Kapitel 4

Eigenschaften von OLS-Schätzfunktionen

"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)

Wir haben bereits im zweiten Kapitel die OLS-Methode kennengelernt und darüber hinaus im letzten Kapitel festgestellt, dass die Anwendung dieser Methode auf eine konkrete Stichprobe *Schätzungen* – also fixe Zahlenwerte (Realisationen) b_1 und b_2 für die Zufallsvariablen $\widehat{\beta}_1$ und $\widehat{\beta}_2$ der SRF $y_i = \widehat{\beta}_1 + \widehat{\beta}_2 x_i + \hat{\varepsilon}_i$ liefert.

Im Abschnitt über die Monte Carlo Simulationen haben wir gesehen, dass die Idee der wiederholten Stichprobenziehungen (*'repeated sampling'*) ganz natürlich zur Idee der Stichprobenkennwertverteilungen führt. Wir interessieren uns im Folgenden vor allem für die Momente[1] solcher Stichprobenkennwertverteilungen.

Die Monte Carlo Simulationen haben uns bereits gezeigt, dass aufgrund des *Gesetzes der Großen Zahl* der Mittelwert einer Stichprobenkennwertverteilung oft 'ziemlich' genau dem wahren Wert der Grundgesamtheit entspricht, und dass aufgrund des *Zentralen Grenzwertsatzes* bei einer genügend großen Anzahl von Ziehungen die Stichprobenkennwertverteilung oft einer Normalverteilung 'ziemlich' ähnlich sieht. Allerdings lieferten uns die Monte Carlo Simulationen nur eine intuitive Vorstellung, keine *'hard facts'* mit denen man rechnen kann.

In diesem Kapitel wollen wir diese Ideen etwas weiter entwickeln und den Erwartungswert und die Varianz der Stichprobenkennwertverteilungen der geschätzten Koeffizienten $\widehat{\beta}_1$ und $\widehat{\beta}_2$ allgemein berechnen. Diese werden uns im nächsten Kapitel schließlich die Durchführung von Hypothesentests erlauben. Allerdings werden wir dies unter der relativ restriktiven Annahme einer deterministischen x Variable und von identisch und unabhängig verteilten (i.i.d.) Störtermen tun. Der Vorteil liegt

[1] Momente sind Kenngrößen einer Zufallsvariablen, bzw. einer Verteilungsfunktion. Das k-te zentrale Moment ist definiert als

$$\mu_k = \mathrm{E}[x - \mathrm{E}(x)]^k$$

Das zentrale Moment erster Ordnung (für $k = 1$) ist stets gleich Null ($\mu_1 = 0$), da $\mu_1 = \mathrm{E}(x-\mu)^1 = \mu - \mu = 0$; das zentrale Moment zweiter Ordnung (für $k = 2$) ist die Varianz ($\mu_2 = \mathrm{E}[x-\mathrm{E}(x)]^2$), das zentrale Moment dritter Ordnung ist die Schiefe, das zentrale Moment vierter Ordnung entspricht der Wölbung bzw. Kurtosis.

darin, dass uns unter diesen Annahmen relativ einfach einen Schätzer für die Standardfehler der Koeffizienten berechnen können. Diese Annahmen werden in späteren Kapiteln gelockert.

Vorher werden wir uns aber noch mit einigen statistischen Eigenschaften des OLS-Schätzers beschäftigen. Wir haben bereits mehrmals erwähnt, dass OLS-Schätzer '*bestmögliche*' Schätzer sind, ohne allerdings genauer zu spezifizieren, was wir darunter verstehen. Dies werden wir in diesem Kapitel nachholen.

Das Konzept der Stichprobenkennwertverteilungen erlaubt es uns nämlich, die Eigenschaften von Schätzfunktionen etwas präziser zu definieren. Konkret wünschen wir uns Schätzfunktionen, die *'im Durchschnitt richtige'* und *'möglichst genaue'* Schätzungen liefern.

Mit *'im Durchschnitt richtig'* meinen wir, dass der *Erwartungswert* der Stichprobenkennwertverteilung gleich dem wahren Wert der Grundgesamtheit sein sollte. In der Sprache der Ökonometrikerinnen wird diese Eigenschaft einer Schätzfunktion *Erwartungstreue* genannt. Mit *'möglichst genau'* meinen wir, dass die Stichprobenkennwertverteilung eine möglichst kleine Varianz haben sollte, oder etwas genauer, dass die Varianz der Stichprobenkennwertverteilung der OLS Schätzer kleiner sein sollte als die Varianz aller *vergleichbaren* alternativen Schätzfunktionen. Eine Schätzfunktion, die diese zweite Eigenschaft erfüllt, wird in der Sprache der Ökonometrikerinnen *'effizient'* genannt.

In diesem Kapitel werden wir zuerst zeigen, dass die OLS-Schätzer unter bestimmten Annahmen tatsächlich *erwartungstreu* und *effizient* sind. Dies ist das Ergebnis des bekannten *Gauss-Markov Theorems*, das in der Ökonometrie eine zentrale Rolle spielt. Tatsächlich wird sich ein großer Teil dieser Veranstaltung mit der Frage beschäftigen, was zu tun ist, wenn eine oder mehrere der *Gauss-Markov Annahmen* verletzt sind. Da das Gauss-Markov Theorem in der Ökonometrie eine derart grundlegende Rolle spielt, werden wir es etwas ausführlicher beweisen.

Die Erwartungstreue und Effizienz der OLS-Schätzer sind sogenannte *'Kleine Stichprobeneigenschaften'*, d.h. sie gelten *auch* in kleinen Stichproben (oder genauer, unabhängig von der Stichprobengröße).

Leider lassen sich diese 'Kleine Stichprobeneigenschaften' in komplizierteren Fällen nicht immer beweisen (z.B. wenn einige der *Gauss-Markov Annahmen* nicht erfüllt sind). Deshalb werden wir im letzten Abschnitt einige '*asymptotische Eigenschaften*' diskutieren. Die wichtigste dieser asymptotischen Eigenschaften ist die *Konsistenz*. Etwas vereinfachend gesprochen ist eine Schätzfunktion *konsistent*, wenn sie mit *zunehmender Stichprobengröße* 'immer genauer' wird.

Schließlich werden wir noch kurz den 'mittleren quadratischen Fehler' (*mean square error*) diskutieren.

Nach dieser etwas ausführlichen Vorschau können wir uns nun an die Arbeit machen. Für alle, denen dieses Kapitel etwas schwierig erscheint, ein kleiner Trost vorab: dieses Kapitel wird in einem späteren Kapitel Schritt für Schritt wiederholt – allerdings unter Verwendung der Matrixschreibweise.

4.1 Kleine Stichprobeneigenschaften

Kleine Stichprobeneigenschaften sind – wie bereits erwähnt – unabhängig von der Stichprobengröße gültig, das heißt, sie gelten *auch* in kleinen Stichproben. Die beiden wichtigsten 'kleine Stichprobeneigenschaften' sind:

1. **Erwartungstreue** (Unverzerrtheit): Eine Schätzfunktion $\widehat{\beta}$ für den wahren Wert β der Grundgesamtheit ist **erwartungstreu** ('*unbiased*'), wenn

$$\boxed{\mathrm{E}(\widehat{\beta}) = \beta}$$

   und zwar für jeden beliebigen Stichprobenumfang n.

   Bei nicht erwartungstreuen Schätzern wird $\mathrm{E}(\widehat{\beta}) - \beta$ Verzerrung (*bias*) genannt

$$\text{Bias} = \mathrm{E}(\widehat{\beta}) - \beta$$

   Erinnern wir uns, der Erwartungswert ist einfach ein mit den Wahrscheinlichkeiten gewichtetes Mittel über alle möglichen Ausprägungen einer Zufallsvariable. Erwartungstreue sagt also nichts über das Ergebnis einer einzelnen Schätzung aus, sondern ist eine Eigenschaft einer Schätzfunktion.

