Inhaltsverzeichnis

4 OLS-Schätzfunktionen: deren Eigenschaften und Standardfehler 1

4.1 Einführung 1

4.1.1 Erwartungstreue, Effizienz und Konsistenz 1

4.2 Erwartungstreue von OLS Schätzfunktionen 4

4.2.1 Deterministische versus stochastische Regressoren 8

4.3 OLS Standardfehler 11

4.3.1 Ein einfaches Beispiel mit Mittelwerten 12

4.3.2 OLS Standardfehler für bivariate Regressionen 15

4.4 Gauss-Markov Theorem 26

4.5 Asymptotische Eigenschaften ('Große Stichprobeneigenschaften') . . . 28

4.5.1 Konsistenz 29

4.5.2 Beispiel: Beweis der Konsistenz des Stichprobenmittelwertes . 30

4.5.3 Beispiel: Unverzerrtheit und Konsistenz von OLS Schätzfunktionen mit stochastischen Regressoren 36

4.5.4 Asymptotische Normalverteilung 39

4.5.5 Asymptotische Effizienz 40

4.6 Der Mittlere Quadratische Fehler (*Mean Square Error*, MSE) 40

4.A Appendix 45

4.A.1 Eine Schätzfunktion für die Varianz der Störterme σ^2 45

4.A.2 Gauss-Markov Beweis 48

4.A.3 R Code 51

Kapitel 4

OLS-Schätzfunktionen: deren Eigenschaften und Standardfehler

"Die Mathematik ist eine Art Spielzeug, welches die Natur uns zuwarf zum Troste und zur Unterhaltung in der Finsternis."
(Jean le Rond d'Alembert, 1717 - 1783)

4.1 Einführung

Worum geht's? Wir haben im letzten Kapitel gesehen, wie wir die gemeinsame Verteilung mehrerer Zufallsvariablen durch eine PRF (*population regression function*) beschreiben können, oder in anderen Worten, durch eine lineare Approximation an die bedingte Erwartungswertfunktion (CEF, *conditional expectation function*). Unser Interesse gilt den Parametern dieser Funktion, z.B. dem Steigungskoeffizienten β_2. Diese Parameter sind unbeobachtbar, aber wir können eine Stichprobe aus der gemeinsamen Verteilung (dem datengenerierenden Prozess) beobachten. In Abbildung 4.1 haben wir intuitiv gezeigt, wie wir von diesen Stichproben-Beobachtungen sowohl zu einer empirischen *Schätzung* (z.B. b_2) als auch zu einer theoretischen *Schätzfunktion* (z.B. $\widehat{\beta}_2$) kommen.[1]

Eine Schätzfunktion ist eine Zufallsvariable und hat als solche eine Dichtefunktion, die wir *Stichprobenkennwertverteilung* (*sampling distribution*) nennen.

Um diese Stichprobenkennwertverteilung wird es in diesem Kapitel hauptsächlich gehen, vor allem um deren ersten beiden Momente, den Erwartungswert und die Varianz. Diese werden wir dann im nächsten Kapitel u.a. für Hypothesentests und Konfidenzintervalle benötigen.

4.1.1 Erwartungstreue, Effizienz und Konsistenz

In diesem Kapitel interessieren wir uns für mögliche Zusammenhänge zwischen Parametern der PRF (*population reression function*) und den Momenten der Stichprobenkennwertverteilung.

Konkret werden wir uns vor allem für drei Fragen interessieren

[1] Die aus der Stichprobe berechnete Schätzung interpretieren wir als eine Realisation der Schätzfunktion, die eine Zufallsvariable ist.

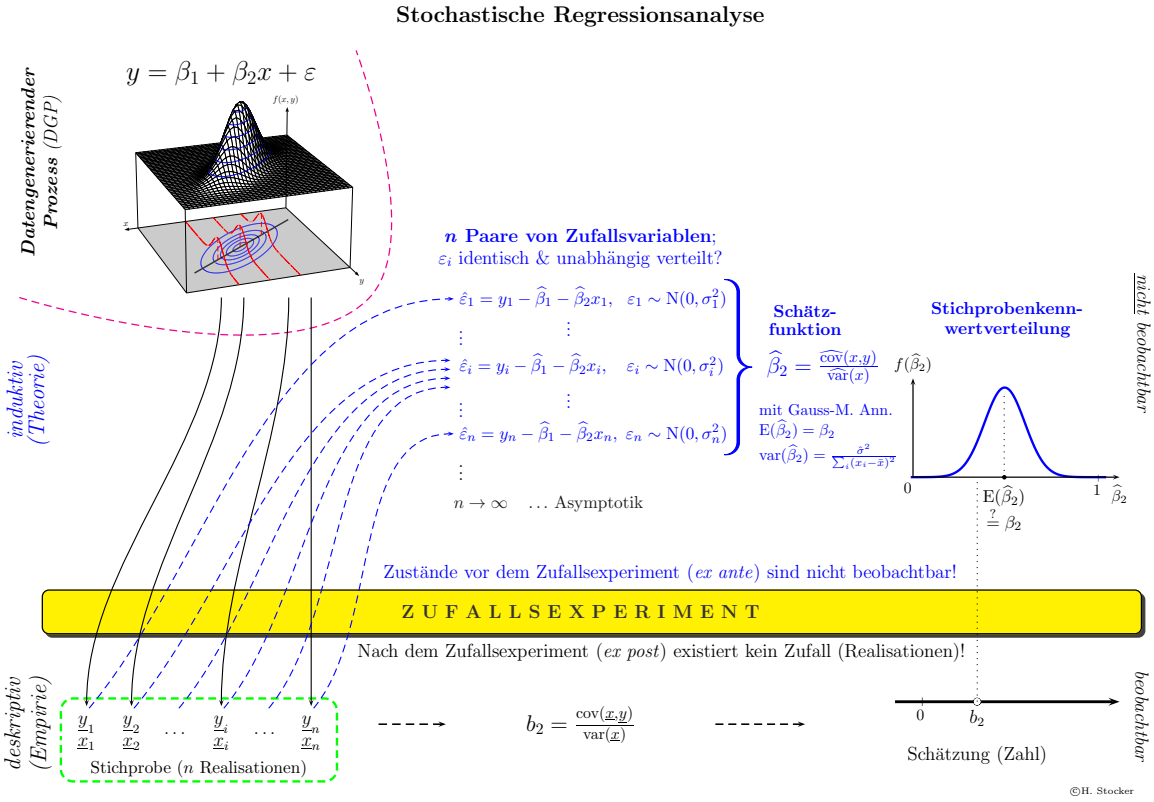

Abbildung 4.1: Vom datengenerierenden Prozess zur Stichprobenkennwertverteilung (vom letzten Kapitel übernommen)

1. liefert die OLS Schätzfunktion *im Durchschnitt* 'richtige' Resultate? Oder etwas präziser formuliert, unter welchen Bedingungen ist der Erwartungswert der Stichprobenkennwertverteilung gleich dem wahren Parameter der linearen CEF (*conditional expectation function*)? Dies ist die Frage nach der *Erwartungstreue* der OLS Schätzfunktion.

   Falls Schätzfunktionen nicht erwartungstreu sind, nennt man sie verzerrt, oder man sagt, sie haben einen *Bias*.

2. Wie 'genau' sind die Schätzungen mit Hilfe von OLS Schätzfunktionen? Oder wieder etwas präziser formuliert, falls die Bedingungen für die Erwartungstreue erfüllt sind, unter welchen zusätzlichen Bedingungen liefern die OLS Schätzfunktionen eine 'größtmögliche' Genauigkeit? Dies ist die Frage nach der *Effizienz* einer Schätzfunktion.

   Unser Indikator für Genauigkeit wird die Standardabweichung der Stichprobenkennwertverteilung sein, und die Standardabweichung einer Schätzfunktion werden wir in Zukunft *Standardfehler* nennen. Um obige Frage beantworten zu können, müssen wir zuerst eine Schätzfunktion für diese Standardfehler von OLS Schätzfunktionen entwickeln.

3. Schließlich werden wir der Frage nachgehen, unter welchen Bedingungen Schätzfunktionen *mit zunehmender Stichprobengröße* immer genauer werden.

   Diese Frage mag auf den ersten Blick etwas befremdlich erscheinen, aber wir werden später sehen, dass diese in der Ökonometrie eine extrem wichtige Rolle

spielt. Etwas salopp formuliert, Schätzfunktionen, die mit zunehmender Stichprobengröße immer genauere Ergebnisse liefern, nennt man *konsistent*. Um die Konsistenz von Schätzfunktioen wird es im letzten Abschnitt dieses Kapitels gehen.

Bevor wir uns an die Arbeit machen werden wir nochmals kurz diese Eigenschaften reflektieren, indem wir in Abbildung 4.2 Schätzfunktionen mit Schießgewehren vergleichen. Auf jede dieser drei Zielscheiben werde mit einem unterschiedlichen Gewehr geschossen, und wir vergleichen die Eigenschaften dieser drei Schießgewehre.

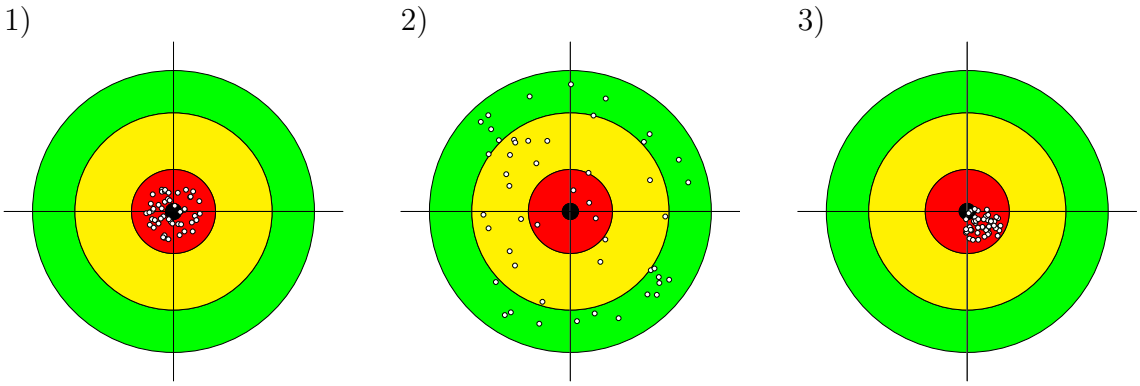

Abbildung 4.2: Eigenschaften von Schätzfunktionen, Vergleich mit Schießgewehr; 1) erwartungstreu und effizient, 2) erwartungstreu, aber nicht effizient, 3) verzerrt, also auch nicht effizient.

Die ersten beiden Gewehre (linke und mittlere Zielscheibe) treffen beide 'im Durchschnitt' richtig, beide Schätzfunktionen sind *erwartungstreu*. Aber offensichtlich trifft das erste Gewehr genauer als das zweite; wir sagen, die erste Schätzfunktion ist *im Verhältnis* zur zweiten *effizient*. Effizienz bezieht sich immer auf einen Vergleich. Später werden wir die Bedingungen ableiten, unter denen OLS Schätzfunktionen genauer sind als *alle anderen linearen und unverzerrten* Schätzfunktionen (dies ist der bekannte Gauss-Markov Beweis).

Das dritte Gewehr hat zwar einen sehr kleinen Streukreis, aber es schießt offensichtlich systematisch daneben (wenngleich in diesem Beispiel nur um ein bisschen). Solche Schätzfunktionen werden *verzerrt* (*'biased'*) genannt, und genießen in der Ökonometrie im Allgemeinen kein sehr hohes Ansehen.

Für die Beweise der *Erwartungstreue* und *Effizienz* wird die Stichprobengröße keine Rolle spielen, sie sollen für jede beliebige Stichprobengröße gelten, also *auch* für kleine Stichproben, und werden deshalb häufig *'Kleine Stichprobeneigenschaften'* genannt.

Im Unterschied dazu wird bei der dritten Eigenschaft der *Konsistenz* untersucht, ob die Genauigkeit der Schätzfunktion bei einer *Vergrößerung* der Stichprobe zunimmt. Für diesen Beweis benötigen wir eine asymptotische Analyse, d.h. wir untersuchen wie sich die Treffsicherheit verändert, wenn die Stichprobengröße gegen Unendlich geht. Deshalb wird die Eigenschaft der Konsistenz eine *asymptotische* Eigenschaft genannt.

Die Eigenschaft der Konsistenz lässt sich im Beispiel mit dem Schießgewehr nicht so gut darstellen, aber von einem Schießgewehr würden wir erwarten, dass dessen Genauigkeit mit abnehmendem Abstand zur Zielscheibe zunimmt. In einem etwas schiefen Vergleich würden wir von einer guten Schätzfunktion erwarten, dass die Genauigkeit mit *zunehmender Stichprobengröße* zunimmt. Diese Eigenschaft der *Konsistenz* lässt sich häufig unter deutlich weniger restriktiven Bedingungen beweisen als die ersten beiden Eigenschaften der Erwartungstreue und Effizienz. Deshalb spielt die Konsistenz in der Ökonometrie eine zentrale Rolle.

Neben diesen drei Eigenschaften der Erwartungstreue, Effizienz und Konsistenz gibt es einige weitere Eigenschaften von Schätzfunktionen (z.B. Suffizienz, asymptotische Effizienz, asymptotische Erwartungstreue), die wir hier allerdings nicht weiter diskutieren werden.

4.2 Erwartungstreue von OLS Schätzfunktionen

Zuerst wollen wir untersuchen, ob OLS Schätzfunktionen den wahren Parameter einer PRF zumindest *im Durchschnitt* richtig treffen, z.B. für den Steigungskoeffizienten[2]

$$\mathrm{E}(\widehat{\beta}_2) \stackrel{?}{=} \beta_2$$

Analoges gilt natürlich auch für andere Parameter, z.B. die Varianz der Störterme σ^2 (d.h. $\mathrm{E}(\hat{\sigma}^2) \stackrel{?}{=} \sigma^2$), aber wir werden uns vorerst auf die OLS Koeffizienten konzentrieren.

Diese Frage scheint auf den ersten Moment unbeantwortbar, erinnern wir uns an Abbildung 4.1, weder die Parameter der PRF noch die Stichprobenkennwertverteilungen der entsprechenden Schätzfunktionen sind beobachtbar, wie sollen wir diese also vergleichen können?

Genaueres Nachdenken zeigt allerdings, dass wir die Frage nur etwas präziser stellen müssen um zu fruchtbaren Schlussfolgerungen zu gelangen. Anstelle zu fragen *ob* dies gilt, sollten wir uns besser fragen, *welche Bedingungen erfüllt sein müssen*, damit dies gilt.

Welche Beziehung herrscht zwischen dem wahren Parameter der PRF (Grundgesamtheit) und dem Erwartungswert der entsprechenden Schätzfunktion, wie können wir diese modellieren?

Wie schon erwähnt wollen wir die Überlegungen am Beispiel des OLS Steigungskoeffizienten erläutern. Die OLS Schätzfunktion dafür ist

$$\widehat{\beta}_2 = \frac{\widehat{\mathrm{cov}}(x, y)}{\widehat{\mathrm{var}}(x)}$$

In dieser Schätzfunktion kommt der wahre Parameter β_2 nicht vor. Aber wenn die PRF eine korrekte Beschreibung des datengenerierenden Prozesses (DGP) darstellt, dann können wir einfach die

$$\text{PRF:}\quad y = \beta_1 + \beta_2 x + \varepsilon$$

[2]Wir erinnern uns, $\widehat{\beta}_2$ ist eine Zufallsvariable und hat eine Stichprobenkennwertverteilung (ein Spezialfall einer Dichtefunktion), und $\mathrm{E}(\widehat{\beta}_2)$ ist das erste Moment dieser Verteilung, also eine fixe, aber unbeobachtbare Zahl!

in die OLS Schätzfunktion einsetzen, und wir erhalten einen Zusammenhang zwischen $\widehat{\beta}_2$ und β_2

$$\widehat{\beta}_2 = \frac{\widehat{\text{cov}}\left(x, (\overbrace{\beta_1 + \beta_2 x + \varepsilon}^{y})\right)}{\widehat{\text{var}}(x)}$$

Damit wir so vorgehen können, müssen zwei Bedingungen erfüllt sein: erstens muss die PRF den datengenerierenden Prozess (DGP) korrekt beschreiben, und zweitens muss für die OLS Schätzfunktion eine eindeutige Lösung existieren!

Im Beispiel mit dem Steigungskoeffizienten ist z.B. offensichtlich, dass für $\widehat{\text{var}}(x) = 0$ keine eindeutige Lösung existiert, weil $\widehat{\beta}_2 = \widehat{\text{cov}}(x, y)/\widehat{\text{var}}(x)$.[3]

Man kann allgemeiner zeigen, dass für die Existenz einer eindeutigen Lösung die Anzahl der Beobachtungen n mindestens so groß sein muss wie die Anzahl der zu schätzenden Koeffizienten k (d.h. $n \geq k$), und dass darüber hinaus keine exakte lineare Abhängigkeit zwischen den Regressoren existieren darf (d.h. kein Regressor darf als Linearkombination der restlichen Regressoren darstellbar sein; näheres dazu im Kapitel zur Matrixschreibweise).

Darüber hinaus wollen wir lediglich zur Vereinfachung annehmen, dass die Elemente der x Variable deterministisch (also keine Zufallsvariablen) sind. Was dies bedeutet und welche Implikationen dies hat werden wir gleich erläutern, im Moment halten wir nur fest, dass die Darstellung für deterministische x deutlich einfacher ist, und dass die meisten der folgenden Aussagen unter einem etwas erweiterten Annahmenset auch für stochastische x gelten.

Um die Erwartungstreue von $\widehat{\beta}_2$ für deterministische x zu beweisen benötigen wir lediglich ein paar einfache Rechenregeln für Kovarianzen. Wir setzen in die Schätzfunktion $\widehat{\beta}_2$ für y die PRF ein

$$\begin{aligned}
\widehat{\beta}_2 &= \frac{\widehat{\text{cov}}(x, y)}{\widehat{\text{var}}(x)} = \frac{\widehat{\text{cov}}\left[x, (\beta_1 + \beta_2 x + \varepsilon)\right]}{\widehat{\text{var}}(x)} \\
&= \frac{1}{\widehat{\text{var}}(x)} \left[\underbrace{\widehat{\text{cov}}(x, \beta_1)}_{=0} + \beta_2 \underbrace{\widehat{\text{cov}}(x, x)}_{=\widehat{\text{var}}(x)} + \widehat{\text{cov}}(x, \varepsilon) \right] \\
&= \beta_2 + \frac{\widehat{\text{cov}}(x, \varepsilon)}{\widehat{\text{var}}(x)}
\end{aligned}$$

und bilden von beiden Seiten den Erwartungswert

$$\text{E}(\widehat{\beta}_2) = \beta_2 + \text{E}\left(\frac{\widehat{\text{cov}}(x, \varepsilon)}{\widehat{\text{var}}(x)}\right) \tag{4.1}$$

Die gibt uns sofort die notwendige Bedingung für die Erwartungstreue des OLS Steigungskoeffizienten $\text{E}(\widehat{\beta}_2) = \beta_2$, nämlich

$$\text{E}(\widehat{\beta}_2) = \beta_2 \quad \text{wenn und nur wenn} \quad \text{E}(\widehat{\text{cov}}(x, \varepsilon)) = 0$$

[3] Wenn $\widehat{\text{var}}(x) = 0$ ist x eine Konstante, und deshalb ein Vielfaches der Regressionskonstante. Wir werden später sehen, dass dies ein Spezialfall *perfekter Multikollinearität* darstellt.

Im multiplen Regressionsmodell muss diese Bedingung für alle Regressoren gelten, wie wir im Kapitel zur Matrixnotation des OLS Modells zeigen werden.

Allerdings haben wir schon vorher zwei Bedingungen benötigt, deshalb benötigen wir *drei* notwendige Bedingungen für die Erwartungstreue des OLS Steigungskoeffizienten. Dies sind auch die ersten drei der insgesamt vier *Gauss-Markov Bedingungen*[4], die uns im Folgenden ständig begleiten werden. Deshalb wiederholen wir sie jetzt etwas ausführlicher auch für den Fall multipler Regressionen:

A1 Linearität der PRF: Die PRF ist linear in den Parametern

$$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

und deren systematischer Teil kann als lineare Approximation an die bedingte Erwartungswertfunktion (CEF, $\mathrm{E}(y|x_1, \ldots, x_k)$) interpretiert werden.

A2 Es existiert keine exakte lineare Abhängigkeit zwischen den Regressoren (x Variablen), und $n \geq k$. Falls diese Annahme verletzt ist existiert keine eindeutige OLS Schätzfunktion.

Um den Grund zu erkennen, warum bei einer exakten linearen Abhängigkeit zwischen den Regressoren keine eindeutige Lösung existiert, sehen wir uns eine einfache Regression mit zwei Dummyvariablen für das Geschlecht an, w für weiblich und m für männlich

$$y_i = \beta_1 + \beta_2 w_i + \beta_3 m_i + \beta_4 x_i + \varepsilon_i$$

wobei x_i ein beliebiger Regressor ist. Man beachte, dass für alle $i = 1, \ldots, n$ gilt $w_i + m_i = 1$, die Summe der beiden Dummyvariablen ist also gleich der Regressionskonstanten, weshalb eine exakte lineare Abhängigkeit zwischen den Regressoren existiert (*Dummyvariablenfalle*).

Stellen wir uns nun vor, wir addieren zum Interzept eine beliebige konstante Zahl c und subtrahieren die gleiche Zahl c von den beiden Koeffizienten der Dummyvariablen

$$\begin{aligned} y_i &= (\beta_1 + c) + (\beta_2 - c)w_i + (\beta_3 - c)m_i + \beta_4 x_i + \varepsilon_i \\ &= c(1 - w_i - m_i) + \beta_1 + \beta_2 w_i + \beta_3 m_i + \beta_4 x_i + \varepsilon_i \\ &= \beta_1 + \beta_2 w_i + \beta_3 m_i + \beta_4 x_i + \varepsilon_i \end{aligned}$$

Die Koeffizienten sind also nicht eindeutig bestimmt, es existieren unendlich viele Lösungen und die Koeffizienten können nicht geschätzt werden. Man beachte, dass für dieses Modell auch der Koeffizient von x, β_4, nicht geschätzt werden kann![5]

Dies ist ein Spezialfall eines Identifikationsproblems, deshalb wird diese Annahme manchmal auch *Identifikationsbedingung* genannt.

[4] Die Darstellung und Reihenfolge der einzelnen Bedingungen unterscheidet sich zwischen Lehrbüchern.

[5] Aber natürlich können alle Koeffizienten des Modells $y_i = \beta_1 + \beta_2 w_i + \beta_3 x_i + \varepsilon_i$ problemlos geschätzt werden.

A3 Die Störterme ε sind linear unabhängig von den Regressoren

$$\mathrm{E}(\widehat{\mathrm{cov}}(x, \varepsilon)) = 0$$

Wir werden diese Bedingung im Folgenden meist sogar noch etwas strenger formulieren, nämlich dass die Störterme ε und die Regressoren *stochastisch unabhängig* sein sollen

$$\mathrm{E}(\varepsilon_i | x_1, \ldots, x_n) = \mathrm{E}(\varepsilon_i) = 0$$

oder etwas ausführlicher

$$\text{Alle } (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n) \text{ sind stochast. unabhängig von allen } (x_1, x_2, \ldots, x_n)$$

Diese etwas strengere Annahme ist insbesondere für stochastische Regressoren von Bedeutung.

