mehr Koeffizienten, d.h. sie legt für zwei oder mehr Koeffizienten eine Restriktion auf.
- Im Allgemeinen kann eine gemeinsame Hypothese q Restriktionen beinhalten.
  - Im obigen Beispiel ist $q = 2$: $\beta_1 = 0$ und $\beta_2 = 0$!
- Die *F*-**Statistik** testet alle Teile einer gemeinsamen Hypothese auf einmal.
- Die Formel für den Spezialfall der gemeinsamen Hypothese $\beta_1 = \beta_{1,0}$ und $\beta_2 = \beta_{2,0}$ in einer Regression mit zwei Regressoren:

$$F = \frac{1}{2}\left[\frac{t_1^2 + t_2^2 - 2\widehat{\rho}_{t_1,t_2}\, t_1\, t_2}{1 - \widehat{\rho}_{t_1,t_2}^2}\right],$$

  wobei $\widehat{\rho}_{t_1,t_2}$ die Korrelation zwischen t_1 und t_2 schätzt.
- Verwirf, wenn F groß ist.
- Wie groß? ...

Die F-Statistik

- Die F-Statistik für β_1 und β_2:

$$F = \frac{1}{2}\left[\frac{t_1^2 + t_2^2 - 2\widehat{\rho}_{t_1,t_2}\, t_1 t_2}{1 - \widehat{\rho}^2_{t_1,t_2}}\right]. \tag{6.1}$$

- Die F-Statistik ist groß, wenn t_1 und/oder t_2 groß sind.
- Die F-Statistik korrigiert (auf genau die richtige Weise) für die Korrelation zwischen t_1 und t_2.
- Die Formel für mehr als zwei β's ist ohne Matrixalgebra sehr hässlich.
- Dies führt zur Verteilung der F-Statistik in großen Stichproben...

Verteilung der F-Statistik in großen Stichproben

- Betrachte den Spezialfall, dass t_1 und t_2 unabhängig sind, so dass $\widehat{\rho}_{t_1,t_2} \xrightarrow{p} 0$; in großen Stichproben wird die Formel

$$F = \frac{1}{2}\left[\frac{t_1^2 + t_2^2 - 2\widehat{\rho}_{t_1,t_2} t_1 t_2}{1 - \widehat{\rho}_{t_1,t_2}^2}\right] \cong \frac{1}{2}(t_1^2 + t_2^2).$$

- Unter der Nullhypothese sind t_1 und t_2 standardnormalverteilt, und sind, in diesem Spezialfall, unabhängig.
- Die Verteilung der F-Statistik in großen Stichproben ist die Verteilung des Durchschnitts zweier unabhängiger quadrierter standardnormalverteilter Zufallsvariablen.

Verteilung der F-Statistik in großen Stichproben

- Die χ^2_q-**Verteilung** mit q Freiheitsgraden ist definiert als die Verteilung der Summe von q unabhängigen quadrierten standardnormalverteilten Zufallsvariablen.
- In großen Stichproben: $F \overset{approx}{\sim} \chi^2_q/q$.
- Ausgewählte Quantile (kritische Werte) von χ^2_q/q:

q	5% critical value	
1	3.84	(warum?)
2	3.00	(der Fall $q = 2$ von oben)
3	2.60	
4	2.37	
5	2.21	

F-Statistik bei Homoskedastie

- **Wenn** die Fehler homoskedastisch sind, gibt es einen einfachen Weg, die homoskedastische Version der F-Statistik auszurechnen:

1. Führe zwei Regressionen durch:
   (a) Eine unter der Nullhypothese (die "restringierte" Regression) und
   (b) eine unter die Alternativhypothese (die "unrestringierte" Regression).
2. Vergleiche die Anpassung der Regressionen mit R^2: Wenn das "unrestringierte" Modell eine hinreichend bessere Anpassung hat, verwirf die Nullhypothese.

F-Statistik bei Homoskedastie

Die "restringierte" und "unrestringierte" Regression

Beispiel 6.2: Fortsetzung.

- Sind die Koeffizienten von *STR* und *Expn* Null?
- **Unrestringierte Regression (unter H_1)**:

$$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$$

- **Restringierte Regression (unter H_0)**:

$$TestScore_i = \beta_0 + \beta_3 PctEL_i + u_i \quad \text{(warum?)}$$

- Anzahl der Restriktionen unter H_0: $q = 2$ (warum?).
- Klar, dass $R^2_{restringiert} < R^2_{unrestringiert}$ (warum?).
- Um wie viel muss $R^2_{unrestringiert}$ größer sein als $R^2_{restringiert}$, um die Koeffizienten von *STR* und *Expn* als statistisch signifikant zu bezeichnen?

F-Statistik bei Homoskedastie

Einfache Formel

- Nur bei Homoskedastie gültige Formel:

$$F = \frac{(R^2_{unrestringiert} - R^2_{restringiert})/q}{(1 - R^2_{unrestringiert})/(n - k_{unrestringiert} - 1)},$$

  wobei:

  - $R^2_{restringiert}$ = das R^2 der restringierten Regression,
  - $R^2_{unrestringiert}$ = das R^2 der unrestringierten Regression,
  - q = die Anzahl der Restriktionen unter der Nullhypothese,
  - $k_{unrestringiert}$ = die Anzahl der Regressoren in der unrestringierten Regression.

- Je größer die Differenz zwischen dem restringierten und unrestringierten R^2, desto größer die Verbesserung der Anpassung durch Hinzufügen der fraglichen Variablen, desto größer F.

F-Statistik bei Homoskedastie

Beispiel 6.2: Fortsetzung.

$$(1) \quad \widehat{TestScore} = \underset{(1.0)}{644.7} - \underset{(0.032)}{0.671} \cdot PctEL, \quad R^2_{restringiert} = 0.4149$$

$$(2) \quad \widehat{TestScore} = \underset{(15.5)}{649.6} - \underset{(0.48)}{0.29} \cdot STR + \underset{(1.59)}{3.87} \cdot Expn - \underset{(0.032)}{0.656} \cdot PctEL,$$

$$R^2_{unrestringiert} = 0.4366,\ k_{unrestringiert} = 3,\ q = 2$$

- Daher:

$$F = \frac{(R^2_{unrestringiert} - R^2_{restringiert})/q}{(1 - R^2_{unrestringiert})/(n - k_{unrestringiert} - 1)}$$
$$= \frac{(0.4366 - 0.4149)/2}{(1 - 0.4366)/(420 - 3 - 1)} = 8.01$$

- Die Heteroskedastie-robuste Statistik aus (6.1) ist hier $F = 5.43$.

Beispiel I

Kalifornische Schulbezirke

```
# estimate the multiple regression model
model <- lm(score ~ STR + english + expenditure, data = CASchools)

# homoskedasticity-only F-test
linearHypothesis(model, c("STR=0", "expenditure=0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## STR = 0
## expenditure = 0
##
## Model 1: restricted model
## Model 2: score ~ STR + english + expenditure
##
##   Res.Df   RSS Df Sum of Sq    F  Pr(>F)
## 1    418 89000
## 2    416 85700  2      3300 8.01 0.00039 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Beispiel II

Kalifornische Schulbezirke

```
# heteroskedasticity-robust F-test
linearHypothesis(model, c("STR=0", "expenditure=0"), white.adjust = "hc1")
```

```
## Linear hypothesis test
##
## Hypothesis:
## STR = 0
## expenditure = 0
##
## Model 1: restricted model
## Model 2: score ~ STR + english + expenditure
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df    F Pr(>F)
## 1    418
## 2    416  2 5.43 0.0047 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F-Statistik bei Homoskedastie

Zusammenfassung

$$F = \frac{(R^2_{unrestringiert} - R^2_{restringiert})/q}{(1 - R^2_{unrestringiert})/(n - k_{unrestringiert} - 1)}$$

- Die F-Statistik bei Homoskedastie verwirft, wenn Hinzufügen der Variablen das R^2 "genügend" erhöht – d.h. wenn Hinzufügen der Variablen die Anpassung der Regression "genug" verbessert.
- Wenn die Fehler homoskedastisch: $F \sim \chi^2_q/q$ für große n.
- Wenn die Fehler jedoch heteroskedastisch sind: Verteilung von F kompliziert und nicht χ^2_q/q.

Zusammenfassung

- Die Heteroskedastie-robuste F-Statistik ist in den meisten Software-Paketen implementiert – man muss sie aber in der Regel aktiv auswählen.
- Für großes n: $F \sim \chi^2_q/q$.
- Die F-Statistik bei Homoskedastie ist historisch wichtig (und daher in der Praxis noch oft genutzt) und intuitiv einleuchtend, aber nicht valide bei Heteroskedastie.

Überblick

- Tests und Konfidenzintervalle für einzelne Koeffizienten
- Testen gemeinsamer Hypothesen
- **Präsentation von Regressionsergebnissen**

Präsentation von Regressionsergebnissen

- Wir haben verschiedene Regressionen durchgeführt und wollen davon berichten.
- Es ist umständlich und schwierig, viele in Gleichungsform ausgeschriebene Regressionen zu lesen.
- Es ist daher üblich, sie in einer Tabelle mit Folgendem abzubilden:
  - geschätzte Regressionskoeffizienten
  - Standardfehler
  - Gütemaße
  - die Anzahl von Beobachtungen
  - ggf. relevante F-Statistiken
  - Andere wichtige Informationen

Präsentation von Regressionsergebnissen

Regressor	(1)	(2)	(3)	(4)	(5)
STR (X_1)	-2.28^{**} (0.52)	-1.10^{*} (0.43)	-1.00^{**} (0.27)	-1.31^{**} (0.34)	-1.01^{**} (0.27)
PctEL (X_2)		-0.650^{**} (0.031)	-0.122^{**} (0.033)	-0.488^{**} (0.030)	-0.130^{**} (0.036)
Anteil mit subventioniertem Essen (X_3)			-0.547^{**} (0.024)		-0.529^{**} (0.038)
Anteil auf Sozialhilfe (X_4)				-0.790^{**} (0.068)	0.048 (0.059)
Achsenabschnitt	698.9^{**} (10.4)	686.0^{**} (8.7)	700.2^{**} (5.6)	698.0^{**} (6.9)	700.4^{**} (5.5)
Sonstige Statistiken					
SER	18.58	14.46	9.08	11.65	9.08
$\bar{R}^2$	0.049	0.424	0.773	0.626	0.773
n	420	420	420	420	420

Tabelle 2: Ergebnisse einer Regression von Test Scores auf STR und Kontrollvariablen für kalifornische Schulbezirke. Signifikanz zum Niveau 5%/1% gekennzeichnet durch */**.