2. **Effizienz:** Eine Schätzfunktion heißt **effizient**, wenn sie

   1. erwartungstreu ist, und
   2. varianzminimal unter allen vergleichbaren erwartungstreuen Schätzfunktionen ist:

$$\boxed{\mathrm{var}(\widehat{\beta}) \leq \mathrm{var}(\widehat{\beta}^*)}$$

      wobei $\widehat{\beta}^*$ jede beliebige lineare und erwartungstreue Schätzfunktion für β sein kann. Effizienz bezieht sich immer auf einen Vergleich der Varianz von Schätzfunktionen, ist also ein relatives Konzept. Deshalb muss stets angeben werden, innerhalb welcher Klasse von Schätzfunktionen ein Schätzer effizient ist. In diesem Kapitel werden wir zeigen, dass der OLS-Schätzer unter einer Reihe von Annahmen innerhalb der Klasse aller unverzerrten linearen Schätzfunktionen effizient ist.

      Diese Eigenschaft wird *Effizienz* genannt, weil effiziente Schätzer die verfügbare Information in der Stichprobe effizient nützen, und deshalb genauer sind als alternative Schätzfunktionen.

4.1.1 Einführung und Wiederholung

Zur Erläuterung starten wir mit einem bekannten Fall aus der einführenden Statistik, einer *univariaten* Zufallsvariable y. Dabei wird üblicherweise angenommen, dass alle Beobachtungen aus der gleichen Verteilung gezogen wurden (also identisch verteilt sind), und dass die einzelnen y_i untereinander statistisch unabhängig sind. Dies wird üblicherweise mit i.i.d. abgekürzt für *'independent and identically-distributed'*. Zudem nehmen wir an, dass der Erwartungswert von y in der Grundgesamtheit

μ sei (d.h. $\mathrm{E}(y) = \mu$), und dass die Varianz der Grundgesamtheit eine endliche Zahl σ^2 sei (d.h. $\mathrm{var}(y) = \sigma^2$). Durch die Verwendung der griechischen Symbole μ und σ^2 wird zum Ausdruck gebracht, dass diese Parameter sind, also unbekannte Zahlen der Grundgesamtheit. Man beachte, dass die empirische Varianz, die auf Grundlage von Realisationen berechnet wird, immer eine endliche Zahl ist, dies muss für die Varianz der Zufallsvariable – die für ein beliebig oft wiederholbares Zufallsexperiment definiert ist – nicht gelten.

Dieser datengenerierende Prozess (DGP) wird kompakt angeschrieben als

$$y_i \sim \text{i.i.d.}(\mu, \sigma^2)$$

Aus der Statistik wissen wir, dass unter diesen Annahmen der *Mittelwert einer Zufallsstichprobe* $\bar{y} = 1/n \sum_i y_i$ ein unverzerrter Schätzer für den Mittelwert der Grundgesamtheit μ ist[2]

$$\mathrm{E}(\bar{y}) = \mu$$

Die Verteilung dieser Stichprobenmittelwerte $\bar{y}$, die man bei wiederholten Stichprobenziehungen erhält, ist eine Stichprobenkennwertverteilung (*sampling distribution*).

In der einführenden Statistik wird gezeigt, dass die *Varianz der Mittelwerte* gleich der Varianz der Grundgesamtheit ($\mathrm{var}(y) := \sigma^2$) dividiert durch die Stichprobengröße n ist[3]

$$\mathrm{var}(\bar{y}) := \sigma^2_{\bar{y}} = \frac{\sigma^2}{n}$$

Da die Varianz der Grundgesamtheit σ^2 üblicherweise ebenso wenig beobachtbar ist wie der Mittelwert μ der Grundgesamtheit, muss die 'wahre' Varianz σ^2 ebenfalls aus der Stichprobe geschätzt werden. Den Schätzer für die Varianz der Grundgesamtheit σ^2 bezeichnen wir mit $\hat{\sigma}^2$.

In der einführenden Statistik wird gezeigt, dass im Fall univariater Verteilungen die Schätzfunktion

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

ein erwartungstreuer Schätzer für die Varianz der Grundgesamtheit σ^2 ist.

Genau das gleiche wollen wir nun für den bivariaten Fall

$$y_i = \widehat{\beta}_1 + \widehat{\beta}_2 x_i + \hat{\varepsilon}_i$$

zeigen, nur untersuchen wir anstelle der Stichprobenkennwertverteilung des Mittelwertes $\bar{y}$ (der als Schätzer für μ verwendet wird) die Stichprobenkennwertverteilungen von $\widehat{\beta}_1$ und $\widehat{\beta}_2$, die als Schätzer für β_1 und β_2 dienen.

---

[2] $\mathrm{E}(\bar{y}) = \mathrm{E}(1/n \sum_i y_i) = 1/n \sum_i \mathrm{E}(y_i) = 1/n(n\mu) = \mu$.

[3] $\mathrm{var}(\bar{y}) = \mathrm{var}(1/n \sum_i y_i) = \mathrm{E}[1/n \sum_i y_i - \mathrm{E}(1/n \sum_i y_i)]^2 = 1/n^2 \,\mathrm{E}[\sum_i (y_i - \mathrm{E}(y_i))]^2 \overset{1)}{=} 1/n^2 \sum_i \mathrm{E}[(y_i - \mathrm{E}(y_i))]^2 \overset{2)}{=} 1/n^2(n\sigma^2) = \sigma^2/n$ wenn $y_i \sim \text{i.i.d.}(\mu, \sigma^2)$, wobei das $\overset{1)}{=}$ Unabhängigkeit (d.h. $\mathrm{cov}(y_i, y_j) = 0$ für $i \neq j$) und $\overset{2)}{=}$ Homoskedastizität (d.h. $\mathrm{var}(y_i) = \sigma^2$) erfordert.

Wir werden uns in diesem Kapitel auf deterministische Regressoren beschränken, die x seien also *'fixed in repeated sampling'*. Dies erleichtert nicht nur die folgenden Herleitungen, sondern auch die Notation ganz beträchtlich.

Die Schlussfolgerungen gelten weitgehend auch für stochastische Regressoren, zumindest wenn die Regressoren nicht mit den Störtermen korreliert sind. Dazu gleich mehr.

4.1.2 Erwartungstreue der geschätzten OLS-Koeffizienten

Wir werden nun zeigen, dass der OLS Schätzer für den Steigungskoeffizienten[4]

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

(mit $i = 1, \ldots, n$) tatsächlich erwartungstreu ist.

Dazu ist es wichtig zu erkennen, dass diese Schätzfunktion für $\widehat{\beta}_2$ linear in den y_i ist, d.h. der Schätzer $\widehat{\beta}_2$ kann auch geschrieben werden als

$$\widehat{\beta}_2 = \sum_{i=1}^{n} w_i y_i \tag{4.1}$$

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.

Dies ist unproblematisch, da die x_i annahmegemäß deterministisch sind (*'fixed in repeated sampling*). Offensichtlich ist $\widehat{\beta}_2$ also eine *lineare Schätzfunktion*; der geschätzte Parameter $\widehat{\beta}_2$ ist eine Linearkombination der stochastischen y_i, wobei die w_i die (deterministischen) Gewichte darstellen, die eine Funktion der x sind.

Die Gewichte w_i haben **drei wichtige Eigenschaften**, die wir gleich benötigen werden:

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. $\sum (x_i - \bar{x}) = n\bar{x} - n\bar{x} = 0$, weil $\bar{x} := \frac{1}{n} \sum_i x_i$!

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

[4] Das dritte Gleichheitszeichen folgt, weil $\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i (x_i - \bar{x}) y_i - \sum_i (x_i - \bar{x}) \bar{y} = \sum_i (x_i - \bar{x}) y_i - \bar{y} \sum_i (x_i - \bar{x}) = \sum_i (x_i - \bar{x}) y_i$ weil $\sum_i (x_i - \bar{x}) = 0$.

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 $\sum_i w_i = 0$, es bleibt also nur zu zeigen, dass $\sum_i w_i x_i = 1$

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

Mit diesen drei Eigenschaften 'bewaffnet' können wir uns nun an den eigentlichen Beweis für die Erwartungstreue machen.

Beweis der Erwartungstreue: Um die Unverzerrtheit (Erwartungstreue) von $\widehat{\beta}_2$ zu zeigen müssen wir einen Zusammenhang zwischen der Schätzfunktion $\widehat{\beta}_2$ und dem entsprechenden Wert β_2 der Grundgesamtheit herstellen, und davon den Erwartungswert bilden.

