Kapitel 2

Das klassische lineare Regressionsmodell

- **Modellannahmen**
- **Kleinste-Quadrate- und Maximum-Likelihood-Schätzung der Modellparameter**
- **Konfidenzintervalle und Parametertests**

2.1 Annahmen des klassischen linearen Regressionsmodells

Wir betrachten im Folgenden die Variablen

Y (Bezeichnungen: abhängige Variable, Regressand, endogene Variable, Zielvariable, Responsevariable)

$X_1,...,X_p$ (Bezeichnungen: unabhängige Variablen, Regressoren, exogene Variablen, Kovariablen, Kontrollvariablen)

zwischen denen ein (zumeist a priori unbekannter) funktionaler Zusammenhang

(1) $$Y = f\left(X_1,...,X_p\right)+\varepsilon$$

bestehe. ε ist eine zufällige *Stör- oder Fehlervariable*, die den *systematischen Term* $f(X_1,...,X_p)$ (die sogenannte *Regressionsfunktion*) überlagert. Letzterer sei

$$f\left(X_1,...,X_p\right)= E\left(Y \mid X_1,...,X_p\right)$$

die bedingte Erwartung von Y für gegebene Werte der Regressoren. *Ziel der Regressionsanalyse* ist die Annäherung der bedingten Erwartung auf der Basis empirischer Daten.

Spezifizieren wir die Variablen $Y, X_1, ..., X_p, \varepsilon$ und die Funktion f durch geeignete Annahmen näher, dann entsteht aus der unspezifischen Gleichung (1) ein spezifisches Regressionsmodell.

Funktionale Form

Wir unterstellen im Folgenden eine lineare Funktion

$$f(z_1, ..., z_p) = \beta_0 + \beta_1 z_1 + ... + \beta_p z_p \ ,$$

so dass der Zusammenhang

(2) $$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \varepsilon$$

gilt. Die Konstanten $\beta_0, \beta_1, ..., \beta_p$ sind unbekannte Modellparameter der *linearen Regressionsfunktion*, die *Regressionskoeffizienten*.

Als Spezialfall der *linearen Mehrfachregression* oder *linearen multiplen Regression* mit $p > 1$ folgt für $p = 1$ die *lineare Einfachregression*

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

B–2.1 *Makroökonomische Konsumfunktion.*

Im Rahmen der Keynesianischen Theorie wird ein linearer Zusammenhang

$$Konsum = \beta_0 + \beta_1 \cdot Einkommen$$

zwischen dem Konsum der privaten Haushalte einer Volkswirtschaft und deren verfügbaren Einkommen unterstellt $(\beta_0, \beta_1 > 0)$. In obiger Gleichung kennzeichnet β_0 den autonomen Konsum und

$$\beta_1 = \frac{d\ Konsum}{d\ Einkommen}$$

die marginale Konsumquote. Aus der Konsumfunktion folgt für ein bestimmtes Einkommensniveau deterministisch ein bestimmtes Konsumniveau. Der Determinismus der Hypothese wird in der Theorie mit Hilfe der *ceteris paribus – Klausel* gerechtfertigt.

Will man den Zusammenhang von Konsum und Einkommen statistisch schätzen, dann schlägt sich in den empirischen Daten die Variation ökonomischer Größen nieder, die in der Theorie als konstant unterstellt werden. Bei der statistischen Modellierung geht man sinnvollerweise von einem systematischen Zusammenhang aus, der durch einen unsystematischen zufälligen Störterm ε überlagert wird:

$$Konsum = \beta_0 + \beta_1 \cdot Einkommen + \varepsilon \quad \boxtimes$$

Skalenniveau der Variablen

Den *Regressanden* Y setzen wir als eine *stetige metrisch skalierte Variable* voraus. Die *Regressoren* $X_1,...,X_p$ können *metrische oder kategoriale* (nominal oder ordinal) *Variablen* sein.

Im Falle metrischer Variablen lassen wir auch Transformationen

$$Y = h(U) \qquad \text{bzw.} \qquad X_i = h(V_i)$$

der ursprünglichen Variablen U bzw. V_i zu, beispielsweise in der Form

$$h(z) = z^2, \quad h(z) = z^{-1}, \quad h(z) = ln\, z \quad \text{usw.}$$

Kategoriale Regressoren mit k Kategorien $1,...,k$ werden durch $m = k - 1$ sogenannte *Dummy-Variablen*

$$X_i^{(1)},...,X_i^{(m)}$$

kodiert.

Kodierung kategorialer Variablen V_i

(a) *Binär-Kodierung* (binäre Dummies)

$$X_i^{(v)} = \begin{cases} 1 & \text{falls } V_i \text{ die Kategorie } v \text{ annimmt} \\ 0 & \text{sonst} \end{cases} \qquad (v = 1,\ldots,m)$$

Im Falle $V_i = k$ gilt hier also $X_i^{(1)} = \ldots = X_i^{(m)} = 0$.

(b) *Effekt-Kodierung* (nicht binäre Dummies)

$$X_i^{(v)} = \begin{cases} 1 & \text{falls } V_i \text{ die Kategorie } v < k \text{ annimmt} \\ -1 & \text{falls } V_i \text{ die Kategorie } k \text{ annimmt} \\ 0 & \text{sonst} \end{cases} \qquad (v = 1,\ldots,m)$$

Im Falle $V_i = k$ ist hier $X_i^{(1)} = \ldots = X_i^{(m)} = -1$.

B–2.2 *Mikroökonomisches Mietpreismodell* (*Binär-Kodierung kategorialer Variablen*).