Zusammenfassung
Multiple Regression

- Multiple Regressionen erlauben es, den Effekt einer Änderung von X_1 auf Y zu schätzen, wobei $X_2, \ldots, X_k$ konstant gehalten werden – damit kommen wir einer kausalen Interpretation des Effektes von X_1 auf Y näher.
- Aber kommen wir ihr auch nah? Darüber müssen Sie weiterhin gründlich nachdenken!
- Wenn eine Variable erhoben werden kann, kann Verzerrung durch diese ausgelassene Variable durch Hinzufügen zur Regression vermieden werden.
- Es gibt kein einfaches Rezept, um zu entscheiden, welche Variablen in die Regression gehören – Sie müssen hier Ihre Urteilskraft benutzen.
- Und seien Sie sich anderer möglicher Probleme beim Interpretieren der Ergebnisse bewusst...

Überblick

Nichtlineare Regressionsfunktionen

(SW Kapitel 8)

Überblick

- **Nichtlineare Modelle**
- Aufspüren von Nichtlinearität

Linear?

R Beispiel

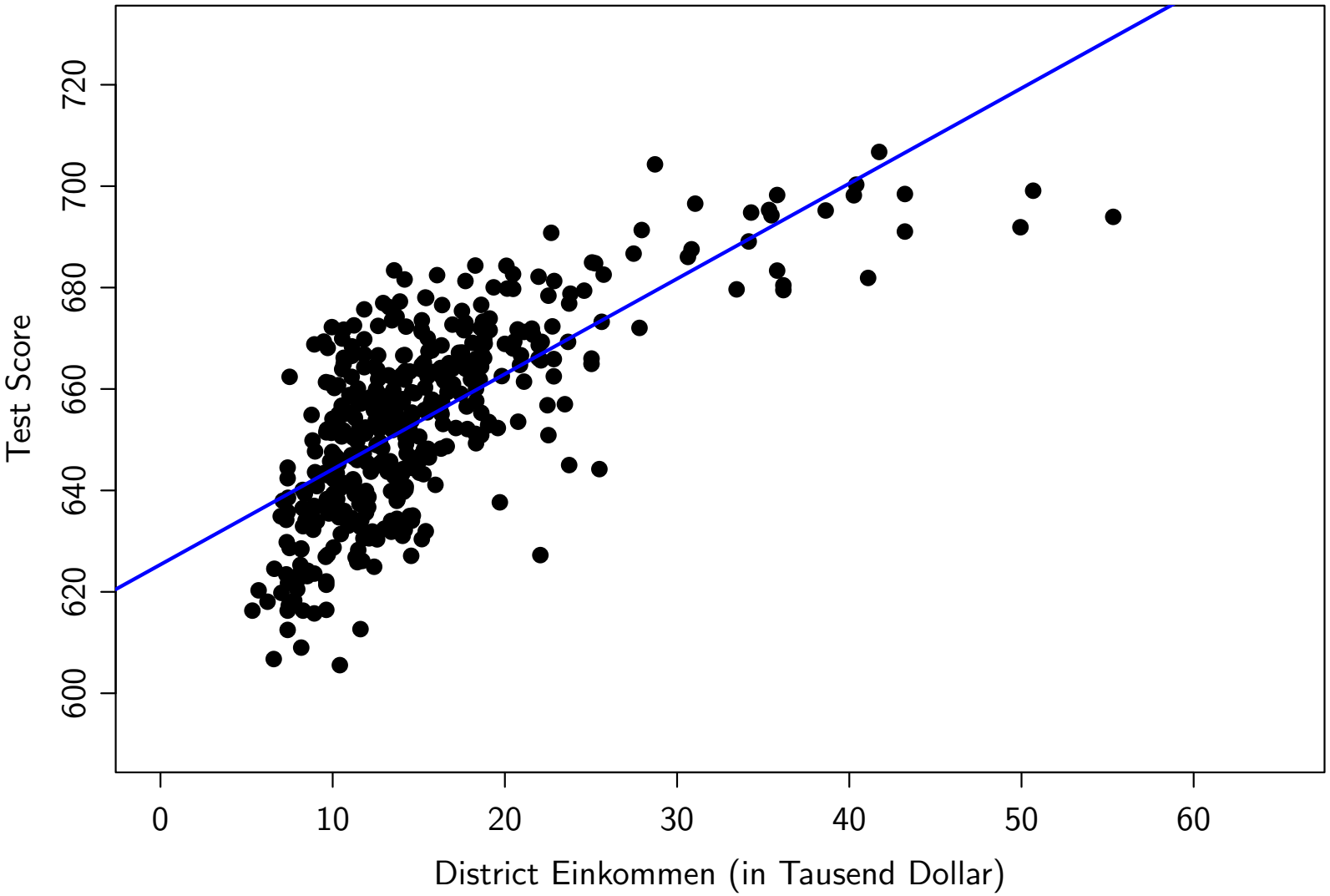


Abbildung 12: Punktediagramm von TestScore und Einkommen.

Nichtlineare Modelle

- Eine Verallgemeinerung des linearen Modells ist das nichtlineare Modell mit additivem Fehler:

$$Y_i = f(X_{i1}, \ldots, X_{ik}) + u_i, \qquad i = 1, \ldots, n.$$

- Hier:

$$\begin{aligned} \Delta Y &= f(X_1, \ldots, X_j + \Delta X_j, \ldots, X_k) - f(X_1, \ldots, X_j, \ldots, X_k) \\ &\approx \frac{\partial f}{\partial x_j} \Delta X_j, \end{aligned}$$

  wobei $\frac{\partial f}{\partial x_j}$ möglicherweise von allen Regressoren abhängt.

- Daher bei Nichtlinearität möglich, dass Änderung in Y durch Änderung in X_j

  1. abhängt vom Niveau von X_j,
  2. abhängt vom Niveau der anderen Regressoren X_i $(i \neq j)$,
  3. abhängt von externen Einflüssen.

Nichtlineare Modelle

$$Y_i = f(X_{i1}, \ldots, X_{ik}) + u_i, \qquad i = 1, \ldots, n.$$

- **Idee**: Rückführung auf lineares Modell und dann Anwenden der bisherigen Theorie.
- Dazu: Benutze **Taylorapproximation** zweiter Ordnung (um 0):

$$f(X_1, \ldots, X_k) \approx f_0 + \sum_{j=1}^{k} \frac{\partial f}{\partial x_j} X_j + \frac{1}{2} \sum_{j=1}^{k} \sum_{l=1}^{k} \frac{\partial^2 f}{\partial x_j \partial x_l} X_j X_l.$$

- Lineare Approximation des nichtlinearen Modells:

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_k X_{ij} + \sum_{j=1}^{k} \sum_{l=1}^{k} \beta_{kl} (X_j X_l)_i + u_i.$$

Immer noch zu allgemein?

- Problem: Anzahl Parameter wächst schnell in k.
- Deswegen gibt es spezielle Modelle für nichtlineare Einflüsse, die
  1. nur vom Niveau von X_j abhängen (Polynome),
  2. nur von den Niveaus anderer Regressoren X_i abhängen (Interaktionsterme),
  3. nur von "externen" Einflüssen abhängen (Dummyvariablen).

1. Polynome

- Für den ersten Fall, nehme einige Potenzen eines (oder mehrerer) Regressor(en), aber keine Kreuzterme.
- Für den Fall eines einzelnen Regressors erhalten wir

$$Y_i = \beta_0 + \sum_{r=1}^{R} \beta_r X_{i1}^r + u_i.$$

- Man erhält bspw. für $R = 2$ ($R = 3$) das quadratische (kubische) Regressionsmodell.
- Im quadratischen Regressionsmodell $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + u$:

$$\frac{\Delta Y}{\Delta X_1} = \beta_1 + 2\beta_2 X_1.$$

- Daher ist $\beta_1 + 2\beta_2 x_1$ der marginale Effekt von X_1 auf Y, wenn $X_1 = x_1$.