Hier ist wichtig zu betonen, dass sich diese Annahme auf die unbeobachtbaren Störterme der Grundgesamtheit bezieht, *nicht* auf die Residuen!

Für die Residuen ist diese Bedingung aufgrund der Bedingungen erster Ordnung (also per Konstruktion) immer erfüllt, aber die Bedingungen erster Ordnung gelten nur für die Residuen, *nicht* notwendigerweise für die Störterme!

Wann immer diese Annahme verletzt ist liefert die OLS Schätzfunktion *verzerrte* Ergebnisse!!!

Im wesentlichen verlangt diese Annahme, dass die Störterme keine verwertbare Informationen der Regressoren enthalten. Stellen wir uns hypothetisch vor, wir würden die x auf die *Störterme* regressieren

$$\varepsilon_i = \gamma_1 + \gamma_2 x_i + u_i \quad \rightarrow \quad \gamma_2 = \frac{\mathrm{cov}(x, \varepsilon)}{\mathrm{var}(x)}$$

(u_i der Störterm dieser Regression). Der Steigungskoeffizient γ_2 ist nur Null, wenn $\mathrm{cov}(x, \varepsilon) = 0$, also wenn die x keinen Erklärungsbeitrag für die ε leisten!

Sollte die Annahme $\mathrm{cov}(x, \varepsilon) = 0$ verletzt sein können wir uns vorstellen, dass die Störterme ε eine Funktion der x sind, also $\varepsilon(x)$.

Der marginale Effekt von

$$y = \beta_1 + \beta_2 x + \varepsilon(x)$$

ist einfach die Ableitung

$$\frac{dy}{dx} = \beta_2 + \frac{d\varepsilon}{dx}$$

Wir interessieren uns für den Wert β_2, aber eine Regression liefert uns nur *die Summe* $\beta_2 + \frac{d\varepsilon}{dx}$, aber daraus können wir den interessierenden Koeffizienten β_2 nicht isolieren!

Dies ist ein weiteres Beispiel für ein *Identifikationsproblem*. Wir halten fest: wann immer irgend eine Form von Abhängigkeit zwischen Störtermen und Regressoren besteht erhalten wir systematisch verzerrte Ergebnisse!

Eine solche Abhängigkeit zwischen Störtermen und Regressoren wird in der Ökonometrie *Endogenität* genannt (oder genauer, *endogene Regressoren*). Man beachte, dass sich diese Definition von Endogenität etwas von dem z.B. in der Mikroökonomik gebräuchlichen Endogenitätsbegriff unterscheidet (wenngleich mikroökonomische Endogenität häufig zu ökonometrischer Endogenität führt).

Endogenität ist eines der Kernprobleme der Ökonometrie, und wird uns später im Kapitel zu *Kausalität und endogene Regressoren* noch ausführlich beschäftigen!

Hier sei nur vorausgeschickt, dass eine Verletzung dieser Annahme eine *kausale Interpretation* von Regressionsergebnissen verunmöglicht, und dass diese Annahme leider ziemlich häufig verletzt ist.

Die wichtigsten Fälle, die zu einer stochastischen Abhängigkeit zwischen Störtermen und Regressoren – und deshalb zu verzerrten Schätzungen der Koeffizienten – führen, sind:

1. fehlenden relevanten Regressoren (*omitted variables*), mit den Spezialfällen von unbeobachteter Heterogenität und Selektionsproblemen,
2. Simultaner Abhängigkeit in interdependenten Systemen (*feed-back Mechanismen*). Im Kern tritt dieses Problem immer auf, wenn zur Beschreibung eines Systems mehr als eine Gleichung benötigt wird, und diese Gleichungen interdependent sind (z.B. Angebots- und Nachfragefunktion). Analog zum ersten Problem könnte man auch von einem *omitted equation bias* sprechen.
3. Messfehler in den erklärenden Variablen. Alle diese Fälle werden wir später natürlich ausführlich diskutieren!

Die vierte Gauss-Markov Annahme benötigen wir schließlich erst später für den Beweis der *Effizienz* von OLS Schätzfunktionen, der Vollständigkeit halber sei sie hier vorausgeschickt; sie betrifft die Störterme der PRF

A4 Die Störterme ε_i sind *identisch und unabhängig* verteilt, kurz geschrieben als $\varepsilon_i \sim \text{i.i.d.}(0, \sigma^2)$. Dies bedeutet, dass die Störterme weder heteroskedastisch noch autokorreliert sein dürfen.

Hier sei nur angemerkt, dass wir diese Annahme nicht für den Beweis der Erwartungstreue benötigt haben, diese Annahme werden wir erst für die Berechnung der OLS Standardfehler und für den Beweis der Effizienz (Gauss-Markov Theorem) benötigen. Näheres dazu folgt im nächsten Abschnitt.

Alle diese Annahmen werden wir später noch ausführlicher diskutieren, vorerst wollen wir aber noch einmal auf den Unterschied zwischen deterministischen und stochastischen Regressoren zurück kommen.

4.2.1 Deterministische versus stochastische Regressoren

Um den Unterschied zwischen deterministischen und stochastischen Regressoren zu verstehen ist es zweckmäßig sich ein klassisches Experiment vorzustellen, z.B. die

berühmten 'Feldexperimente' von R.A. Fisher. Dabei wird eine kontrollierte Menge von Dünger auf verschiedene Versuchsflächen ausgebracht und untersucht, wie der Ertrag (von z.B. von Kartoffeln) von dieser *kontrolliert* ausgebrachten Düngermenge Dünger abhängt. Der Experimentator wählt und fixiert die Menge, deshalb gibt es in Bezug auf die eingesetzte Menge kein Element der Unsicherheit. Für verschiedene, aber jeweils fixe Mengen von x sammelt er jeweils wiederholte Beobachtungen von y.

Dies kann man anhand der vom letzten Kapitel übernommenen Abbildung 4.2.1 vor Augen führen.

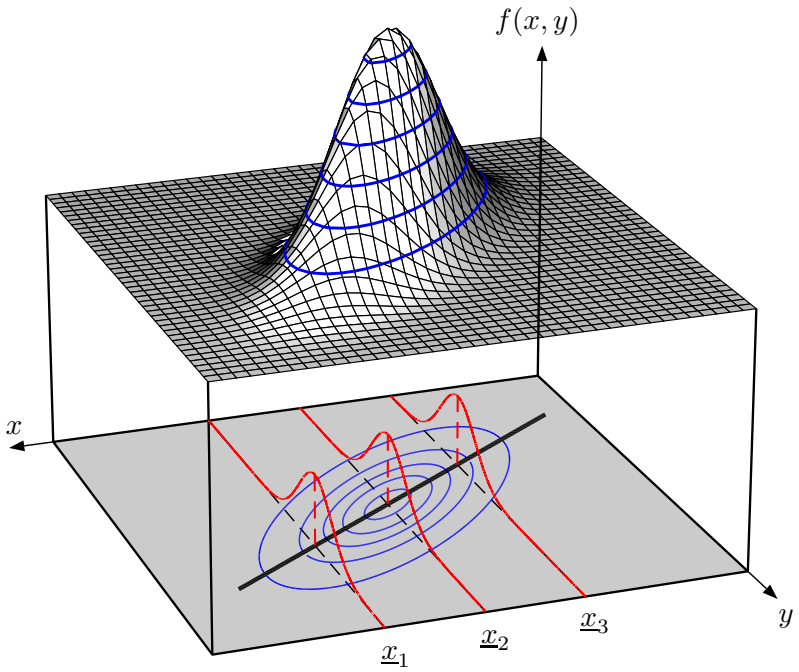

Abbildung 4.3: Datengenerierender Prozess und PRF (aus dem letzten Kapitel übernommen)

Bei deterministischen Regressoren stellen wir uns vor, dass der Regressor x, z.B. die Düngermenge, bei bestimmten Werten fixiert wird, z.B. in Abbildung 4.2.1 bei den Mengen $\underline{x}_1, \underline{x}_2$ und $\underline{x}_3$ (die x sind wieder unterstrichen um anzudeuten, dass dies fixe Zahlen sind). Auch bei einer fest vorgegebenen Düngermenge wird der (Kartoffel-)Ertrag aufgrund anderer Einflüsse und von Zufallsereignissen etc. schwanken; diese Unsicherheit wird durch die (rot eingezeichneten) bedingten Dichtefunktionen der Störterme ε_i abgebildet.

Die Menge wird vom Experimentator fix vorgegeben, dies ist gemeint, wenn man von *'fixed in repeated sampling'* spricht, und dies ist auch der Grund, warum Regressoren oft *Kontrollvariablen* genannt werden. Für experimentelle Untersuchungen ist diese Vorstellung durchaus vernünftig und angebracht.

Allerdings müssen wir auch dabei annehmen, dass der Experimentator wirklich exogen handelt, d.h. zum Beispiel, dass er seine Aktionen nicht in Abhängigkeit vom Experiment anpasst. In der Sprache der Ökonometrikerinnen sagt man, die Regressoren sollten *'as good as randomly assigned'* sein. Modernere Versuchsprotokolle verlangen häufig auch bei Experimenten eine randomisierte Wahl der vorgegebenen Mengen.

In sozialwissenschaftlichen Zusammenhängen haben Forscher nur in Ausnahmefällen die Möglichkeit ihre Regressoren zu 'kontrollieren', vielmehr werden Paare von x und y beobachtet, oder genauer, es werden Stichproben von $(\underline{x}_i, \underline{y}_i)$ Paaren gezogen.

In diesem Fall sind die x nicht mehr *'fixed in repeated sampling'* (also deterministisch), sondern sind ebenso wie die y Zufallsvariablen, d.h. stochastisch.

Für den Fall mit stochastischen x benötigen wir zwei weitere Annahmen, nämlich

AS 1: Die (x_i, y_i) Paare für $i = 1, \ldots n$ sind *identisch* und *unabhängig* verteilt. Dies ist eine Annahme über das Verfahren der Stichprobenziehungen: wenn der DGP eine echte *Zufallsstichprobe* liefert sollte die Verteilung jedes (x_i, y_i) Paares der Verteilung in der Grundgesamtheit entsprechen.

Die Annahme einer echten Zufallsstichprobe ist natürlich ziemlich restriktiv, echte Zufallsstichproben existieren hauptsächlich in Lehr- und Märchenbüchern!

AS 2: Große Ausreißer sind unwahrscheinlich: diese Annahme soll verhindern, dass auch in einer sehr großen Stichprobe eine einzelne Beobachtung das Ergebnis determiniert.

Technisch wird dies ausgedrückt über die vierten Momente (Kurtosis) der Zufallsvariablen: $0 < \mathrm{E}(x_i^4) < \infty$ und $0 < \mathrm{E}(y_i^4) < \infty$. Diese Annahme dürfte in den meisten Fällen weniger restriktiv sein wie die vorhergehende Annahme einer echten Zufallsziehung, sie wird v.a. für die Herleitung der asymptotischen Eigenschaften von OLS Schätzfunktionen benötigt.

Der Beweis für die Erwartungstreue von OLS Schätzfunktionen mit *stochastischen Regressoren* funktioniert dann ziemlich ähnlich wie für deterministische Regressoren, nur dass wir die Regressoren nicht physisch 'kontrollieren' können, sondern sie implizit durch Konditionierung, d.h. die Bildung *bedingter* Erwartungswerte fixieren. Wenn es um *Kausalitätsfragen* geht ist dies natürlich etwas völlig anderes als die tatsächliche Kontrolle durch einen Experimentator, und deshalb eignen sich Regressionen alleine im allgemeinen *nicht* für Kausalaussagen, aber für die Herleitung der allgemeinen Bedingungen für die Erwartungstreue reicht dies.

Allerdings benötigen wir dazu *bedingte Erwartungswerte in Bezug auf Zufallsvariablen*, z.B. $\mathrm{E}(\varepsilon_i|x_1, x_2, \ldots, x_n)$ wobei jedes einzelne x_i eine Zufallsvariable ist. Die Mathematik dahinter ist deutlich aufwändiger, aber Mathematiker haben bewiesen, dass eine *stochastische bedingte Erwartungswertfunktion* unter wenig restriktiven Bedingungen existiert und eindeutig ist.[6]

In diesem Fall sind auch die bedingten Erwartungswerte Zufallsvariablen, aber da auch für stochastische Regressoren das Gesetz iterierten Erwartungen gilt kann man daraus den unbedingten Erwartungswert bestimmen.

Die Kernannahme für die Erwartungstreue ist auch für stochastische x_i, dass die Störterme ε_i stochastisch unabhhängig von *allen* x_i sind, d.h. die Annahme A3

$$\mathrm{E}(\varepsilon_i|x_1, \ldots, x_n) = 0$$

[6] Für diesen Beweis mit Hilfe des Radon–Nikodym Theorems benötigt man u.a. Maßtheorie.

Man beachte, dass dies für alle ε_i mit $i = 1, \ldots, n$ gelten muss (im multiplen Regressionsmodell muss ε_i stochastisch unabhängig von allen x_{jh} mit $j = 1, \ldots n$ und $h = 1, \ldots, k$ sein)!
Falls die Annahme AS 1 erfüllt ist und die (x_i, y_i) identisch und unabhängig verteilt sind gilt $\mathrm{E}(\varepsilon_i|x_1, \ldots, x_n) = \mathrm{E}(\varepsilon_i|x_i)$
Für Gleichung 4.1 (Seite 5) schreiben wir also

$$\begin{aligned}
\mathrm{E}(\widehat{\beta}_2) &= \beta_2 + \mathrm{E}\left[\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \\
&= \beta_2 + \mathrm{E}\left[\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\,\mathrm{E}(\varepsilon_i|x_1, \ldots, x_n)}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \\
&= \beta_2
\end{aligned}$$

wenn $\mathrm{E}(\varepsilon_i|x_i) = 0$!

Falls $\mathrm{E}(\widehat{\beta}_2) = \beta_2$ kann die Erwartungstreue des Interzepts $\widehat{\beta}_1 = \bar{y} - \widehat{\beta}_2\bar{x}$ einfach gezeigt werden, wir setzen wieder die PRF für y in der Schätzfunktion $\widehat{\beta}_1$ ein

$$\mathrm{E}(\widehat{\beta}_1) = \mathrm{E}\left[(\beta_1 + \beta_2\bar{x}) - \widehat{\beta}_2\bar{x}\right] = \beta_1 + \beta_2\bar{x} - \beta_2\bar{x} = \beta_1$$

Wir halten also fest: die Kernannahme für die Erwartungstreue der OLS Schätzfunktionen $\widehat{\beta}_2$ und $\widehat{\beta}_1$ ist auch für stochastische Regressoren, dass alle ε_i stochastisch unabhängig von allen x_i (für $i = 1, \ldots, n$) sind! Fälle, in denen diese Annahme verletzt ist, werden uns später noch intensiv beschäftigen.

Kehren wir nochmals zurück zu Abbildung 4.2 (Seite 3) mit den Zielscheiben und Schießgewehren. Das Kriterium der Erwartungstreue erlaubt uns eine Entscheidung zwischen dem ersten und dritten Schießgewehr, im allgemeinen bevorzugen Ökonometrikerinnen erwartungstreue Schätzfunktionen (linkes Panel) gegenüber verzerrten Schätzfunktionen (rechtes Panel).
Aber dieses Kriterium ermöglicht uns keine Entscheidung zwischen dem ersten und zweiten Schießgewehr, sowohl die erste als auch die hinter der zweiten Zielscheibe (mittleres Panel) stehende Schätzfunktion sind erwartungstreu.
Um zwischen diesen beiden entscheiden zu können benötigen wir ein weiteres Kriterium, die *Effizienz* (Varianzminimalität).
Doch um diese zeigen zu können benötigen wir zuerst eine Schätzfunktion für die Standardabweichung der Stichprobenkennwertverteilung von $\widehat{\beta}_2$, das heißt, die OLS Standardfehler.

4.3 OLS Standardfehler

Bevor wir die OLS Standardfehler herleiten werden wir die grundlegenden Überlegungen anhand eines viel einfacheren Beispiels erläutern, nämlich anhand eines Mittelwertes über drei Beobachtungen. Wir werden später sehen, dass diese Überlegungen genauso für die OLS Standardfehler (d.h. die Standardabweichung der OLS Stichprobenkennwertverteilung) gelten.

4.3.1 Ein einfaches Beispiel mit Mittelwerten

Angenommen wir beobachten drei Zahlen, z.B. $\{5, 1, 3\}$; das könnten z.B. Antworten vom Likert-Typ eines Fragebogens sein[7], die Gewichte in Gramm einer gerade neu entdeckten Insektenart, oder das Resultat eines Würfelspiels. In der deskriptiven Statistik würden wir als *Kennwerte* dieser drei Beobachtungen z.B. den Mittelwert 3 und die Varianz 8/3 berechnen.

In der induktiven Statistik liegt unser Fokus anders, wir interessieren uns für die zugrunde liegenden Grundgesamtheit (d.h. den datengenerierenden Prozess) und dessen Parameter (z.B. Mittelwert μ und Varianz σ^2), z.B. die durchschnittliche Meinung der Befragten oder das unbekannte Durchschnittsgewicht der neuen Insektenart und dessen Streuung.

Wie früher schon ausgeführt stellen wir uns hinter jeder dieser drei Beobachtungen eine Zufallsvariable vor, z.B. y_1, y_2 und y_3, und interpretieren die drei beobachteten Zahlen als *Realisationen* dieser drei Zufallsvariablen.

Zufallsvariablen sind durch Dichtefunktionen charakterisiert und diese durch deren Momente. Wir wollen annehmen, dass die ersten beiden Momente, Erwartungswert μ und Varianz σ^2, existieren.

Wenn wir nichts über den datengenerierenden Prozess wissen müssen wir davon ausgehen, dass alle drei Zufallsvariablen unterschiedliche Erwartungswerte μ_1, μ_2, μ_3 und eine unterschiedliche Varianz $\sigma_1^2, \sigma_2^2, \sigma_3^2$ haben.

Wir interessieren uns für den Mittelwert in der unbeobachtbaren Grundgesamtheit, und als mögliche Schätzfunktionen betrachten wir zwei unterschiedlich gewichtete Summen dieser drei Zufallsvariablen (auch der Mittelwert der deskriptiven Statistik ist eine gewichtete Summe, allerdings der Realisationen).

Die zwei konkreten Schätzfunktionen sind

1. $$\hat{\theta} = \frac{1}{2}\, y_1 + \frac{1}{3}\, y_2 + \frac{1}{6} y_3$$
2. $$\widehat{\psi} = \frac{1}{3}\, y_1 + \frac{1}{3}\, y_2 + \frac{1}{3} y_3$$

Beide Schätzfunktionen sind gewichtete Summen, nur deren Gewichte unterscheiden sich; im ersten Fall $\hat{\theta}$ (gesprochen theta Dach) sind die Gewichte unterschiedlich groß, im zweiten Fall $\widehat{\psi}$ (gesprochen psi Dach) sind die Gewichte für alle drei Beobachtungen gleich groß $(1/n)$.

Erwartungstreue

Welche der beiden Schätzfunktionen ist 'besser'? Wir können den Erwartungswert bilden um die Erwartungstreue zu überprüfen, und da $\mathrm{E}(y_i) = \mu_i$ (mit $i = 1, 2, 3$)

[7] Natürlich sind Likert Fragen ordinal skaliert, weshalb der Mittelwert in diesem Fall keine sehr geeignete Kennzahl ist, aber das spielt im Moment keine Rolle.

können wir schreiben

$$\begin{aligned}
\mathrm{E}(\hat{\theta}) &= \frac{1}{2}\,\mathrm{E}(y_1) + \frac{1}{3}\,\mathrm{E}(y_2) + \frac{1}{6}\,\mathrm{E}(y_3) = \frac{1}{2}\,\mu_1 + \frac{1}{3}\,\mu_2 + \frac{1}{6}\mu_3 \\
\mathrm{E}(\widehat{\psi}) &= \frac{1}{3}\,\mathrm{E}(y_1) + \frac{1}{3}\,\mathrm{E}(y_2) + \frac{1}{3}\,\mathrm{E}(y_3) = \frac{1}{3}\,\mu_1 + \frac{1}{3}\,\mu_2 + \frac{1}{3}\mu_3
\end{aligned}$$

Damit sind wir offensichtlich nicht viel klüger geworden. Um hier weiter zu kommen benötigen wir zusätzliches Wissen, welches wir in Form einer *Annahme* über den *sampling* Prozess verwenden können.

In diesem Fall vermuten wir, dass die drei Beobachtungen durch den selben datengenerierenden Prozess erzeugt wurden, z.B. Realisationen beim Würfeln oder Zufallsstichprobenziehungen aus der gleichen Grundgesamtheit sind. In diesem Fall ist es vernünftig anzunehmen, dass alle drei Zufallsvariablen y_i den gleichen Erwartungswert μ haben, also den (unbekannten) Mittelwert der Grundgesamtheit.

Mit dieser Annahme $\mathrm{E}(y_1) = \mathrm{E}(y_2) = \mathrm{E}(y_3) = \mu$ erhalten wir

$$\begin{aligned}
\mathrm{E}(\hat{\theta}) &= \frac{1}{2}\,\mu + \frac{1}{3}\,\mu + \frac{1}{6}\mu = \mu \\
\mathrm{E}(\widehat{\psi}) &= \frac{1}{3}\,\mu + \frac{1}{3}\,\mu + \frac{1}{3}\mu = \mu
\end{aligned}$$

In diesem Fall sind also beide Schätzfunktionen *erwartungstreu*, und es ist nicht schwer zu erkennen, dass in diesem einfachen Fall alle Schätzfunktionen, deren Gewichte sich auf Eins ergänzen, erwartungstreu sind.

Man beachte, dass die Annahme $\mathrm{E}(y_i) = \mu$ voraussetzt, dass es sich um eine echte Zufallsstichprobe handelte; im Fall von *Selektionsproblemen* oder *unbeobachteter Heterogenität* (d.h. wenn sich die Beobachtungen durch unbekannte Charakteristika wesentlich unterscheiden) können die daraus gezogenen Schlussfolgerungen sehr irreführend sein.

Effizienz

Für welche der beiden Schätzfunktionen sollen wir uns also entscheiden? Ein naheliegendes Kriterium wäre, die *genauere* Schätzfunktion zu wählen, also die Schätzfunktion mit der kleineren Varianz.

Die Varianz der Schätzfunktion $\hat{\theta}$ ist[8]

$$\begin{aligned}
\mathrm{var}(\hat{\theta}) &= \mathrm{var}\left(\frac{1}{2}\,y_1 + \frac{1}{3}\,y_2 + \frac{1}{6}\,y_3\right) \\
&= \frac{1}{4}\,\mathrm{var}(y_1) + \frac{1}{9}\,\mathrm{var}(y_2) + \frac{1}{36}\,\mathrm{var}(y_3) + \\
&\quad \frac{2}{2\times 3}\,\mathrm{cov}(y_1, y_2) + \frac{2}{2\times 6}\,\mathrm{cov}(y_1, y_3) + \frac{2}{3\times 6}\,\mathrm{cov}(y_2, y_3)
\end{aligned}$$

[8]Erinnern Sie sich an $(a+b+c)^2 = a^2 + b^2 + c^2 + 2ab + 2ac + 2bc$ die Varianz ist das zweite Moment, also quadratisch.