Der logarithmische Mietpreis von Wohnungen soll durch die jeweilige Wohnfläche und Lage der Wohnung statistisch erklärt werden. Im Gegensatz zu den metrisch skalierten Variablen *Miete* und *Fläche* ist hierbei *Lage* eine kategoriale Variable mit drei möglichen Ausprägungen „*normal*“, „*gut*“ und „*sehr gut*“. Die Wohnungslage können wir beispielsweise mittels der beiden Dummies

$$Lage^{+} = \begin{cases} 1 & \text{falls Wohnung in guter Lage} \\ 0 & \text{sonst} \end{cases}, \quad Lage^{++} = \begin{cases} 1 & \text{falls Wohnung in sehr guter Lage} \\ 0 & \text{sonst} \end{cases}$$

kodieren und folgenden Regressionsansatz formulieren:

$$log(Miete) = \beta_0 + \beta_1 \cdot Fläche + \beta_2 \cdot Lage^{+} + \beta_3 \cdot Lage^{++} + \varepsilon$$

bzw.

$$Miete = \begin{cases} exp(\beta_0 + \beta_1 \cdot Fläche + \varepsilon) & \text{falls Lage normal} \\ exp([\beta_0 + \beta_2] + \beta_1 \cdot Fläche + \varepsilon) & \text{falls Lage gut} \\ exp([\beta_0 + \beta_3] + \beta_1 \cdot Fläche + \varepsilon) & \text{falls Lage sehr gut .} \end{cases}$$

Mit diesem Regressionsansatz unterstellen wir einen konstanten Mietaufschlag bei guter Lage (falls $\beta_2 > 0$) und einen stärkeren konstanten Mietaufschlag bei sehr guter Lage (falls $\beta_3 > \beta_2 > 0$) gegenüber der Normallage.

Vermuten wir hingegen, dass *Fläche* und *Lage* einen interaktiven Einfluss auf *log*(*Miete*) ausüben, dann können wir dies z.B. wie folgt berücksichtigen:

$$log(Miete) = \beta_0 + \beta_1 \cdot Fläche + \beta_2 \cdot Fläche \cdot Lage^{+} + \beta_3 \cdot Fläche \cdot Lage^{++} + \varepsilon$$

bzw.

$$Miete = \begin{cases} exp(\beta_0 + \beta_1 \cdot Fläche + \varepsilon) & \text{falls Lage normal} \\ exp(\beta_0 + (\beta_1 + \beta_2) \cdot Fläche + \varepsilon) & \text{falls Lage gut} \\ exp(\beta_0 + (\beta_1 + \beta_3) \cdot Fläche + \varepsilon) & \text{falls Lage sehr gut.} \end{cases}$$

Die Produkte $Fläche \cdot Lage^{+}$ und $Fläche \cdot Lage^{++}$ bezeichnet man als *Interaktionen* zweier Regressoren ☒

Daten

Zum Zwecke der Modellschätzung beobachten wir die Variablen

$$Y, X_1, \ldots, X_p$$

n-mal. Bei dem Datenmaterial kann es sich um Querschnittsdaten oder Längsschnittsdaten handeln.

Querschnittsdaten liegen vor, wenn Werte der Variablen simultan an n verschiedenen statistischen Einheiten gemessen werden. Es handelt sich i. d. R. um Stichprobenmaterial.

Beispiel: Wählen wir in einer Stadt n Wohnungen aus und erheben an einem bestimmten Stichtag deren Quadratmetermieten, Wohnflächen und Lagen, so erhalten wir einen Querschnittsdatensatz.

Längsschnitts- oder Zeitreihendaten liegen vor, wenn Werte der Variablen zu unterschiedlichen Zeitpunkten an einer statistischen Einheit gemessen werden und deren chronologische Anordnung erhalten bleibt. Die Messzeitpunkte orientieren sich i. d. R. am Kalender und sind (näherungsweise) äquidistant.

Beispiel: Erheben wir das verfügbare Einkommen und den Konsum aller privaten Haushalte in der Bundesrepublik jährlich im Zeitablauf, so erhalten wir einen Zeitreihendatensatz.

Die *konkreten Daten* fassen wir als Realisierungen von *Stichprobenvariablen* (potenzielle Beobachtungen) auf. Zur Vereinfachung der Notation schreiben wir Daten wie Stichprobenvariablen gleich:

$$y_\nu, x_{\nu 1}, \ldots, x_{\nu p} \qquad (\nu = 1, \ldots, n) \ .$$

Im Falle von Querschnittsdaten kennzeichnet der Index ν die statistischen Einheiten, an denen die Werte gemessen wurden. Liegen Zeitreihen vor, dann ist ν als Zeitindex zu interpretieren, der für verschiedene Messzeitpunkte steht.

Hinweis

Eine Mischform von Querschnitts- und Längsschnittsdaten sind die *Paneldaten*. Sie liegen vor, wenn Werte der Variablen an n verschiedenen statistischen Einheiten wiederholt im Zeitablauf gemessen werden und die chronologische Anordnung der Werte erhalten bleibt. Paneldaten erfordern spezifische Modellformulierungen und werden im Rahmen des klassischen linearen Regressionsmodells *nicht* betrachtet.

Beispiel: Wählen wir in einer Stadt n Wohnungen aus und erheben im Zeitablauf mit einjährigem Zeitabstand deren Quadratmetermieten, Wohnflächen und Wohnlagen, so erhalten wir einen Paneldatensatz.

Die Stichprobenvariablen $y_\nu, x_{\nu 1}, \ldots, x_{\nu p}$ $(\nu = 1, \ldots, n)$ erfüllen die n Gleichungen

(3) $$y_\nu = \beta_0 + \beta_1 x_{\nu 1} + \ldots + \beta_p x_{\nu p} + \varepsilon_\nu \qquad (\nu = 1, \ldots, n)\,,$$

wobei die Störungen ε_ν nicht direkt beobachtbar sind. Mit

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

können wir die n Gleichungen (3) als eine Matrixgleichung (*lineare Regressionsgleichung*)

(4) $$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

schreiben.