1. Polynome

Beispiel: Quadratisches Regressionsmodell

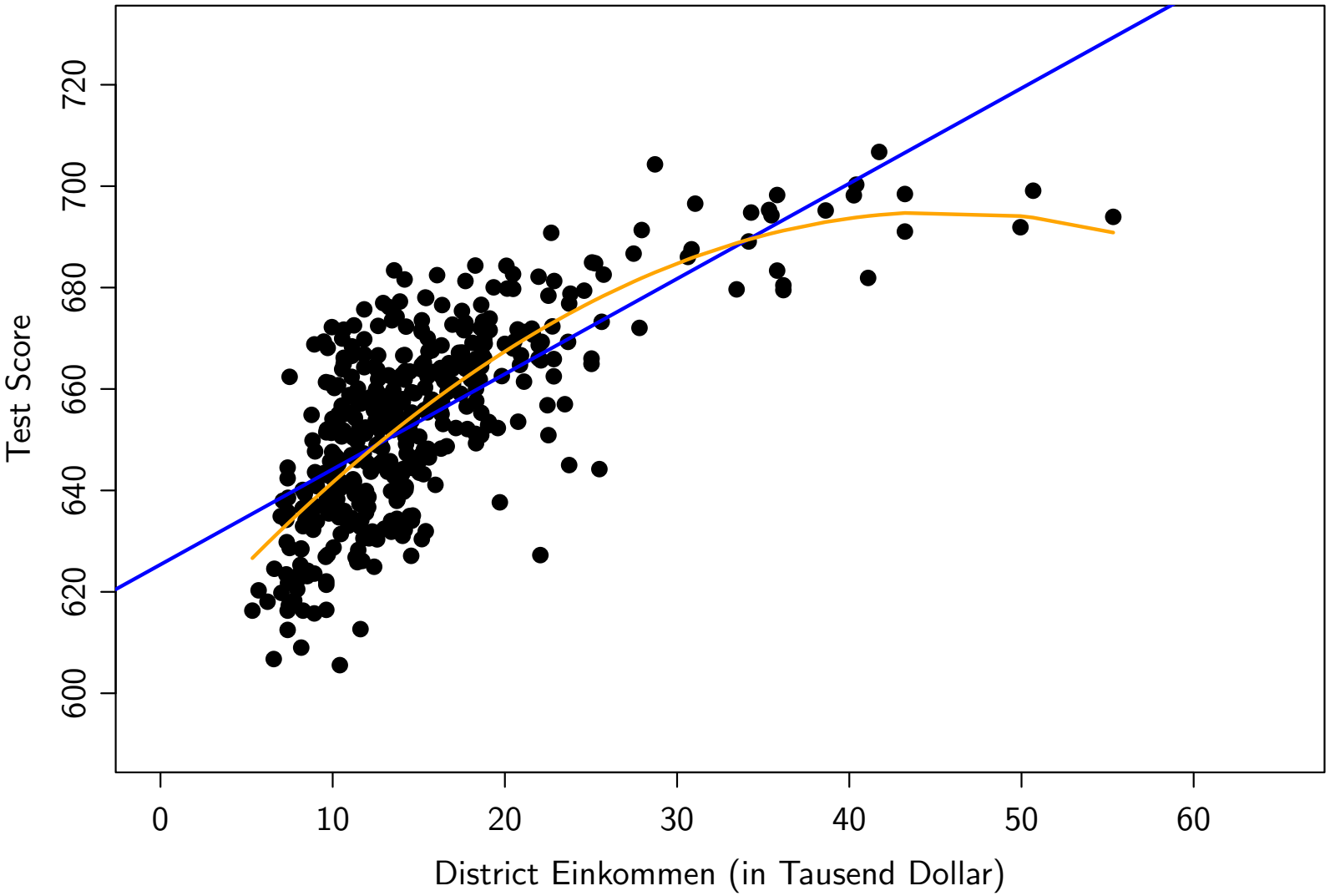


Abbildung 13: Punktediagramm von TestScore und Einkommen.

2. Interaktionsterme

- Im zweiten Fall, wähle einige sinnvolle Kreuzprodukte $X_i X_l$.
- Dies modelliert Interaktionen zwischen Regressoren:
  - Der Effekt einer Änderung in X_i auf Y hängt ab von X_l (und umgekehrt).
- Im Fall von zwei Regressoren:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + u.$$

- Der Einfluss von X_1 variiert mit X_2 (und vice versa): Durch X_2 kann X_1 effektiver sein (oder weniger effektiv – abhängig vom Vorzeichen von β_3).
- Über die 1. Ableitung ist der Einfluss von X_1

$$\frac{\Delta Y}{\Delta X_1} = \beta_1 + \beta_3 X_2.$$

- Daher ist $\beta_1 + \beta_3 x_2$ der marginale Effekt von X_1 auf Y, wenn $X_2 = x_2$.

3. Dummyvariablen

Änderung des Steigungsparameters

- Dummyvariable (annehmend, passend sortierte Daten):

$$D_{i,N} = \begin{cases} 0\,, & i = 1, \ldots, N, \\ 1\,, & i = N+1, \ldots, n. \end{cases}$$

- Änderung des Steigungsparameters:

$$Y_i = \beta_0 + \beta_1 X_{i1} + (D_{i,N}\beta_2) X_{i1} + u_i\,, \qquad i = 1, \ldots, n.$$

- Wir haben also praktisch einen zusätzlichen Regressor $X_{i2} = X_{i1} D_{i,N}$.
- Anders formuliert,

$$Y_i = \begin{cases} \beta_0 + \beta_1 X_{i1} + u_i\,, & i = 1, \ldots, N, \\ \beta_0 + (\beta_1 + \beta_2) X_{i1} + u_i\,, & i = N+1, \ldots, n. \end{cases}$$

- Auch dies ist eine Form von Interaktion (zwischen einer stetigen und einer binären Variable).

3. Dummyvariablen
Änderung des Achsenabschnitts

- Oft wird angenommen, dass Gruppen denselben Steigungsparameter, aber unterschiedliche Achsenabschnitte haben.
- Dies führt zum Modell mit Konstante und step-Dummy:

$$Y_i = \begin{cases} \beta_0 + \sum_{k=1}^{K} \beta_k X_{ik} + u_i\,, & i = 1, \ldots, N, \\ \beta_0 + \beta_0^* + \sum_{k=1}^{K} \beta_k X_{ik} + u_i\,, & i = N+1, \ldots, n. \end{cases}$$

- Oder auch:

$$Y_i = \beta_0 + \beta_0^* D_{i,N} + \sum_{k=1}^{K} \beta_k X_{ik} + u_i, \qquad i = 1, \ldots, n.$$

Logarithmen von Y und X

- Es gibt noch mehr Tricks nichtlineare auf lineare Modelle zu transformieren.
- Z.B. nutze nichtlinear transformierte Regressoren (X_i) und Regressanden (Y).
- Der Logarithmus ist eine oft genutzte Transformation (siehe `earthquakes.R`) mit interessanten Eigenschaften:
- **Log-lineares Modell**: $\log(Y) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + u$
  - Änderung von 1% in X_1 bewirkt Änderung von β_1% in Y.
- **Linear-log Modell**: $Y = \beta_0 + \beta_1 \log(X_1) + \ldots + \beta_k X_k + u$
  - Änderung von 1% in X_1 bewirkt Änderung von $0.01 \cdot \beta_1$ in Y.
- **Log-log Modell**: $\log(Y) = \beta_0 + \beta_1 \log(X_1) + \ldots + \beta_k X_k + u$
  - Änderung von 1% in X_1 bewirkt Änderung von β_1% in Y.

Überblick

- Nichtlineare Modelle
- **Aufspüren von Nichtlinearität**

Woher weiß man, dass Nichtlinearitäten vorliegen?

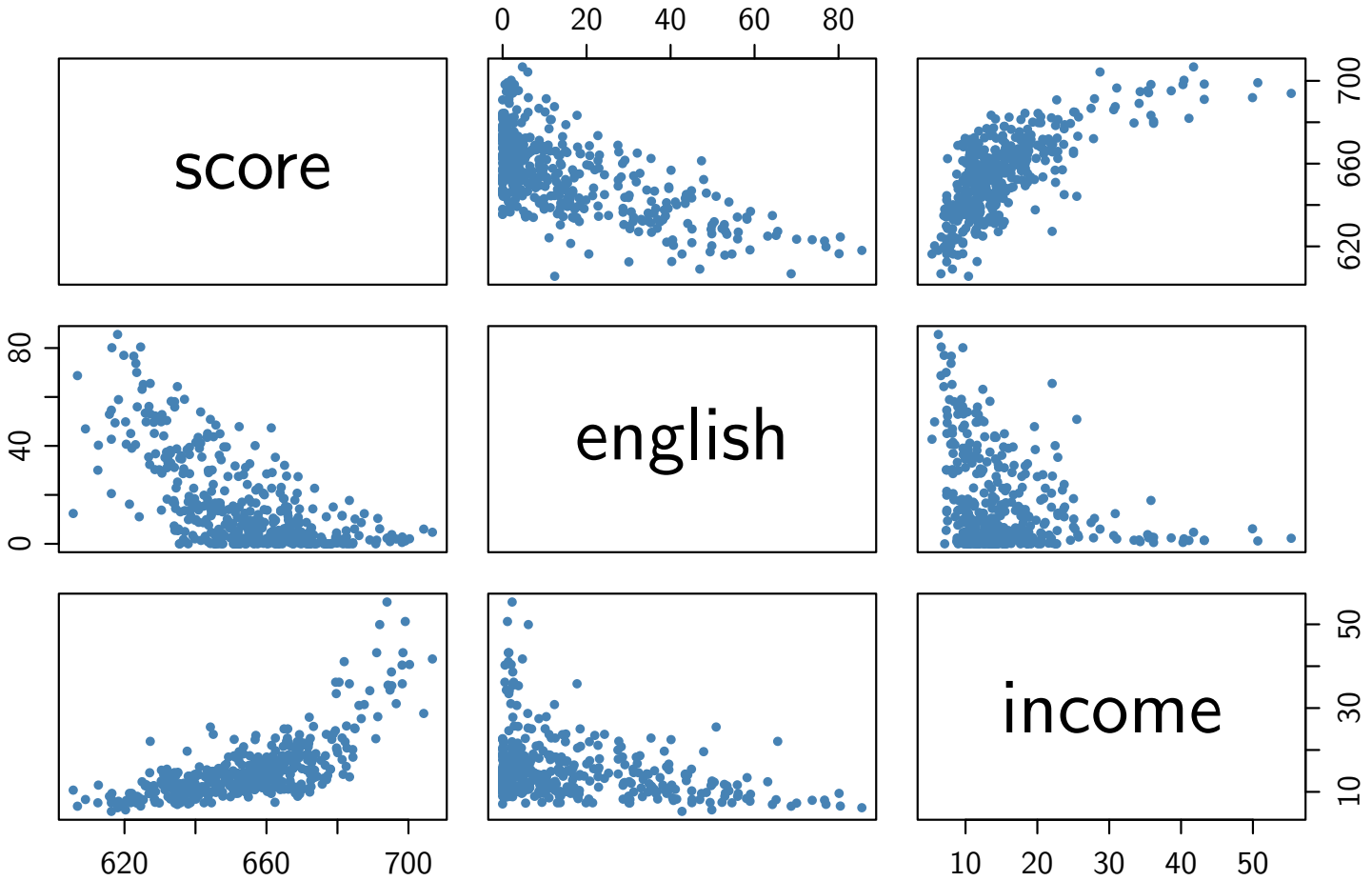
- Grafische Analyse
  - Erstelle alle möglichen Punktediagramme;
  - dies liefert auch Hinweise auf ein möglicherweise geeignetes nichtlineares Modell.
- Tests
  - Spezifiziere ein allgemeineres nichtlineares Modell (mit Polynomen und Kreuzprodukten);
  - und teste dann die Signifikanz der hinzugefügten Regressoren.
- Ökonomische Theorie
  - z.B. ob man eher prozentuale Änderungen erwarten sollte usw.