Dies sieht ziemlich unappetitlich aus, aber wir können wieder überlegen, wie sich Annahmen über den datengenerierenden Prozess auswirken würden.

Wenn es sich um eine echte Zufallstichprobe handelt würden wir erwarten, dass alle drei Zufallsvariablen die gleiche Varianz σ^2 haben, d.h. wir *nehmen an*, dass $\text{var}(y_1) = \text{var}(y_2) = \text{var}(y_3) = \sigma^2$.

Wenn alle Zufallsvariablen die gleiche Verteilung haben dann sind auch die ersten beiden Momente gleich, wir sagen, die Zufallsvariablen sind *identisch* verteilt

$$\begin{aligned}\text{var}(\hat{\theta}) &= \left(\frac{1}{4}+\frac{1}{9}+\frac{1}{36}\right)\sigma^2 + \\ &\quad \frac{2}{2\times 3}\,\text{cov}(y_1,y_2) + \frac{2}{2\times 6}\,\text{cov}(y_1,y_3) + \frac{2}{3\times 6}\,\text{cov}(y_2,y_3)\end{aligned}$$

Auch wenn die Zufallsvariablen identisch verteilt sind können die Kovarianzen ungleich Null sein. Wenn z.B. aus einer Urne *ohne Zurücklegen* gezogen wird ändern sich die Wahrscheinlichkeiten in Abhängigkeit davon, was *vorher* gezogen wurde (ähnlich bei vielen Kartenspielen). In diesem Fall wären die Zufallsvariablen nicht unabhängig, und die Kovarianz *zwischen* einzelnen Zufallsvariablen wäre ungleich Null!

Deshalb benötigen wir für die Kovarianzen eine zweite Annahme, die *Unabhängigkeit*, oder im Spezialfall der linearen Unabhängigkeit, $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$ für $i, j = 1, \ldots, n)$ und $i \neq j$.

Wenn wir z.B. aus einer Urne *mit Zurücklegen* ziehen, oder wenn es sich um das Ergebnis wiederholter Würfe mit einem Würfel handelt, dann ist die Annahme vermutlich gerechtfertigt, dass die Kovarianzen gleich Null sind.

In diesem Fall sagen wir, die Zufallsvariablen sind *unabhängig* verteilt. Wenn beide Annahmen erfüllt sind nennen wir sie *identisch und unabhängig* verteilt, und wenn wir den Erwartungswert mit μ und die Varianz σ^2 bezeichnen schreiben wir dies

$$y_i \sim \text{i.i.d.}(\mu, \sigma^2) \qquad \text{für } i = 1, 2, 3$$

wobei i.i.d. für *independent and identically distributed* steht (wir erinnern uns, dass wir diese Bedingung bereits weiter oben als Annahme A4 eingeführt haben).

Nur wenn die drei Zufallsvariablen *identisch und unabhängig* verteilt sind gilt

$$\text{var}(\hat{\theta}) = \frac{14}{36}\,\sigma^2$$

Analog können wir unter dieser i.i.d. Annahme die Varianz der zweiten Schätzfunktion $\widehat{\psi} = \frac{1}{3}(y_1 + y_2 + y_3)$ berechnen

$$\text{var}(\widehat{\psi}) = \frac{1}{9}\,(\sigma^2 + \sigma^2 + \sigma^2) = \frac{1}{3}\,\sigma^2$$

Also ist

$$\text{var}(\widehat{\psi}) = \frac{6}{18}\,\sigma^2 < \frac{7}{18}\,\sigma^2 = \text{var}(\hat{\theta})$$

Die Schätzfunktion $\widehat{\psi}$ mit den gleich großen Gewichten 1/3 hat eine kleinere Varianz, ist also genauer als die erste Schätzfunktion, oder in der Sprache der Statistiker, die Schätzfunktion $\widehat{\psi}$ ist *effizienter* als die ebenso erwartungstreue Schätzfunktion $\hat{\theta}$.

Man beachte aber, dass die beiden Ausdrücke für die Varianz den unbekannten Parameter σ^2 enthalten, deshalb ist dies noch keine Schätzfunktion für die Varianz.
In der 'Einführung Statistik' haben Sie vermutlich gelernt, dass Sie diese Varianz mit

$$\hat{\sigma}^2 = \frac{\sum_i (y_i - \bar{y})}{n-1}$$

erwartungstreu schätzen können.
Analoges werden wir für die OLS Schätzfunktion für die Standardfehler machen.
Mit ein bisschen Rechnerei kann man sogar zeigen, dass keine andere lineare und erwartungstreue Schätzfunktion existiert, die eine kleinere Varianz als $\widehat{\psi}$ hat, bei der alle Gewichte gleich groß sind und den Wert $\frac{1}{n}$ haben, also das arithmetische Mittel. Dies ist der berühmte *Gauss-Markov* Beweis, den wir im übernächsten Abschnitt etwas allgemeiner für OLS Schätzfunktionen demonstrieren werden (wir erinnern uns, dass der einfache Mittelwert ein Spezialfall einer OLS Regression nur auf die Regressionskonstante ist).
Allerdings ist die Annahme, dass die einzelnen Zufallsvariablen ε_i i.i.d.$(0, \sigma^2)$ verteilt sind, keineswegs harmlos oder selbstverständlich, sie ist in vielen höchst relevanten Fällen mit hoher Wahrscheinlichkeit verletzt. Wir werden in späteren Kapiteln zur Heteroskedastizität und Autokorrelation zeigen, wie man in solchen realistischeren Fällen vorgehen kann.
Aber vorher wollen wir endlich die Herleitung der OLS Standardfehler demonstrieren, die wir im nächsten Kapitel für die Hypothesentests und Konfidenzintervalle benötigen werden.

4.3.2 OLS Standardfehler für bivariate Regressionen

Wir beginnen wieder mit dem bivariaten Modell $y_i = \widehat{\beta}_1 + \widehat{\beta}_2 x_i + \hat{\varepsilon}_i$ und erinnern uns, dass die *Schätzfunktion* $\widehat{\beta}_2$ eine Zufallsvariable ist.
Wir werden zuerst zeigen, dass die OLS Schätzfunktion für den Steigungskoeffizienten linear in den y_i ist. Die Schätzfunktion $\widehat{\beta}_2$ ist

$$\widehat{\beta}_2 = \frac{\widehat{\mathrm{cov}}(x,y)}{\widehat{\mathrm{var}}(x)} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{\sum_i (x_i - \bar{x}) y_i}{\sum_i (x_i - \bar{x})^2}$$

mit $i = 1, \ldots, n$.
Das dritte Gleichheitszeichen folgt, weil $\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i (x_i - \bar{x}) y_i - \sum_i (x_i - \bar{x})\bar{y} = \sum_i (x_i - \bar{x}) y_i - \bar{y} \sum_i (x_i - \bar{x}) = \sum_i (x_i - \bar{x}) y_i$, da $\sum_i (x_i - \bar{x}) = 0$.
Deshalb können wir die Schätzfunktion für $\widehat{\beta}_2$ auch schreiben als

$$\widehat{\beta}_2 = \sum_{i=1}^{n} w_i y_i \tag{4.2}$$

mit den Gewichten

$$w_i := \frac{(x_i - \bar{x})}{\sum_j (x_j - \bar{x})^2}$$

d.h. $\widehat{\beta}_2$ ist eine gewichtete Summe der y_i!
Diese Gewichte w_i sind Funktionen der einzelnen x_i und haben drei wichtige Eigenschaften, wie man einfach zeigen kann:

1. $\boxed{\sum_i w_i = 0}$ (die Summe der Gewichte ist Null)

   da

   $$\sum_i w_i = \sum_i \left(\frac{(x_i - \bar{x})}{\sum_j (x_j - \bar{x})^2} \right) = \frac{\sum_i (x_i - \bar{x})}{\sum_j (x_j - \bar{x})^2} = 0$$

   mit $i, j = 1, \ldots, n$, weil die Summe der Abweichungen vom Mittelwert immer Null ist, d.h. $\sum (x_i - \bar{x}) = n\bar{x} - n\bar{x} = 0$. Dies folgt aus $\bar{x} := \frac{1}{n} \sum_i x_i$!

2. $\boxed{\sum_i w_i^2 = 1 / \sum_i (x_i - \bar{x})^2}$

   da

   $$\sum_i w_i^2 = \sum_i \left(\frac{(x_i - \bar{x})}{\sum_j (x_j - \bar{x})^2} \right)^2 = \frac{\sum_i (x_i - \bar{x})^2}{\left(\sum_i (x_i - \bar{x})^2 \right)^2} = \frac{1}{\sum (x_i - \bar{x})^2}$$

   mit $i, j = 1, \ldots, n$.

3. $\boxed{\sum_i w_i (x_i - \bar{x}) = \sum_i w_i x_i = 1}$

   Das erste '=' gilt, weil $\bar{x} \sum_i w_i = 0$, es bleibt also nur zu zeigen, dass $\sum_i w_i x_i = 1$

   $$\begin{aligned}
   \sum w_i x_i &= \frac{\sum (x_i - \bar{x}) x_i}{\sum (x_i - \bar{x})^2} \\
   &= \frac{\sum x_i^2 - \bar{x} \sum x_i}{\sum x_i^2 - 2\bar{x} \sum x_i + n\bar{x}^2} \\
   &= \frac{\sum x_i^2 - n\bar{x}^2}{\sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2} \qquad (\text{da } \sum x_i = n\bar{x}) \\
   &= \frac{\sum x_i^2 - n\bar{x}^2}{\sum x_i^2 - n\bar{x}^2} \\
   &= 1
   \end{aligned}$$

Bewaffnet mit diesen drei Eigenschaften der Gewichte w_i können wir wieder gleich wie früher vorgehen, d.h. wir setzen die PRF $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, in die Schätzfunktion $\widehat{\beta}_2 = \widehat{\text{cov}}(x, y) / \widehat{\text{var}}(x) = \sum_i w_i y_i$ ein um den Zusammenhang zwischen der Schätzfunktion und dem Parameter der Grundgesamtheit herzustellen

$$\begin{aligned}
\widehat{\beta}_2 &= \sum_i w_i y_i = \sum_i w_i (\beta_1 + \beta_2 x_i + \varepsilon_i) \\
&= \beta_1 \sum_i w_i + \beta_2 \sum_i w_i x_i + \sum_i w_i \varepsilon_i \\
&= \beta_2 + \sum_i w_i \varepsilon_i \qquad (4.3)
\end{aligned}$$

da wir schon gezeigt haben, dass $\sum w_i = 0$ und $\sum w_i x_i = 1$.

Hinweis: wenn wir von (4.3) den Erwartungswert bilden erhalten wir natürlich die gleiche Bedingung für die Erwartungstreue wie früher

$$\begin{aligned}
\mathrm{E}(\widehat{\beta}_2) &= \mathrm{E}\left(\beta_2 + \sum_i w_i\varepsilon_i\right) \\
&= \beta_2 + \sum_i \mathrm{E}(w_i\varepsilon_i) \qquad (\text{weil } \mathrm{E}(\beta_2) = \beta_2) \\
&= \beta_2 + \mathrm{E}\left(\frac{\sum_i (x_i - \overline{x})\varepsilon_i}{\sum_i (x_i - \overline{x})^2}\right) \\
&= \beta_2 + \mathrm{E}\left(\frac{\mathrm{cov}(x,\varepsilon)}{\mathrm{var}(x)}\right) \qquad (4.4)
\end{aligned}$$

Daraus folgt wieder, dass die Schätzfunktion $\widehat{\beta}_2$ nur dann erwartungstreu ist, wenn a) die PRF linear und richtig spezifiziert war, b) die Schätzfunktion $\widehat{\beta}_2$ existiert und eindeutig ist, und c) die erklärende Variable x und die Störterme unkorreliert sind, bzw. wenn $\mathrm{cov}(x,\varepsilon) = 0$. ■

Nach diesen allgemeinen Vorbemerkungen können wir nun endlich die Varianzen und Kovarianz der Schätzfunktionen $\widehat{\beta}_2$ und $\widehat{\beta}_1$ berechnen.

Die Varianz der Zufallsvariable $\widehat{\beta}_2$ ist definiert als

$$\begin{aligned}
\mathrm{var}(\widehat{\beta}_2) &= \mathrm{E}[\widehat{\beta}_2 - \mathrm{E}(\widehat{\beta}_2)]^2 \\
&= \mathrm{E}[\widehat{\beta}_2 - \beta_2]^2 \qquad (\text{wenn } \mathrm{E}(\widehat{\beta}_2) = \beta_2, \text{ siehe oben}) \\
&= \mathrm{E}\left(\sum_i w_i\varepsilon_i\right)^2 \qquad (\text{da } \widehat{\beta}_2 = \beta_2 + \textstyle\sum w_i\varepsilon_i; \text{ s. Gleichung (4.3)}) \\
&= \mathrm{E}\left(w_1^2\varepsilon_1^2 + w_2^2\varepsilon_2^2 + \cdots + w_n^2\varepsilon_n^2 + \cdots\right. \\
&\quad \left. \cdots + 2w_1w_2\varepsilon_1\varepsilon_2 + \cdots + 2w_{n-1}w_n\varepsilon_{n-1}\varepsilon_n\right) \\
&= \underbrace{\mathrm{E}\left(\sum_{i=1}^{n} w_i^2\varepsilon_i^2\right)}_{\substack{= \sigma^2 \sum_i w_i^2 \text{ wenn} \\ \text{homoskedastisch}}} + \underbrace{\mathrm{E}\left(\sum_{i=1}^{n}\sum_{\substack{j=2 \\ j>i}}^{n} 2w_iw_j\varepsilon_i\varepsilon_j\right)}_{\substack{= 0 \text{ wenn keine} \\ \text{Autokorrelation}}} \qquad (4.5)
\end{aligned}$$

Dieser letzte Ausdruck ist mit all den Kreuztermen etwas 'unappetitlich' lang. Außerdem enthält er weit mehr unbekannte (Kreuz-)Produkte von Störtermen als Beobachtungen (n), es wäre also völlig aussichtslos diese Varianz aus einer Stichprobe schätzen zu wollen. Man beachte, dass dies sehr ähnlich aussieht wie in unserem einführenden Beispiel mit (y_1, y_2, y_3), vgl. Seite 13.

Um hier weiter zukommen benötigen wir wieder zusätzliche Annahmen, diesmal über die Störterme ε_i.

Eine radikale (und nicht immer realistische) Annahme, die das Problem massiv vereinfacht, ist

$$\varepsilon_i \sim \text{i.i.d.}\left(0, \sigma^2\right)$$

Dies ist eine sehr kompakte Schreibweise für ε_i ist unabhängig und identisch verteilt (i.i.d. steht für *'independent and identically distributed'*) mit $\mathrm{E}(\varepsilon_i) = 0$ und $\mathrm{var}(\varepsilon_i) =$

σ^2; das heißt, vor der Klammer steht die Art der Verteilung, das erste Argument in der Klammer ist der Erwartungswert, das zweite Argument die Varianz (generell werden in der Klammer die Parameter der Verteilung angegeben, in diesem Fall sind dies Erwartungswert und Varianz).

Im einzelnen umfasst dies folgende Annahmen:

1. alle Störterme ε_i sind identisch verteilt (d.h. jedes einzelne ε_i hat die gleiche Verteilung, und deshalb auch die gleichen Momente).

   Dies kommt im zweiten i von i.i.d. (*identically distributed*) zum Ausdruck. Dies impliziert auch, dass die Varianz aller Störterme ε_i gleich groß ist, also einfach eine reelle Zahl σ^2 ist. Anders ausgedrückt, alle ε_i haben die gleiche endliche Varianz σ^2. Dies kommt in den *bedingten Dichtefunktionen* der Störterme in Abbildung 4.1 (Seite 2) zum Ausdruck. Dass wie in dieser Abbildung alle Dichtefunktionen exakt die gleiche Form haben ist keineswegs selbstverständlich und wird wohl nur in Ausnahmefällen gelten. Einer der wenigen Spezialfälle, bei denen die bedingten Erwartungswerte exakt auf einer Gerade liegen und die bedingten Varianzen konstant sind (d.h. unabhängig von x sind), bilden gemeinsam normalverteilte Zufallsvariablen (die in Abbildung 4.1 dargestellt sind).

   Wenn die Annahme, dass die bedingte Varianz der Störterme konstant und unabhängig von x ist, also $\operatorname{var}(\varepsilon_i|x_1,\ldots,x_n) = \sigma^2$, erfüllt ist, spricht man von *homoskedastischen* Störtermen, wenn die Annahme verletzt ist spricht man von *heteroskedastischen* Störtermen (oder einfach von Heteroskedastizität).

2. Unabhängigkeit der Störterme, d.h. $\mathrm{E}(\varepsilon_i\varepsilon_j|x_1,\ldots,x_n) = 0$ für $i,j = 1,\ldots,n$ und $i \neq j$ (dies impliziert auch $\operatorname{cov}(\varepsilon_i,\varepsilon_j) = 0$ für $i \neq j$); dies kommt im ersten i von i.i.d. (*independent*) zum Ausdruck. Wenn diese Annahme verletzt ist spricht man von *Autokorrelation* der Störterme.

3. $\mathrm{E}(\varepsilon_i) = 0$: Diese Annahme haben wir bereits für den Beweis der Erwartungstreue benötigt (wenn die x stochastisch sind wird die wesentlich strengere Annahme A3: $\mathrm{E}(\varepsilon_i|x_1,\ldots,x_n) = 0$ benötigt, d.h. alle bedingten Erwartungswert der ε_i müssen Null sein. Wenn $\mathrm{E}(\varepsilon_i|x_1,\ldots,x_n) = 0$ folgt aus dem Gesetz der iterierten Erwartungen automatisch $\mathrm{E}(\varepsilon_i) = 0$).

Um Gleichung (4.5) zu vereinfachen benötigen wir die ersten zwei dieser drei Annahmen, d.h. $\mathrm{E}(\varepsilon_i^2) = \sigma^2$ (Homoskedastizität) und $\mathrm{E}(\varepsilon_i\varepsilon_j) = 0$ für $i \neq j$ (Unabhängigkeit).

Wenn die Annahme $\mathrm{E}(\varepsilon_i\varepsilon_j) = 0$ erfüllt ist (d.h. keine Autokorrelation vorliegt) fallen die Kreuzterme in Gleichung (4.5) weg, deshalb gilt in diesem Fall

$$\operatorname{var}(\widehat{\beta}_2) = \mathrm{E}\left(\sum_i w_i^2\varepsilon_i^2\right)$$

Wenn die x_i (und damit automatisch auch die w_i) deterministisch sind können die w_i vor den Erwartungswertoperator gezogen werden

$$\operatorname{var}(\widehat{\beta}_2) = \sum_i w_i^2\,\mathrm{E}(\varepsilon_i^2)$$

Wenn zusätzlich die erste Annahme $\mathrm{E}(\varepsilon_i^2) = \sigma^2$ (keine Heteroskedastizität) erfüllt ist gilt schließlich

$$\mathrm{var}(\widehat{\beta}_2) = \sum_i w_i^2 \sigma^2 = \sigma^2 \sum_i w_i^2$$

da σ^2 ein fixer Parameter der Grundgesamtheit ist.
Nun haben wir bereits vorhin gezeigt (Seite 16), dass $\sum w_i^2 = \frac{1}{\sum (x_i - \bar{x})^2}$.

Deshalb erhalten wir unter den obigen Annahmen

A1: Linearität der PRF,

A2: keine perfekte Multikollinearität,

A3: stochastische Unabhängigkeit von Regressoren und Störtermen, $\mathrm{E}(\varepsilon_i|x) = \mathrm{E}(\varepsilon_i)$

A4: $\varepsilon_i \sim \text{i.i.d.}(0, \sigma^2)$

den folgenden Ausdruck für die **Varianz des OLS Steigungskoeffizienten** $\widehat{\beta}_2$ gleich

$$\mathrm{var}(\widehat{\beta}_2) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}$$

Der aufmerksamen Leserin wird nicht entgangen sein, dass diese Varianz (bei deterministischen Regressoren) keine Zufallsvariable ist (also auch keine Schätzfunktion), und dass dieser Ausdruck darüber hinaus auch wenig hilfreich ist, da er die unbekannte Varianz der Störterme enthält.

Unsere nächste Aufgabe wird es deshalb sein, eine Schätzfunktion $\hat{\sigma}^2$ für den unbekannten Parameter σ^2 zu finden, die es uns erlaubt, aus den Stichprobendaten eine Schätzung für die Varianz der Störterme zu berechnen.

Vorher wollen wir aber noch kurz die entsprechenden Ausdrücke für die restlichen Parameter angeben.

Die **Varianz des Interzepts** $\widehat{\beta}_1$ kann ähnlich, wenngleich etwas mühsamer, hergeleitet werden

$$\mathrm{var}(\widehat{\beta}_1) = \mathrm{E}[\widehat{\beta}_1 - \mathrm{E}(\widehat{\beta}_1)]^2 = \sigma^2 \frac{\sum x_i^2}{n \sum \ddot{x}_i^2}$$

Da $\widehat{\beta}_1$ und $\widehat{\beta}_2$ Zufallsvariablen sind kann man auch die **Kovarianz** zwischen den beiden Schätzfunktionen berechnen. Diese ist definiert

$$\begin{aligned} \mathrm{cov}(\widehat{\beta}_1, \widehat{\beta}_2) &= \mathrm{E}\{[\widehat{\beta}_1 - \mathrm{E}(\widehat{\beta}_1)][\widehat{\beta}_2 - \mathrm{E}(\widehat{\beta}_2)]\} \\ &= \mathrm{E}[(\widehat{\beta}_1 - \beta_1)(\widehat{\beta}_2 - \beta_2)] \end{aligned}$$

Wir erinnern uns, dass $\widehat{\beta}_1 = \bar{y} - \widehat{\beta}_2 \bar{x}$ und bei Erwartungstreue von $\widehat{\beta}_2$ gilt $\mathrm{E}(\widehat{\beta}_1) = \bar{y} - \beta_2 \bar{x}$. Daraus folgt $\widehat{\beta}_1 - \mathrm{E}(\widehat{\beta}_1) = -\bar{x}(\widehat{\beta}_2 - \beta_2)$.

Wenn wir dies oben einsetzen erhalten wir

$$\begin{aligned}\operatorname{cov}(\widehat{\beta}_1, \widehat{\beta}_2) &= \mathrm{E}[(\widehat{\beta}_1 - \beta_1)(\widehat{\beta}_2 - \beta_2)] \\ &= -\bar{x}\,\mathrm{E}(\widehat{\beta}_2 - \beta_2)^2 \\ &= -\bar{x}\,\operatorname{var}(\widehat{\beta}_2)\end{aligned}$$

Die Kovarianzen zwischen Schätzfunktionen werden wir später für Tests von gemeinsamen Hypothesen (*'joint hypothesis'*) benötigen.
Wir fassen zusammen: unter den bisher getroffenen Annahmen A1 – A4 gilt

$$\begin{aligned}\mathrm{E}(\widehat{\beta}_2) &= \beta_2 & \qquad \operatorname{var}(\widehat{\beta}_2) &= \tfrac{\sigma^2}{\sum[x_i-\bar{x}]^2} \\ \mathrm{E}(\widehat{\beta}_1) &= \beta_1 & \qquad \operatorname{var}(\widehat{\beta}_1) &= \tfrac{\sigma^2 \sum x_i^2}{n\sum[x_i-\bar{x}]^2} \\ & & \qquad \operatorname{cov}(\widehat{\beta}_1, \widehat{\beta}_2) &= \tfrac{-\bar{x}\sigma^2}{\sum[x_i-\bar{x}]^2}\end{aligned}$$

Aber wie schon erwähnt helfen uns diese Formeln für die OLS Varianzen der Koeffizienten noch nicht wirklich weiter, da sie den unbeobachtbaren Parameter σ^2 enthalten, d.h. wir können damit die eigentlich interessierenden Standardfehler der Koeffizienten noch nicht schätzen. Dazu benötigen wir zuerst eine Schätzfunktion für die Varianz der Störterme σ^2.