Dazu wird in der Regel nach dem folgenden Muster vorgegangen: man setzt den wahren Zusammenhang der Grundgesamtheit, d.h. $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, in die Schätzfunktion $\widehat{\beta}_2 = \widehat{\text{cov}}(x, y)/\widehat{\text{var}}(x) = \sum_i w_i y_i$ (siehe (4.1)) ein, und bildet anschließend den Erwartungswert davon.

Einsetzen des 'wahren' Zusammenhangs in die Schätzfunktion gibt

$$\begin{aligned}
\widehat{\beta}_2 &= \sum_i w_i y_i = \sum_i w_i (\beta_1 + \beta_2 x_i + \varepsilon_i) \\
&= \beta_1 \sum_i w_i + \beta_2 \sum_i w_i x_i + \sum_i w_i \varepsilon_i \\
&= \beta_2 + \sum_i w_i \varepsilon_i \qquad (4.2)
\end{aligned}$$

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

Nun bilden wir davon den Erwartungswert

$$\begin{aligned} \mathrm{E}(\widehat{\beta}_2) &= \mathrm{E}\left(\beta_2 + \sum_i w_i \varepsilon_i\right) \\ &= \beta_2 + \sum_i \mathrm{E}(w_i \varepsilon_i) \qquad (\text{weil } \mathrm{E}(\beta_2) = \beta_2) \\ &= \beta_2 + \mathrm{E}\left(\frac{\sum_i (x_i - \overline{x})\varepsilon_i}{\sum_i (x_i - \overline{x})^2}\right) \\ &= \beta_2 + \mathrm{E}\left(\frac{\mathrm{cov}(x, \varepsilon)}{\mathrm{var}(x)}\right) \end{aligned} \tag{4.3}$$

Daraus folgt, dass der Schätzer $\widehat{\beta}_2$ nur dann erwartungstreu ist, wenn die erklärende Variable x und die Störterme unkorreliert sind, bzw. wenn $\mathrm{cov}(x, \varepsilon) = 0$.

Vorsicht, dies bezieht sich auf die Störterme ε des datengenerierenden Prozesses, nicht auf die Residuen $\hat{\varepsilon}$! Aufgrund der Mechanik des OLS Schätzers (d.h. aus den Bedingungen 1. Ordnung) folgt immer $\sum_i x_i \hat{\varepsilon}_i = 0$, d.h. dass die erklärenden Variablen und die *Residuen* unkorreliert sind, dies gilt aber *nur* für die Stichprobe, nicht notwendigerweise für die Störterme ε_i der Grundgesamtheit!

Man beachte, dass der zweite Term von Gleichung (4.3), d.h. $\mathrm{cov}(x, \varepsilon)/\mathrm{var}(x)$, auch als Steigungskoeffizient einer Regression von ε auf x interpretiert werden kann, d.h. wenn $\varepsilon_i = \alpha_1 + \alpha_2 x_i + \nu_i$ ist $\alpha_2 = \mathrm{cov}(x, \varepsilon)/\mathrm{var}(x)$, deshalb können wir auch schreiben $\mathrm{E}(\widehat{\beta}_2) = \beta_2 + \mathrm{E}(\alpha_2)$.

Offensichtlich ist OLS-Schätzer $\widehat{\beta}_2$ nur dann erwartungstreu, d.h. $\mathrm{E}(\widehat{\beta}_2) = \beta_2$, *wenn* die x mit den Störtermen ε der Grundgesamtheit im Erwartungswert unkorreliert sind.

Übung: Zeigen Sie, dass man das selbe Ergebnis erhält, wenn man den wahren Zusammenhang $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ in die OLS Schätzfunktion $\widehat{\beta}_2 = \mathrm{cov}(x, y)/\mathrm{var}(x)$ einsetzt und den Erwartungswert bildet. □

Sollten die Störterme mit den erklärenden Variablen korreliert sein, ist der OLS Schätzer verzerrt! Eine solche Korrelation ist leider nicht so selten, tatsächlich handelt es sich dabei um eines der gravierendsten Probleme der Ökonometrie. Regressoren, die mit den Störtermen ε korreliert sind, werden *endogene Regressoren* genannt, oder man spricht einfach von einem *Endogenitätsproblem*).

Es gibt mehrere Ursachen die zu einer stochastischen Abhängigkeit zwischen Störtermen ε und erklärenden Variablen x führen. Der vermutlich häufigste Fall sind irrtümlich nicht berücksichtigte erklärende Variablen, die sowohl mit y als auch mit einer oder mehreren berücksichtigten x Variablen korreliert sind (*'omitted variables'*). Weitere Beispiele sind Fälle, in denen der datengenerierenden Prozess durch ein simultanes Gleichungssystem beschrieben wird und im System *feed-backs* auftreten, oder auch einfache Messfehler in den x-Variablen.

In solchen Fällen von *Endogenität* sind – wie wir soeben gesehen haben – die OLS-Schätzer systematisch verzerrt! Diese 'tieferen' Probleme werden wir erst in späteren Kapiteln ausführlich diskutieren.

Im Moment wollen wir uns das Leben aber noch einfach machen und deterministische x annehmen.[5] Wenn x_i deterministisch ist, ist natürlich auch w_i deterministisch (*'fixed in repeated sampling'*), also können die w_i vor den Erwartungswertoperator gezogen werden. Für deterministische x reicht die wesentlich weniger strenge Annahme $\mathrm{E}(\varepsilon_i) = 0$, damit der Schätzer unverzerrt ist, denn

$$\begin{aligned} \mathrm{E}(\widehat{\beta}_2) &= \beta_2 + \sum w_i \,\mathrm{E}(\varepsilon_i) \\ &= \beta_2 \qquad \text{wenn } \mathrm{E}(\varepsilon_i) = 0 \end{aligned}$$

Viel einfacher lässt sich zeigen, dass $\widehat{\beta}_1 = \bar{y} - \widehat{\beta}_2 \bar{x}$ ebenfalls ein unverzerrter Schätzer für β_1 ist

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

wenn $\mathrm{E}(\widehat{\beta}_2) = \beta_2$.

Wir fassen zusammen: $\widehat{\beta}_2 = \widehat{\mathrm{cov}}(y, x)/\widehat{\mathrm{var}}(x)$ ist ein erwartungstreuer (unverzerrter) Schätzer für β_2, wenn die Störterme *der Grundgesamtheit* ε_i mit den x_i unkorreliert sind. Bei deterministischen x reicht die wesentlich weniger strenge Annahme $\mathrm{E}(\varepsilon_i) = 0$ für den Beweis der Erwartungstreue von $\widehat{\beta}_2$.

4.1.3 Die Varianz und Kovarianz der OLS Schätzer

Den Erwartungswert der Schätzfunktionen $\widehat{\beta}_1$ und $\widehat{\beta}_2$ haben wir bereits berechnet und dabei festgestellt, dass $\mathrm{cov}(x_i, \varepsilon_i) = 0$ eine notwendige Bedingung für die Erwartungstreue der OLS-Schätzer ist.

Als nächstes wollen wir die Varianzen der Schätzfunktionen $\widehat{\beta}_1$ und $\widehat{\beta}_2$ berechnen. Schätzungen für diese Varianzen werden es uns schließlich erlauben statistische Tests durchzuführen.