Woher weiß man, dass Nichtlinearitäten vorliegen?

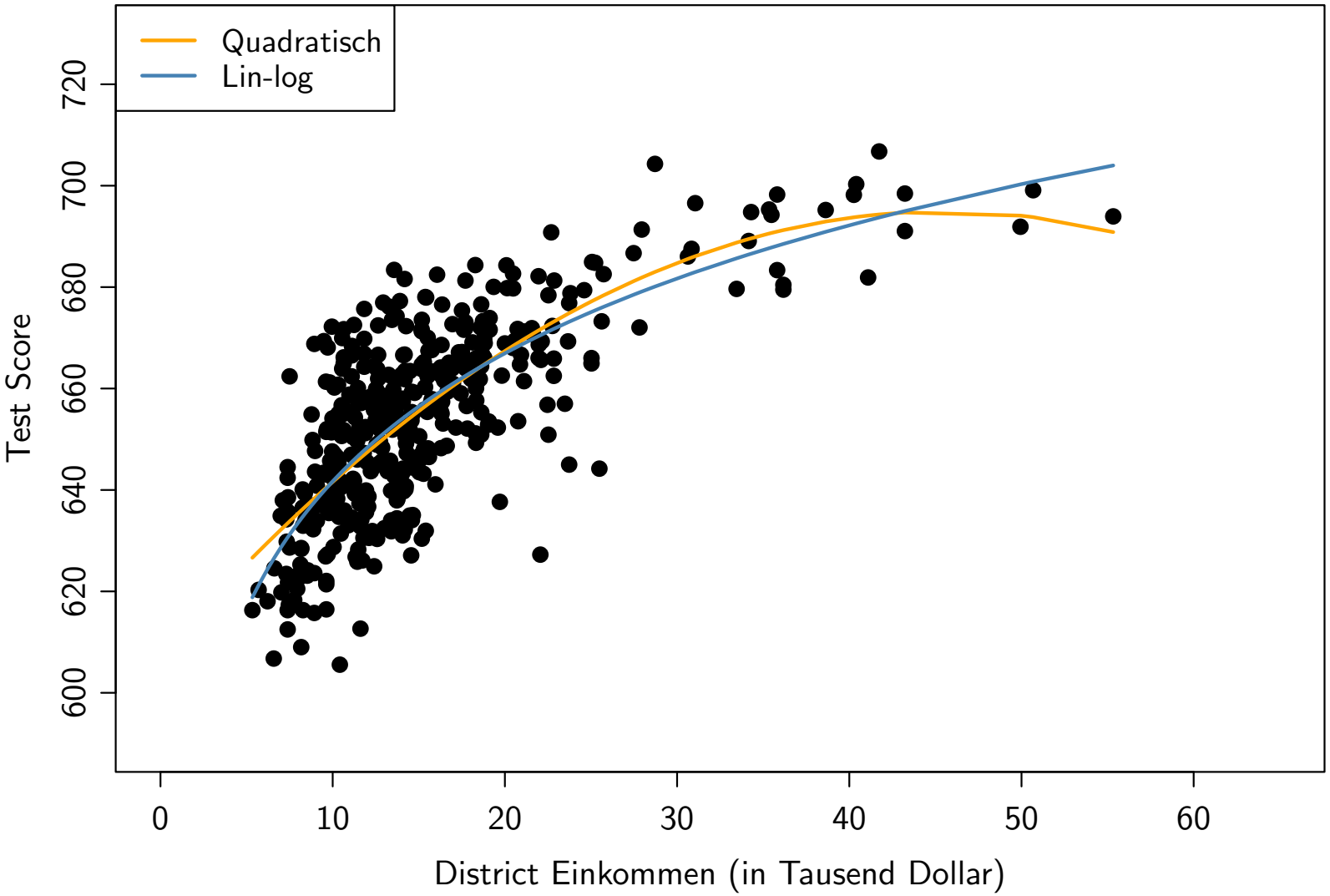
- Die Ansätze müssen nicht isoliert voneinander genutzt werden.
- Tests sind am einfachsten zu benutzen.
- Ökonomische Argumente können subjektiv sein.
- Grafik ist sehr intuitiv:
  - trage die abhängige Variable *und* die Residuen
  - gegen alle Regressoren *und* Kreuzprodukte ab.
- Eine Grafik weist oft die richtige Richtung;
  - ... aber der Leser erwartet oft trotzdem noch einen formalen Test.

Viele Grafiken

```
pairs(CASchools[, c("score","english","income")], col="steelblue", cex=0.6)
```



Erwarte nicht DAS perfekte Modell zu finden!



Überblick

Bewerten von multiplen Regressionen

(SW Kapitel 9)

Bewerten von multiplen Regressionen

- Machen wir einen Schritt zurück und werfen einen allgemeinen Blick auf Regressionen.
- Gibt es systematische Wege, Regressionsuntersuchungen kritisch zu bewerten? Wir kennen ihre Stärken – was aber sind Fallstricke bei multiplen Regressionen?
- Zentrale Stärken einer multiplen Regression:
  - Sie liefert eine Schätzung des Effekts einer beliebigen Änderung von X auf Y.
  - Sie löst das Problem der Verzerrung durch ausgelassene Variablen, wenn diese erhoben werden können.
  - Sie kann mit nichtlinearen Effekten umgehen (Effekte die von X abhängen).
- KQ kann trotzdem noch *verzerrte* Schätzer des wahren *kausalen* Effekts liefern – also "invalide" Inferenz...

Interne und externe Validität

- **Interne Validität**: Die statistische Inferenz bezüglich kausaler Effekte ist für die betrachtete Population valide ("gute Statistik").
- **Externe Validität**: Die statistische Inferenz kann hin zu anderen Populationen/Situationen verallgemeinert werden, wobei die "Situationen" sich etwa beziehen auf
  - die rechtliche, politische, zeitliche oder physische Umgebung.

Überblick

- **Gefahren für interne Validität**
- Gefahren für externe Validität
- Interne und Externe Validität bei Prognosen

Gefahren für interne Validität

- Interne Validität: "Funktionieren die Methoden korrekt?"
- Fünf Gefahren für die interne Validität:

  1. Ausgelassene Variablen
  2. Fehlerhafte funktionale Form
  3. Messfehler in den Variablen
  4. Verzerrung durch selektive Stichproben
  5. Simultane Kausalität

- All diese führen dazu, dass $\mathrm{E}(u|X_1, \ldots, X_k) \neq 0$, so dass KQ verzerrt und inkonsistent ist.

1. Ausgelassene Variablen

Verzerrung durch ausgelassene Variablen

- Verzerrung durch ausgelassene Variablen entsteht, wenn eine ausgelassene Variable

  1. eine Determinante von Y ist und
  2. mit mindestens einem einbezogenen Regressor korreliert.

- Wir haben das Problem bisher für Einfachregressionen diskutiert.
- Die Verzerrung tritt aber auch bei mehreren Regressoren auf, wenn die ausgelassene Variable 1. und 2. erfüllt.

Lösungen für OVB

1. Wenn die Variable erhoben werden kann, füge sie als zusätzlichen Regressor in die Regression ein.
2. Benutze *Paneldaten*: Dort wird jede/s Einheit/Individuum mehr als einmal beobachtet.
3. Wenn die Variable nicht erhoben werden kann, nutze eine *Instrumentvariablenregression*.
4. Führe ein randomisiertes Experiment durch.
   - Warum funktioniert das? Wenn X randomisiert ist, ist X automatisch unabhängig von u, d.h. $\mathrm{E}(u|X) = 0$.

Auswahl der Variablen und Modellspezifikation

- Man kann nicht beliebig viele Variablen in die Regression stecken, selbst wenn man sie erheben kann (warum?).
- Also:

Zusätzliche Variable: Ja oder Nein?

- Spezifiziere ein "Basismodell".
- Spezifiziere einige plausible alternative Modelle, die zusätzliche potenziell relevante Regressoren beinhalten – nutze Urteilskraft, kein mechanisches Rezept ...
- Sind die potenziellen Variablen statistisch signifikant?
- Prüfe, ob die potenziellen Variablen die Schätzung des uns interessierenden Koeffizienten β_1 (wesentlich) ändern.

Beispiel

Beispiel 8.1: Kalifornische Schulbezirke.