Eine Schätzfunktion für die Varianz der Störterme σ^2

Da die Ausdrücke für die OLS Standardfehler noch die unbekannte Varianz σ^2 der Störterme enthalten benötigen wir als nächstes eine erwartungstreue Schätzfunktion $\hat{\sigma}^2$ für das wahre σ^2 der Grundgesamtheit.

Es liegt nahe, eine solche Schätzfunktion aus den beobachtbaren Stichprobenresiduen zu berechnen. Leider sind diese Berechnungen etwas umständlich und nicht sehr intuitiv, deshalb habe ich die detaillierten Herleitungen in den Appendix 4.A.1 verbannt.

Dort zeigen wir, dass es tatsächlich einen engen Zusammenhang zwischen der Quadratsumme der Residuen und der Varianz der Störterme gibt, wir müssen lediglich die Quadratsumme der Residuen $\sum_i \hat{\varepsilon}_i^2$ durch die Anzahl der Freiheitsgrade $n-2$ dividieren, oder etwas allgemeiner auch für multiple Regressionen mit k erklärenden Variablen (inkl. Interzept)

$$\hat{\sigma}^2 = \frac{\sum_i \hat{\varepsilon}_i^2}{n-k}$$

Die Wurzel dieser erwartungstreuen Schätzfunktion wird in der Literatur **Standardfehler der Regression** ('*standard error of regression*' oder '*standard error of estimate*') genannt

$$\boxed{\hat{\sigma} = \sqrt{\frac{\sum_i \hat{\varepsilon}_i^2}{n-k}}} \tag{4.6}$$

Wie man im Appendix 4.A.1 sehen kann, werden auch für diese Herleitung wiederholt die Annahmen A1 – A4 benötigt. Ist auch nur eine dieser vier Annahmen verletzt liefert obige Formel für den *Standardfehler der Regression* $\hat{\sigma}^2$ falsche Ergebnisse, d.h. dann ist die Schätzfunktion für σ^2 verzerrt!

Standardfehler der Koeffizienten

Unser eigentliches Interesse gilt ja den Standardfehlern der Koeffizienten. Diese können wir nun einfach berechnen, indem wir die Schätzfunktion für die Standardfehler der Regression in die Formeln für die Varianzen der Koeffizienten einsetzen. Dies gibt uns die gesuchten *Schätzfunktionen für die Varianzen der Koeffizienten*

$$\widehat{\text{var}}(\widehat{\beta}_2) := \hat{\sigma}^2_{\widehat{\beta}_2} = \frac{\hat{\sigma}^2}{\sum[x_i - \bar{x}]^2}, \qquad \widehat{\text{var}}(\widehat{\beta}_1) := \hat{\sigma}^2_{\widehat{\beta}_1} = \frac{\hat{\sigma}^2 \sum x_i^2}{n \sum[x_i - \bar{x}]^2}$$

und die Wurzeln daraus sind die *Standardfehler der Koeffizienten*

$$\widehat{\text{se}}(\widehat{\beta}_2) := \hat{\sigma}_{\widehat{\beta}_2} = \sqrt{\frac{\hat{\sigma}^2}{\sum[x_i - \bar{x}]^2}}, \qquad \widehat{\text{se}}(\widehat{\beta}_1) := \hat{\sigma}_{\widehat{\beta}_1} = \sqrt{\frac{\hat{\sigma}^2 \sum x_i^2}{n \sum[x_i - \bar{x}]^2}}$$

Freiheitsgrade: Wir haben gesehen, dass wir für die Berechnung einer erwartungstreuen Schätzfunktion für σ^2 die Quadratsumme der Stichprobenresiduen $\sum_i \hat{\varepsilon}_i^2$ durch $n - k$ dividieren müssen, nicht durch n, wie man das ad hoc erwarten würde. Warum ist das so?

Die Schätzung von Parametern ist eng verbunden mit der jeweils zur Verfügung stehenden Information. Für eine intuitive Erklärung erinnern wir uns an die Herleitung der OLS-Schätzfunktionen. Dazu haben wir folgenden Ausdruck minimiert

$$\min_{\widehat{\beta}_1, \widehat{\beta}_2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \min_{\widehat{\beta}_1, \widehat{\beta}_2} \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_1 - \widehat{\beta}_2 x_i \right)^2$$

Für jeden zu schätzenden Parameter erhalten wir eine Bedingungen erster Ordnung

$$\begin{aligned} \frac{\partial \sum \hat{\varepsilon}_i^2}{\partial \widehat{\beta}_1} &= -2 \sum \underbrace{\left(y_i - \widehat{\beta}_1 - \widehat{\beta}_2 x_i \right)}_{\hat{\varepsilon}_i} = 0 \quad \Rightarrow \quad \sum \hat{\varepsilon}_i = 0 \\ \frac{\partial \sum \hat{\varepsilon}_i^2}{\partial \widehat{\beta}_2} &= -2 \sum \underbrace{\left(y_i - \widehat{\beta}_1 - \widehat{\beta}_2 x_i \right)}_{\hat{\varepsilon}_i} x_i = 0 \quad \Rightarrow \quad \sum x_i \hat{\varepsilon}_i = 0 \end{aligned}$$

Diese beiden Gleichungen legen eine Restriktion auf die Residuen.

Wenn wir z.B. nur die Residuen $\hat{\varepsilon}_1, \hat{\varepsilon}_2, \ldots, \hat{\varepsilon}_{n-2}$ einer bivariaten Regression kennen würden, könnten wir die beiden fehlenden Residuen $\hat{\varepsilon}_{n-1}$ und $\hat{\varepsilon}_n$ mit Hilfe dieser beiden Bedingungen 1. Ordnung $\sum_i \hat{\varepsilon}_i = 0$, $\sum_i x_i \hat{\varepsilon}_i = 0$ berechnen.

Am einfachsten kann man sich dies mit einer Regression nur auf die Regressionskonstante und 3 Beobachtungen vorstellen. Angenommen wir kennen von den drei Residuen nur zwei, z.B. $\hat{\varepsilon}_1 = -3$ und $\hat{\varepsilon}_2 = +1$. Wir wissen, dass die Residuen die Bedingung erster Ordnung $\sum_{i=1}^{3} \hat{\varepsilon}_i = 0$ erfüllen, deshalb folgt aus unmittelbar das dritte Residuum $\hat{\varepsilon}_3 = 2$.

Nicht alle der Residuen sind deshalb 'frei', sondern manche sind durch die Bedingungen erster Ordnung determiniert, und enthalten deshalb 'keine Information' über die Störterme der Grundgesamtheit ε_i. Da wir für jeden zu schätzenden Parameter

eine Bedingung erster Ordnung haben, verlieren wir mit jedem geschätzten Parameter einen Freiheitsgrad. In bivariaten Modell haben wir zwei Parameter geschätzt ($\widehat{\beta}_1$ und $\widehat{\beta}_2$), deshalb verlieren wir zwei Freiheitsgrade. In der multiplen Regression benötigen wir k Bedingungen erster Ordnung, deshalb verlieren wir k Freiheitsgrade.

Mit Hilfe der Schätzfunktion $\hat{\sigma}$ (Standardfehler der Regression) können wir nun die erwartungstreuen Schätzfunktionen für die *Varianz der Schätzfunktionen* $\widehat{\beta}_1$ und $\widehat{\beta}_2$, d.h. $\hat{\sigma}^2_{\widehat{\beta}_1}$ und $\hat{\sigma}^2_{\widehat{\beta}_2}$, berechnen, und natürlich auch deren Wurzeln, die *Standardfehler der Koeffizienten*.

Realisationen daraus (Schätzungen), die wir durch Einsetzen der Stichprobendaten erhalten, werden uns später die Durchführung statistischer Tests ermöglichen.

Wir fassen nochmals zusammen:

$$
\begin{aligned}
\widehat{\beta}_2 &= \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} \\
\widehat{\mathrm{se}}(\widehat{\beta}_2) := \hat{\sigma}_{\widehat{\beta}_2} &= \sqrt{\frac{\hat{\sigma}^2}{\sum(x_i-\bar{x})^2}} \\
\\
\widehat{\beta}_1 &= \bar{y} - \widehat{\beta}_2\bar{x} \\
\widehat{\mathrm{se}}(\widehat{\beta}_1) := \hat{\sigma}_{\widehat{\beta}_1} &= \sqrt{\frac{\hat{\sigma}^2\,\sum x_i^2}{n\sum(x_i-\bar{x})^2}} \\
\\
\mathrm{cov}(\widehat{\beta}_1,\widehat{\beta}_2) &= \frac{-\bar{x}\hat{\sigma}^2}{\sum(x_i-\bar{x})^2} \\
\\
\hat{\sigma}^2 &= \frac{\sum\hat{\varepsilon}_i^2}{n-2}
\end{aligned}
$$

Damit haben wir die wesentlichen Elemente beisammen. Die Standardfehler der Koeffizienten sind ein Maß für die 'Genauigkeit' der Schätzfunktionen, d.h. eine Schätzfunktion ist ceteris paribus umso genauer, je kleiner deren Standardfehler ist. Man beachte, dass die Standardfehler und die Koeffizienten die gleiche Dimension haben, deshalb ist nicht die absolute Größe der Standardfehler entscheidend, sondern deren Größe *im Verhältnis* zu den Koeffizienten!

Determinanten der Standardfehler der Koeffizienten

Wovon hängt die Größe der Standardfehler der Koeffizienten nun ab? Wir wollen uns vorerst auf den Standardfehler des Steigungskoeffizienten $\widehat{\beta}_2$ im bivariaten Modell beschränken.

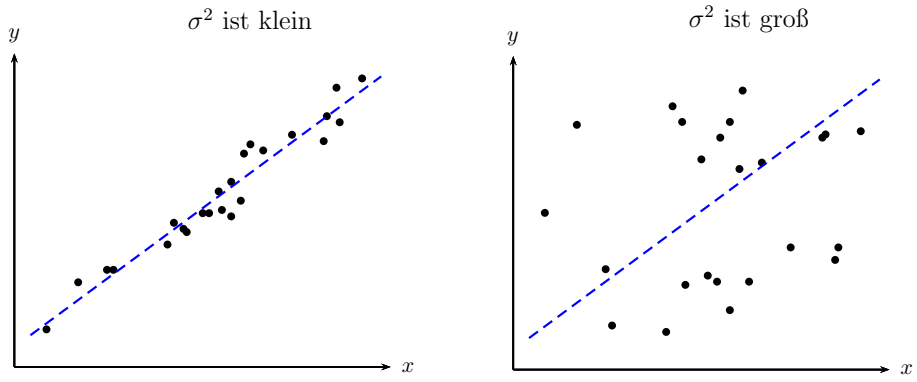

Abbildung 4.4: Regressionen mit unterschiedlicher Varianz von ε (σ^2).

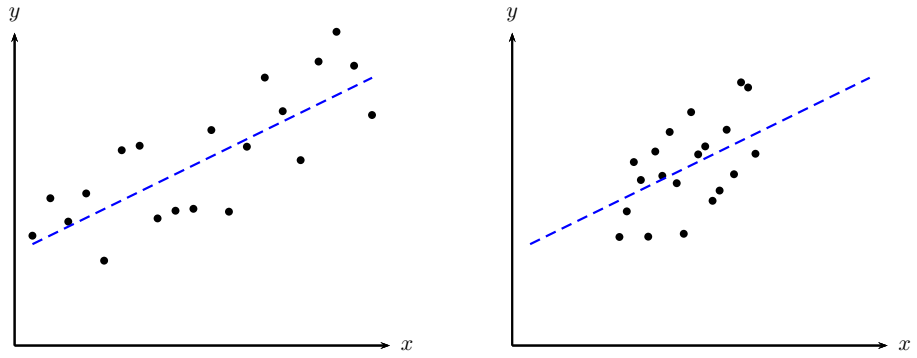

Abbildung 4.5: Unterschiedliche Streuung der erklärenden x Variable. In der linken Abbildung streut x stark, in der rechten Abbildung ist die Streuung von x deutlich kleiner.

Ceteris paribus ist der Standardfehler

$$\widehat{\text{se}}(\widehat{\beta}_2) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

umso kleiner, ...

1. ...je kleiner die Varianz der Störterme der Grundgesamtheit σ^2 ist (zur Erinnerung, $\hat{\sigma}^2$ ist eine erwartungstreue Schätzfunktion für σ^2). Abbildung 4.4 zeigt zwei Stichproben, die sich nur in der Varianz der Grundgesamtheit σ^2 unterscheiden.

2. ...je *größer* die Streuung der x ist, d.h., je größer $\sum_i (x_i - \bar{x})^2$) ist. Abbildung 4.5 zeigt zwei Stichproben mit gleichem $\hat{\sigma}^2$, die sich nur in der Streuung der x unterscheiden. Es ist offensichtlich, dass die Schätzung umso genauer , je größer die Streuung der x ist!

3. ...je größer der Stichprobenumfang n ist, da der Nenner $\sum_{i=1}^{n}(x_i - \bar{x})^2$ mit dem Stichprobenumfang n zunimmt. Offensichtlich können wir $\widehat{\beta}_2$ umso genauer schätzen, je größer die Stichprobe ist.

4. Im multiplen Regressionsmodell mit mehreren Regressoren kommt noch eine vierte Determinante dazu; ceteris paribus ist der Standardfehler eines Koeffizienten umso kleiner, je weniger der entsprechende Regressor mit allen anderen Regressoren dieser Regression korreliert ist. Dies wird im Folgenden etwas näher erläutert.

Standardfehler der Koeffizienten im multiplen Regressionsmodell

Die Herleitung der Standardfehler für Regressionsmodelle mit mehreren erklärenden x Variablen ist in Summennotation etwas umständlich, deshalb wird hier nur das Ergebnis vorweggenommen, die Details folgen im Kapitel zur Matrixschreibweise des OLS Modells.

Für das multiple Regressionsmodell

$$y_i = \widehat{\beta}_1 + \widehat{\beta}_2 x_{i2} + \widehat{\beta}_3 x_{i3} + \cdots + \widehat{\beta}_h x_{ih} + \cdots + \widehat{\beta}_k x_{ik} + \hat{\varepsilon}_i$$

kann man zeigen, dass eine Schätzfunktion für den Standardfehler eines beliebigen Steigungskoeffizienten $\widehat{\beta}_h$ durch den folgenden Ausdruck gegeben ist

$$\widehat{\text{se}}(\widehat{\beta}_h) = \sqrt{\frac{\hat{\sigma}^2}{(1 - R_h^2) \sum_i (x_{ih} - \bar{x}_h)^2}} \tag{4.7}$$

Dieser Standardfehler unterscheidet sich nur durch den Term $(1 - R_h^2)$ im Nenner vom Standardfehler für den bivariaten Fall.

Das Bestimmtheitsmaß R_h^2 wird aus folgender Hilfsregression berechnet

$$x_{ih} = \widehat{\alpha}_1 + \widehat{\alpha}_2 x_{i2} + \cdots + \widehat{\alpha}_{h-1} x_{ih-1} + \widehat{\alpha}_{h+1} x_{ih+1} + \cdots + \widehat{\alpha}_k x_{ik} + \nu_i \quad \rightarrow \quad R_h^2$$

das heißt, der interessierende Regressor x_h wird *auf alle anderen Regressoren* regressiert. Offensichtlich ist das Bestimmtheitsmaß R_h^2 dieser Hilfsregression umso größer, je stärker der Regressor x_h mit allen anderen Regressoren korreliert ist.

Aus Gleichung (4.7) ist ersichtlich, dass der Standardfehler eines Koeffizienten $\widehat{\text{se}}(\widehat{\beta}_h)$ auch davon abhängt, wie stark er mit den anderen Regressoren korreliert ist.

Im Extremfall, wenn der Regressor x_h überhaupt nicht mit den restlichen Regressoren korreliert ist, ist $R_h^2 = 0$, und wir erhalten exakt den gleichen Standardfehler wie in einer bivariaten Regression; dieser Fall untereinander exakt unkorrelierter Regressoren ist aber eher unrealistisch.

Im anderen Extremfall, wenn eine exakte lineare Abhängigkeit zwischen den Regressoren existiert, ist $R_h^2 = 1$ und der Nenner Null, da $1 - R_h^2 = 0$. Dieser Fall wird *perfekte Multikollinearität* genannt und der Standardfehler ist in diesem Fall also ebenso wenig definiert wie der Koeffizient (siehe Annahme A2).

Wenn $R_h^2 < 1$ kann der Standardfehler zwar berechnet werden, aber der Standardfehler ist umso größer, je näher R_h^2 bei Eins liegt, d.h., je größer die lineare Abhängigkeit zwischen den Regressoren ist. Dies nennen wir *Multikollinearität*, und wir werden dieses Problem später noch ausführlich diskutieren.

Robuste Standardfehler

Wie nützlich sind die OLS Standardfehler? Zuerst einmal müssen wir festhalten, dass auch die Standardfehler Schätzfunktionen sind, und wir in der empirischen Analyse nur Realisationen dieser Zufallsvariablen beobachten.

Die Schätzfunktionen für die OLS Standardfehler sind nur erwartungstreu, wenn *alle* Annahmen A1 – A4 erfüllt sind (im Fall stochastischer Regressoren müssen natürlich zusätzlich AS1 und AS2 erfüllt sein).

Aber selbst wenn die Annahmen A1 – A3 erfüllt sind, die Koeffizienten also erwartungstreu geschätzt werden können, sind die OLS Schätzfunktionen für die Standardfehler häufig verzerrt.

Es zeigt sich nämlich, dass die Annahme A4: $\varepsilon_i \sim \text{i.i.d.}(0, \sigma^2)$ in vielen Fällen kaum zu rechtfertigen ist.

Insbesondere ist in *Querschnittsdaten* die Annahme der *Homoskadastizität* häufig verletzt

$$\text{var}(\varepsilon_i | x_1, \ldots, x_n) = \sigma_i^2 \neq \sigma^2$$

d.h. die bedingten Varianzen sind nicht konstant.

In *Zeitreihendaten* ist sehr oft die Annahme der *Unabhängigkeit* verletzt

$$\text{cov}(\varepsilon_i, \varepsilon_j | x_1, \ldots, x_n) \neq 0 \qquad \text{für } i, j = 1, \ldots, n \text{ und } i \neq j$$

d.h. die Störterme sind autokorreliert.

In beiden Fällen sind die OLS Standardfehler verzerrt, liefern also falsche Ergebnisse. Wir werden diese beiden Fälle in den Kapiteln zur *Heteroskedastizität* und *Autokorrelation* ausführlich diskutieren, hier wollen wir nur vorausschicken, dass für diese Fälle sogenannte *robuste Standardfehler* entwickelt wurden.

Diese beruhen im wesentlichen auf Schätzungen der Haupt- und Kreuzterme in Gleichung (4.5). Für Querschnittsdaten muss man die insgesamt n Terme $\sum_i w_i^2 \varepsilon_i^2$ aus der Stichprobe schätzen.

Dies scheint auf den ersten Blick unmöglich, aber Forscher wie Eicker (1963), Huber (1967) und White (1980) haben gezeigt, dass man solche Standardfehler tatsächlich schätzen kann, allerdings kann man nur deren Konsistenz beweisen, nicht deren Erwartungstreue und Effizienz.

Deshalb werden diese *Heteroskedastie-konsistente* (*heteroscedasticity-consistent*, HC) Standardfehler oder nach ihren Entdeckern Eicker-Huber-White (oder eine Teilmenge dieser Namen) Standardfehler genannt.

Standardfehler, die zusätzlich mögliche Autokorrelation berücksichtigen, werden meist HAC (*heteroscedasticity- and autocorrelation-consistent*) Standardfehler genannt.

Solche robusten Standardfehler sind heute in allen statistischen/ökonometrischen Softwarepaketen implementiert und es stellt sich die Frage, welche Standardfehler man verwenden soll.

Wir werden dies in späteren Kapiteln ausführlicher diskutieren, hier seien nur ein paar allgemeine Bemerkungen vorausgeschickt.

- Robuste Standardfehler sind in der Regel deutlich 'ungenauer', d.h., die Varianz ihrer Stichprobenkennwertverteilung ist deutlich größer als die der OLS Standardfehler.

  Allerdings sind sie auch konsistent wenn die Annahme A4 verletzt ist, während die OLS Standardfehler in diesem Fall verzerrt sind.

- Falls die Stichprobe 'ziemlich groß' ist (was man unter 'groß' versteht hängt von der Beschaffenheit der Daten ab, aber einige hundert Beobachtungen sollten es in der Regel schon sein) werden heute oft routinemäßig robuste Standardfehler verwendet.

- Wenn die Stichprobe nicht so groß ist, ist die Entscheidung schwieriger, es existiert ein klarer *trade-off*: falls die Störterme tatsächlich homoskedastisch sind verzichten wir bei robusten Standardfehlern unnotwendigerweise auf Genauigkeit, falls die Störterme aber heteroskedastisch sind, sind die OLS Standardfehler verzerrt. Wir werden später statistische Tests auf Heteroskedastizität kennen lernen, die unter gewissen Umständen die Entscheidung erleichtern können.

Mehr zu diesem Thema später ...

4.4 Gauss-Markov Theorem

"Beweisen muss ich diesen Käs',
sonst ist die Arbeit unseriös."
(F. Wille)

Bisher haben wir uns ausschließlich mit der Erwartungstreue von OLS-Schätzfunktionen und mit der Schätzung von deren Standardfehlern beschäftigt. In diesem Abschnitt werden wir nun die *Effizienz* von OLS-Schätzfunktionen beweisen, oder etwas genauer, wir werden untersuchen, *unter welchen Annahmen* OLS-Schätzfunktionen effizient sind. Dies ist der Inhalt des berühmten *Gauss-Markov Theorems*, welches besagt, dass OLS-Schätzfunktionen unter den bereits früher getroffenen Annahmen A1 – A4 von allen möglichen *linearen und erwartungstreuen Schätzfunktionen* die kleinste Varianz haben.

> Unter den (Gauss'schen) Annahmen des 'klassischen linearen Regressionsmodells' haben OLS-Schätzfunktionen innerhalb der Klasse aller linearen und erwartungstreuen Schätzfunktionen die kleinste Varianz, oder in anderen Worten, sie sind **BLUE**, d.h. OLS ist ein **B**est **L**inear **U**nbiased **E**stimator.

Die OLS-Schätzfunktion ist – wie wir bereits gesehen haben – linear in y_i, da $\widehat{\beta}_2 = \sum w_i y_i$.
Man kann zeigen, dass – wenn die Gauss-Markov Annahmen A1 – A4 erfüllt sind – OLS-Schätzfunktionen effizient sind, d.h.

$$\text{var}(\widehat{\beta}_2^{\text{OLS}}) \leq \text{var}(\tilde{\beta}_2^{\,*})$$

wobei $\tilde{\beta}_2^*$ jede beliebige lineare und erwartungstreue Schätzfunktion für β_2 sein kann.