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

$$\begin{aligned} \mathrm{var}(\widehat{\beta}_2) &= \mathrm{E}[\widehat{\beta}_2 - \mathrm{E}(\widehat{\beta}_2)]^2 \\ &= \mathrm{E}[\widehat{\beta}_2 - \beta_2]^2 \qquad (\text{wenn } \mathrm{E}(\widehat{\beta}_2) = \beta_2, \text{ siehe oben}) \\ &= \mathrm{E}\left(\sum_i w_i \varepsilon_i\right)^2 \qquad (\text{da } \widehat{\beta}_2 = \beta_2 + \sum w_i \varepsilon_i; \text{ s. Gleichung (4.2)}) \\ &= \mathrm{E}\left(w_1^2\varepsilon_1^2 + w_2^2\varepsilon_2^2 + \cdots + w_n^2\varepsilon_n^2 + \cdots \right. \\ &\quad \left. \cdots + 2w_1 w_2 \varepsilon_1 \varepsilon_2 + \cdots + 2w_{n-1} w_n \varepsilon_{n-1} \varepsilon_n\right) \\ &= \underbrace{\mathrm{E}\left(\sum_{i=1}^{n} w_i^2 \varepsilon_i^2\right)}_{\substack{= \sigma^2 \sum_i w_i^2 \text{ wenn} \\ \text{homoskedastisch}}} + \underbrace{\mathrm{E}\left(\sum_{i=1}^{n} \sum_{\substack{j=2 \\ j>i}}^{n} 2 w_i w_j \varepsilon_i \varepsilon_j\right)}_{\substack{= 0 \text{ wenn keine} \\ \text{Autokorrelation}}} \end{aligned} \tag{4.4}$$

[5] Man beachte, dass die Kovarianz zwischen einer Zufallsvariable und einer deterministischen Variable immer Null ist. Die Kovarianz ist definiert $\mathrm{cov}(y, x) = \mathrm{E}[y_i - \mathrm{E}(y_i)][x_i - \mathrm{E}(x_i)]$, wobei *über alle möglichen* Ausprägungen aufsummiert wird. Eine deterministische Variable kann als eine degenerierte Zufallsvariable angesehen werden, die nur eine Ausprägung mit Wahrscheinlichkeit Eins annimmt; für deterministische x_i gilt also $x_i = \mathrm{E}(x_i)$, weshalb $\mathrm{cov}(y, x) = 0$.

Dieser letzte Ausdruck ist mit all den Kreuztermen etwas 'unappetitlich' lang. Außerdem enthält er weit mehr als n Unbekannte (Kreuz-)Produkte von Störtermen, es wäre also völlig aussichtslos diese Varianz aus einer Stichprobe schätzen zu wollen.

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

Eine radikale Annahme, die das Problem allerdings 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 $\mathrm{var}(\varepsilon_i) = \sigma^2$; 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. werden aus der gleichen Verteilung gezogen); 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. Ist diese Annahme erfüllt spricht man von *homoskedastischen* Störtermen, ist die Annahme verletzt spricht man von *heteroskedastischen* Störtermen (oder einfach von Heteroskedastizität). Genauer genommen ist bei Heteroskedastizität die *bedingte Varianz* $\mathrm{var}(\varepsilon_i|x)$ eine Funktion der erklärenden Variable, d.h. $\mathrm{var}(\varepsilon_i|x) = \sigma^2(x)$.

2. Unabhängigkeit der Ziehungen, d.h. $\mathrm{E}(\varepsilon_i\varepsilon_j) = 0$ für $i \neq j$ (dies impliziert auch $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$ 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 $\mathrm{E}(\varepsilon_i|x_i) = 0$ benötigt, d.h. der bedingte Erwartungswert der ε_i muss Null sein. Damit werden wir uns erst später beschäftigen.)

Um Gleichung (4.4) 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 (4.4) weg, deshalb gilt in diesem Fall

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

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

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

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 5), dass $\sum w_i^2 = \frac{\sum \ddot{x}_i^2}{\left(\sum \ddot{x}_i^2\right)^2} = \frac{1}{\sum (x_i - \bar{x})^2}$.

Deshalb ist die **Varianz des OLS-Schätzers** $\widehat{\beta}_2$ gleich

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

Man beachte, dass dies nur gilt, wenn alle Annahmen erfüllt sind, die wir zur Herleitung benötigt haben; konkret sind dies Exogenität der Regressoren ($\mathrm{cov}(x, \varepsilon) = 0$) und $\varepsilon_i \sim \text{i.i.d.}(0, \sigma^2)$, wobei die zweite Annahme impliziert $\mathrm{E}(\varepsilon_i) = 0$, $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$ für $i \neq j$ (keine Autokorrelation) und $\mathrm{var}(\varepsilon_i) = \sigma^2$ (keine Heteroskedastizität).

Ähnlich (wenngleich etwas mühsamer) kann man zeigen, dass die **Varianz des Interzepts** $\hat{\boldsymbol{\beta}}_1$ folgendermaßen berechnet werden kann:

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

Da $\widehat{\beta}_1$ und $\widehat{\beta}_2$ Zufallsvariablen sind kann man auch die **Kovarianz** zwischen den beiden Schätzern 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

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

Die Kovarianzen zwischen Schätzern werden wir später für Tests von gemeinsamen Hypothesen (*'joint hypothesis'*) benötigen.

Wir fassen zusammen: unter den bisher getroffenen Annahmen deterministischer x und $\varepsilon_i \sim \text{i.i.d.}(0, \sigma^2)$ gilt

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

4.1.4 Ein Schätzer für die Varianz der Störterme σ^2

Nun haben wir zwar einen Schätzer für $\widehat{\beta}_1$ und $\widehat{\beta}_2$ sowie eine Formel für deren Varianzen, aber in diesen Formeln für die Varianzen kommt die unbekannte Varianz der Störterme der Grundgesamtheit σ^2 vor.

Da wir diese nicht beobachten können müssen wir als nächstes eine *erwartungstreue Schätzfunktion* $\hat{\sigma}^2$ für das wahre σ^2 der Grundgesamtheit herleiten.

Leider kommt das 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 vor, deshalb müssen wir im folgenden einen indirekten und teilweise etwas mühsamen Weg gehen, um einen Schätzer für σ^2 zu erhalten.[6]

Wir erinnern uns, das wahre Modell der Grundgesamtheit ist

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

und für die Mittelwerte gilt[7]

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

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 einem Schätzer 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

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

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

[7] $\sum_i y_i = n\beta_1 + \beta_2 \sum_i x_i + \sum_i \varepsilon_i$. 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 die OLS Schätzer 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 letzten 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 einen erwartungstreuen Schätzer für die Varianz der Grundgesamtheit σ^2 bestimmen, denn aus der letzten Gleichung folgt

$$\frac{\mathrm{E}(\sum \hat{\varepsilon}_i^2)}{n-2}=\sigma^2 \tag{4.5}$$

Wir definieren nun ein

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

denn aufgrund Gleichung (4.5) gilt $\mathrm{E}(\hat{\sigma}^2)=\sigma^2$, also ist $\hat{\sigma}^2$ ein erwartungstreuer Schätzer für σ^2!

Wir können also tatsächlich aus den Stichprobenresiduen $\hat{\varepsilon}_i$ einen erwartungstreuen Schätzer $\hat{\sigma}^2$ für die Varianz der Störterme der Grundgesamtheit σ^2 berechnen, indem wir die Quadratsumme der Residuen $\sum_i \hat{\varepsilon}_i^2$ durch die Anzahl der Freiheitsgrade $n-2$ dividieren.

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

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

Man beachte aber, dass wir für die Herleitung wiederholt die Annahme gemacht haben, dass die Varianz der Störterme konstant ist, $\mathrm{E}(\varepsilon_i^2) = \sigma^2$ (d.h. keine Heteroskedastizität vorliegt), und dass die Störterme untereinander unkorreliert sind, $\mathrm{E}(\varepsilon_i \varepsilon_j) = 0$ für $i \neq j$ (d.h. keine Autokorrelation vorliegt), und dass der Regressor x exogen ist (d.h. $\mathrm{cov}(x, \varepsilon) = 0$).

Ist mindestens eine dieser Annahmen verletzt wird der nach obiger Formel berechnete *Standardfehler der Regression* falsche Ergebnisse liefern, d.h. ein verzerrter Schätzer für σ^2 sein.[8]

Standardfehler der Koeffizienten: Durch Einsetzen dieses Schätzers für den Standardfehler der Regression in die Formeln für die Varianzen der Koeffizienten erhalten wir *Schätzer* für die Varianzen der Koeffizienten

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

und die Wurzeln daraus sind die Standardfehler der Koeffizienten

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

Freiheitsgrade: Wir haben gesehen, dass wir zur Berechnung eines erwartungstreuen Schätzers für σ^2 die Quadratsumme der Stichprobenresiduen $\sum_i \hat{\varepsilon}_i^2$ durch $n-2$ 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 des OLS-Schätzers. 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

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

[8] Man beachte aber, dass wir die beiden ersten Annahmen nicht benötigt haben, um die *Erwartungstreue* der Schätzer $\widehat{\beta}_1$ und $\widehat{\beta}_2$ zu zeigen.