Über die bisherigen Voraussetzungen hinaus sind weitere Annahmen bezüglich der Variablen in Gleichung (4) notwendig.

Statistische Eigenschaften der Regressoren und Störungen

Sind die *Regressoren* im Modell rein exogen bestimmt, dann kann man die Regressorenwerte auch als fest vorgegeben und die Stichprobenvariablen $x_{\nu 1},\ldots,x_{\nu p}$ *als deterministische Größen* betrachten. Die $(n, p+1)$-Regressormatrix $\boldsymbol{X}$ in Gleichung (4) ist nun eine feste, nicht-stochastische Matrix.

Wir setzen ferner voraus, dass die Werte der Regressoren keine Redundanzen in Form exakter linearer Abhängigkeiten aufweisen und fordern den vollen Spaltenrang $rg(\boldsymbol{X}) = p+1$ der Regressormatrix.

Die *Störungen* ε_ν werden *als Zufallsvariablen mit spezifischen Charakteristika* angenommen. Für alle $\nu,\kappa = 1,\ldots,n$ und $\nu \neq \kappa$ soll gelten:

$$E(\varepsilon_\nu) = 0 \ ,$$

$$Var(\varepsilon_\nu) = E\left(\{\varepsilon_\nu - E(\varepsilon_\nu)\}^2\right) = E\left(\varepsilon_\nu^2\right) = \sigma^2 = \text{const.} \qquad (\textit{Homoskedastizität}) \ ,$$

$$Cov(\varepsilon_\nu,\varepsilon_\kappa) = E\left(\{\varepsilon_\nu - E(\varepsilon_\nu)\}\{\varepsilon_\kappa - E(\varepsilon_\kappa)\}\right) = E(\varepsilon_\nu\varepsilon_\kappa) = 0 \qquad (\textit{Unkorreliertheit}) \ .$$

Aufgrund obiger Annahmen üben die Störungen keinen systematischen Einfluss auf die endogene Variable Y aus.

Erwartungswerte, Varianzen und Kovarianzen der Störungen fassen wir in einem *n-Erwartungswert-vektor*

$$E(\boldsymbol{\varepsilon}) = E\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n\end{pmatrix} = \begin{pmatrix}E(\varepsilon_1)\\ E(\varepsilon_2)\\ \vdots\\ E(\varepsilon_n)\end{pmatrix} = \begin{pmatrix}0\\ 0\\ \vdots\\ 0\end{pmatrix} = \mathbf{0}$$

und einer (*n,n*)-*Varianz-Kovarianz-Matrix*

$$Cov(\boldsymbol{\varepsilon}) = E\left(\{\boldsymbol{\varepsilon} - E(\boldsymbol{\varepsilon})\}\{\boldsymbol{\varepsilon} - E(\boldsymbol{\varepsilon})\}'\right) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = E\begin{pmatrix}\varepsilon_1\varepsilon_1 & \varepsilon_1\varepsilon_2 & \cdots & \varepsilon_1\varepsilon_n\\ \varepsilon_2\varepsilon_1 & \varepsilon_2\varepsilon_2 & \cdots & \varepsilon_2\varepsilon_n\\ \vdots & \vdots & \ddots & \vdots\\ \varepsilon_n\varepsilon_1 & \varepsilon_n\varepsilon_2 & \cdots & \varepsilon_n\varepsilon_n\end{pmatrix}$$

$$= \begin{pmatrix}E(\varepsilon_1^2) & E(\varepsilon_1\varepsilon_2) & \cdots & E(\varepsilon_1\varepsilon_n)\\ E(\varepsilon_2\varepsilon_1) & E(\varepsilon_2^2) & \cdots & E(\varepsilon_2\varepsilon_n)\\ \vdots & \vdots & \ddots & \vdots\\ E(\varepsilon_n\varepsilon_1) & E(\varepsilon_n\varepsilon_2) & \cdots & E(\varepsilon_n^2)\end{pmatrix} = \begin{pmatrix}\sigma^2 & 0 & \cdots & 0\\ 0 & \sigma^2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \sigma^2\end{pmatrix} = \sigma^2\begin{pmatrix}1 & 0 & \cdots & 0\\ 0 & 1 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 1\end{pmatrix} = \sigma^2\boldsymbol{I}$$

kompakt zusammen. Hierbei ist **0** der *n*-Nullvektor und $\boldsymbol{I}$ die (*n,n*)-Einheitsmatrix.

Die potentiellen Werte $y_1,...,y_n$ der endogene Variable Y sind aufgrund obiger Modellannahmen Linearkombinationen deterministischer und zufälliger Größen und folglich Zufallsvariablen. Es gilt:

$$E(\boldsymbol{y}) = E(\boldsymbol{X\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{X\beta} + E(\boldsymbol{\varepsilon}) = \boldsymbol{X\beta} \ ,$$

$$Cov(\boldsymbol{y}) = Cov(\boldsymbol{X\beta} + \boldsymbol{\varepsilon}) = Cov(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I} \ .$$

Der systematische Term $\boldsymbol{X\beta}$ in der Regressionsgleichung (4) entspricht somit dem n-Erwartungswertvektor von $\boldsymbol{y}$. Die Streuung der endogenen Variable wird auf die Störungen zurückgeführt.