Das Gauss-Markov Theorem und die zugrunde liegenden Gauss-Markov Annahmen spielen in der Ökonometrie eine ähnlich fundamentale Rolle wie das Modell vollständiger Konkurrenz in der Mikroökonomik, sie stellen das Referenzmodell schlechthin dar. Einen Großteil der restlichen Veranstaltung werden wir uns mit Fällen beschäftigen, wenn die Gauss-Markov Annahmen nicht erfüllt sind.

Der Beweis der Effizienz von OLS-Schätzfunktionen war früher einmal der Höhepunkt jeder einführenden Ökonometrie-Veranstaltung. In der modernen Ökonometrie spielt der Gauss-Markov Beweis vielleicht nicht mehr diese Rolle, aber trotzdem nimmt er immer noch einen zentralen Stellenwert ein, nicht zuletzt wegen der Gauss-Markov Annahmen, die für diesen Beweis benötigt werden.

Es gibt mehrere Möglichkeiten diesen Beweis zu führen, die Grundidee des im Appendix angeführten Beweises kann man folgendermaßen skizzieren:

1. Wir gehen von einer beliebigen linearen Schätzfunktion aus (z.B. $\tilde{\beta} = \sum_i c_i\, y_i$).
2. Wir ermitteln die notwendigen Bedingungen, unter denen diese lineare Schätzfunktion erwartungstreu ist.

   ($\mathrm{E}(\tilde{\beta}) = \beta$ wenn $\sum_i c_i = 0$ und $\sum_i c_i x_i = 1$, siehe Appendix S. 48).
3. Wir minimieren die Varianz dieser beliebigen linearen Schätzfunktion unter der Nebenbedingung, dass diese lineare Schätzfunktion erwartungstreu ist. Dies kann z.B. mit Hilfe einer Lagrange Funktion erfolgen (Minimierung unter Nebenbedingungen).

   (min: $\mathrm{var}(\tilde{\beta})$ unter Nebenbedingung: $\sum_i c_i = 0$ und $\sum_i c_i x_i = 1$; $\rightarrow$ Lagrange)
4. Wir zeigen, dass die aus der Minimierung resultierende – also varianzminimale – Schätzfunktion genau die OLS-Schätzfunktion ist (d.h. $c_i = w_i$). Deshalb ist die OLS Schätzfunktion *varianzminimal*.

   Allerdings benötigen wir für die Beweisführung wieder die gleichen A1 – A4 Annahmen wie früher, die sogenannten Gauss-Markov Annahmen. Selbstverständlich gilt der Beweis nur, falls diese Annahmen tatsächlich gültig sind. Deshalb ist es ratsam genau darauf zu achten, an welcher Stelle welche Annahmen getroffen werden müssen.

Der komplette Beweis findet sich im Appendix 4.A.2.

Wir haben für den Gauss-Markov Beweis eine Reihe von Annahmen benötigt, die wir auch schon für die Herleitung der Schätzfunktion für σ^2 verwendet haben.

4.5 Asymptotische Eigenschaften ('Große Stichprobeneigenschaften')

Wir haben bisher *Schätzfunktionen* $\widehat{\beta}_1$ und $\widehat{\beta}_2$ für die Parameter β_1 und β_2 hergeleitet, die es uns erlauben aus den beobachtbaren Daten einer Stichprobe *Schätzungen* für die interessierende Parameter einer unbekannten Grundgesamtheit zu berechnen. Um die Anwendbarkeit dieser Schätzfunktionen unter verschiedenen Bedingungen beurteilen zu können, müssen wir deren Eigenschaften und die zugrunde liegenden Annahmen beurteilen können.

Bisher haben wir zwei Eigenschaften von Schätzfunktionen untersucht, nämlich *Unverzerrtheit* und *Effizienz*. Diese Eigenschaften gelten unabhängig von der Stichprobengröße, also *auch* in kleinen Stichproben. Deshalb werden diese Eigenschaften häufig 'Kleine-Stichproben Eigenschaften' genannt. In manchen Fällen können auch die Stichprobenkennwertverteilungen von solchen Schätzfunktionen allgemein ermittelt werden, zum Beispiel die Verteilung des Mittelwertes bei wiederholten i.i.d. Zufallsstichprobenziehungen.

Aber oft sind wesentliche Annahmen verletzt, die zur Herleitung der 'Kleine Stichprobeneigenschaften' benötigt wurden, und häufig können in komplizierteren Fällen Eigenschaften wie Erwartungstreue oder Effizienz nicht bewiesen werden.

In solchen Fällen wird meist auf sogenannte 'Große-Stichproben Eigenschaften' (*asymptotische* Eigenschaften) zurückgegriffen, die häufig unter weniger restriktiven Annahmen bewiesen werden können.

Am einfachsten können die grundlegenden asymptotischen Konzepte anhand der Verteilung des Mittelwertes von insgesamt n Zufallsvariablen veranschaulicht werden. Wir gehen von einer unbeobachtbaren Grundgesamtheit aus, die durch einen Mittelwert μ und Varianz σ^2 charakterisiert ist.

Wir stellen uns vor, dass aus aus dieser Grundgesamtheit eine Stichprobe der Größe n gezogen wird, woraus wir den Stichprobenmittelwert $\bar{x}_n$ berechnen; das tiefgestellte n gibt die Anzahl der Beobachtungen an, auf denen der Stichprobenmittelwert beruht.

Im Folgenden untersuchen wir eine *Folge von Schätzfunktionen* $\hat{\mu}_n$ die entsteht, wenn sequentiell eine weitere Beobachtung dazukommt. Für den einfachen Stichprobenmittelwert ist eine solche Folge von Schätzfunktionen z.B.

$$\{\hat{\mu}_i\} = \left\{ x_1, \frac{x_1 + x_2}{2}, \frac{x_1 + x_2 + x_3}{3}, \ldots, \frac{x_1 + x_2 + \cdots + x_n}{n} \right\}$$

mit $i = 1, \ldots, n$. Diese Mittelwerte sind natürlich selbst wieder Zufallsvariablen mit Dichtefunktionen $f(\hat{\mu}_i)$. Die asymptotische Theorie untersucht unter anderem, wie sich eine Folge solcher Zufallsvariablen $\hat{\mu}_n$ und deren Verteilung verhält, wenn die Stichprobengröße n gegen Unendlich geht, d.h. $n \to \infty$.

Wir würden natürlich hoffen, dass die Schätzungen umso genauer werden, umso größer die Stichprobe wird. Diese Überlegungen führen zu einer der wichtigsten Eigenschaften von Schätzfunktionen, nämlich zur *Konsistenz*.

Da die folgenden Ausführungen ziemlich allgemein gehalten sind schreiben wir θ für einen beliebigen Parameter einer Verteilung, und mit $\hat{\theta}$ bezeichnen wir wie üblich

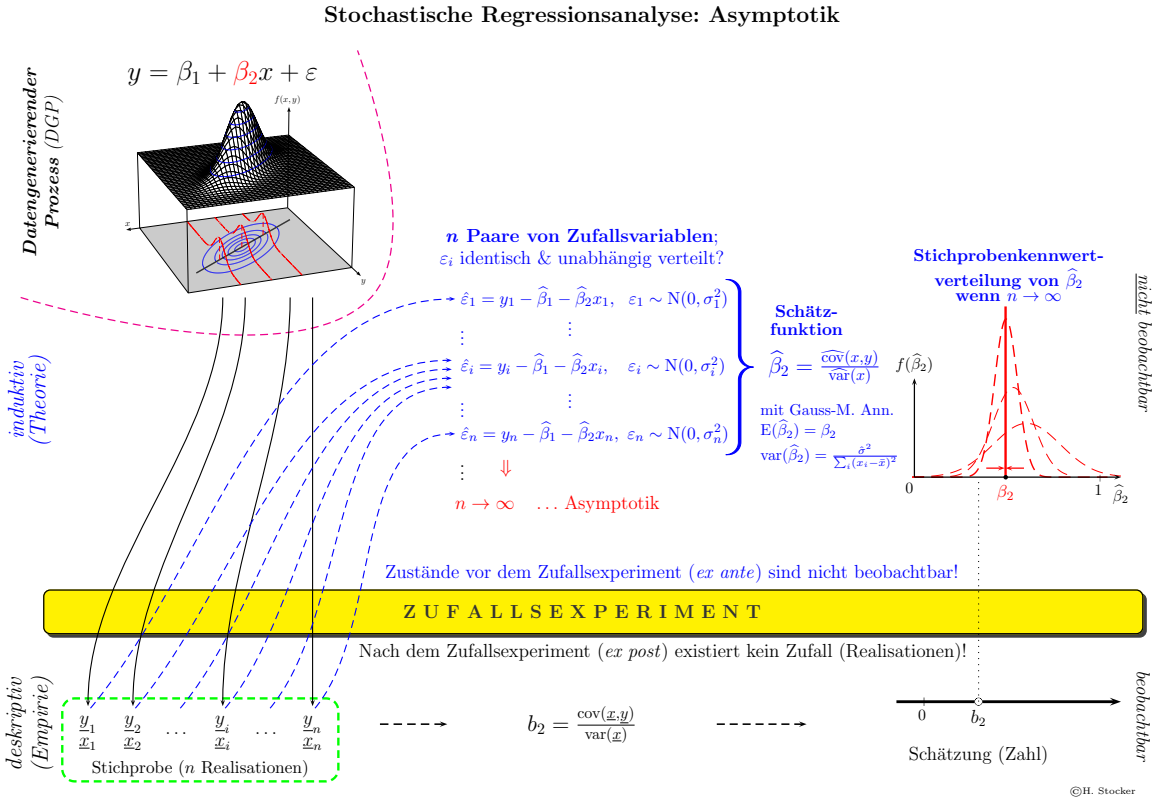

Abbildung 4.6: Konsistente Schätzfunktionen konvergieren mit zunehmender Stichprobengröße gegen den wahren Wert.

die Schätzfunktion für diesen Parameter (θ könnte zum Beispiel der Mittelwert μ oder der Steigungskoeffizient β_2 aus unserem früheren Beispiel sein).

Wie schon erwähnt sind *asymptotische Eigenschaften* vor allem in Fällen von Bedeutung,

- in denen sich 'kleine Stichprobeneigenschaften' nicht ermitteln lassen, oder
- wenn man wissen möchte, ob sich der Erwartungswert einer verzerrten Schätzfunktion $\hat{\theta}$ wenigstens mit zunehmender Stichprobengröße (d.h. für $n \to \infty$) dem wahren Parameter θ zubewegt.

4.5.1 Konsistenz

Die Konsistenz (*'consistency'*) ist vermutlich die wichtigste asymptotische Eigenschaft. Die Grundidee ist ziemlich einfach, konsistente Schätzfunktionen werden mit zunehmender Stichprobengröße immer genauer.

Die formale Definition sieht zunächst etwas schwierig aus:

Sei θ ein interessierender Parameter und $\hat{\theta}_n$ eine Schätzfunktion für θ, die auf einer Stichprobe $x_1, x_2, \ldots, x_n$ der Größe n beruht, dann ist $\hat{\theta}_n$ eine *konsistente* Schätzfunktion für θ wenn für jedes $\delta > 0$ gilt

$$\lim_{n \to \infty} \Pr\left[|\hat{\theta}_n - \theta| < \delta\right] = 1 \qquad \delta > 0$$

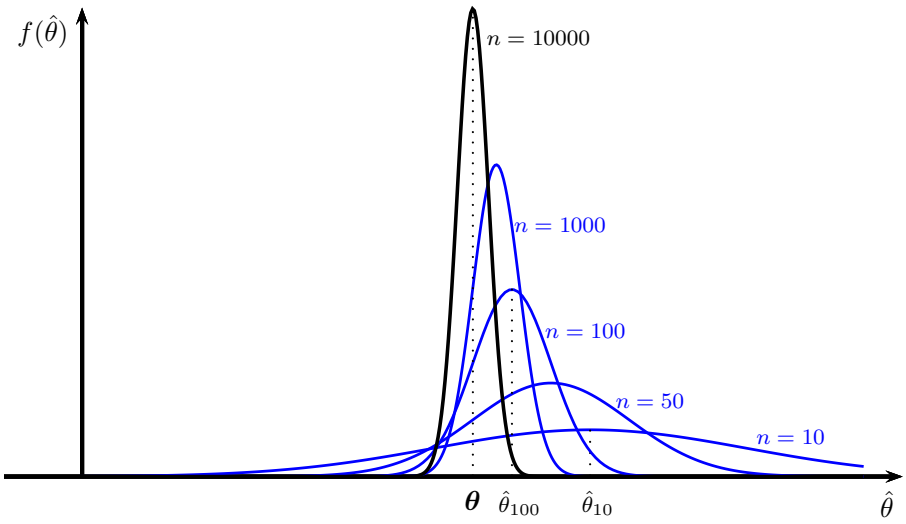

Abbildung 4.7: Konsistente Schätzfunktionen können in kleinen Stichproben verzerrt sein, konvergieren aber mit steigendem Stichprobenumfang der Wahrscheinlichkeit nach gegen den wahren Wert θ.

das heißt, dass die Wahrscheinlichkeit, dass mit steigendem Stichprobenumfang der Absolutbetrag der Differenz zwischen $\hat{\theta}_n$ und θ kleiner als eine beliebig kleine Zahl δ wird, mit zunehmendem Stichprobenumfang gegen 1 konvergiert.

Etwas ungenau lässt sich dies folgendermaßen ausdrücken: wenn der Stichprobenumfang sehr sehr groß wird, wird es sehr wahrscheinlich, dass die Schätzfunktion sehr nahe beim wahren Wert θ der Grundgesamtheit liegt.

Wenn der Stichprobenumfang n unendlich groß wird "kollabiert" die Dichtefunktion einer konsistenten Schätzfunktion $\hat{\theta}_n$ im Punkt θ (siehe Abb. 4.7).

Eine hinreichende, aber nicht notwendige Bedingung für Konsistenz ist, dass

$$\lim_{n\to\infty} \mathrm{E}(\hat{\theta}_n) = \theta \quad \text{und} \quad \lim_{n\to\infty} \mathrm{var}(\hat{\theta}_n) = 0$$

d.h. wenn die Schätzfunktion *asymptotisch unverzerrt*[9] ist und die Varianz gegen Null geht.

4.5.2 Beispiel: Beweis der Konsistenz des Stichprobenmittelwertes

Zur Demonstration beschränken uns auf den allereinfachsten Fall. Gegeben seien n identisch und unabhängig verteilte Zufallsvariablen

$$x_i \sim \text{i.i.d.}(\mu, \sigma^2) \qquad \text{mit } i = 1, \ldots, n$$

Eine Schätzfunktion für den Mittelwert dieser Zufallsvariablen $\widehat{\mu}$ ist selbst wieder eine Zufallsvariable

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

[9] Asymptotische Erwartungstreue (*Asymptotic Unbiasedness*): $\hat{\theta}_n$ ist eine asymptotisch erwartungstreue Schätzfunktion für θ wenn gilt: $\lim_{n\to\infty} \mathrm{E}(\hat{\theta}_n) = \theta$.

Der Erwartungswert und die Varianz von $\widehat{\mu}$ können einfach berechnet werden

$$\mathrm{E}(\widehat{\mu}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(x_i) = \frac{1}{n}\,n\mu = \mu$$

die Schätzfunktion $\widehat{\mu}$ ist also erwartungstreu.

Ähnlich kann die Varianz von $\widehat{\mu}$ berechnet werden

$$\begin{aligned}
\mathrm{var}(\widehat{\mu}) &= \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i - \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\right]^2 \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}(x_i) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\mathrm{cov}(x_i, x_j) \\
&= \frac{\sigma^2}{n} \qquad \text{wenn } \mathrm{var}(x_i) = \sigma^2 \text{ und } \mathrm{cov}(x_i, x_j) = 0, \text{ d.h. wenn } x_i \sim \text{i.i.d.}(\mu, \sigma^2)
\end{aligned}$$

Wir wollen nun mit Hilfe von *Chebychev's Ungleichung* zeigen, dass die Schätzfunktion $\widehat{\mu}$ auch *konsistent* ist.

Chebychevs (Tschebyscheffsche) Ungleichung besagt in diesem Fall

$$\Pr\left(|\widehat{\mu} - \mu| \geq \delta\right) \leq \frac{\mathrm{var}(\widehat{\mu})}{\delta^2} \qquad \text{für } \delta > 0$$

Beweis: Um dies zu zeigen definieren wir eine Zufallsvariable $W = \widehat{\mu} - \mu$, und $f(W)$ sei die Dichtefunktion von W. Man beachte, dass $\mathrm{E}(W^2) = \mathrm{E}[(\widehat{\mu} - \mu)^2] = \mathrm{var}(\widehat{\mu})$. δ sei eine beliebige positive Konstante.

Dann gilt

$$\begin{aligned}
\mathrm{E}(W^2) &= \int_{-\infty}^{+\infty} w^2 f(w)\,dw \\
&= \int_{-\infty}^{-\delta} w^2 f(w)\,dw + \int_{-\delta}^{+\delta} w^2 f(w)\,dw + \int_{+\delta}^{+\infty} w^2 f(w)\,dw \\
&\geq \int_{-\infty}^{-\delta} w^2 f(w)\,dw + \int_{+\delta}^{+\infty} w^2 f(w)\,dw \\
&\geq \delta^2\left[\int_{-\infty}^{-\delta} f(w)\,dw + \int_{+\delta}^{+\infty} f(w)\,dw\right] \qquad \text{(siehe Abb. 4.8)} \\
&= \delta^2 \int_{|W|>\delta} f(w)\,dw = \delta^2 \Pr(|W| \geq \delta) \\
&= \delta^2 \Pr(|\widehat{\mu} - \mu| \geq \delta)
\end{aligned}$$

Die erste Gleichung folgt aus der Definition des Erwartungswertes, die zweite Gleichung folgt, weil die Bereiche, über die integriert wird, die gesamten reellen Zahlen umfasst, und die erste Ungleichung folgt weil der weggelassene Term positiv ist. Die zweite Ungleichung folgt, weil für den Bereich der Integration $w^2 \geq \delta^2$ gilt; dies folgt

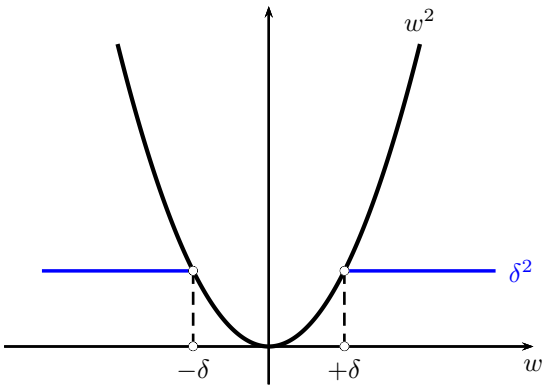

Abbildung 4.8: Für $w \leq -\delta$ und $w \geq +\delta$ ist $w^2 \geq \delta^2$.

aus der quadratischen Funktionsform, wie man anhand von Abbildung 4.8 einfach erkennen kann.

Die vorletzte Gleichung folgt aus der Definition von $\Pr(|W| \geq \delta)$. Unter Berücksichtigung von

$$\mathrm{E}(W^2) = \mathrm{E}[(\widehat{\mu} - \mu)^2] = \mathrm{var}(\widehat{\mu}) \geq \delta^2 \Pr(|\widehat{\mu} - \mu| \geq \delta)$$

folgt daraus Chebychev's Ungleichung (für $\delta > 0$)

$$\boxed{\Pr\left(|\widehat{\mu} - \mu| \geq \delta\right) \leq \frac{\mathrm{var}(\widehat{\mu})}{\delta^2}}$$

Da wir angenommen haben $x_i \sim$ i.i.d.(μ, σ^2) ist die Varianz $\mathrm{var}(\widehat{\mu}) = \sigma^2/n$ (siehe oben), deshalb folgt für die rechte Seite von Chebychev's Ungleichung

$$\frac{\mathrm{var}(\widehat{\mu})}{\delta^2} = \frac{\sigma^2}{n\delta^2}$$

Nun betrachten wir wieder eine Folge von Schätzfunktionen $\{\widehat{\mu}\}_n$, die wir verkürzt $\widehat{\mu}_n$ schreiben. Wenn die Stichprobengröße n gegen Unendlich geht folgt

$$\lim_{n\to\infty} \frac{\sigma^2}{(n\delta^2)} \to 0$$

und deshalb auch für die linke Seite von Chebychev's Ungleichung

$$\lim_{n\to\infty} \Pr\left(|\widehat{\mu}_n - \mu| > \delta\right) \to 0$$

Insbesondere kann $\delta > 0$ auch beliebig klein sein.

Man sagt in diesem Fall auch, dass die zentrierten Zufallsvariablen $\widehat{\mu}_n - \mu$ *in Wahrscheinlichkeit* gegen Null konvergieren, bzw. dass $\widehat{\mu}$ eine *konsistente* Schätzfunktion für μ ist.

Dies kann man dahingehend interpretieren, dass sich die empirischen Mittelwerte $\bar{x}_n$ mit zunehmendem n um den 'wahren' Wert μ stabilisieren (dies impliziert allerdings nicht, dass diese Annäherung monoton erfolgen muss, es kann immer wieder Ausreißer geben).

Erst wenn der Stichprobenumfang n unendlich groß ist kollabiert die Stichprobenkennwertverteilung von $\widehat{\mu}$ im 'wahren' μ.

Offensichtlich gilt dies nicht nur für die Konsistenz von $\widehat{\mu}$, sondern z.B. auch für die Konsistenz des Steigungskoeffizienten $\widehat{\beta}_2$ mit $\text{var}(\widehat{\beta}_2) = \sigma^2/[\sum_{i=1}^{n}(x_i - \bar{x})^2]$.

Dies ist ein einfaches Beispiel für ein *schwaches Gesetz der großen Zahl*. Wir haben dies unter der relativ strengen Annahme gezeigt, dass die einzelnen Zufallsvariablen x_i alle i.i.d.-verteilt sind, d.h. identisch verteilt und untereinander stochastisch unabhängig sind, sowie endliche Erwartungswerte und Varianzen haben.

Es gibt zahlreiche weitere Gesetze der großen Zahl, die teilweise mit weniger strengen Annahmen auskommen, z.B. dass die Zufallsvariablen nur paarweise unkorreliert und die Folge ihrer Varianzen beschränkt sein muss.

Generell sind 'Gesetze der großen Zahlen' meist Aussagen über das Verhalten von Kenngrößen (z.B. Momenten) einer *großen* Zahl von Zufallsvariablen. Beweise von 'Gesetzen der großen Zahlen' erfolgen meist mit Hilfe des Konzepts der Konvergenz der Wahrscheinlichkeit nach (*'Convergence in Probability'*, auch Stochastische Konvergenz genannt), die eine Beschreibung des Verhaltens von Zufallsvariablen bei wachsendem Stichprobenumfang erlaubt.

Beispiel 1: Würfeln Stellen Sie sich vor, Sie würfeln sehr oft mit einem fairen Würfel. Wir erwarten, dass im Durchschnitt jede Zahl in einem 1/6 der Fälle erscheint, und wir wissen bereits, dass der Erwartungswert der Augenzahl 3.5 ist (das mit Wahrscheinlichkeiten gewichtete Mittel über alle möglichen Ausprägungen).