Diese beiden Gleichungen legen eine Restriktion auf die Residuen. Wenn wir z.B. nur die Residuen $\hat{\varepsilon}_1, \hat{\varepsilon}_2, \ldots, \hat{\varepsilon}_{n-2}$ 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.[9] Zwei der Residuen sind deshalb nicht 'frei', sondern 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 diesem Fall haben wir zwei Parameter geschätzt ($\widehat{\beta}_1$ und $\widehat{\beta}_2$), deshalb verlieren wir zwei Freiheitsgrade.

Mit Hilfe des Schätzers $\hat{\sigma}$ (Standardfehler der Regression) können wir nun die erwartungstreuen Schätzer für die *Varianz der Parameter* $\widehat{\beta}_1$ und $\widehat{\beta}_2$, d.h. $\hat{\sigma}^2_{\widehat{\beta}_1}$ und $\hat{\sigma}^2_{\widehat{\beta}_2}$ aus den Stichprobendaten berechnen, die uns später die Durchführung statistischer Tests ermöglichen wird.

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 Schätzer sind ein Maß für die 'Genauigkeit' der Schätzer, d.h. ein Schätzer ist ceteris paribus umso genauer, je kleiner sein Standardfehler ist. Man beachte, dass die Standardfehler die gleiche Dimension haben wie die Koeffizienten, deshalb ist nicht die absolute Größe des Standardfehlers entscheidend, sondern seine Größe *im Verhältnis* zum Koeffizienten. Als Faustregel kann man sich merken, dass der Standardfehler höchstens halb so groß sein sollten wie der dazugehörige Koeffizient; mehr dazu im Kapitel über Hypothesentests.

[9] Stellen Sie sich vor, Sie kennen von drei Residuen nur zwei, z.B. $\hat{\varepsilon}_1 = -3$ und $\hat{\varepsilon}_2 = +1$. Wenn Sie wissen, dass die Residuen die Bedingung $\sum_{i=1}^{3} \hat{\varepsilon}_i = 0$ erfüllen, können Sie daraus sofort schließen, dass $\hat{\varepsilon}_3 = 2$ sein muss.

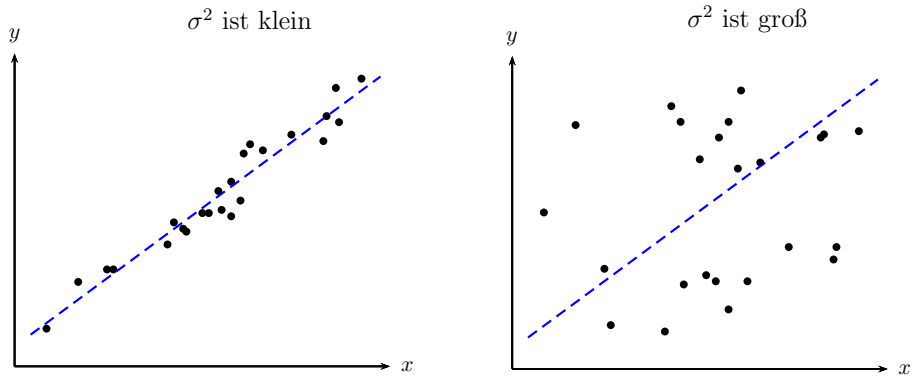


Abbildung 4.1: Regressionen mit unterschiedlicher Varianz von ε (σ^2).

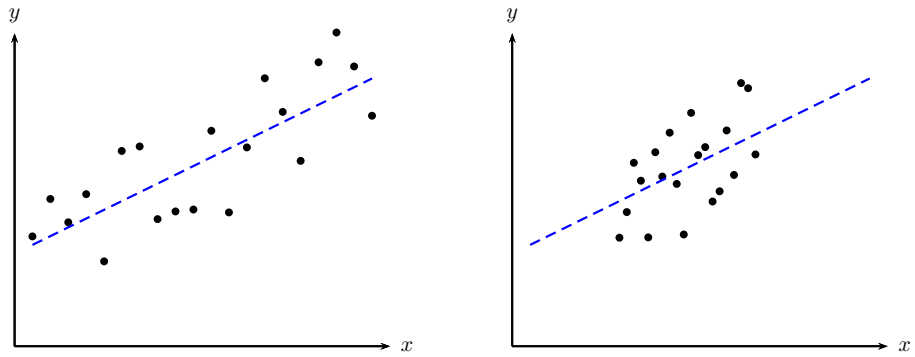


Abbildung 4.2: Unterschiedliche Streuung der erklärenden x Variable. In der linken Abbildung streut x stark, in der rechten Abbildung ist die Streuung von x deutlich kleiner.

Wovon hängt nun die Größe des Standardfehlers ab? Wir wollen uns vorerst auf den meist interessierenden Standardfehler des Steigungskoeffizienten $\widehat{\beta}_2$ beschränken.

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 ein Schätzer für σ^2). Abbildung 4.1 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 $\sum_i (x_i - \bar{x})^2$). Abbildung 4.2 zeigt zwei Stichproben mit gleichem σ^2, die sich nur in der Streuung der x unterscheiden. Offensichtlich ist 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 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, wenn wir die Matrixschreibweise einführen.

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 ein Schätzer für den Standardfehler eines beliebigen Steigungskoeffizienten $\widehat{\beta}_h$ durch den folgenden Ausdruck gegeben ist

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

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

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

Dabei wird der interessierenden Regressor x_h 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 $\widehat{\text{se}}(\widehat{\beta}_h)$ ceteris paribus umso größer ist, je näher das R_h^2 bei Eins liegt. In anderen Worten, der Schätzer $\widehat{\beta}_h$ ist ceteris paribus umso ungenauer, je größer die lineare Abhängigkeiten zwischen x_h und allen anderen Regressoren ist. Diesen Fall werden wir später unter dem Begriff *Multikollinearität* ausführlich diskutieren.

Im Extremfall, wenn es eine lineare Abhängigkeit zwischen den Regressoren gibt, ist das R_h^2 exakt gleich Eins, deshalb ist der Nenner des Standarsfehlers (4.7) gleich Null, d.h. der Standardfehler ist nicht definiert. Dieser Extremfall wird *perfekte Multikollinearität* genannt.

4.2 Gauss-Markov Theorem

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

Bisher haben wir uns ausschließlich mit der Erwartungstreue des OLS-Schätzers und mit der Schätzung von dessen Varianz beschäftigt. In diesem Abschnitt werden wir nun die *Effizienz* des OLS-Schätzers beweisen. Das Gauss-Markov Theorem besagt nämlich, dass der OLS-Schätzer unter bestimmten Annahmen von allen möglichen *linearen und erwartungstreuen Schätzfunktionen* die kleinste Varianz hat, bzw.

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

Die OLS-Schätzfunktion ist – wie wir bereits gesehen haben – linear, da $\widehat{\beta}_2 = \sum w_i y_i$.

Wir werden nun zeigen, dass – wenn die unten angeführten Gauss-Markov Annahmen erfüllt sind – der OLS-Schätzer effizient ist, d.h. $\text{var}(\widehat{\beta}_2^{\text{OLS}}) \leq \text{var}(\widehat{\beta}_2^*)$ wobei $\widehat{\beta}_2^*$ 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.

4.2.1 Beweis für die Effizienz des OLS-Schätzers (Gauss-Markov Theorem)

Der Beweis der Effizienz des OLS-Schätzers ist einer der Höhepunkte jeder einführenden Ökonometrie-Veranstaltung, genießen Sie also das Folgende.[10] Die Grundidee dieses Beweises funktioniert folgendermaßen:

1. Wir gehen von einer beliebigen linearen Schätzfunktion aus.
2. Wir ermitteln die notwendigen Bedingungen, unter denen diese lineare Schätzfunktion erwartungstreu ist.
3. Wir minimieren die Varianz dieser beliebigen linearen Schätzfunktion unter der Nebenbedingung, dass diese lineare Schätzfunktion erwartungstreu ist.

[10] Wer mit dem 'Genießen' Probleme hat sei getröstet, Sie werden in der Veranstaltung auch noch 'Anwendungsorientierteres' erleben.