Als Zusatzannahme werden häufig *normalverteilte Störungen* $\varepsilon_\nu \sim N(0,\sigma^2)$ bzw. $\boldsymbol{\varepsilon} \sim N_n(\boldsymbol{0},\sigma^2 \boldsymbol{I})$ unterstellt. N_n steht hierbei abkürzend für n-dimensionale Normalverteilung. Unter Gültigkeit der Zusatzannahme folgen die Stichprobenvariablen $y_1,...,y_n$ als Linearkombinationen der Störungen ebenfalls einer n-dimensionale Normalverteilung:

$$\boldsymbol{y} \sim N_n(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}) \ .$$

Können die Störungen sinnvoll als eine Überlagerung zahlreicher voneinander unabhängiger Einflussfaktoren auf die endogene Variable Y aufgefasst werden, dann liefert der *Zentrale Grenzwertsatz* eine Rechtfertigung der Normalverteilungsannahme.

Die Modellannahmen fassen wir in einer Definition zusammen.

D–2.1 Klassisches lineares Modell mit festen Regressoren

(A1) $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ (Linearität).

(A2) $\boldsymbol{X}$ ist eine nicht-stochastische (feste) Matrix. Es gilt $rg(\boldsymbol{X}) = p+1$; d.h. $\boldsymbol{X}$ hat vollen Spaltenrang und ist daher spaltenregulär.

(A3) Die Störungen $\boldsymbol{\varepsilon}$ sind zufällige Größen mit

$$E(\boldsymbol{\varepsilon}) = \mathbf{0} \quad \text{und} \quad Cov(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}\,.$$

(A4) Zusatzannahme (optional)

$$\boldsymbol{\varepsilon} \sim N_n\left(\mathbf{0}, \sigma^2 \boldsymbol{I}\right).$$

Die Annahme deterministischer Regressoren ist in vielen, aber nicht in allen Situationen plausibel. Dies wird im nachfolgenden Beispiel deutlich. In der statistischen Literatur betrachtet man deshalb neben dem linearen Modell mit festen Regressoren auch einen *Modellansatz mit stochastischen Regressoren*. Da beide Modellansätze allerdings im Wesentlichen zu denselben Ergebnissen führen, beschränken wir unsere Betrachtungen vereinfachend auf den oben vorgestellten Modelltyp.

B–2.1 *Makroökonomische Konsumfunktion.*

Die statistische Schätzung einer makroökonomischen Konsumfunktion für eine Volkswirtschaft erfolgt auf der Basis von Zeitreihendaten. In der Regel liegen jährliche oder quartalsweise Messwerte des Konsums und des verfügbaren Einkommens der privaten Haushalte zugrunde (vgl. Kapitel 2.2.). Der Keynesianischen Ansatz führt zu den linearen Regressionsgleichungen

$$konsum_\nu = \beta_0 + \beta_1 \cdot einkommen_\nu + \varepsilon_\nu \quad (\nu = 1,\ldots,n).$$

Das Einkommen in der Zeitperiode ν, also $einkommen_\nu$, betrachten wir als eine *exogen vorgegebene feste Größe*. Der Konsum $konsum_\nu$ in der Zeitperiode ν wird als Überlagerung des systematischen nicht-zufälligen Terms $E(konsum_\nu) = \beta_0 + \beta_1 \cdot einkommen_\nu$ und der zufälligen Störung ε_ν erklärt. Die Zufallsvariable $konsum_\nu$ steht für die potentiellen Werte des Regressanden *Konsum* bei gegebenem Wert $einkommen_\nu$ des Regressors *Einkommen.* Folglich dürfen wir $E(konsum_\nu)$ als bedingten Erwartungswert von *Konsum* gegeben $Einkommen = einkommen_\nu$ interpretieren. Wir schreiben dies symbolisch:

$$E(Konsum \mid Einkommen = einkommen_\nu) = \beta_0 + \beta_1 \cdot einkommen_\nu \ .$$

Die Störungen ε_ν fassen unsystematisch auf den Konsum wirkende Einflussfaktoren zusammen. Im

klassischen lineare Modell wir vorausgesetzt, dass die Variabilität der Einflüsse konstant (und damit unabhängig vom Regressor) ist. Folglich ist auch die Variabililtät des Konsums konstant:

$$Var(konsum_\nu) = Var(\beta_0 + \beta_1 \cdot einkommen_\nu + \varepsilon_\nu) = Var(\varepsilon_\nu) = \sigma^2 \qquad (\nu = 1,\ldots,n).$$

Man bezeichnet dies als *Homoskedastizität* im Gegensatz zur *Heteroskedastizität*, welche bei nicht konstanten Varianzen der Störungen vorliegt. Wirken sehr viele unabhängige Einflussfaktoren auf den Konsum, so kann ggf. unter Berufung auf den Zentralen Grenzwertsatz die *Normalverteilung* der Störungen angenommen werden. Eine graphische Darstellung der Zusammenhänge findet sich in Abbildung 2.1.

Die Annahme eines deterministischen Regressors lässt sich in obigem Beispiel einfach begründen. Es gibt jedoch auch Modellformulierungen, bei denen die Annahme deterministischer Regressoren unplausibel ist. Unterstellt man beispielsweise, dass der Periodenkonsum $konsum_\nu$ nicht nur vom Periodeneinkommen $einkommen_\nu$, sondern auch vom Konsumniveau der Vorperiode abhängt, d. h.

$$konsum_\nu = \beta_0 + \beta_1 \cdot einkommen_\nu + \beta_2 \cdot konsum_{\nu-1} + \varepsilon_\nu \qquad (\nu = 1,\ldots,n),$$

dann ist der *Regressor* $konsum_{\nu-1}$ *endogener und folglich stochastischer Natur*.