Wir können den Computer z.B. 300 Mal würfeln lassen und nach jedem Wurf den neuen Anteil und den Mittelwert berechnen lassen. Bei den ersten sechs Würfen erhalten Wir die Sequenz: 5, 5, 3, 3, 3, 2, ...

Die dicke blaue Linie in Abbildung 4.9 zeigt die fortlaufenden Mittelwerte über 300 Würfe. Für den ersten Wurf ist die Augenzahl zugleich der Mittelwert 5, nach zwei Würfen erhalten Sie wieder den Mittelwert $(5 + 5)/2 = 5$, nach dem 3. Wurf $(5 + 5 + 3)/3 = 4.33$, ..., für den 6. Wurf $(5 + 5 + 3 + 3 + 3 + 2)/6 = 3.5$, usw.

Die grau eingezeichneten Linen zeigen die Konvergenz der Mittelwerte für andere Folgen von Realisationen.

Abbildung 4.9 kann z.B. mit folgendem R-Code erzeugt werden. Dabei wird die kumulierte Summe (`cumsum()`) durch den Trend `i` dividiert, wir erhalten also Mittelwerte für zunehmende Stichprobengrößen

```
n <- 300
i <- 1:n
set.seed(123456)
means <- cumsum(sample(x = 1:6, size = n, replace = TRUE))/i

x11()
plot(means, type = "l", ylim = c(1,6),
     main = "Mittelwerte bei zunehmender Stichprobengröße",
     ylab = "Mittelwert", xlab = "Stichprobengröße n",col="blue", lwd=2)
for (j in 1:7) { # weitere Folgen
```

```
  lines(cumsum(sample(x = 1:6, size = n, replace = TRUE))/i, col="gray")
}
abline(h=3.5, col = "red", lty=2, lwd=2)
```

Wiederum, über umso mehr Würfe gemittelt wird, umso mehr nähern sich diese Mittelwerte dem Erwartungswert 3.5 an. Dies ist kein Zufall, sondern ist eine Konsequenz des Gesetzes der Großen Zahl.

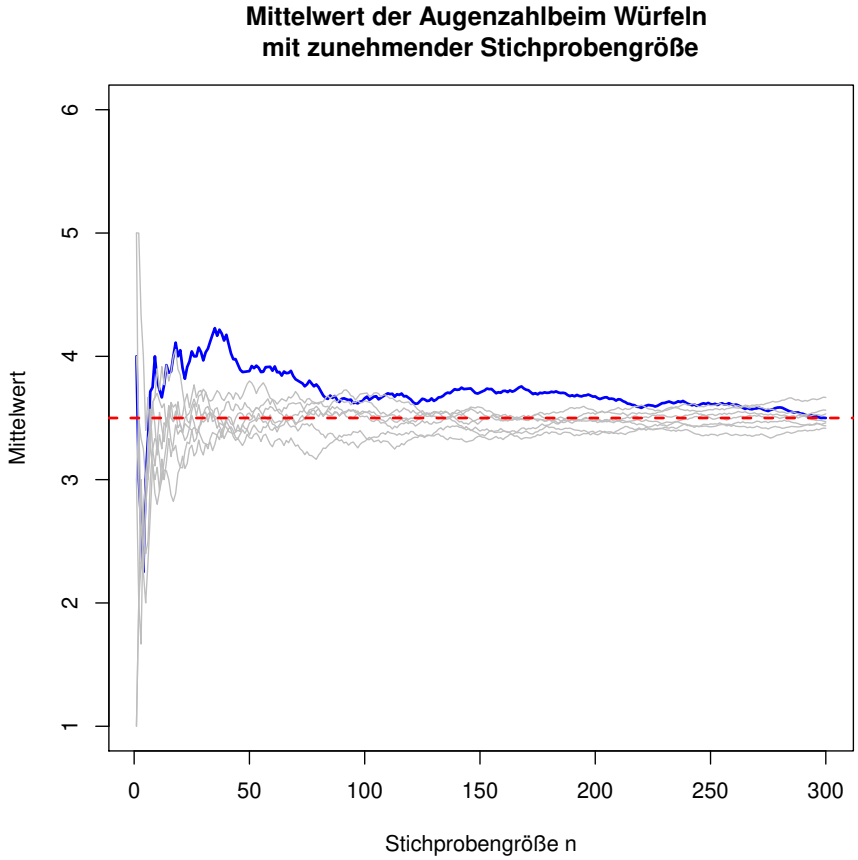

Abbildung 4.9: Gesetz der Großen Zahl: Bei sehr häufiger Wiederholung des Zufallsexperiments unter identischen Bedingungen nähert sich die Folge der empirischen Mittelwerte mit steigendem Stichprobenumfang dem theoretischen Erwartungswert 3.5 an.
R Code: siehe Appendix Seite 51

Beispiel: Stundenlöhne Für dieses Beispiel verwenden wir Brutto-Stundenlöhne (StdL) von 4809 unselbständig Beschäftigten in Österreich (EU-Silc 2015). Der durchschnittliche Stundenlohn über alle Beobachtungen beträgt $\overline{\text{StdL}} = 16.12$ Euro.

Wir betrachten diese 4809 Beobachtungen als Grundgesamtheit, lassen den Computer daraus Stichproben ziehen (mit Zurücklegen), und berechnen für jede dieser Stichproben den Mittelwert.

Abbildung 4.10 zeigt das Ergebnis, wiederum nähern sich die Mittelwerte der Stichproben mit zunehmender Stichprobengröße dem wahren Wert (d.h. dem durchschnittlichen Stundenlohn der Grundgesamtheit $\overline{\text{StdL}} = 16.12$) an.

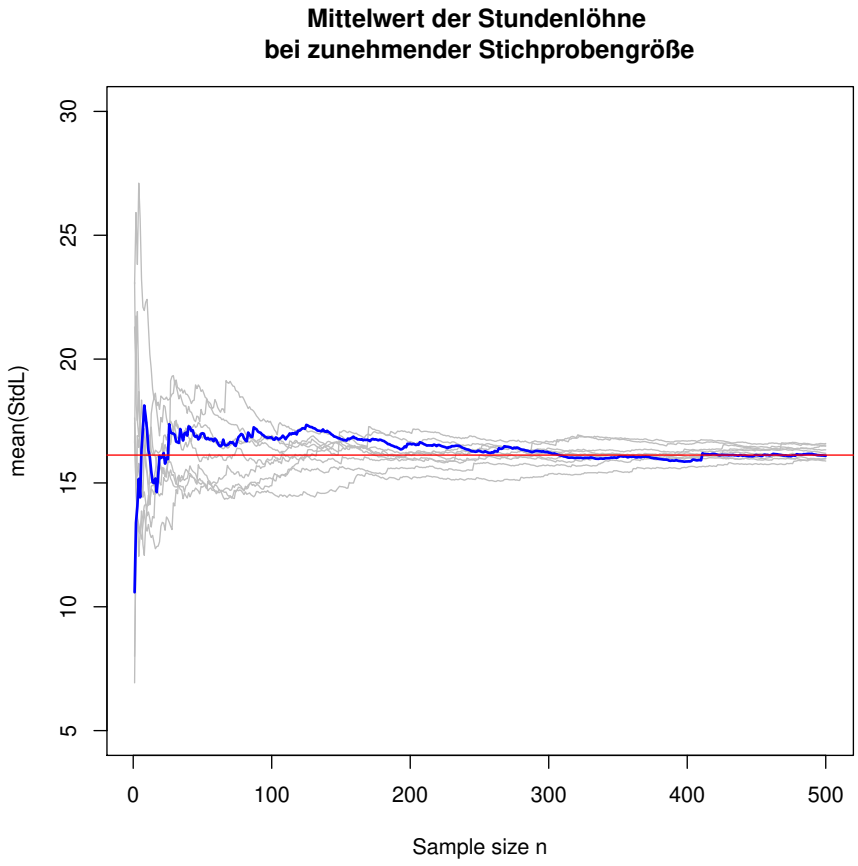

Abbildung 4.10: Gesetz der Großen Zahl: Aus 4809 beobachteten Stundenlöhnen werden mehrere Stichproben gezogen. Für jede Stichprobe werden für $i = 1, 2, \ldots, 500$ die Mittelwerte über jeweils die ersten i Beobachtungen gerechnet. Mit zunehmender Stichprobengröße n nähern sich die Mittelwerte aus diesen Stichproben dem wahren Wert $\overline{\text{StdL}} = 16.12$ an.

Konsistenz einer Schätzfunktion bedeutet allgemein, dass eine Folge von Schätzfunktionen $\hat{\theta}_n$ stochastisch gegen das wahre θ konvergiert, also ein Gesetz der großen Zahl erfüllt ist; in anderen Worten, bei Konsistenz konvergiert eine Folge von Schätzfunktionen $\hat{\theta}_n$ in Wahrscheinlichkeit gegen den wahren Wert θ.

Dies wird oft kürzer geschrieben als

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

Dafür hat sich auch die Notation des sogenannten probability-limits (plim) eingebürgert

$$\operatorname{plim} \hat{\theta}_n = \theta$$

dies ist einfach nur eine andere Schreibweise für $\hat{\theta} \xrightarrow{p} \theta$, was wiederum nur eine Kurzschreibweise für

$$\lim_{n\to\infty} \Pr\left[|\hat{\theta}_n - \theta| < \delta\right] = 1 \qquad \delta > 0$$

ist, wobei δ beliebig klein gewählt werden kann.

Man beachte, dass es *keine* einfache Beziehung zwischen Effizienz und Konsistenz einer Schätzfunktion gibt. Eine Schätzfunktion kann zwar effizient und erwartungstreu sein, aber trotzdem *nicht* konsistent sein (z.B. wenn die Schätzfunktion nicht von n abhängt). Häufiger ist der Fall, dass eine Schätzfunktion zwar konsistent, aber nicht erwartrungstreu ist! Natürlich können Schätzfunktionen auch konsistent und effizient, oder weder konsistent noch effizient sein.

Die Bedeutung der Konsistenz resultiert ganz wesentlich daraus, dass das Rechnen mit 'probability-limits' relativ einfach ist. Intuitiv können wir uns vorstellen, dass – wenn das 'probability-limit' existiert und gegen einen festen Wert konvergiert ($\operatorname{plim} \hat{\theta}_n = \theta$) – wir mit 'probability-limits' wie mit 'normalen Zahlen' rechnen können.

Regeln für das Rechnen mit 'probability-limits'

1. Für eine beliebige Konstante c gilt

$$\operatorname{plim} c = c$$

2. Seien $\hat{\theta}_n$ und $\widehat{\vartheta}_n$ Zufallsvariablen (z.B. Schätzfunktionen) mit $\operatorname{plim} \hat{\theta}_n = \theta$ und $\operatorname{plim} \widehat{\vartheta}_n = \vartheta$ (θ wird *theta* und ϑ *vartheta* gesprochen) dann gilt

$$\begin{aligned} \operatorname{plim}(\hat{\theta}_n + \widehat{\vartheta}_n) &= \operatorname{plim} \hat{\theta}_n + \operatorname{plim} \widehat{\vartheta}_n = \theta + \vartheta \\ \operatorname{plim}(\hat{\theta}_n \widehat{\vartheta}_n) &= \operatorname{plim} \hat{\theta}_n \operatorname{plim} \widehat{\vartheta}_n = \theta\vartheta \\ \operatorname{plim}\left(\frac{\hat{\theta}_n}{\widehat{\vartheta}_n}\right) &= \frac{\operatorname{plim} \hat{\theta}_n}{\operatorname{plim} \widehat{\vartheta}_n} = \frac{\theta}{\vartheta} \qquad (\text{für } \vartheta \neq 0) \end{aligned}$$

   Man beachte, dass die letzten beiden Eigenschaften für den Erwartungswertoperator nur dann gelten, wenn $\hat{\theta}$ und $\widehat{\vartheta}$ stochastisch unabhängig sind. Aus diesen Gründen ist Konsistenz üblicherweise deutlich einfacher zu beweisen als Erwartungstreue oder Effizienz.

3. Wenn $\hat{\theta}$ eine konsistente Schätzfunktion für θ ist und $h(\hat{\theta})$ eine stetige Funktion von $\hat{\theta}$ ist gilt

$$\operatorname{plim} h(\hat{\theta}) = h(\theta)$$

   Man sagt auch, dass sich die Konsistenz 'überträgt'. Wenn $\hat{\theta}$ eine konsistente Schätzfunktion für θ ist, dann ist z.B. $1/\hat{\theta}$ auch eine konsistente Schätzfunktion für $1/\theta$ (für $\hat{\theta} \neq 0$); oder $\ln \hat{\theta}$ ist eine konsistente Schätzfunktion für für $\ln \theta$ (für $\hat{\theta} > 0$). Dies gilt nicht für den Erwartungswertoperator!

4.5.3 Beispiel: Unverzerrtheit und Konsistenz von OLS Schätzfunktionen mit stochastischen Regressoren

Bisher haben wir angenommen, dass die erklärende Variable x deterministisch ist, das heißt, dass bei wiederholten Stichprobenziehungen nur verschiedene y gezogen werden, aber die x (z.B. durch einen Experimentator) 'fest gehalten' werden.

In diesem Unterabschnitt interessieren uns die Eigenschaften von OLS-Schätzfunktionen, wenn die erklärende Variable x ebenso stochastisch ist. Die OLS-Schätzfunktion für den Steigungskoeffizienten ist bekanntlich

$$\widehat{\beta}_2 = \frac{\widehat{\text{cov}}(x, y)}{\widehat{\text{var}}(x)} = \frac{\sum \ddot{x}_i \ddot{y}_i}{\sum \ddot{x}_i^2}$$

wobei $\ddot{x}_i := x_i - \bar{x}$ und $\ddot{y}_i := y_i - \bar{y}$.

Wie schon früher erwähnt benötigen wir für den Fall stochastischer Regressoren einige zusätzliche Annahmen

1. Die (y_i, x_i) für $i = 1, \ldots, n$ sind über die Beobachtungen i identisch und unabhängig verteilt (i.i.d.), es handelt sich also um eine echte *Zufallsstichprobe*. Jede einzelne Ziehung aus einer gemeinsamen Wahrscheinlichkeits- oder Dichtefunktion liefert ein Paar von zwei Zufallsvariablen (y_i, x_i), und deren gemeinsame Wahrscheinlichkeiten entsprechen den Wahrscheinlichkeiten der Grundgesamtheit. Dies bedeutet, dass sich die Grundgesamtheit (bzw. der Datengenerierende Prozess) zwischen den Ziehungen nicht ändert (alle (y_i, x_i) sind identisch verteilt), und das Ergebnis einer Ziehung hat keinen direkten Einfluss auf irgend eine andere Ziehung (Unabhängigkeit).

2. Die Erwartungswerte und Varianzen von y_i und x_i sind nicht unendlich groß, und 'große Ausreißer sind unwahrscheinlich'. Etwas technischer kann dies mit Hilfe der vierten Momente geschrieben werden als

   $$0 < \text{E}(y_i^4) < \infty \quad \text{und} \quad 0 < \text{E}(x_i^4) < \infty$$

   Dies bedeutet, dass die Kurtosis nicht unendlich groß sein darf. Wir wissen, dass die Varianz das zweite Moment einer Zufallsvariable ist, intuitiv können wir uns deshalb vorstellen, dass die 'Varianz der Varianzen' nicht unendlich groß werden darf. Diese Annahme wird benötigt, damit die asymptotischen Approximationen gültig sind. Für die meisten Anwendungen ist diese Annahme nicht sonderlich streng.

3. $\text{E}(\varepsilon_i | x_1, \ldots, x_n) = 0$ oder bei Gültigkeit der ersten Annahme $\text{E}(\varepsilon_i | x_i) = 0$: Der auf alle x_i bedingte Erwartungswert der Störterme ist gleich Null. Wir haben bereits gesehen, dass dies *stochastische Unabhängigkeit* ε_i und x_i impliziert. Dies ist eine strengere Annahme als $\text{cov}(\varepsilon_i, x_i) = 0$, da die Kovarianz nur ein Maß für die *lineare* Abhängigkeit ist. In anderen Worten, $\text{E}(\varepsilon_i | x_i) = 0$ impliziert $\text{cov}(\varepsilon_i, x_i) = 0$, aber nicht umgekehrt!

   Wir haben bereits gesehen, dass diese Bedingung erforderlich ist für die Erwartungstreue der OLS Schätzfunktion.

Um die Erwartungstreue zu überprüfen setzen wir wieder den wahren Zusammenhang $\ddot{y}_i = \beta_2 \ddot{x}_i + \varepsilon_i$ ein und bilden den Erwartungswert

$$\text{E}[\widehat{\beta}_2] = \beta_2 + \text{E}\left[\frac{\sum \ddot{x}_i \varepsilon_i}{\sum \ddot{x}_i^2}\right]$$

Wenn nun die $\ddot{x}_i$ stochastisch sind hängt die Erwartungstreue von der gemeinsamen Wahrscheinlichkeitsverteilung von $\ddot{x}_i$ und ε_i ab (man beachte, dass bei stochastischen

Regressoren auch der Nenner $\sum_i \ddot{x}_i^2$ eine Zufallsvariable ist, und bekanntlich ist $\mathrm{E}(\sum_i \ddot{x}_i\varepsilon_i / \sum_i \ddot{x}_i^2) \neq \mathrm{E}(\sum_i \ddot{x}_i\varepsilon_i) / \mathrm{E}(\sum_i \ddot{x}_i^2)$!).

Die *Erwartungstreue* der Schätzfunktion $\widehat{\beta}_2$ können wir nur zeigen wenn wir annehmen, dass alle $\ddot{x}_i$ (d.h. $\ddot{x}_1, \ddot{x}_2, \ldots \ddot{x}_n$) stochastisch unabhängig von allen ε_i (d.h. $\varepsilon_1, \varepsilon_2, \ldots \varepsilon_n$) sind, oder für i.i.d. Stichproben $\mathrm{E}(\varepsilon_i|x_i) = 0$. In diesem Fall gilt

$$\begin{aligned} \mathrm{E}\left[\frac{\sum \ddot{x}_i\varepsilon_i}{\sum \ddot{x}_i^2}\right] &= \sum\left[\mathrm{E}\left(\frac{\ddot{x}_i}{\sum \ddot{x}_i^2}\,\varepsilon_i\right)\right] \\ &= \sum\left[\mathrm{E}\left(\frac{\ddot{x}_i}{\sum \ddot{x}_i^2}\right)\mathrm{E}(\varepsilon_i|x_i)\right] = 0 \end{aligned}$$

da $\mathrm{E}(\varepsilon_i|x_i) = 0$.[10]

Um die Konsistenz zu zeigen bilden wir das probability-limit und wenden die entsprechenden Rechenregeln an

$$\begin{aligned} \operatorname{plim} \widehat{\beta}_2 &= \operatorname{plim} \beta_2 + \operatorname{plim}\left[\frac{\sum \ddot{x}_i\varepsilon_i}{\sum \ddot{x}_i^2}\right] \\ &= \beta_2 + \left[\frac{\operatorname{plim}\sum \ddot{x}_i\varepsilon_i}{\operatorname{plim}\sum \ddot{x}_i^2}\right] \\ &= \beta_2 + \frac{\operatorname{plim}\left[\frac{1}{n}\sum \ddot{x}_i\varepsilon_i\right]}{\operatorname{plim}\left[\frac{1}{n}\sum \ddot{x}_i^2\right]} \end{aligned}$$

Man beachte, dass $\sum_{i=1}^n \ddot{x}_i^2$ eine Summe von n positiven Zufallsvariablen ist. Wenn n gegen Unendlich geht würden wir deshalb erwarten, dass $\sum_{i=1}^n \ddot{x}_i^2$ unendlich groß wird. Deshalb dividieren wir Zähler und Nenner in der dritten Zeile durch n und erhalten damit eine konsistente Schätzfunktion für die Varianz und Kovarianz der Grundgesamtheit. Unter den vorher getroffenen Annahmen können wir davon ausgehen, dass diese gegen einen festen Wert konvergieren.

Die Schätzfunktion $\widehat{\beta}_2$ ist also *konsistent*, wann immer die Störterme der Grundgesamtheit ε_i und die erklärenden Variablen $\ddot{x}_i$ stochastisch unabhängig sind, d.h. wenn

$$\operatorname{plim}\left[\frac{1}{n}\sum \ddot{x}_i\varepsilon_i\right] = 0 \quad \text{und} \quad \operatorname{plim}\left[\frac{1}{n}\sum \ddot{x}_i^2\right] = \sigma_{\ddot{x}}^2 > 0$$

da in diesem Fall

$$\operatorname{plim} \widehat{\beta}_2 = \beta_2 + \frac{0}{\sigma_{\ddot{x}}^2} = \beta_2$$

Im Unterschied zum Beweis für die Erwartungstreue müssen für Konsistenz nicht *alle* $x_1, x_2, \ldots x_n$ mit allen $\varepsilon_1, \varepsilon_2, \ldots \varepsilon_n$ unkorreliert sein, sondern es genügt für Konsistenz, wenn die x_i einer Beobachtung oder Zeitperiode mit den entsprechenden ε_i der gleichen Beobachtung oder Periode unkorreliert sind.

Wichtig ist aber nach wie vor die Annahme, dass die Störterme der Grundgesamtheit ε_i mit dem Regressor x_i unkorreliert sind. Ist diese Annahme nicht erfüllt sind auch OLS-Schätzfunktionen nicht konsistent!

[10] Das zweite Gleichheitszeichen folgt aus dem Gesetz der iterierten Erwartungen $\mathrm{E}(\varepsilon_i) = \mathrm{E}_x[\mathrm{E}(\varepsilon_i|x_i)]$.

Im wesentlichen verlangen wir von den Regressoren x also, dass sie nur über den spezifizierten Zusammenhang $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ mit den y verknüpft sind, und dass es keine anderen nicht spezifizierten Zusammenhänge zwischen x und y gibt, d.h. dass die x_i und ε_i stochastisch unabhängig sind. Nicht spezifizierte Zusammenhänge erzeugen in der Regel eine Korrelation zwischen den ε und x, was dazu führt, dass OLS Schätzfunktionen weder erwartungstreu, effizient noch konsistent sind! Wie schon erwähnt tritt dies z.B. bei simultanen Gleichungssystemen oder 'omitted variables' auf.

4.5.4 Asymptotische Normalverteilung

Eine Schätzfunktion ist *asymptotisch normalverteilt*, wenn deren normierte Stichprobenkennwertverteilung mit zunehmender Stichprobengröße gegen die Normalverteilung konvergiert. Das dahinter liegende stochastische Konzept ist eine *Konvergenz hinsichtlich der Verteilung* (*'Convergence in Distribution'*). Vereinfacht gesprochen bedeutet dies, dass die Verteilung einer Folge von (normierten) Schätzfunktionen $\hat{\theta}_n$ aus Stichproben des Umfangs n, die alle derselben Grundgesamtheit entnommen wurden, mit zunehmendem Stichprobenumfang in eine Normalverteilung übergeht, und dies unabhängig von der Verteilung der Grundgesamtheit! Beweise der Konvergenz hinsichtlich der Verteilung führen zu *Zentralen Grenzwertsätzen*.

Bei den Zentralen Grenzwertsätzen handelt es sich um eine Familie schwacher Konvergenzaussagen aus der Wahrscheinlichkeitstheorie. Allen gemeinsam ist die Aussage, dass die (normierte) Summe einer großen Zahl von unabhängigen, identisch verteilten Zufallsvariablen annähernd (standard-)normalverteilt ist. Dies erklärt zum Teil die Sonderstellung der Normalverteilung.