- Welche Variablen möchte man *idealerweise* zur Schätzung des kausalen Effekts von STR mit Distriktdaten nutzen?
- *Tatsächlich* vorhandene Variablen in dem kalifornischen Datensatz sind:
  - Schüler/Lehrer-Verhältnis (STR)
  - Anteil Englischlernende im Distrikt ($PctEL$)
  - Ausgaben der Schulen pro Schüler
  - Name des Distrikts (so dass etwa Regenniederschlag nachgeschlagen werde könnte)
  - Anteil mit subventioniertem Essen
  - Anteil auf Sozilahilfe
  - Durchschnittliches Einkommen im Distrikt
- Welche dieser Variablen würden Sie einbeziehen wollen?

2. Falsche funktionale Form

Linearität?

- Was passiert, wenn der Einfluss eines Regressors nichtlinear ist?
- Mögliche Lösungen für falsche funktionale Form:

  1. Stetige abhängige Variable: nutze eine "geeignete" nichtlineare Spezifikation für X (Logarithmen, Interaktionen, etc.). Vgl. Kapitel 7.
  2. Diskrete (bspw. binäre) abhängige Variable: wir brauchen eine Erweiterung multipler Regressionsmethoden ("probit" oder "logit" für binäre Abhängige).

3. Verzerrung durch Messfehler

Verzerrung durch Messfehler

- Bisher haben wir angenommen, dass X fehlerfrei erhoben werden kann.
- In der Praxis weisen ökonomische Daten oft Messfehler auf:
  - Fehler beim Verarbeiten administrativer Daten
  - Erinnerungslücken in Umfragen ("Seit wann arbeiten Sie in ihrem Beruf?")
  - Unklare Fragen ("Was war im letzten Jahr ihr Einkommen?")
  - Absichtlich falsche Antworten ("Wie hoch sind ihre finanziellen Anlagen? Wie oft fahren Sie betrunken Auto?")

Verzerrung durch Messfehler

- Messfehler für einen Regressor resultiert im Allgemeinen in Verzerrung.
- Nehme zur Illustration an, dass

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \qquad i = 1, \ldots, n,$$

  "korrekt" ist, in dem Sinne, dass die drei KQ-Annahmen erfüllt sind (insbesondere $\mathrm{E}(u \mid X) = 0$).
- Definiere:

$$X_i = \text{wahrer (aber nicht erhobener) Wert von } X;$$
$$\widetilde{X}_i = \text{ungenau gemessene Version von } X.$$

Das ist ein ernstes Problem

- Dann gilt

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + u_i \\ &= \beta_0 + \beta_1 \tilde{X}_i + [\beta_1 (X_i - \tilde{X}_i) + u_i] \\ &= \beta_0 + \beta_1 \tilde{X}_i + \tilde{u}_i, \qquad \text{wobei } \tilde{u}_i = \beta_1 (X_i - \tilde{X}_i) + u_i. \end{aligned}$$

- Nun ist aber $\tilde{X}_i$ typischerweise korreliert mit $\tilde{u}_i$, so dass $\widehat{\beta}_1$ verzerrt ist:

$$\begin{aligned} \text{Cov}(\tilde{X}_i, \tilde{u}_i) &= \text{Cov}(\tilde{X}_i, \beta_1 (X_i - \tilde{X}_i) + u_i) \\ &= \beta_1 \, \text{Cov}(\tilde{X}_i, X_i - \tilde{X}_i) + \text{Cov}(\tilde{X}_i, u_i) \\ &= \beta_1 [\text{Cov}(\tilde{X}_i, X_i) - \text{Var}(\tilde{X}_i)] + 0 \neq 0, \end{aligned}$$

da im Allgemeinen gilt, dass $\text{Cov}(\tilde{X}_i, X_i) \neq \text{Var}(\tilde{X}_i)$.

Das ist ein ernstes Problem

$$Y_i = \beta_0 + \beta_1 \tilde{X}_i + \tilde{u}_i, \quad \text{wobei} \quad \tilde{u}_i = \beta_1 (X_i - \tilde{X}_i) + u_i$$

- Wenn X_i mit Fehler gemessen wird, ist $\tilde{X}_i$ i.A. korreliert mit $\tilde{u}_i$, so dass $\widehat{\beta}_1$ verzerrt und inkonsistent ist.
- Es ist möglich, die Verzerrung unter spezifischen Annahmen über den Messfehler genauer anzugeben (z.B., dass $\tilde{u}_i$ und X_i unkorreliert sind).
- Diese Formeln sind speziell, aber das Ergebnis, dass Messfehler in X zu Verzerrung führt, gilt allgemein.

Mögliche Lösungen für Messfehler

1. Bessere Daten.
2. Entwickle ein spezielles Modell für den Messfehlerprozess.
   - Das ist nur möglich, wenn man viel über den Messfehler weiß – z.B. wenn eine Teilstichprobe der Daten mit administrativen Daten abgeglichen wurde und die Unterschiede modelliert wurden. (Sehr speziell, wir verfolgen das nicht weiter.)
3. Instrumentvariablenregressionen (vgl. Kapitel 9).

4. Stichprobenverzerrung/Sample selection bias

Sample selection-Verzerrung

- Bisher haben wir einfache Zufallsstichproben aus der Population angenommen.
- Manchmal ist das nicht plausibel, wenn sich die Stichprobe "selbst auswählt".
- **Sample selection-Verzerrung** entsteht, wenn der Auswahlprozess:
  i. das Vorhandensein der Daten beeinflusst und
  ii. dieser Prozess mit der abhängigen Variablen in Verbindung steht.

Beispiel

Beispiel 8.2: Krankenhäuser und Gesundheit.

- Sind Krankenhäuser gut für Ihre Gesundheit?
  - Einerseits: Sie werden von Experten überwacht (also JA).
  - Andererseits: Jeder dort ist krank, was die Sache u.U. schlimmer macht (also NEIN).
- Man beschaffe sich Daten zur Gesundheit derer, die zur Genesung nach **Hause** geschickt wurden, und derer, die ins **Krankenhaus** eingewiesen wurden.
- Nun sind aber die Eingewiesenen typischerweise in schlechterem Zustand, so dass Effekt des Krankenhauses nur für Teilstichprobe beobachtet werden kann – für die der Effekt reduziert sein wird!
- Ähnlich ist das Problem gelagert bei "survival Verzerrung" (`MutualFunds.R`), der Gehaltsschere für arbeitende Mütter, etc.

Lösungen für sample selection-Verzerrung

- Sammle Daten so, dass sample selection vermieden wird.
  - Mutual funds-Beispiel: ändere die Stichprobe von denen, die nach 10 Jahren verfügbar sind, zu denen, die zu Beginn verfügbar sind (berücksichtige auch gescheiterte Fonds).
  - Ausbildungsrenditen-Beispiel: erhebe College-/Uniabsolventen, nicht Arbeitnehmer (berücksichtige auch Arbeitslose).
- Randomisiertes kontrolliertes Experiment.
- Konstruiere ein Modell des sample selection-Problems und schätze dieses Modell (wird hier nicht behandelt).

5. Verzerrung durch simultane Kausalität

Seien Sie vorsichtig mit Kausalität

- Bisher haben wir angenommen, dass X kausal für Y ist.
- Was ist, wenn auch Y kausal für X ist?

Beispiel 8.3: Kalifornische Schulbezirke.

- Niedrige STR führt zu besseren Ergebnissen.
- Was aber, wenn Distrikte mit niedrigen Testergebnissen zusätzliche Ressourcen erhalten? Als Ergebnis eines solchen politischen Prozesses haben Sie auch niedrige STR.
- Was bedeutet das für eine Regression von $TestScore$ auf STR?

Verzerrung durch simultane Kausalität

$$\text{Kausaler Effekt für } Y \text{ von } X: \quad Y_i = \beta_0 + \beta_1 X_i + u_i.$$
$$\text{Kausaler Effekt für } X \text{ von } Y: \quad X_i = \gamma_0 + \gamma_1 Y_i + v_i.$$

- Ein großes u_i bedeutet großes Y_i, was großes X_i impliziert (wenn $\gamma_1 > 0$).
- Also ist $\text{Corr}(X_i, u_i) \neq 0$.
- Also ist $\widehat{\beta}_1$ verzerrt und inkonsistent.
- Beispiel: Ein Distrikt mit besonders schlechten Testergebnissen gegeben die STR (negatives u_i) erhält mehr Geld, womit die STR reduziert wird. Also sind STR_i und u_i korreliert.

Lösungen Verzerrung durch simultane Kausalität

1. Randomisiertes kontrolliertes Experiment: Da X_i randomisiert ist, gibt es keinen Einfluss der abhängigen Variablen Y_i (wenn sich alle ans Experiment halten).
2. Entwickle und schätze ein vollständiges Modell beider Kausalitätsrichtungen.
   - Das ist die Idee hinter vielen großen Makromodellen (bspw. Federal Reserve Bank-US). In der Praxis sehr schwierig.
3. Nutze Instrumentvariablenregressionen, um den kausalen Effekt (Effekt von X auf Y, ohne den Effekt von Y auf X) zu schätzen
   - Siehe Kapitel 9

Überblick

- Gefahren für interne Validität
- **Gefahren für externe Validität**
- Interne und Externe Validität bei Prognosen

Extrapolieren?

- Wie weit kann man die Ergebnisse für Klassengrößen aus Kalifornien verallgemeinern?
- Unterschiede in Populationen
  - Kalifornien in 2005?
  - Massachusetts in 2005?
  - Mexico in 2005?
- Unterschiede in den Rahmenbedingungen
  - unterschiedliche rechtliche Voraussetzungen (Sonderschulen o.Ä.).
  - unterschiedliche Behandlung bilingualer Erziehung.
  - Unterschiede in Lehrercharakteristiken.

Überblick

- Gefahren für interne Validität
- Gefahren für externe Validität
- **Interne und Externe Validität bei Prognosen**

Interne und Externe Validität bei Prognosen

Weniger Grund zur Sorge

- Prognose und Schätzung von kausalen Effekten sind unterschiedliche Ziele.
- Für kausale Effekt ist interne Validität sehr wichtig.
- Für Prognosen ist externe Validität ist sehr wichtig.
  - Das aus historischen Daten geschätzte Modell muss in der (zumindest nahen) Zukunft gelten.