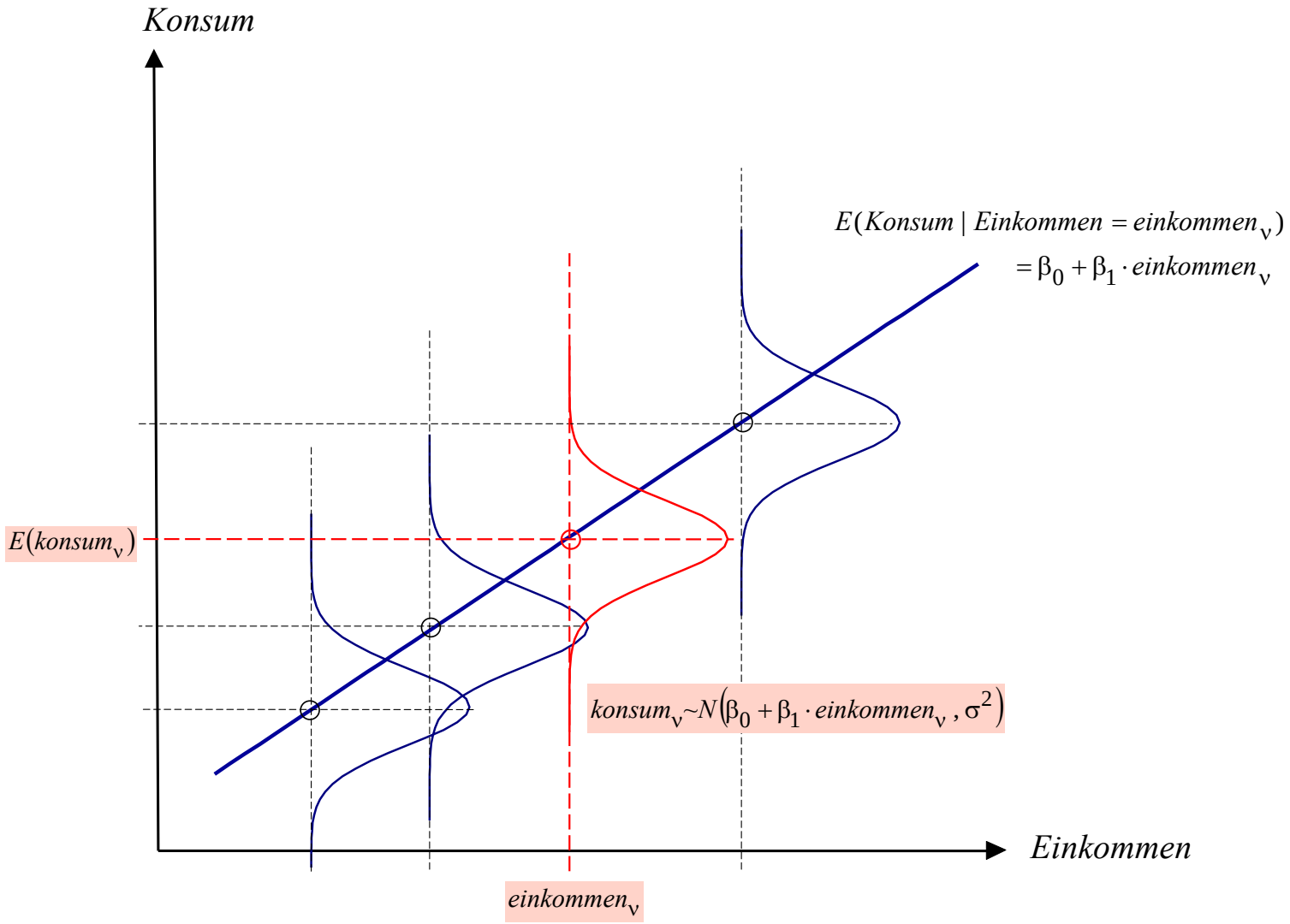


Abb. 2.1: Veranschaulichung der linearen Einfachregression mit homoskedastischen normalverteilten Störungen ☒

2.2 Parameterschätzung: Kleinste-Quadrate-Methode

Das Modell 2.1 enthält die Regressionskoeffizienten und die Varianz der Störungen

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \quad \text{sowie} \quad \sigma^2$$

als unbekannten Modellparameter, die es im Folgenden auf der Datenbasis $y_\nu, x_{\nu 1}, \ldots, x_{\nu p}$ $(\nu = 1, \ldots, n)$ zu schätzen gilt. Die Regressionskoeffizienten $\boldsymbol{\beta}$ schätzen wir zunächst mit Hilfe der von CARL FRIEDRICH GAUSS [1777-1855] entwickelten *Methode der kleinsten Quadrate* (englisch: *Ordinary Least Squares* - *OLS*).

Zur Rechtfertigung der OLS-Methode werden die Annahmen A1 – A3 des Modells in D-2.1 – nicht aber die Normalverteilungsannahme A4 – benötigt. Der Vektor $\boldsymbol{\beta}$ der Regressionskoeffizienten wird im Rahmen des Ansatzes so bestimmt, dass die Quadratsumme der Abweichungen der Werte des Regressanden Y von der Regressionsfunktion (Quadratsumme der Residuen) minimal wird.