4. Wir werden sehen, dass die aus der Minimierung resultierende – also varianzminimale – Schätzfunktion genau der OLS-Schätzer ist. Deshalb ist der OLS Schätzer *varianzminimal*.

   Allerdings werden wir im Laufe der Beweisführung einige Annahmen benötigen, 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.

Also los, wir beginnen mit dem Steigungsparameter $\widehat{\beta}_2$. Um die Effizienz des OLS-Schätzers $\widehat{\beta}_2$ 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.

Der Beweis soll außerdem nur für erwartungstreue Schätzfunktionen gelten, d.h. wir müssen 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 des obigen Schätzers 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 $\tilde{\beta}_2 = \sum c_i y_i$ ein unverzerrter Schätzer für β_2 ist müssen die Bedingungen $\sum c_i = 0$ und $\sum c_i x_i = 1$ erfüllt sein.[11]

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

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

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

[11] Man beachte, dass die Gewichte $w_i = \ddot{x}_i / \sum_j \ddot{x}_j^2$ auf Seite 5 diese Bedingungen erfüllten.

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 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 des OLS-Schätzers. Damit haben wir gezeigt, dass der OLS-Schätzer tatsächlich die minimale Varianz unter allen linearen erwartungstreuen Schätzfunktionen hat, *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 des OLS-Schätzers.

Ähnlich kann ein BLU[12] 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.

Wir haben für den Gauss-Markov Beweis eine Reihe von Annahmen benötigt, die wir teilweise auch schon für die Herleitung des Schätzers für σ^2 verwendet haben. Wir werden nun diese Annahmen noch einmal ausführlich und etwas übersichtlicher zusammenfassen.

4.2.2 Annahmen des 'klassischen linearen Regressionsmodells' (CLRM)

Hier fassen wir noch einmal die Annahmen zusammen, die für den Beweis der Effizienz des OLS Schätzers benötigt wurden:

A1 Der datengenerierende Prozess wird durch die lineare Funktion

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

korrekt beschrieben, oder in anderen Worten, die 'wahre' Beziehung zwischen der erklärenden Variablen x und der zu erklärenden Variablen y (d.h. die *'Population Regression Function'*) ist linear in den Parametern.

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

A2 Die Anzahl der zur Verfügung stehenden Beobachtungen ist mindestens so groß wie die Anzahl der zu schätzenden Koeffizienten, und es gibt keine exakte lineare Abhängigkeit zwischen den erklärenden Variablen.

Für das bivariate Modell $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ bedeutet dies, dass $n \geq 2$ und dass x kein Vielfaches der Regressionskonstanten sein darf, d.h., dass x mindestens zwei verschiedene Ausprägungen haben muss.

Für das multiple Regressionsmodell mit k erklärenden Variablen impliziert dies, dass $n \geq k$ sein muss, und dass keine erklärende Variable sich als Linearkombination der restlichen erklärenden Variablen darstellen lässt (*keine perfekte Multikollinaearität*).

Falls diese Annahme verletzt ist existiert keine eindeutige Lösung für den OLS Schätzer!

A3 Die Störterme ε sind nicht mit den erklärenden Variablen x korreliert, bzw. $\text{cov}(x, \varepsilon) = 0$.

Für stochastische x wird wird stochastische Unabhängigkeit gefordert, d.h.

$$\text{E}(\varepsilon_i|x) = \text{E}(\varepsilon_i) = 0$$

Im Folgenden werden wir darüber hinaus häufig annehmen, dass die erklärenden Variablen x *deterministisch* sind, d.h. die x werden bei wiederholten Stichprobenziehungen ('*repeated sampling*') als fest gegebene (deterministische) Größen angenommen. In diesem Fall – und wenn die PRF korrekt spezifiziert wurde (Annahme 1) – reicht die wesentlich weniger strenge Annahme $\text{E}(\varepsilon_i) = 0$.

Wenn diese und alle vorhergehenden Annahmen erfüllt sind ist der OLS Schätzer unverzerrt; eine Verletzung dieser Annahme führt zu einem Bias des OLS Schätzers.

A4 Die insgesamt n Störterme ε_i haben alle die gleiche Verteilung mit Erwartungswert Null und Varianz σ^2, und sind darüber hinaus untereinander unkorreliert. Dies wird kompakt geschrieben als

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

wobei i.i.d. für *'independent and identically distributed'* steht.

Die Annahme $\text{E}(\varepsilon_i) = 0$ haben wir bereits vorher getroffen, darüber hinaus fordert diese Annahme

1. Homoskedastizität: alle ε_i haben die gleiche konstante Varianz σ^2. Etwas genauer fordert diese Annahme, dass die *bedingten* Varianzen der Störterme gleich sein müssen, d.h.

   $$\text{var}(\varepsilon_i|x) = \text{E}(\varepsilon_i|x)^2 = \sigma^2$$

   Wenn die Störterme diese Annahme verletzen spricht man von *Heteroskedastizität*. Abbildung 4.3 zeigt zwei Beispiele für Heteroskedastizität, in der linken Abbildung nimmt die Varianz der Störterme mit x zu, in der rechten Abbildung nimmt sie ab.

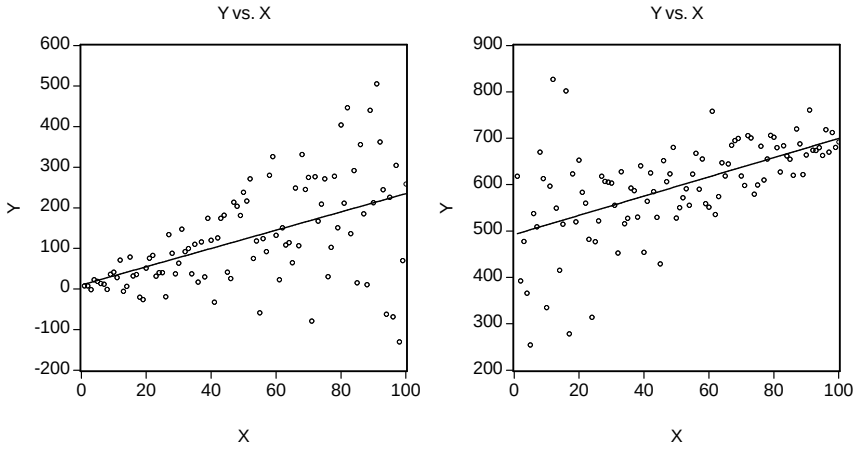


Abbildung 4.3: Heteroskedastische Störterme: Die Varianz der Störterme σ_i^2 ist nicht konstant, sondern hängt von x ab.

2. Die Störterme ε_i der Grundgesamtheit sind *nicht autokorreliert*, d.h. die (bedingte) Korrelation zwischen zwei beliebigen Störtermen ε_i und ε_j für $i \neq j$ ist gleich Null:

$$\mathrm{E}\left(\varepsilon_i \varepsilon_j | x\right) = 0 \quad \text{für } i \neq j$$

Wie bereits mehrfach erwähnt impliziert diese Annahme auch $\mathrm{cov}(\varepsilon_i, \varepsilon_j | x) = 0$ (Achtung: der Umkehrschluss gilt nicht, aus einer Kovarianz von Null folgt *nicht* notwendigerweise stochastische Unabhängigkeit, da die Kovarianz nur lineare Abhängigkeiten misst). Abbildung 4.4 zeigt zwei Fälle mit autokorrelierten Störtermen.

Falls die Annahme $\varepsilon_i \sim \text{i.i.d.}(0, \sigma^2)$ verletzt ist, aber alle vorhergehenden Annahmen erfüllt sind, ist der OLS Schätzer zwar unverzerrt, aber nicht effizient.

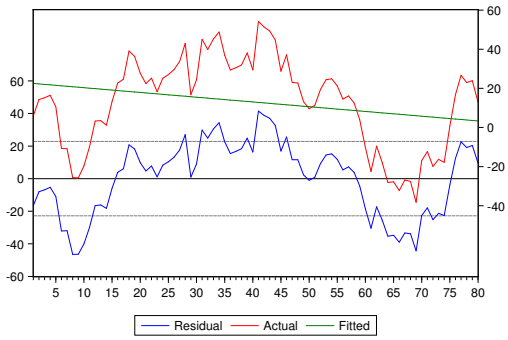
Nur wenn alle vier vorhergehenden Annahmen erfüllt sind ist der OLS Schätzer effizient!