Beispiel 8.4: Kalifornische Schulbezirke.

- Eltern-Problem (Prognose): Angenommen wir ziehen von einem Schulbezirk in einen anderen mit ($STR \to STR + 1$), was ist meine Prognose für den Unterschied im Testergebnis meines Kindes?
- Bildungsminister-Problem (Kausal): Alles andere gleichhaltend, was ist der Effekt einer Erhöhung ($STR \to STR + 1$) auf die Testergebnisse?

Überblick

Instrumentvariablen-Regressionen

(SW Kapitel 12)

Instrumentvariablen-Regression

- Drei bedeutende Bedrohungen der internen Validität sind:
  1. Verzerrung durch eine ausgelassene Variable, die mit X korreliert ist, die wir nicht beobachten, so dass wir sie nicht in die Regression einfügen können;
  2. Verzerrung durch simultane Kausalität (X ist kausal für Y, Y ist kausal für X);
  3. Verzerrung durch Messfehler in X.
- Instrumentvariablen-Regression kann diese Verzerrung aufgrund von $\mathrm{E}(u|X) \neq 0$ mit einer Instrumentvariable Z vermeiden.

Überblick

- **Instrumentvariablen-Regression**
- Eine Geschichte über Nachfrage und Angebot
- Das allgemeine IV Regressionsmodell

Etwas Terminologie

- "Endogen" bedeutet "innerhalb des Systems determiniert", das heißt eine endogene Variable wird gemeinsam mit Y determiniert, bzw. durch simultane Kausalität beeinflusst.
- Diese Definition ist jedoch eng: IV-Regression kann auch gegen Verzerrung durch ausgelassene Variablen und bei Messfehlern verwendet werden.
- Deshalb definieren wir allgemeiner:

Definition 9.1: Endogen & Exogen.

- Eine **endogene** Variable ist eine, die korreliert mit u ist;
- Eine **exogene** Variable ist eine, die unkorreliert mit u ist.

Wie funktionieren Instrumente?

- **Beispiel**: Einfaches lineares Regressionsmodell

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \quad i = 1, \ldots, n,$$

  wobei X und u möglicherweise korrelieren.
- Egal, wo sie herkommt, die Korrelation von X und u wird von der KQ-Prozedur "verwechselt" mit Korrelation zwischen Y und X.
- Daher ist $\widehat{\beta}_1$ verzerrt.

Wie funktionieren Instrumente?

- Die Idee hinter IV ist es, X in zwei Teile aufzuteilen:
  1. ein Teil, der mit u korreliert sein könnte, und
  2. ein Teil, der es nicht ist.
- Die Schätzung wird wieder glaubwürdig, wenn wir nur den "unkorrelierten Teil" von X als Regressor nutzen.
- Dies wird mit einer Instrumentvariable Z gemacht, die...
  - ... *unkorreliert* ist mit u,
  - ... aber *korreliert* ist mit X.
- Die Instrumentvariable zeigt die Bewegungen in X auf, die unkorreliert sind mit u, und nutzt diese zum Schätzen von β_1.

Zwei Bedingungen für ein valides Instrument

- Für das Beispiel mit einem Regressor und einem Instrument,

$$Y_i = \beta_0 + \beta_1 X_i + u_i,$$

muss eine valide **Instrumentvariable (IV)** Z die beiden folgenden Bedingungen erfüllen:

1. $\text{Corr}(Z, u) = 0$ (Exogenität des Instruments);
2. $\text{Corr}(Z, X) \neq 0$ (Relevanz des Instruments).

- Nehme für den Augenblick an, dass wir ein valides Instrument haben. (Beispiele dafür später.)
- Wie nutzt man nun Z konkret zur Schätzung von β_1?

Der IV-Schätzer

Two Stage Least Squares (TSLS, 2SLS)

1. Die erste KQ-Regression isoliert den Teil von X, der unkorreliert mit u ist:

$$X_i = \pi_0 + \pi_1 Z_i + v_i.$$

   - Wir berechnen die prognostizierten Werte $\widehat{X} = \widehat{\pi}_0 + \widehat{\pi}_1 Z$.
   - In diesen steckt der mit u unkorrelierte Teil von X, da $\text{Corr}(\widehat{X}, u) \approx \text{Corr}(\pi_0 + \pi_1 Z, u) = \text{Corr}(Z, u) = 0$.

2. Ersetze X durch $\widehat{X}$ in der uns interessierenden KQ-Regression:

$$Y_i = \beta_0 + \beta_1 \widehat{X}_i + u_i. \tag{9.1}$$

   Das liefert

$$\widehat{\beta}_1^{TSLS} = \frac{\sum_{i=1}^n (Z_i - \overline{Z})(Y_i - \overline{Y})}{\sum_{i=1}^n (Z_i - \overline{Z})(X_i - \overline{X})} = \frac{s_{ZY}}{s_{ZX}}.$$

Der IV-Schätzer

- Da für große n, $\text{Corr}(\widehat{X}, u) = 0$, ist KQ-Annahme 1, $\text{E}(u \mid \widehat{X}) = 0$, plausibel.
- Also kann β_1 mit der KQ-Regression (9.1) geschätzt werden.
- Der resultierende Schätzer $\widehat{\beta}_1^{TSLS}$ heißt **Two Stage Least Squares (TSLS) Schätzer**. `2SLS.R`

Inferenz für TSLS

- In großen Stichproben ist der TSLS Schätzer konsistent und seine Stichprobenverteilung ist normal.
- Inferenz (Hypothesentests, Konfidenzintervalle) verläuft (fast) wie gehabt,
- ... fast, weil die üblichen Standardfehler der Regression des zweiten Schrittes den ersten Schritt nicht berücksichtigen.
- Benutze also die in vielen Software-Paketen implementierte TSLS Prozedur, um korrekte Standardfehler und damit korrekte Inferenz zu bekommen.
- Heteroskedastie-robuste Standardfehler sind aus den bekannten Gründen ebenfalls eine gute Idee.

Zusammenfassung bisher

- Nehme an, dass ein valides (d.h. exogenes und relevantes) Instrument Z zur Verfügung steht.
  1. Regressiere X auf Z: Speichere die angepassten Werte $\widehat{X}$.
  2. Regressiere Y auf $\widehat{X}$: Koeffizient von $\widehat{X}$ ist der TSLS Schätzer, $\widehat{\beta}_1^{TSLS}$.
- Die Kernidee ist, dass der erste Schritt den Teil der Variation in X extrahiert, der unkorreliert mit u ist.
- $\widehat{\beta}_1^{TSLS}$ ist ein konsistenter Schätzer von β_1, ...
- ...und hat eine asymptotische Normalverteilung
- Man vergesse aber nicht, die korrekten Standardfehler zum Testen zu benutzen.

Überblick

- Instrumentvariablen-Regression
- **Eine Geschichte über Nachfrage und Angebot**
- Das allgemeine IV Regressionsmodell

Ein Blick in die Historie...

- Die IV Regression wurde ursprünglich zum Schätzen von Nachfrageelastizitäten für landwirtschaftliche Güter entwickelt.
- Zum Beispiel Butter:

$$\log(Q_i^{Butter}) = \beta_0 + \beta_1 \log(P_i^{Butter}) + u_i.$$

- β_1 = Preiselastizität von Butter = die prozentuale Änderung der Menge bei einer 1%igen Änderung des Preises.
  - (Man erinnere sich an die Diskussion der log-log-Spezifikation).
- Daten: Beobachtungen zu Preis und Menge von Butter für verschiedene Jahre.
- Die KQ-Regression von $\log(Q)$ auf $\log(P)$ leidet jedoch an Verzerrung aufgrund von simultaner Kausalität ...

Warum Verzerrung durch simultane Kausalität?

- Verzerrung durch simultane Kausalität in der KQ-Regression von $\log(Q)$ auf $\log(P)$ entsteht, weil sowohl Preis als auch Menge durch die *Interaktion* von Nachfrage und Angebot determiniert werden.