D–2.2 OLS-Schätzer für die Regressionskoeffizienten im klassischen linearen Modell

Der OLS-Schätzer $\hat{\boldsymbol{\beta}}_{OLS}$ für $\boldsymbol{\beta}$ ist definiert als Minimumstelle der Quadratsummenfunktion

(5a) $$Q(\boldsymbol{\beta}) = \sum_{\nu=1}^{n} \varepsilon_\nu^2 = \sum_{\nu=1}^{n} \left(y_\nu - \beta_0 - \beta_1 x_{\nu 1} - \ldots - \beta_p x_{\nu p}\right)^2 = \sum_{\nu=1}^{n} (y_\nu - \boldsymbol{x}_\nu' \boldsymbol{\beta})^2 ,$$

wobei $\boldsymbol{x}_\nu' = (x_{\nu 1}, \ldots, x_{\nu n})$ die ν-te Zeile der Regressormatrix $\boldsymbol{X}$ bezeichnet. In Matrixschreibweise notieren wir die Quadratsummenfunktion gemäß

(5b)
$$\begin{aligned} Q(\boldsymbol{\beta}) &= \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= (\boldsymbol{y}' - \boldsymbol{\beta}'\boldsymbol{X}')(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \end{aligned}$$

mit $\boldsymbol{\varepsilon} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$.

Zur Herleitung des OLS-Schätzers bedient man sich der Differentialrechnung. *Notwendige Bedingung* für ein Extremum der Funktion $Q(\boldsymbol{\beta})$ in der Stelle $\hat{\boldsymbol{\beta}}_{OLS}$ ist, dass die partiellen Ableitungen erster Ordnung der Zielfunktion in der Stelle $\hat{\boldsymbol{\beta}}_{OLS}$ verschwinden. Die partiellen Ableitungen von

$$Q(\boldsymbol{\beta}) = \sum_{\nu=1}^{n} \left(y_\nu - \beta_0 - \beta_1 x_{\nu 1} - \ldots - \beta_p x_{\nu p} \right)^2$$

lauten

$$\begin{aligned}
\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_0} &= \sum_{\nu=1}^{n} 2\left(y_\nu - \beta_0 - \beta_1 x_{\nu 1} - \ldots - \beta_p x_{\nu p} \right) \cdot (-1) &&= -2 \sum_{\nu=1}^{n} \left(y_\nu - \beta_0 - \beta_1 x_{\nu 1} - \ldots - \beta_p x_{\nu p} \right) \\
\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_1} &= \sum_{\nu=1}^{n} 2\left(y_\nu - \beta_0 - \beta_1 x_{\nu 1} - \ldots - \beta_p x_{\nu p} \right) \cdot (-x_{\nu 1}) &&= -2 \sum_{\nu=1}^{n} \left(y_\nu x_{\nu 1} - \beta_0 x_{\nu 1} - \beta_1 x_{\nu 1} x_{\nu 1} - \ldots - \beta_p x_{\nu 1} x_{\nu p} \right) \\
&\vdots && \vdots \\
\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_p} &= \sum_{\nu=1}^{n} 2\left(y_\nu - \beta_0 - \beta_1 x_{\nu 1} - \ldots - \beta_p x_{\nu p} \right) \cdot (-x_{\nu p}) &&= -2 \sum_{\nu=1}^{n} \left(y_\nu x_{\nu p} - \beta_0 x_{\nu p} - \beta_1 x_{\nu 1} x_{\nu p} - \ldots - \beta_p x_{\nu p} x_{\nu p} \right).
\end{aligned}$$

Wir vernachlässigen den Faktor 2 und setzen die Ableitungen gleich Null

$$\begin{aligned}
\sum_{\nu=1}^{n}\left(-y_\nu+\hat\beta_0+\hat\beta_1 x_{\nu 1}+\ldots+\hat\beta_p x_{\nu p}\right) &\overset{!}{=} 0\\
\sum_{\nu=1}^{n}\left(-y_\nu x_{\nu 1}+\hat\beta_0 x_{\nu 1}+\hat\beta_1 x_{\nu 1}x_{\nu 1}+\ldots+\hat\beta_p x_{\nu 1}x_{\nu p}\right) &\overset{!}{=} 0\\
&\vdots\\
\sum_{\nu=1}^{n}\left(-y_\nu x_{\nu p}+\hat\beta_0 x_{\nu p}+\hat\beta_1 x_{\nu 1}x_{\nu p}+\ldots+\hat\beta_p x_{\nu p}x_{\nu p}\right) &\overset{!}{=} 0\ .
\end{aligned}$$

Umformen liefert die sog. *Normalgleichungen*, ein inhomogenes lineares Gleichungssystem:

(6)
$$\begin{aligned}
\hat\beta_0\sum_{\nu=1}^{n}1 \;+\; \hat\beta_1\sum_{\nu=1}^{n}x_{\nu 1} \;+\ldots+\; \hat\beta_p\sum_{\nu=1}^{n}x_{\nu p} &= \sum_{\nu=1}^{n}y_\nu\\
\hat\beta_0\sum_{\nu=1}^{n}x_{\nu 1} \;+\; \hat\beta_1\sum_{\nu=1}^{n}x_{\nu 1}x_{\nu 1} \;+\ldots+\; \hat\beta_p\sum_{\nu=1}^{n}x_{\nu 1}x_{\nu p} &= \sum_{\nu=1}^{n}y_\nu x_{\nu 1}\\
\vdots\qquad\qquad\qquad\vdots\qquad\qquad\qquad\vdots\qquad\qquad &\\
\hat\beta_0\sum_{\nu=1}^{n}x_{\nu p} \;+\; \hat\beta_1\sum_{\nu=1}^{n}x_{\nu 1}x_{\nu p} \;+\ldots+\; \hat\beta_p\sum_{\nu=1}^{n}x_{\nu p}x_{\nu p} &= \sum_{\nu=1}^{n}y_\nu x_{\nu p}\ .
\end{aligned}$$