4.3 Asymptotische Eigenschaften ('Große Stichprobeneigenschaften')

Wir haben bisher *Schätzfunktionen* für $\widehat{\beta}_1$ und $\widehat{\beta}_2$ 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 Schätzer unter verschiedenen Bedingungen beurteilen zu können, müssen deren Eigenschaften beurteilt werden 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

Positive Autokorrelation:



Negative Autokorrelation:

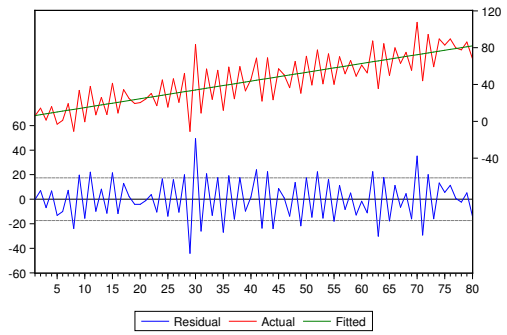


Abbildung 4.4: Autokorrelierte Störterme: $\text{cov}(\varepsilon_i, \varepsilon_j) \neq 0$ für $i \neq j$. Im Fall positiver Autokorrelation ist ein Vorzeichenwechsel der Störterme *seltener* als für i.i.d. Störterme, im Fall negativer Autokorrelation ist ein Vorzeichenwechsel der Störterme *häufiger* als für i.i.d. Störterme.

die Stichprobenkennwertverteilungen von solchen Schätzern allgemein ermittelt werden, zum Beispiel die Verteilung der Mittelwerte aus wiederholten Zufallsstichprobenziehungen, die aus einer normalverteilten Grundgesamtheit gezogen wurden.

Aber oft sind wesentliche Annahmen verletzt, die zur Herleitung der 'Kleine Stichprobeneigenschaften' benötigt wurden, und deshalb können in komplizierteren Fällen Eigenschaften wie Erwartungstreue oder Effizienz häufig 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 einer Zufallsvariablen veranschaulicht werden. Sei X eine Zufallsvariable mit unbekannter Dichtefunktion, von der aber bekannt ist, dass deren Momente Mittelwert μ und Varianz σ^2 fixe Zahlen sind (d.h. *nicht* unendlich groß sind). Wir stellen uns vor, dass aus dieser Verteilung n Zahlen gezogen werden und daraus der Stichprobenmittelwert $\bar{x}_n$ berechnet wird, wobei das tiefgestellte n angibt, auf wievielen Beobachtungen der Stichprobenmittelwert beruht.

Im Folgenden untersuchen wir eine *Folge von Schätzfunktionen* $\hat{\mu}_n$ untersuchen, denn wenn zusätzliche Beobachtungen dazukommen, ändert sich in der Regel auch die Schätzfunktion. Für den einfachen Stichprobenmittelwert ist eine solche Folge von Schätzfunktionen z.B.

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

Diese Mittelwerte sind natürlich selbst wieder Zufallsvariablen mit Dichtefunktionen $f(\hat{\mu}_n)$. 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 werden uns zur vermutlich wichtigsten Eigenschaft von Schätzfunktionen führen, nämlich zur *Konsistenz*; mehr dazu etwas später.

Da die folgenden Ausführungen ziemlich allgemein gehalten sind schreiben wir θ für einen beliebigen Parameter einer Verteilung, und mit $\widehat{\theta}$ bezeichnen wir wie üblich die Schätzfunktion für diesen Parameter (θ könnte zum Beispiel der Mittelwert μ oder der Steigungskoeffizient β_2 aus unserem früheren Beispiel sein).

Asymptotische Eigenschaften sind 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 $\widehat{\theta}$ wenigstens mit steigender Stichprobengröße (d.h. für $n \to \infty$) dem wahren Parameter θ zubewegt.

4.3.1 Konsistenz (*Consistency*)

Die Konsistenz ist die für uns wichtigste asymptotische Eigenschaft. Die Grundidee ist ziemlich einfach, konsistente Schätzer werden mit zunehmender Stichprobengröße immer genauer.

Die formale Definition sieht zunächst etwas schwierig aus:

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

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

das heißt, dass die Wahrscheinlichkeit, dass mit steigendem Stichprobenumfang der Absolutbetrag der Differenz zwischen $\widehat{\theta}_n$ 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 der Schätzer sehr nahe beim wahren Wert θ der Grundgesamtheit liegt.

Wenn der Stichprobenumfang n unendlich groß wird "kollabiert" die Dichtefunktion einer konsistenten Schätzfunktion $\widehat{\theta}_n$ im Punkt θ (siehe Abb. 4.5).

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

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

d.h. wenn der Schätzer *asymptotisch unverzerrt*[13] ist und die Varianz gegen Null geht.

Um die tiefere Bedeutung der Konvergenz zu verstehen benötigt man einige Begriffe aus der Stochastik, die hier nur ganz kurz gestreift werden.

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

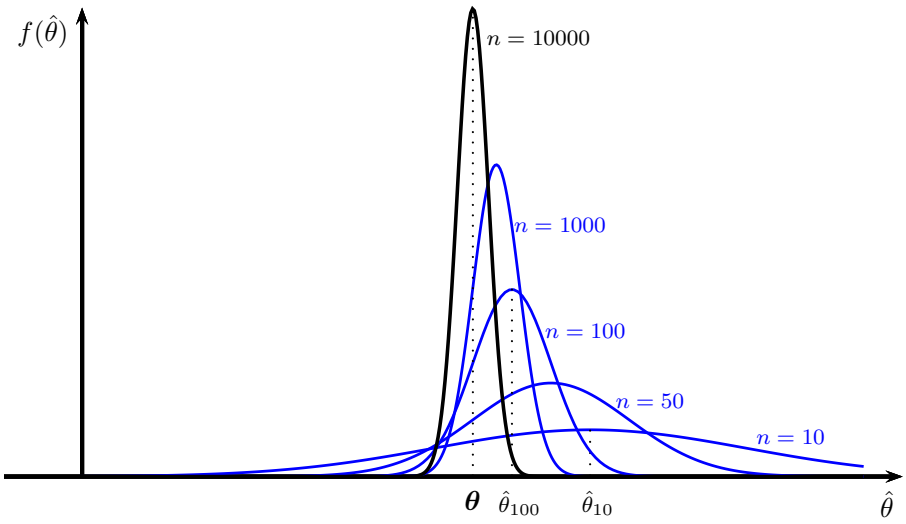


Abbildung 4.5: Konsistente Schätzer können in kleinen Stichproben verzerrt sein, konvergieren aber mit steigendem Stichprobenumfang der Wahrscheinlichkeit nach gegen den wahren Wert θ.

Konvergenz der Wahrscheinlichkeit nach (*'Convergence in Probability'*, auch Stochastische Konvergenz genannt) ist ein zentrales Konzept zur Klärung des Verhaltens von Zufallsvariablen bei wachsendem Stichprobenumfang. Sie gibt – vereinfacht gesprochen – an, in welchem Bereich sich im Falle unendlich vieler Experimente die Zufallsvariable befindet. Das Konzept der stochastischen Konvergenz wird benötigt um *'Gesetze der großen Zahl'* zu beweisen.

Gesetze der großen Zahl Generell sind 'Gesetze der großen Zahlen' meist Aussagen über das Verhalten von Kenngrößen (z.B. Momenten) einer *großen* Zahl von Zufallsvariablen.

Beispiel: Für eine unendliche Folge von Zufallsvariablen $x_1, x_2, \ldots$, die alle denselben Erwartungswert μ besitzen, wird folgende Konvergenzaussage als (ein) schwaches Gesetz der großen Zahlen bezeichnet:

Das arithmetische Mittel von n Zufallsvariablen $\hat{\mu}_n = (x_1 + x_2 + \cdots + x_n)/n$ konvergiert stochastisch gegen μ wenn n gegen Unendlich geht; das bedeutet, für jede positive Zahl δ (beliebig klein) gilt

$$\lim_{n\to\infty} \Pr\left(|\hat{\mu}_n - \mu| < \delta\right) = 1$$

Dieses schwache Gesetz der großen Zahl gilt beispielsweise, wenn die Zufallsvariablen $x_1, x_2, x_3, \ldots$ endliche Varianzen $\sigma_1^2, \sigma_2^2, \ldots$ haben, die zudem durch eine gemeinsame obere Grenze beschränkt sind, sowie untereinander unkorreliert sind (d.h., $\mathrm{cov}(x_i, x_j) = 0$ für $i \neq j$).