Die bekannteste Aussage wird auch einfach als "Der Zentrale Grenzwertsatz" bezeichnet und befasst sich mit unabhängigen, identisch verteilten Zufallsvariablen, deren Erwartungswert und Varianz existieren und nicht unendlich groß sind.

$$\lim_{n\to\infty} \Pr\left\{ \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sigma} \leq y \right\} = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2} dz$$

oder einfacher

$$\sqrt{n}\,(\hat{\theta} - \theta) \xrightarrow{d} \mathrm{N}(0, \sigma^2)$$

d.h. wenn $x_i \sim$ i.i.d.(θ, σ^2) und $0 < \sigma^2 < \infty$, dann konvergiert die Verteilung von $\sqrt{n}(\hat{\theta} - \theta)$ gegen die Normalverteilung mit Mittelwert Null und Varianz σ^2.

Die Multiplikation mit $\sqrt{n}$ ist erforderlich, da aufgrund der Konsistenz der Schätzfunktion die Varianz von $(\hat{\theta}_n - \theta)$ mit zunehmendem n immer kleiner wird. Durch die Multiplikation mit $\sqrt{n}$ wird diese Abnahme der Varianz exakt ausgeglichen und wir erhalten eine stabile *Grenzverteilung*.

Es gibt eine ganze Reihe von zentralen Grenzwertsätzen, die teilweise deutlich allgemeiner und unter weniger strengen Annahmen gelten.

Man kann zeigen, dass auch die *OLS Schätzfunktionen* unter den obigen Annahmen asymptotisch normalverteilt sind. Die Beweise dazu finden sich in jedem fortgeschrittenen Lehrbuch zur Ökonometrie, hier werden wir die Auswirkungen des zentralen Grenzwertsatzes nur anhand zweier einfacher Monte Carlo Simulationen vorführen.

Beispiel 1: Würfeln Siehe Abbildung 4.11.

Beispiel 2: Stundenlöhne Siehe Abbildung 4.12.

4.5.5 Asymptotische Effizienz

$\hat{\theta}$ sei eine Schätzfunktion für θ. Die Varianz der asymptotischen Verteilung von $\hat{\theta}$ heißt asymptotische Varianz von $\hat{\theta}$. Wenn $\hat{\theta}$ konsistent ist und die asymptotische Varianz kleiner ist als die aller anderen konsistenten Schätzfunktionen, dann heißt $\hat{\theta}$ *asymptotisch effizient*. Man kann zeigen, dass OLS Schätzfunktionen bei stochastischen Regressoren asymptotisch effizient sind.

4.6 Der Mittlere Quadratische Fehler (*Mean Square Error*, MSE)

Kehren wir noch einmal zurück zu Abbildung 4.2 (Seite 3) mit den drei Zielscheiben zurück.

Die Wahl zwischen dem ersten und dritten Schießgewehr haben wir mit Hilfe des Kriteriums der Erwartungstreue gerechtfertigt, häufig werden erwartungstreue Schätzfunktionen gegenüber verzerrten Schätzfunktionen bevorzugt.

Die Wahl zwischen dem ersten und zweiten Schießgewehr, also zwei erwartungstreuen Schätzfunktionen, trafen wir anhand des Kriteriums der Effizienz, ceteris paribus werden wir effiziente (genauere) Schätzfunktionen gegenüber nicht effizienten bevorzugen.

Was aber, wenn wir uns zwischen dem zweiten und dritten Schießgewehr entscheiden müssen? Offensichtlich ist die zweite Schätzfunktion zwar erwartungstreu, aber sehr ungenau. Die dritte Schätzfunktion ist zwar 'ein bisschen' verzerrt, aber viel genauer als die zweite Schätzfunktion.

In solchen Fällen ist die Entscheidung schwieriger, es gibt einen klaren *trade-off*. Um eine Wahl zwischen den beiden rechtfertigen zu können würden wir entsprechende Verlustfunktionen benötigen, also zusätzliche Information.

Ein viel einfacheres Kriterium, das allerdings keine Entscheidung 'rechtfertigen' kann, aber immerhin manchmal nützlich ist den prinzipiellen *trade-off* deutlich zu machen, ist der *'Mean Square Error'* (MSE), der gewissermaßen Varianz und Verzerrung in einer Kennzahl zusammenfasst (vgl. Abbildung 4.13).

Wir beginnen wieder ganz allgemein und bezeichnen einen interessierenden Parameter einer Verteilung mit θ, und die Schätzfunktion für diesen Parameter mit $\hat{\theta}$. Eine konkrete Schätzung erhält man, wenn man die Stichprobenbeobachtungen in die Formel für $\hat{\theta}$ einsetzt.

Folgende Konzepte sind im folgenden von Bedeutung:

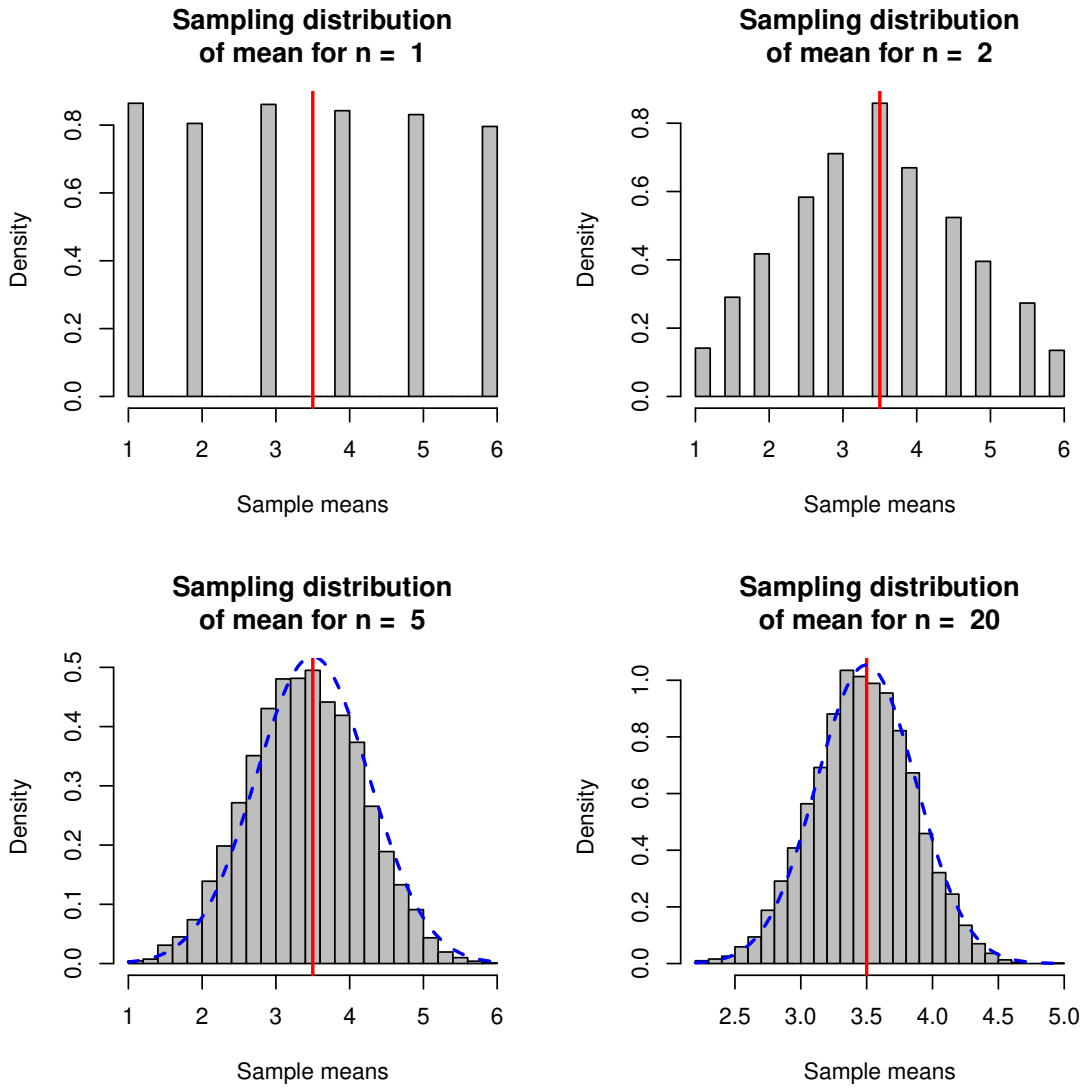

Abbildung 4.11: Empirische Simulation von Stichprobenkennwertvertelungen und des *Zentralen Grenzwertsatzes*: Wir betrachten die unendliche Folge aller Zahlen, die wir beim Werfen eines fairen Würfels erhalten können. Aus dieser Folge ziehen wir sehr oft (in diesem Beispiel 10000 mal) Stichproben der Größe n, und berechnen für jede dieser Stichproben das arithmetische Mittel.
Die Grafik links oben zeigt die relative Häufigkeitsverteilung des arithmetischen Mittels für die Stichprobengröße $n = 1$; in diesem Fall ist jeder Mittelwert die Beobachtung selbst, und jede Zahl erscheint gleich oft.
Die restlichen Grafiken zeigen, dass sich die *Verteilung der Stichprobenmittelwerte* (d.h. die Stichprobenkennwertverteilung) mit zunehmenden Stichprobenumfang einer Normalverteilung annähert. Dies demonstriert die Gültigkeit des Zentralen Grenzwertsatzes.
In diesem Beispiel wird bereits für kleine Stichprobengrößen eine relativ gute Approximation erreicht.
R Code: siehe Appendix Seite 51

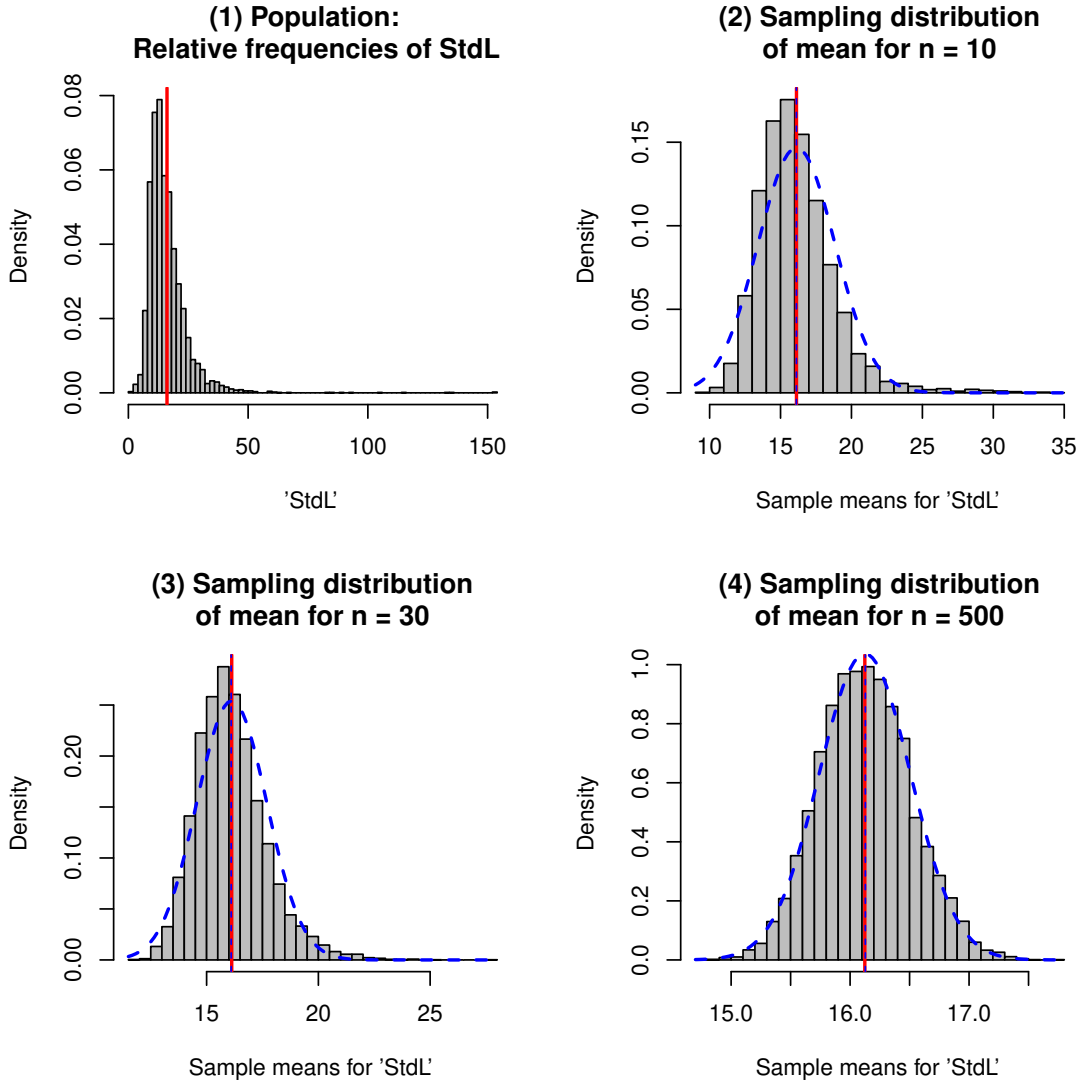

Abbildung 4.12: (1) Grafik links oben: relative Häufigkeiten der Brutto-Stundenlöhne von insgesamt über 4809 österreichischen ArbeitnehmerInnen (EU-Silc, 2015). Diese repräsentieren für uns die Grundgesamtheit.
Zentraler Grenzwertsatz: Werden aus dieser Grundgesamtheit viele i.i.d. Stichproben der Größe n gezogen und die jeweiligen Mittelwerte berechnet, so nähert sich die *Verteilung dieser Mittelwerte* mit zunehmendem Stichprobenumfang n einer Normalverteilung an. Da die Grundgesamtheit eine sehr (rechts-)schiefe Verteilung aufweist werden ziemlich große Stichproben benötigt um eine gute Approximation durch die Normalverteilung zu erhalten.
R Code: siehe Appendix Seite 51

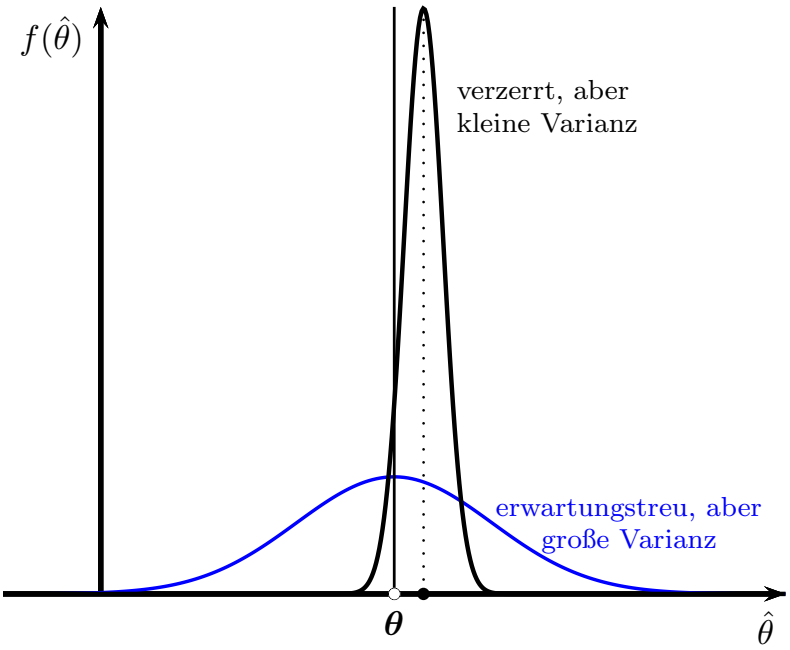

Abbildung 4.13: Mean Square Error Abwägung zwischen erwartungstreuen Schätzfunktionen mit großer Varianz und verzerrten Schätzfunktionen mit kleiner Varianz.

$$\begin{aligned} \text{Stichprobenfehler } (\textit{sampling error}) &= \hat{\theta} - \theta \\ \text{Verzerrung (Bias)} &= \mathrm{E}(\hat{\theta}) - \theta \\ \text{Mean Square Error} &= \mathrm{E}(\hat{\theta} - \theta)^2 \\ \text{Varianz} &= \mathrm{E}\left[\hat{\theta} - \mathrm{E}(\hat{\theta})\right]^2 \end{aligned}$$

Der Stichprobenfehler ist einfach der Unterschied zwischen der Schätzfunktion und dem wahren Wert der Grundgesamtheit (also unbeobachtbar). Die Größe des Stichprobenfehlers wird sich üblicherweise von Stichprobe zu Stichprobe unterscheiden. Die Verzerrung ist die Differenz zwischen dem Erwartungswert der Stichprobenkennwertverteilung einer Schätzfunktion und dem wahren Wert der Grundgesamtheit. Diese ist für eine Schätzfunktion ein fester Wert der Null oder ungleich Null sein kann, sich aber nicht zwischen Stichproben unterscheidet.

Der Mean Square Error misst die Streuung der Verteilung einer Schätzfunktion um den wahren Wert. Diese ähnelt der Varianz, aber während die Varianz die Streuung um den Erwartungswert der Verteilung misst, gibt der MSE die Streuung um den wahren Wert an. Für erwartungstreue Schätzfunktionen sind Varianz und MSE natürlich gleich, aber für nicht erwartungstreue Schätzfunktionen müssen sie unterschieden werden.

Dies kann folgendermaßen gezeigt werden:

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= \mathrm{E}(\hat{\theta}-\theta)^2 \\
&= \mathrm{E}[\hat{\theta}-\mathrm{E}(\hat{\theta})+\mathrm{E}(\hat{\theta})-\theta]^2 \qquad (\mathrm{E}(\hat{\theta}) \text{ addieren und subtr.}) \\
&= \mathrm{E}[\hat{\theta}-\mathrm{E}(\hat{\theta})]^2+\mathrm{E}[\mathrm{E}(\hat{\theta})-\theta]^2+2\,\mathrm{E}[\hat{\theta}-\mathrm{E}(\hat{\theta})][\mathrm{E}(\hat{\theta})-\theta] \\
&= \mathrm{E}[\hat{\theta}-\mathrm{E}(\hat{\theta})]^2+\mathrm{E}[\mathrm{E}(\hat{\theta})-\theta]^2+ \\
&\quad +2\{[\mathrm{E}(\hat{\theta})]^2-[\mathrm{E}(\hat{\theta})]^2-\theta\,\mathrm{E}(\hat{\theta})+\theta\,\mathrm{E}(\hat{\theta})\} \\
&= \mathrm{E}[\hat{\theta}-\mathrm{E}(\hat{\theta})]^2+\mathrm{E}[\mathrm{E}(\hat{\theta})-\theta]^2 \\
&= \mathrm{var}(\hat{\theta})+[\mathrm{Bias}(\hat{\theta})]^2
\end{aligned}$$

Dieser Zusammenhang gilt für alle Schätzfunktionen. Akademische Forscher neigen oft dazu, unverzerrte Schätzfunktionen selbst auf Kosten eines größeren MSE zu bevorzugen, da sie ihre Studie als eine von vielen Studien wahrnehmen und hoffen, dass sich die größere Streuung über die vielen Studien mittelt. In vielen praktischen Anwendungen gibt es allerdings nur eine einzige Schätzung (Studie), und da spielt es keine Rolle, ob der Fehler aus einer systematischen Verzerrung oder einer größeren Varianz resultiert – Fehler ist Fehler. Für Prognosen kann ein kleiner MSE manchmal wichtiger sein als die Erwartungstreue.

Es gibt auch eine enge Beziehung zwischen dem MSE und der Konsistenz einer Schätzfunktion

$$\hat{\theta} \text{ ist konsistent, wenn } \mathrm{E}(\hat{\theta}_n-\theta)^2 \to 0 \text{ für } n \to \infty$$

Daraus folgt, dass eine Schätzfunktion $\hat{\theta}$ nur konsistent ist, wenn für $n \to \infty$ der Bias *und* die Varianz gegen Null konvergieren.

Literaturverzeichnis

Eicker, F. (1963), 'Asymptotic normality and consistency of the least squares estimators for families of linear regressions', *The Annals of Mathematical Statistics* **34**(2), 447–456.

Huber, P. J. (1967), The behavior of maximum likelihood estimates under nonstandard conditions, *in* 'Proceedings of the fifth Berkeley symposium on mathematical statistics and probability', Vol. 1, University of California Press, pp. 221–233.

White, H. (1980), 'A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity', *Econometrica* **48**(4), 817–838.

4.A Appendix

4.A.1 Eine Schätzfunktion für die Varianz der Störterme σ^2

Da σ^2 in dem nach der OLS Methode zu minimierenden Ausdruck $\min \sum_i (y_i - \widehat{\beta}_1 - \widehat{\beta}_2 x_i)^2$ nicht vorkommt müssen wir im folgenden einen indirekten und teilweise etwas mühsamen Weg gehen, um eine Schätzfunktion für σ^2 zu erhalten.[11]

Wir erinnern uns, das wahre Modell der Grundgesamtheit ist

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

und für die Mittelwerte gilt[12]

$$\bar{y} = \beta_1 + \beta_2 \bar{x} + \bar{\varepsilon}$$

Das Modell mit mittelwerttransformierten Daten (d.h. in Abweichungsform) ist also

$$y_i - \bar{y} = \beta_2 (x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon})$$

Man beachte, dass das Interzept β_1 bei der Differenzenbildung wegfällt.

Wir sind an einer Schätzfunktion für die Varianz der unbeobachtbaren Störterme der Grundgesamtheit ε_i interessiert. Da wir diese nicht kennen ist es naheliegend, dazu von den beobachtbaren Stichprobenresiduen $\hat{\varepsilon}$ auszugehen. Deshalb versuchen wir einen Zusammenhang zwischen den Störtermen der Grundgesamtheit ε und den Stichprobenresiduen $\hat{\varepsilon}$ herzustellen (bzw. zwischen deren Varianzen).

Um die Schreibweise etwas zu Vereinfachen kennzeichnen wir im Folgenden mittelwerttransformierte Daten mit zwei Punkten über der Variable, d.h.

$$\ddot{x}_i := (x_i - \bar{x})$$

Wir beginnen damit, den wahren Zusammenhang der Grundgesamtheit $\ddot{y}_i = \beta_2 \ddot{x}_i + (\varepsilon_i - \bar{\varepsilon})$ in den Stichproben-Zusammenhang $\hat{\varepsilon}_i = \ddot{y}_i - \widehat{\beta}_2 \ddot{x}_i$ einzusetzen und erhalten

$$\hat{\varepsilon}_i = \beta_2 \ddot{x}_i + (\varepsilon_i - \bar{\varepsilon}) - \widehat{\beta}_2 \ddot{x}_i = (\beta_2 - \widehat{\beta}_2)\ddot{x}_i + (\varepsilon_i - \bar{\varepsilon})$$

Wir sind letztendlich an einer Varianz interessiert, deshalb quadrieren wir diesen Ausdruck

$$\hat{\varepsilon}_i^2 = (\widehat{\beta}_2 - \beta_2)^2 \ddot{x}_i^2 + (\varepsilon_i - \bar{\varepsilon})^2 - 2(\widehat{\beta}_2 - \beta_2)\ddot{x}_i(\varepsilon_i - \bar{\varepsilon})$$

und summieren über alle n Beobachtungen auf (beachte, dass $\sum_{i=1}^n \ddot{x}_i = 0$)

$$\sum \hat{\varepsilon}_i^2 = (\widehat{\beta}_2 - \beta_2)^2 \sum \ddot{x}_i^2 + \sum (\varepsilon_i - \bar{\varepsilon})^2 - 2(\widehat{\beta}_2 - \beta_2) \sum \ddot{x}_i \varepsilon_i$$

und nehmen von beiden Seiten den Erwartungswert

$$\mathrm{E}\left[\sum \hat{\varepsilon}_i^2\right] = \underbrace{\mathrm{E}(\widehat{\beta}_2 - \beta_2)^2 \sum \ddot{x}_i^2}_{A} + \underbrace{\mathrm{E}\left[\sum (\varepsilon_i - \bar{\varepsilon})^2\right]}_{B} - \underbrace{2\,\mathrm{E}\left[(\widehat{\beta}_2 - \beta_2) \sum \ddot{x}_i \varepsilon_i\right]}_{C}$$

[11] Die folgenden Ausführungen halten sich eng an Gujarati 1995.