Die Normalgleichungen

$$
\begin{array}{lcl}
\hat{\beta}_0 \sum_{\nu=1}^{n} 1 + \hat{\beta}_1 \sum_{\nu=1}^{n} x_{\nu 1} + \ldots + \hat{\beta}_p \sum_{\nu=1}^{n} x_{\nu p} = \sum_{\nu=1}^{n} y_\nu \\
\hat{\beta}_0 \sum_{\nu=1}^{n} x_{\nu 1} + \hat{\beta}_1 \sum_{\nu=1}^{n} x_{\nu 1} x_{\nu 1} + \ldots + \hat{\beta}_p \sum_{\nu=1}^{n} x_{\nu 1} x_{\nu p} = \sum_{\nu=1}^{n} y_\nu x_{\nu 1} \\
\vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
\hat{\beta}_0 \sum_{\nu=1}^{n} x_{\nu p} + \hat{\beta}_1 \sum_{\nu=1}^{n} x_{\nu 1} x_{\nu p} + \ldots + \hat{\beta}_p \sum_{\nu=1}^{n} x_{\nu p} x_{\nu p} = \sum_{\nu=1}^{n} y_\nu x_{\nu p}
\end{array}
\tag{6}
$$

können wir kompakt als eine Matrixgleichung notieren:

$$
\begin{pmatrix}
\sum_{\nu=1}^{n} 1 & \sum_{\nu=1}^{n} x_{\nu 1} & \cdots & \sum_{\nu=1}^{n} x_{\nu p} \\
\sum_{\nu=1}^{n} x_{\nu 1} & \sum_{\nu=1}^{n} x_{\nu 1} x_{\nu 1} & \cdots & \sum_{\nu=1}^{n} x_{\nu 1} x_{\nu p} \\
\vdots & \vdots & \ddots & \vdots \\
\sum_{\nu=1}^{n} x_{\nu p} & \sum_{\nu=1}^{n} x_{\nu 1} x_{\nu p} & \cdots & \sum_{\nu=1}^{n} x_{\nu p} x_{\nu p}
\end{pmatrix}
\cdot
\begin{pmatrix}
\hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p
\end{pmatrix}
=
\begin{pmatrix}
\sum_{\nu=1}^{n} y_\nu \\ \sum_{\nu=1}^{n} y_\nu x_{\nu 1} \\ \vdots \\ \sum_{\nu=1}^{n} y_\nu x_{\nu p}
\end{pmatrix}.
$$

Mit

$$\begin{pmatrix} \sum_{\nu=1}^{n} 1 & \sum_{\nu=1}^{n} x_{\nu 1} & \cdots & \sum_{\nu=1}^{n} x_{\nu p} \\ \sum_{\nu=1}^{n} x_{\nu 1} & \sum_{\nu=1}^{n} x_{\nu 1} x_{\nu 1} & \cdots & \sum_{\nu=1}^{n} x_{\nu 1} x_{\nu p} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{\nu=1}^{n} x_{\nu p} & \sum_{\nu=1}^{n} x_{\nu 1} x_{\nu p} & \cdots & \sum_{\nu=1}^{n} x_{\nu p} x_{\nu p} \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{pmatrix} \cdot \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} = \boldsymbol{X'X}$$

$$\begin{pmatrix} \sum_{\nu=1}^{n} y_{\nu} \\ \sum_{\nu=1}^{n} y_{\nu} x_{\nu 1} \\ \vdots \\ \sum_{\nu=1}^{n} y_{\nu} x_{\nu p} \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \boldsymbol{X'y} \qquad \text{und} \qquad \hat{\boldsymbol{\beta}}_{OLS} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}$$

folgt

(7) $$\boldsymbol{X'X\hat{\beta}}_{OLS} = \boldsymbol{X'y} \ .$$

Das Lösen des Gleichungssystems liefert den *OLS-Schätzer* für $\boldsymbol{\beta}$:

$$(\boldsymbol{X'X})^{-1}\boldsymbol{X'X\hat{\beta}}_{OLS} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y} \quad \Rightarrow \quad \boldsymbol{I}\cdot\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y} \quad \Rightarrow$$

(8) $$\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y} \ .$$

Mittels *Vektordifferentiation* (siehe Anhänge II und III) lässt sich der OLS-Schätzer kompakter herleiten. Der Vektor der partiellen Ableitungen erster Ordnung der Quadratsummenfunktion

$$Q(\boldsymbol{\beta}) = \boldsymbol{y'y} - 2\boldsymbol{y'X\beta} + \boldsymbol{\beta'X'X\beta}$$

ist

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{0} - 2\boldsymbol{X'y} + 2\boldsymbol{X'X\beta} \ .$$

Setzen wir die Ableitungen gleich Null

$$-2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} \stackrel{!}{=} \mathbf{0} \tag{6}$$

folgt die Matrix-Normalgleichung

$$\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{X}'\boldsymbol{y} \tag{7}$$

mit der Lösung

$$\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \quad . \tag{8}$$

Zweifaches partielles Differenzieren von $Q(\boldsymbol{\beta})$ liefert

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = 2\boldsymbol{X}'\boldsymbol{X} \, .$$

Hinreichende Bedingung für ein Minimum der Funktion $Q(\boldsymbol{\beta})$ in der Stelle $\hat{\boldsymbol{\beta}}_{OLS}$ ist, dass die *Kreuzproduktmatrix* $\boldsymbol{X}'\boldsymbol{X}$ positiv definit ist, d.h. es muss gelten

$$\boldsymbol{c}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{c} > 0 \qquad \text{für alle Vektoren } \boldsymbol{c} \in IR^{p+1} \text{ mit } \boldsymbol{c} \neq \mathbf{0} \, .$$