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

Dies wird oft kürzer geschrieben als

$$\widehat{\theta} \xrightarrow{p} \theta$$

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

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

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

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

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.

Regeln für das Rechnen mit 'probability-limits'

1. Für eine beliebige Konstante c gilt

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

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

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

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

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

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

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

4.3.2 Beispiel: Unverzerrtheit und Konsistenz des OLS-Schätzers bei stochastischen Regressoren (x)

Bisher haben wir angenommen, dass die erklärende Variable x deterministisch ist, d.h. dass bei wiederholten Stichprobenziehungen nur verschiedene y gezogen werden, aber die x gewissermaßen 'fest gehalten' werden.

In diesem Unterabschnitt interessieren uns die Eigenschaften des OLS-Schätzers, wenn die erklärende Variable x ebenso stochastisch ist. Der OLS-Schätzer 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 $\ddot{y}_i := y_i - \bar{y}$.

Für den Fall stochastischer Regressoren benötigen wir einige zusätzliche Annahmen

1. $\text{E}(\varepsilon_i|x_i) = 0$: Der auf die x 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 des OLS Schätzers.

2. 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 einfache *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).

   Diese Bedingungen wurden auch schon für den Beweis der Effizienz des OLS Schätzers benötigt.

3. 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.

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

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

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 des Schätzers $\widehat{\beta}_2$ 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, d.h. $\mathrm{E}(\varepsilon_i | x_i) = 0$. 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 $\mathrm{E}(\varepsilon_i | x_i) = 0$.[14]

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 konsistente Schätzer für die Varianz und Kovarianz der Grundgesamtheit. Unter den vorher getroffenen Annahmen können wir davon ausgehen, dass diese gegen einen festen Wert konvergieren.

Der Schätzer $\widehat{\beta}_2$ 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

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

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

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 ist der OLS-Schätzer auch nicht konsistent!

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 spezifizierten Zusammenhänge können z.B. bei simultanen Gleichungssystemen oder 'omitted variables' auftreten. Solche nicht spezifizierte Zusammenhänge können eine Korrelation zwischen den ε und x erzeugen, was dazu führt, dass der OLS Schätzer weder unverzerrt, effizient noch konsistent ist!

4.3.3 Asymptotische Normalverteilung

Ein Schätzer ist asymptotisch normalverteilt, wenn seine 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) Schätzern $\widehat{\theta}_n$ aus Stichproben des Umfangs n, die alle derselben Grundgesamtheit entnommen wurden, mit zunehmendem Stichprobenumfang in eine Normalverteilung übergeht, und das unabhängig von der Verteilung der Grundgesamtheit! Beweise der Konvergenz hinsichtlich der Verteilung führen zu den *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 endlich sind.

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

oder einfacher

$$\sqrt{n}\,(\hat{\mu} - \mu_x) \xrightarrow{d} \mathrm{N}(0, \sigma_x^2)$$

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

Es gibt eine ganze Reihe von zentralen Grenzwertsätzen, die teilweise deutlich allgemeiner gelten.

Man kann zeigen, dass die OLS Schätzer unter den obigen Annahmen asymptotisch normalverteilt sind.

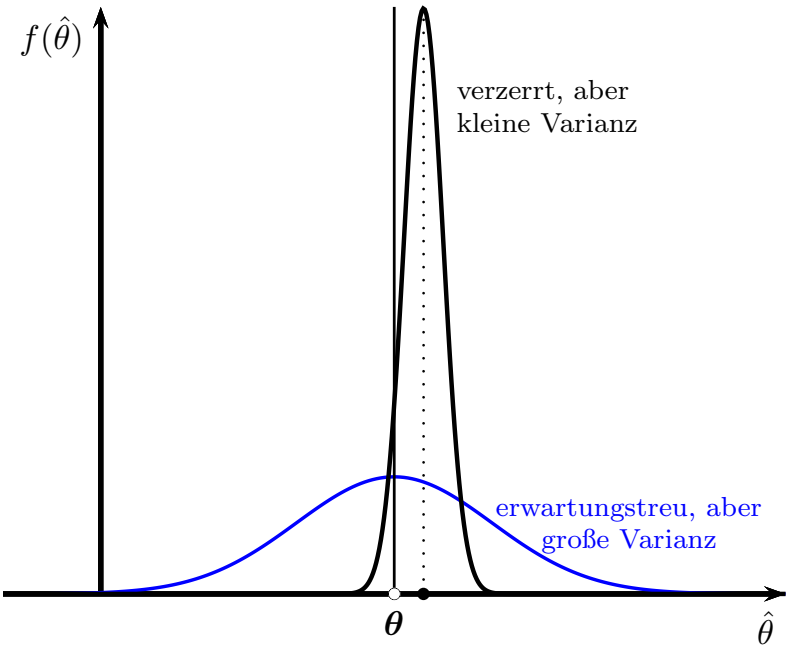


Abbildung 4.6: Mean Square Error Abwägung zwischen erwartungstreuen Schätzfunktionen mit großer Varianz und verzerrten Schätzfunktionen mit kleiner Varianz.

4.3.4 Asymptotische Effizienz

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

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

Wir haben uns bisher nur mit erwartungstreuen Schätzfunktionen beschäftigt. Manchmal ist aber keine erwartungstreue Schätzfunktion verfügbar. In solchen Fällen wird manchmal auf den 'Mean Square Error' (MSE) zurückgegriffen, der Varianz und Verzerrung zusammenfaßt und sich deshalb besonders zur Beurteilung nicht erwartungstreuer Schätzfunktionen eignet (siehe Abb. 4.6).

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

Folgende Konzepte sind im folgenden von Bedeutung:

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

Der Stichprobenfehler ist einfach der Unterschied zwischen dem Schätzer aus der Stichprobe und dem wahren Wert der Grundgesamtheit. Die Größe des Stichprobenfehlers wird sich üblicherweise von Stichprobe zu Stichprobe unterscheiden. Die Verzerrung ist die Differenz zwischen dem Erwartungswert der Stichprobenverteilung eines Schätzers und dem wahren Wert der Grundgesamtheit. Diese ist für einen Schätzer 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 eines Schätzers um den wahren Wert. Er ähnelt darin 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:

$$\begin{aligned}
\mathrm{MSE}(\widehat{\theta}) &= \mathrm{E}(\widehat{\theta} - \theta)^2 \\
&= \mathrm{E}[\widehat{\theta} - \mathrm{E}(\widehat{\theta}) + \mathrm{E}(\widehat{\theta}) - \theta]^2 \\
&= \mathrm{E}[\widehat{\theta} - \mathrm{E}(\widehat{\theta})]^2 + \mathrm{E}[\mathrm{E}(\widehat{\theta}) - \theta]^2 + 2\,\mathrm{E}[\widehat{\theta} - \mathrm{E}(\widehat{\theta})][\mathrm{E}(\widehat{\theta}) - \theta] \\
&= \mathrm{E}[\widehat{\theta} - \mathrm{E}(\widehat{\theta})]^2 + \mathrm{E}[\mathrm{E}(\widehat{\theta}) - \theta]^2 + \\
&\quad +2\{[\mathrm{E}(\widehat{\theta})]^2 - [\mathrm{E}(\widehat{\theta})]^2 - \theta\,\mathrm{E}(\widehat{\theta}) + \theta\,\mathrm{E}(\widehat{\theta})\} \\
&= \mathrm{E}[\widehat{\theta} - \mathrm{E}(\widehat{\theta})]^2 + \mathrm{E}[\mathrm{E}(\widehat{\theta}) - \theta]^2 \\
&= \mathrm{var}(\widehat{\theta}) + [\mathrm{Bias}(\widehat{\theta})]^2
\end{aligned}$$

Dieser Zusammenhang gilt für alle Schätzer. Akademische Forscher neigen oft dazu unverzerrte Schätzer 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 Unverzerrtheit.

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

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

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