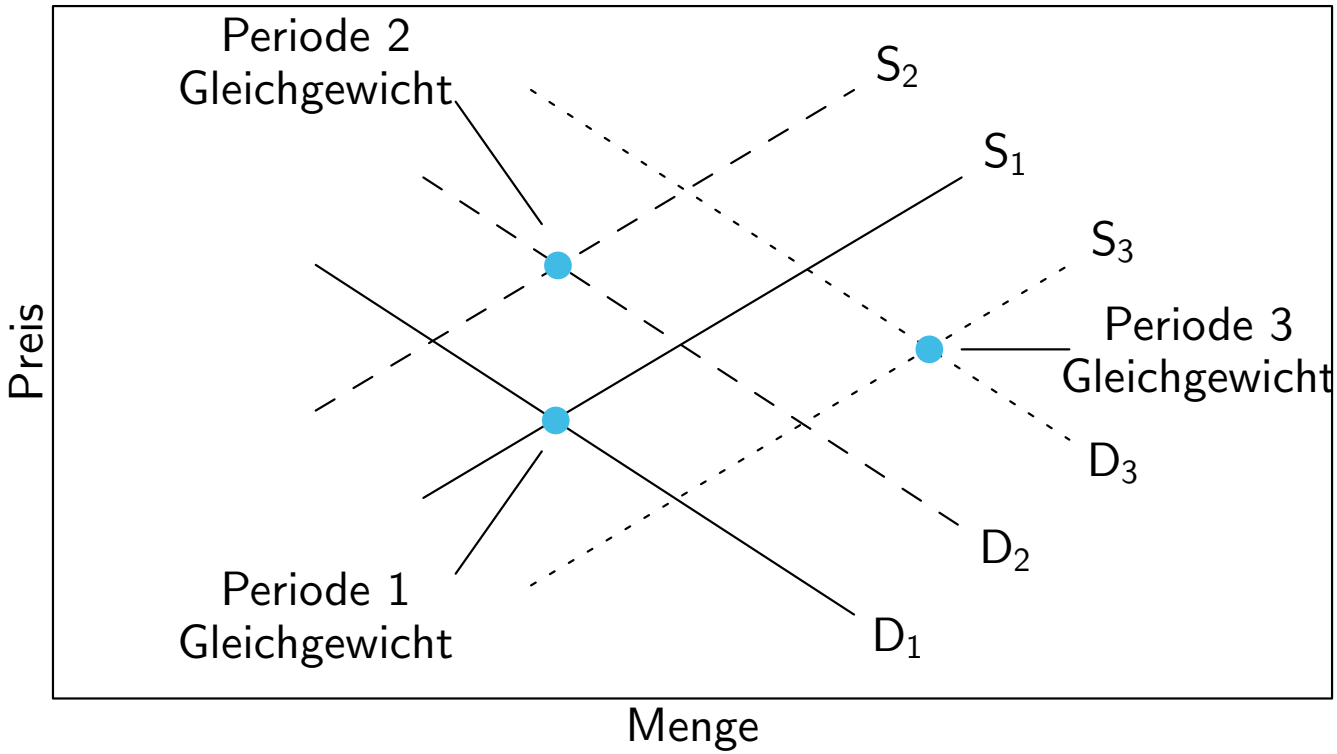


Abbildung 14: (a) Angebot und Nachfrage in drei Zeitperioden.

Warum Verzerrung durch simultane Kausalität?

- Diese Interaktion von Nachfrage und Angebot liefert ...

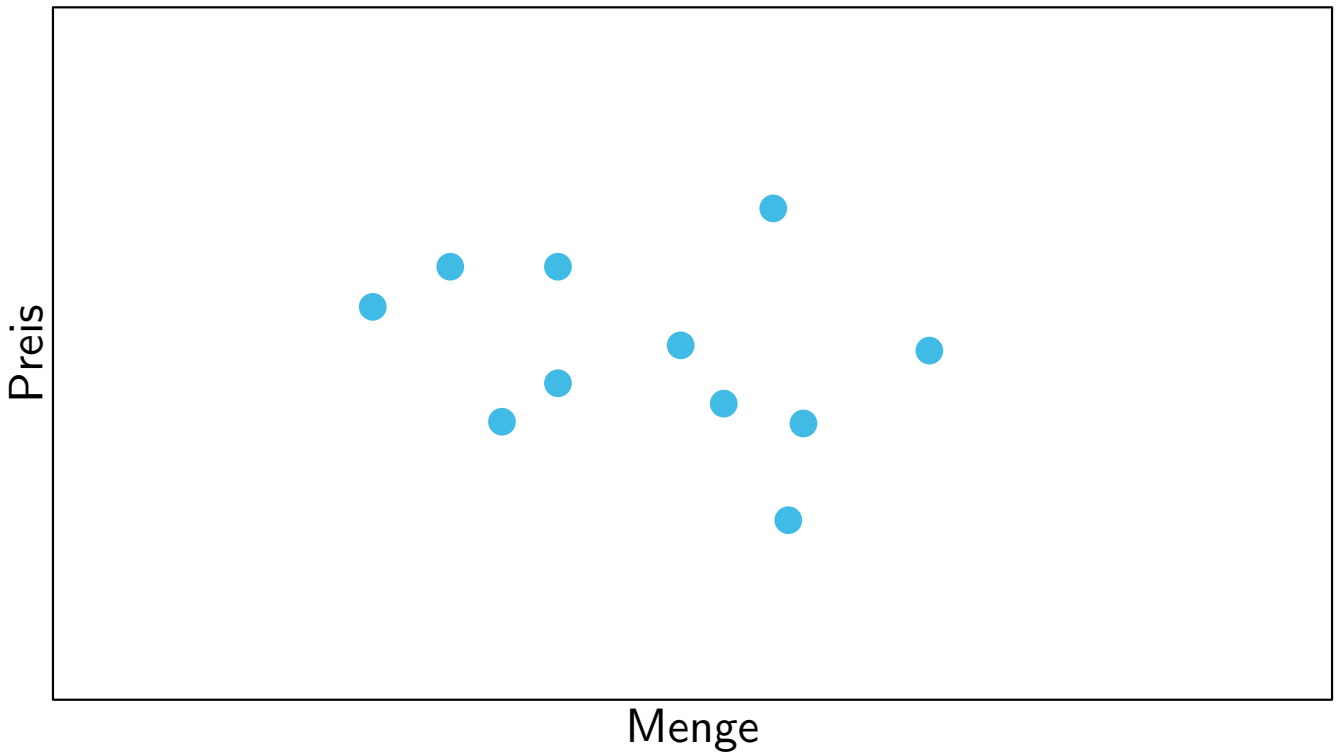


Abbildung 15: (b) Gleichgewichtspreise und -mengen in 11 Zeitperioden.

- Würde eine Regression mit diesen Daten die Nachfragekurve schätzen?

Wie funktioniert TSLS hier?

- Aber was würde man bekommen wenn sich nur die Angebotskurve verschöbe?

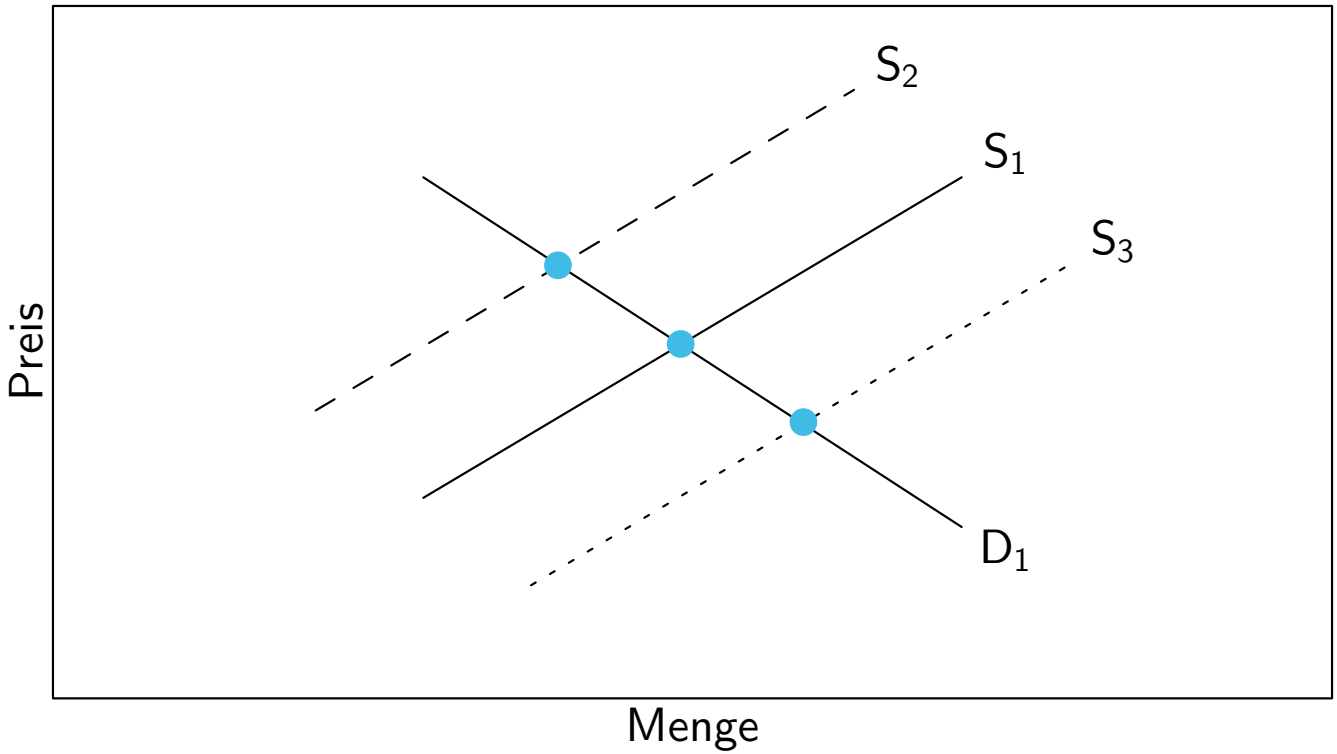


Abbildung 16: (c) Gleichgewichtspreise und -mengen bei verschobener Angebotskurve.

- TSLS schätzt die Nachfragekurve, indem Änderungen in Preis und Menge isoliert werden, die aus Änderungen im Angebot resultieren.

TSLS im Angebot-Nachfrage-Beispiel

- Z wäre eine Variable, die das Angebot ändert, nicht jedoch die Nachfrage.
- Setze Z = Regenmenge in Regionen, die Milchprodukte herstellen.
- Ist Z ein valides Instrument?

1. Exogen? ($\text{Corr}(Z, u) = 0$?)
   Plausibel: Ob es in diesen Regionen regnet oder nicht, sollte die Nachfrage nicht beeinflussen.

2. Relevant? ($\text{Corr}(Z, \log(P)) \neq 0$?)
   Plausibel: Fehlender Regen bedeutet weniger Gras für Kühe und das bedeutet weniger Butter.

TSLS im Angebot-Nachfrage-Beispiel

1. Also regressiere $\log(P)$ auf Z (Regenmenge) und speichere den Teil der Variation in den Preisen, der nicht von Nachfrageänderungen beeinflusst wird.
2. ... regressiere $\log(Q)$ auf $\widehat{\log(P)}$ um die gewünschte Elastizität zu schätzen.