[12] $\sum_i y_i = n\beta_1 + \beta_2 \sum_i x_i + \sum_i \varepsilon_i$. Dividieren durch n gibt $\bar{y} = \beta_1 + \beta_2 \bar{x} + \bar{\varepsilon}$.

Die folgende Rechnerei ist etwas umständlich, sie werden später sehen, dass sich dies in Matrixschreibweise deutlich einfacher darstellen lässt.

Nun aber ans Werk! Wir haben bereits gezeigt dass

$$\text{var}(\widehat{\beta}_2) = \text{E}(\widehat{\beta}_2 - \beta_2)^2 = \frac{\sigma^2}{\sum \ddot{x}_i^2} := \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}$$

Daraus folgt, dass der erste Term $A = \sigma^2$.

Der zweite Term $B = \text{E}\left[\sum_i (\varepsilon_i - \bar{\varepsilon})^2\right] = (n-1)\sigma^2$, wenn die $\varepsilon_i \sim \text{i.i.d.}(0, \sigma^2)$.

Beweis: Zuerst ist zu beachten, dass der Störterm ε_i eine von n verschiedenen Zufallsvariablen ist, da $i = 1 \ldots, n$. Wir haben angenommen, dass $\text{E}(\varepsilon_i) = 0$, d.h. wenn wir *über alle möglichen* – mit den Wahrscheinlichkeiten gewichteten – *Ausprägungen* der einen Zufallsvariable ε_i aufsummieren ist diese gewichtete Summe Null.

Daraus folgt aber nicht, dass $\sum_{i=1}^n \varepsilon_i = 0$, denn hier summieren wir über n *verschiedene* Zufallsvariablen auf! Die Bedingungen erster Ordnung für OLS Schätzfunktionen garantieren zwar, dass für die *Residuen* gilt $\sum_{i=1}^n \hat{\varepsilon}_i = 0$ (sofern die Regression ein Interzept enthält), dies muss aber nicht für die Störterme ε_i gelten!

Weiters erinnern wir uns, dass $\text{var}(\varepsilon_i) := \text{E}[\varepsilon_i - \text{E}(\varepsilon_i)]^2 = \text{E}(\varepsilon_i)^2 = \sigma^2$ wenn die insgesamt n Störterme alle homoskedastisch und nicht autokorreliert sind.

$$\begin{aligned}
\text{E}\left[\sum_i (\varepsilon_i - \bar{\varepsilon})^2\right] &= \text{E}\left[\sum_i (\varepsilon_i^2 - 2\varepsilon_i\bar{\varepsilon} + \bar{\varepsilon}^2)\right] \\
&= \text{E}\sum_i \left[\varepsilon_i^2 - 2\varepsilon_i\left(\frac{1}{n}\sum_j \varepsilon_j\right) + \left(\frac{1}{n}\sum_j \varepsilon_j\right)^2\right] \\
&= \sum_i \text{E}(\varepsilon_i)^2 - 2\,\text{E}\sum_i \left[\frac{1}{n}\left(\varepsilon_i \sum_j \varepsilon_j\right)\right] + \sum_i \left[\text{E}\left(\frac{1}{n}\sum_j \varepsilon_j\right)^2\right] \\
&= n\sigma^2 - \sum_i \frac{2}{n}\,\text{E}(\varepsilon_i)^2 + \sum_i \left(\frac{1}{n^2}\sum_j \text{E}(\varepsilon_j)^2\right) \qquad \begin{array}{l}\text{wenn } \text{E}(\varepsilon_i^2) = \sigma^2 \text{ und} \\ \text{E}(\varepsilon_i\varepsilon_j) = 0 \text{ für } i \neq j\end{array} \\
&= n\sigma^2 - \frac{2}{n}\sum_i \sigma^2 + \sum_i \left(\frac{1}{n^2}\sum_j \sigma^2\right) \\
&= n\sigma^2 - 2\sigma^2 + \sigma^2 \\
&= (n-1)\sigma^2
\end{aligned}$$

mit $i, j = 1, \ldots, n$. Dabei haben wir wiederholt von den Annahmen $\text{E}(\varepsilon_i)^2 = \sigma^2$ und $\text{E}(\varepsilon_i\varepsilon_j) = 0$ für $i \neq j$ (d.h. Homoskedastizität und Unabhängigkeit) Gebrauch gemacht. Das impliziert, dass das folgende Ergebnis nur bei Homoskedastizität und Unabhängigkeit gilt!

Hinweis: Um z.B. zu sehen, dass

$$2\,\text{E}\sum_i \left[\frac{1}{n}\left(\varepsilon_i \sum_j \varepsilon_j\right)\right] = \sum_i \frac{2}{n}\,\text{E}(\varepsilon_i)^2 = 2\sigma^2$$

empfiehlt es sich die innere Summe auszuschreiben

$$\begin{aligned} 2\,\mathrm{E}\sum_i \left[\frac{1}{n}\,\varepsilon_i\,(\varepsilon_1+\varepsilon_2+\cdots+\varepsilon_i+\cdots+\varepsilon_n)\right] &= \\ \frac{2}{n}\sum_i \left[\mathrm{E}(\varepsilon_i\varepsilon_1)+\mathrm{E}(\varepsilon_i\varepsilon_2)+\cdots+\mathrm{E}(\varepsilon_i^2)+\cdots+\mathrm{E}(\varepsilon_i\varepsilon_n)\right] &= \frac{2}{n}\sum_i \mathrm{E}(\varepsilon_i^2) = 2\sigma^2 \end{aligned}$$

wenn (und nur wenn!) $\mathrm{E}(\varepsilon_i^2)=\sigma^2$ und $\mathrm{E}(\varepsilon_i\varepsilon_j)=0$ für $i\neq j$, d.h. wenn die Störterme homoskedastisch und nicht autokorreliert sind. Falls diese Bedingungen verletzt sind gelten die letzten beiden Gleichheitszeichen nicht! Diese Fälle werden wir später in den Kapiteln über Heteroskedastizität und Autokorrelation ausführlicher diskutieren.

Übungsaufgabe: Zeigen Sie, dass $\mathrm{E}(\bar{\varepsilon}^2)=\sigma^2/n$. Welche Annahmen sind dazu erforderlich? □

Für den dritten Term $C=2\,\mathrm{E}\left[(\widehat{\beta}_2-\beta_2)\sum \ddot{x}_i\varepsilon_i\right]$ berücksichtigen wir, dass

$$\widehat{\beta}_2=\frac{\sum_i \ddot{x}_i\ddot{y}_i}{\sum_i \ddot{x}_i^2}=\frac{\sum_i \ddot{x}_i(\beta_2\ddot{x}_i+\varepsilon_i)}{\sum_i \ddot{x}_i^2}=\beta_2+\frac{\sum_i \ddot{x}_i\varepsilon_i}{\sum_i \ddot{x}_i^2}$$

weshalb $\sum_i \ddot{x}_i\varepsilon_i=(\widehat{\beta}_2-\beta_2)\sum_i \ddot{x}_i^2$. Einsetzen in $C=2\,\mathrm{E}\left[(\widehat{\beta}_2-\beta_2)\sum \ddot{x}_i\varepsilon_i\right]$ unter Berücksichtigung von $\mathrm{var}(\widehat{\beta}_2)=\mathrm{E}[\widehat{\beta}_2-\mathrm{E}(\widehat{\beta}_2)]^2=\sigma^2/\sum_i \ddot{x}_i^2$ gibt

$$C=2\,\mathrm{E}\left[(\widehat{\beta}_2-\beta_2)^2\sum_i \ddot{x}_i^2\right]=\frac{2\sigma^2\sum_i \ddot{x}_i^2}{\sum_i \ddot{x}_i^2}=2\sigma^2$$

Wir fassen nun die Terme $A=\sigma^2$, $B=(n-1)\sigma^2$ und $C=2\sigma^2$ zusammen

$$\mathrm{E}\left[\sum \hat{\varepsilon}_i^2\right]=\sigma^2+(n-1)\sigma^2-2\sigma^2=(n-2)\sigma^2$$

Daraus können wir wieder eine erwartungstreue Schätzfunktion für die Varianz der Grundgesamtheit σ^2 bestimmen, denn aus der letzten Gleichung folgt

$$\frac{\mathrm{E}(\sum \hat{\varepsilon}_i^2)}{n-2}=\sigma^2 \tag{4.8}$$

Wir definieren nun ein

$$\hat{\sigma}^2=\frac{\sum \hat{\varepsilon}_i^2}{n-2}$$

denn aufgrund Gleichung (4.8) gilt $\mathrm{E}(\hat{\sigma}^2)=\sigma^2$, also ist $\hat{\sigma}^2$ eine erwartungstreue Schätzfunktion für σ^2!

4.A.2 Gauss-Markov Beweis

Um die Effizienz der OLS-Schätzfunktion $\widehat{\beta}_2$ zu beweisen minimieren wir die Varianz einer beliebigen *linearen* Schätzfunktion $\tilde{\beta}_2$ (sprich β_2 Schlange)

$$\tilde{\beta}_2 = \sum_{i=1}^{n} c_i y_i$$

wobei die c_i (beliebige) deterministische Gewichte sind, die natürlich eine Funktion der x_i sein können.

Wir interessieren uns ausschließlich für erwartungstreue Schätzfunktionen, deshalb müssen wir zuerst die notwendigen Bedingungen ermitteln, unter denen die lineare Schätzfunktion $\tilde{\beta}_2 = \sum_{i=1}^{n} c_i y_i$ erwartungstreu ist.

Erwartungstreue bedeutet

$$\mathrm{E}(\tilde{\beta}_2) = \beta_2$$

Einsetzen der obigen Schätzfunktion gibt:

$$\begin{aligned}
\mathrm{E}(\tilde{\beta}_2) &= \mathrm{E}(\sum c_i y_i) \\
&= \sum c_i \,\mathrm{E}(y_i) \qquad (\text{da } c_i \text{ deterministisch}) \\
&= \sum c_i(\beta_1 + \beta_2 x_i) \qquad [\text{da } y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \text{ und } \mathrm{E}(\varepsilon_i) = 0] \\
&= \beta_1 \sum c_i + \beta_2 \sum c_i x_i \\
&= \beta_2 \qquad \text{wenn } \sum c_i = 0 \quad \text{und } \sum c_i x_i = 1
\end{aligned}$$

Das heißt, damit $\tilde{\beta}_2 = \sum c_i y_i$ eine unverzerrte Schätzfunktion für β_2 ist müssen die Bedingungen $\sum c_i = 0$ und $\sum c_i x_i = 1$ erfüllt sein.[13]

Nun minimieren wir die Varianz von $\tilde{\beta}_2$ unter diesen beiden Nebenbedingungen für Unverzerrtheit.

Die Varianz von $\tilde{\beta}_2$ ist

$$\begin{aligned}
\mathrm{var}(\tilde{\beta}_2) &= \mathrm{var}\left(\sum c_i y_i\right) \\
&= \sum c_i^2 \,\mathrm{var}\,(y_i) \qquad (\text{weil die } y_i \text{ statistisch unabhängig sind}) \\
&= \sum c_i^2 \,\sigma^2 = \sigma^2 \sum c_i^2
\end{aligned}$$

da unter den Annahmen deterministischer x und $\mathrm{E}(\varepsilon_i) = 0$ gilt $\mathrm{var}(y_i) = \mathrm{var}(\varepsilon_i) = \sigma^2$, weil $\mathrm{var}(y_i) := \mathrm{E}[\beta_1 + \beta_2 x_i + \varepsilon_i - \mathrm{E}(\beta_1 + \beta_2 x_i + \varepsilon_i)]^2 = \mathrm{E}[\varepsilon_i - \mathrm{E}(\varepsilon_i)]^2) = \mathrm{E}(\varepsilon_i)^2 = \sigma^2$.

Man beachte, dass wir dabei auch von den Gauss-Markov Annahmen über den Störterm $\varepsilon_i \sim$ i.i.d.$(0, \sigma^2)$ (d.h. unter anderem, keine Autokorrelation und keine Heteroskedastizität) Gebrauch gemacht haben.

Wir suchen nun die Gewichte $c_1, c_2, \ldots, c_n$, die die Varianz von $\tilde{\beta}_2$ unter den Nebenbedingungen $\sum c_i = 0$ und $\sum c_i x_i = 1$ (Erwartungstreue) minimieren. Dies ist eine einfache Minimierungsaufgabe unter Nebenbedingungen und kann z.B. mit der

[13] Man beachte, dass die Gewichte $w_i = \ddot{x}_i / \sum_j \ddot{x}_j^2$ auf Seite 15 diese Bedingungen erfüllten.

Lagrange Methode einfach gelöst werden. Da wir zwei Nebenbedingungen haben benötigen wir zwei Lagrangemultiplikatoren λ_1 und λ_2.

Die Lagrangefunktion ist

$$\mathcal{L}(c_1, \ldots, c_n, \lambda_1, \lambda_2) = \sigma^2 \sum c_i^2 - \lambda_1 \left(\sum c_i \right) - \lambda_2 \left(\sum c_i x_i - 1 \right)$$

und die Bedingungen erster Ordnung für ein Optimum sind

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial c_1} &= 2c_1\sigma^2 - \lambda_1 - \lambda_2 x_1 = 0 \\
\frac{\partial \mathcal{L}}{\partial c_2} &= 2c_2\sigma^2 - \lambda_1 - \lambda_2 x_2 = 0 \\
&\vdots \\
\frac{\partial \mathcal{L}}{\partial c_n} &= 2c_n\sigma^2 - \lambda_1 - \lambda_2 x_n = 0 \\
\frac{\partial \mathcal{L}}{\partial \lambda_1} &= \sum c_i = 0 \\
\frac{\partial \mathcal{L}}{\partial \lambda_2} &= \sum c_i x_i - 1 = 0
\end{aligned}$$

Aus diesen $n+2$ Gleichungen können die Unbekannten $c_1, \ldots, c_n, \lambda_1$ und λ_2 berechnet werden.

Die ersten n Gleichungen können geschrieben werden als

$$\begin{aligned}
c_1 &= \frac{1}{2\sigma^2}(\lambda_1 + \lambda_2 x_1) \\
c_2 &= \frac{1}{2\sigma^2}(\lambda_1 + \lambda_2 x_2) \\
&\vdots \\
c_n &= \frac{1}{2\sigma^2}(\lambda_1 + \lambda_2 x_n)
\end{aligned}$$

Aufsummieren dieser Gleichungen gibt

$$\sum_i c_i = \frac{1}{2\sigma^2}(n\lambda_1 + \lambda_2 \sum_i x_i) = 0$$

da $\sum_i c_i = 0$ eine Bedingung erster Ordnung ist.

Als nächstes können wir die erste Gleichung des obigen Gleichungssystems mit x_1, die zweite mit x_2 usw. multiplizieren

$$\begin{aligned}
c_1 x_1 &= \frac{1}{2\sigma^2}(\lambda_1 x_1 + \lambda_2 x_1^2) \\
c_2 x_2 &= \frac{1}{2\sigma^2}(\lambda_1 x_2 + \lambda_2 x_2^2) \\
&\vdots \\
c_n x_n &= \frac{1}{2\sigma^2}(\lambda_1 x_n + \lambda_2 x_n^2)
\end{aligned}$$

Aufsummieren gibt

$$\sum_i c_i x_i = \frac{1}{2\sigma^2}\left(\lambda_1 \sum_i x_i + \lambda_2 \sum_i (x_i^2)\right) = 1$$

wobei $\sum_i c_i x_i = 1$ wieder eine Bedingung erster Ordnung ist.
Diese beiden Gleichungen können nach λ_1 und λ_2 gelöst werden (nicht so schüchtern, versuchen Sie's ruhig mal!)

$$\begin{aligned}\lambda_1 &= \frac{-2\sigma^2 \sum x_i}{n(\sum x_i^2) - (\sum x_i)^2} \\ \lambda_2 &= \frac{2n\sigma^2}{n(\sum x_i^2) - (\sum x_i)^2}\end{aligned}$$

Diese Gleichungen können schließlich in

$$c_i = \frac{1}{2\sigma^2}(\lambda_1 + \lambda_2 x_i)$$

eingesetzt werden und geben die Lösung

$$c_i = \frac{n x_i - \sum_j x_j}{n(\sum_j x_j^2) - (\sum_j x_j)^2} = \frac{(x_i - \bar{x})}{\sum_j (x_j - \bar{x})^2}$$

Deshalb ist

$$\tilde{\beta}_2 = \sum_{i=1}^{n} c_i y_i = \frac{\sum_i (x_i - \bar{x})\, y_i}{\sum_i (x_i - \bar{x})^2}$$

eine effiziente (d.h. erwartungstreue und varianzminimale) Schätzfunktion. Aber dies ist genau die Gleichung der OLS-Schätzfunktion. Damit haben wir gezeigt, dass OLS-Schätzfunktionen tatsächlich die minimale Varianz unter allen linearen erwartungstreuen Schätzfunktionen haben, *wenn die Gauss-Markov Annahmen erfüllt sind.* qed

Dieser Ansatz liefert auch eine alternative Möglichkeit die Varianz von $\widehat{\beta}_2$ zu berechnen, denn wir haben vorhin gezeigt, dass $\text{var}(\tilde{\beta}_2) = \sigma^2 \sum c_i^2$.
Wir multiplizieren

$$c_i = \frac{n x_i - \sum_j x_j}{n(\sum_j x_j^2) - (\sum_j x_j)^2}$$

mit c_i und Summieren über alle i (für $i, j = 1, \ldots, n$)

$$\sum c_i^2 = \frac{n \sum_i (c_i x_i) - \sum_i c_i \sum_j x_j}{n(\sum_j x_j^2) - (\sum_j x_j)^2}$$

Da

$$\sum c_i = 0 \qquad \text{und} \qquad \sum c_i x_i = 1$$

folgt

$$\sum c_i^2 = \frac{n}{n(\sum x_i^2) - (\sum x_i)^2}$$

also

$$\text{var}(\tilde{\beta}_2) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}$$

Dies ist wiederum exakt die Varianz der OLS-Schätzfunktion.

Ähnlich kann ein BLU[14] Schätzer für $\tilde{\beta}_1$ und dessen Varianz berechnet werden:

$$\begin{aligned} \tilde{\beta}_1 &= \bar{y} - \tilde{\beta}_2 \bar{x} \\ \text{var}(\tilde{\beta}_1) &= \frac{\sigma^2 (\sum x_i^2)}{n \sum \ddot{x}_i^2} \end{aligned}$$

Eine allgemeinere untere Abschätzung der Varianzen einer erwartungstreuen Schätzfunktion erlaubt die **Rao-Cramer'sche Ungleichung** (siehe z.B. Kmenta 1990, S. 160f, Frohn 1995).

Übungsaufgabe: Zeigen Sie, dass $\sum(x_i^2) - \frac{1}{n}(\sum x_i)^2 = \sum(x_i - \bar{x})^2$.
Hinweis: es ist einfacher zu zeigen, dass $\sum(x_i - \bar{x})^2$ gleich $\sum(x_i^2) - \frac{1}{n}(\sum x_i)^2$ ist.

4.A.3 R Code

Grafik 4.9 (Seite 34):

```
#########################################################################
# LLN - Law of Large Numbers
#########################################################################
n <- 300
i <- 1:n
set.seed(123456)
means <- cumsum(sample(x = 1:6, size = n, replace = TRUE))/i

x11()
plot(means, type = "l", ylim = c(1,6),
     main = "Mittelwert der Augenzahlbeim Würfeln \nmit zunehmender Stichprobengröße"
     ylab = "Mittelwert", xlab = "Stichprobengröße n",col="blue", lwd=2)
for (j in 1:7) { # weitere Folgen
  lines(cumsum(sample(x = 1:6, size = n, replace = TRUE))/i, col="gray")
}
abline(h=3.5, col = "red", lty=2, lwd=2)
```

Grafik 4.11 (Seite 41):

[14]BLUE bedeutet *Best Linear Unbiased Estimator*, man spricht also von von einem BLU Schätzer.

```
#########################################################################
# CLT - Central Limit Theorem
#########################################################################
rm(list=ls())
set.seed(123456)

draws <- 10000  # number of samples drawn
mw <- rep(NA, draws) # to be filled with means

x11() # dev.new()
par(mfrow = c(2,2)) # 4 graphs: 2 rows, 2 columns

# n: Sample Size
for (n in c(1, 2, 5, 20)){
  for (i in 1:draws) {
    mw[i] <- mean(sample(x = 1:6, size = n, replace = TRUE))
  }
  hist(mw, col="gray", freq = FALSE, breaks = 30,
       main = paste("Sampling Distribution\nof mean; n = ", n),
       xlab = "Sample means")
  abline(v = 3.5, lwd=2, col="red")
  if (n > 4) {
    curve(dnorm(x, mean=mean(mw), sd=sd(mw)),
          col="blue", lwd=2, lty=2, add=TRUE)
  }
}
#dev.off()
```

Grafik 4.12 (Seite 42):

```
#########################################################################
# CLT - Central Limit Theorem  (StdL, EU-Silc 2015)
#########################################################################
rm(list=ls(all=TRUE))
d <- read.csv("http://www.hsto.info/statistik/dl/stdl2015.csv")
attach(d)

draws <- 10000  # number of samples drawn (rows)
set.seed(123456)
x11() # dev.new()
par(mfrow = c(2,2))

# Population
hist(StdL, col="gray", freq = FALSE, breaks = 60,
     main = "(1) Population:  \n Relative frequencies of StdL",
     xlab = "\'StdL\'")
abline(v = mean(StdL), lwd=2, col="red")
```

```
# n: Sample Size 10, 30, 500
g <- 1
for (n in c(10, 30, 500)){
  my.samples <- matrix(NA, nrow = draws, ncol = n) # for samples
  for (i in 1:draws) {
    # each row is a different sample of size n
    my.samples[i, ] <- sample(x = StdL, size = n, replace = TRUE)
  }
  sample.means <- apply(my.samples, MARGIN = 1, mean) # calculate means
  g <- g+1
  hist(sample.means, col="gray", freq = FALSE, breaks = 30,
       main = paste0("(", g, ") Sampling distribution\nof mean for n = ", n),
       xlab = "Sample means for \'StdL\'")
  abline(v = mean(StdL), lwd=2, col="red")
  curve(dnorm(x, mean=mean(StdL), sd=sd(StdL)/sqrt(n)),
        col="blue", lwd=2, lty=2, add=TRUE)
}
```