- Angebotsänderungen erlauben es uns also, die Nachfragekurve zu identifizieren.
- **Zusammenfassung:** IV funktioniert, aber es ist nicht einfach, ein valides Instrument zu finden.

Überblick

- Instrumentvariablen-Regression
- Eine Geschichte über Nachfrage und Angebot
- **Das allgemeine IV Regressionsmodell**

Mehr Regressoren/Instrumente

- Bisher haben wir IV Regression mit einem einzelnen endogenen Regressor (X) und einem einzelnen Instrument (Z) betrachtet.
- Wir müssen dies verallgemeinern zu:
  - Multiplen endogenen Regressoren ($X_1, \ldots, X_k$);
  - multiplen exogenen Variablen in der Regression ($W_1, \ldots, W_r$)
    (diese brauchen wir aus den bekannten Gründen zur Vermeidung der Verzerrung durch ausgelassene Variablen);
  - multiple Instrumentvariablen ($Z_1, \ldots, Z_m$, $m \geq k$)
    (mehr relevante Instrumente können eine kleinere Varianz von TSLS bedeuten).
- Neue Terminologie: Identifikation und Überidentifikation.

Identifikation

- Ein Parameter ist **identifiziert** wenn unterschiedliche Werte des Parameters unterschiedliche Verteilungen der Daten produzieren.
- Bei IV Regressionen hängt Identifikation der Koeffizienten von der Relation zwischen der Anzahl der Instrumente (m) und der Anzahl der endogenen Regressoren (k) ab.
- Intuitiv: Wenn weniger Instrumente als endogene Regressoren vorliegen, kommen wir nicht weiter...
  - Zum Beispiel, wenn $k = 1$ aber $m = 0$, d.h. keine Instrumente.

Identifikation

Die Koeffizienten $\beta_1, \ldots, \beta_k$ nennen wir:

- **exakt identifiziert**, wenn $m = k$.
  - Es gibt genau genug Instrumente um $\beta_1, \ldots, \beta_k$ zu schätzen.
- **überidentifiziert**, wenn $m > k$.
  - Es gibt mehr Instrumente als nötig, um $\beta_1, \ldots, \beta_k$ zu schätzen – wenn dem so ist, kann man (teilweise) testen, ob die Instrumente valide sind (ein Test der "überidentifizierenden Restriktionen").
- **unteridentifiziert**, wenn $m < k$.
  - Es gibt zu wenige Instrumente zum Schätzen – wir müssen mehr Instrumente beschaffen.

Zusammenfassung: Das allgemeine Modell

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik} + \beta_{k+1} W_{i1} + \ldots + \beta_{k+r} W_{ir} + u_i.$$

- Y ist die abhängige Variable.
- $X_1, \ldots, X_k$ sind die endogenen Regressoren (potenziell korreliert mit u).
- $W_1, \ldots, W_r$ sind die einbezogenen exogenen Variablen/Regressoren (unkorreliert mit u).
- $\beta_0, \ldots, \beta_{k+r}$ sind die unbekannten Regressionskoeffizienten.
- $Z_1, \ldots, Z_m$ sind die m Instrumentvariablen (die ausgelassenen exogenen Variablen).
- Die Koeffizienten sind überidentifiziert, wenn $m > k$; exakt identifiziert, wenn $m = k$; und unteridentifiziert, wenn $m < k$.

TSLS mit einem einzelnen endogenen Regressor

$$\boxed{Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 W_{i1} + \ldots + \beta_{r+1} W_{ir} + u_i.}$$

- m Instrumente: $Z_1, \ldots, Z_m$.
- Erster Schritt
  - Regressiere X_1 per KQ auf *alle* exogenen Regressoren $W_1, \ldots, W_r$ *und* $Z_1, \ldots, Z_m$ – wir wollen keine Verzerrung durch ausgelassene Variablen!
  - Berechne die prognostizierten Werte $\widehat{X}_i$, $i = 1, \ldots, n$.
- Zweiter Schritt
  - Regressiere Y mit KQ auf $\widehat{X}$ und $W_1, \ldots, W_r$.
  - Die Koeffizienten dieses zweiten Schrittes sind die Koeffizienten des TSLS-Schätzers, aber die KQ SE's dieses Schrittes sind verkehrt.
  - Überlasse Software die korrekte Berechnung der SE's.

Beispiel
Lohnregression

Beispiel 9.2: Card (1995).

- Betrachten Sie die Lohnregression von Card (1995)

$$wage = \beta_0 + \beta_1 \cdot educ + u.$$

- u enthält die ausgelassene Variable *ability*, die *educ* beeinflusst.
- Ein Instrument könnte $Z =$ "Entfernung zum nächsten College" sein.
  1. $\text{Cov}(Z, u) = 0$, wenn die Entfernung zum College unkorreliert mit *ability* ist.
  2. $\text{Cov}(Z, X) \neq 0$, wenn Leute, die nahe am College wohnen, eher dorthin gehen.

Beispiel

Lohnregression

```
card <- read.dta("Daten/card.dta")
# see http://fmwww.bc.edu/ec-p/data/wooldridge/card.des for brief description

# Checking for relevance: reduced form
redf <- lm(educ ~ nearc4+exper+I(exper^2)+black+smsa+south+smsa66+
           reg662+reg663+reg664+reg665+reg666+reg667+reg668+reg669, data=card)
# OLS
ols <- lm(log(wage)~educ+exper+I(exper^2)+black+smsa+south+smsa66+
           reg662+reg663+reg664+reg665+reg666+reg667+reg668+reg669, data=card)

# IV estimation
iv <- ivreg(log(wage) ~ educ+exper+I(exper^2)+black+smsa+south+smsa66+
             reg662+reg663+reg664+reg665+reg666+reg667+reg668+reg669
           | nearc4+exper+I(exper^2)+black+smsa+south+smsa66+
             reg662+reg663+reg664+reg665+reg666+reg667+reg668+reg669, data=card)

# 2SLS by hand, incorrect standard errors (although by very little here)
ivbyhand <- lm(log(wage)~fitted(redf)+exper+I(exper^2)+black+smsa+south+smsa66+
              reg662+reg663+reg664+reg665+reg666+reg667+reg668+reg669, data=card)
```

Beispiel I

Lohnregression

```
# Pretty regression table of selected coefficients
stargazer(redf, ols, iv, ivbyhand, type="text",
          keep=c("ed", "near", "exp", "bl"),keep.stat=c("n", "rsq"))
```

```
##
## ================================================
##                         Dependent variable:
##              ------------------------------------
##                educ              log(wage)
##                 OLS       OLS    instrumental    OLS
##                                      variable
##                 (1)       (2)        (3)         (4)
## ------------------------------------------------
## nearc4       0.320***
##               (0.088)
##
## educ                    0.075***    0.132**
##                          (0.003)    (0.055)
##
## fitted(redf)                                     0.132**
##                                                  (0.057)
##
```

Beispiel II

Lohnregression

```
## exper          -0.413*** 0.085***    0.108***   0.108***
##                 (0.034)   (0.007)     (0.024)    (0.024)
##
## I(exper2)        0.001   -0.002***  -0.002***   -0.002***
##                 (0.002)  (0.0003)    (0.0003)   (0.0003)
##
## black          -0.936*** -0.199***  -0.147***   -0.147***
##                 (0.094)   (0.018)     (0.054)    (0.055)
##
## ------------------------------------------------------------
## Observations    3,010     3,010       3,010      3,010
## R2              0.477     0.300       0.238      0.195
## ============================================================
## Note:                          *p<0.1; **p<0.05; ***p<0.01
```

TSLS mit mehreren endogenen Regressoren

- Wiederhole einfach obige Prozedur für jeden einzelnen endogenen Regressor.
- Man erinnere sich dabei immer, dass Instrumente
  1. exogen und
  2. relevant sein müssen.

TSLS mit mehreren endogenen Regressoren

Die IV-Regressionsannahmen

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik} + \beta_{k+1} W_{i1} + \ldots + \beta_{k+r} W_{ir} + u_i.$$

1. $\mathrm{E}(u \mid W_1, \ldots, W_r) = 0$, d.h. "die exogenen Regressoren sind exogen".
2. $(Y_i, X_{i1}, \ldots, X_{ik}, W_{i1}, \ldots, W_{ir}, Z_{i1}, \ldots, Z_{im}, u_i)$ sind eine u.i.v. Stichprobe.
3. Die X's, W's und Z's haben endliche vierte Momente.
4. Die Instrumente $(Z_1, \ldots, Z_m)$ sind valide.

- Unter den IV-Regressionsannahmen 1–4 sind der TSLS-Schätzer und seine t-Statistik normalverteilt.
- (Man erinnere sich aber an das Problem mit den Standardfehlern.)

Prüfen der Validität von Z?

- Instrumente müssen relevant sein.
  - Prüfe die 1.-Schritt Schätzer auf ihre Signifikanz!
  - Wenn die 1.-Schrittregression X nicht erklärt, sind die Instrumente entweder "schwach" oder sogar irrelevant.
- Instrumente müssen exogen sein.
  - Dies kann für $m > k$ *teilweise* mit ökonometrischen Methoden geprüft werden, mit so genannten Tests für überidentifizierende Restriktionen, *aka* J-Test.
  - Man benötigt immer auch ökonomische Argumente, um Exogenität begründen zu können!
    - Z.B. "kein Effekt von Regen auf Nachfrage".
    - Z.B. "kein Effekt von Nähe zum College auf Fähigkeit".

Referenzen I

Card, D. 1995. "Using Geographic Variation in College Proximity to Estimate the Return to Schooling." In *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp*, edited by L. N. L. N. Christofides, E. K. Grant, and R. Swidinsky, 201–22. Toronto: University of Toronto Press.

Galton, F. 1886. "Regression Towards Mediocrity in Hereditary Stature." *The Journal of the Anthropological Institute of Great Britain and Ireland* 15: 246–63.