Wir überprüfen die Aussage. Mit $\hat{\boldsymbol{\beta}} \equiv \hat{\boldsymbol{\beta}}_{OLS}$ gilt

$$\begin{aligned} Q(\boldsymbol{\beta}) &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{X}[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}])'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{X}[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}]) \\ &= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + 2(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \end{aligned}$$

wegen

$$2(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = 2(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{X}'\hat{\boldsymbol{\varepsilon}} = 0 \qquad \text{(siehe Satz 2.3).}$$

Ist nun

$$\boldsymbol{c}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{c} > 0 \qquad \text{für alle } \boldsymbol{c} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \in IR^{p+1} \text{ mit } \boldsymbol{c} \neq \boldsymbol{0}\ ,$$

folgt stets

$$Q(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) > (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = Q\left(\hat{\boldsymbol{\beta}}\right) \quad \text{für alle } \boldsymbol{\beta} \neq \hat{\boldsymbol{\beta}}\,.$$

S–2.1 Eigenschaften der Kreuzproduktmatrix

(*a*) $\boldsymbol{X'X}$ ist stets nicht nichtnegativ definit, also $\boldsymbol{c'X'Xc} \geq 0$ für alle $\boldsymbol{c} \in IR^{p+1}$ mit $\boldsymbol{c} \neq \boldsymbol{0}$.

(*b*) $rg(\boldsymbol{X}) = p+1 \Rightarrow rg(\boldsymbol{X'X}) = p+1$

$\Rightarrow \boldsymbol{X'X}$ regulär

$\Rightarrow (\boldsymbol{X'X})^{-1}$ existiert

$\Rightarrow \boldsymbol{X'X}$ positiv definit, also $\boldsymbol{c'X'Xc} > 0$ für alle $\boldsymbol{c} \in IR^{p+1}$ mit $\boldsymbol{c} \neq \boldsymbol{0}$.

(ohne Beweis)

Folgerung

Die in A2 formulierte Annahme $rg(\boldsymbol{X}) = p+1$ des klassischen linearen Modells stellt sicher, dass

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

existiert und der gesuchte OLS-Schätzer $\hat{\boldsymbol{\beta}}_{OLS}$ für $\boldsymbol{\beta}$ ist.

Die Funktionswerte der geschätzten Regressionsfunktion

(9) $$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} \qquad \text{mit} \qquad \hat{\boldsymbol{y}}' = (\hat{y}_1,...,\hat{y}_n)$$

können wir als geschätzte Werte (*fitted values*) des Regressanden interpretieren. Die Differenz

(10) $$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} \qquad \text{mit} \qquad \hat{\boldsymbol{\varepsilon}}' = (\hat{\varepsilon}_1,...,\hat{\varepsilon}_n)$$

liefert den sog. *Residuenvektor* des geschätzten Modells. Das *Residuum* $\hat{\varepsilon}_\nu$ ist ein Schätzwert der unbeobachtbaren Störung ε_ν. Es gilt

(11) $$Q(\hat{\boldsymbol{\beta}}_{OLS}) = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \sum_{\nu=1}^{n} \hat{\varepsilon}_\nu^2 \quad .$$

Die Summe nennen wir im Folgenden *Residualquadratsumme*.

B–2.3 *Lineare Einfachregression.*

Beispielhaft betrachten wir den Spezialfall der Einfachregression

$$y_\nu = \beta_0 + \beta_1 x_\nu + \varepsilon_\nu \qquad (\nu = 1,\ldots,n)$$

bzw.

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \qquad \text{mit} \qquad \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \text{und} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Hier gilt

$$\boldsymbol{X'X} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \cdot \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n & \sum_{\nu=1}^{n} x_\nu \\ \sum_{\nu=1}^{n} x_\nu & \sum_{\nu=1}^{n} x_\nu^2 \end{pmatrix}$$

und

$$\boldsymbol{X'y} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_{\nu=1}^{n} y_\nu \\ \sum_{\nu=1}^{n} x_\nu y_\nu \end{pmatrix}.$$

Die Normalgleichung $\boldsymbol{X'X\hat{\beta}}_{OLS} = \boldsymbol{X'y}$ besitzt die Form

$$\begin{pmatrix} n & \sum_{\nu=1}^{n} x_\nu \\ \sum_{\nu=1}^{n} x_\nu & \sum_{\nu=1}^{n} x_\nu^2 \end{pmatrix} \cdot \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \sum_{\nu=1}^{n} y_\nu \\ \sum_{\nu=1}^{n} x_\nu y_\nu \end{pmatrix} \quad \text{bzw.} \quad \begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{\nu=1}^{n} x_\nu &= \sum_{\nu=1}^{n} y_\nu \\ \hat{\beta}_0 \sum_{\nu=1}^{n} x_\nu + \hat{\beta}_1 \sum_{\nu=1}^{n} x_\nu^2 &= \sum_{\nu=1}^{n} y_\nu x_\nu \end{aligned}$$

Dividieren wir die linke und die rechte Seite der ersten skalaren Gleichung durch n, so erhalten wir

$$\hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y} \quad \Rightarrow \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \text{mit} \qquad \bar{x} = \frac{1}{n}\sum_{\nu=1}^{n} x_\nu \ \text{ und } \ \bar{y} = \frac{1}{n}\sum_{\nu=1}^{n} y_\nu \,.$$

Setzen wir dies in die zweite Gleichung, so folgt nach einigen Umformungen

$$\left(\bar{y} - \hat{\beta}_1 \bar{x}\right)\sum_{\nu=1}^{n} x_\nu + \hat{\beta}_1 \sum_{\nu=1}^{n} x_\nu^2 = \sum_{\nu=1}^{n} y_\nu x_\nu \quad \Rightarrow \quad \hat{\beta}_1 = \frac{\sum_{\nu=1}^{n} x_\nu y_\nu - n\bar{x}\bar{y}}{\sum_{\nu=1}^{n} x_\nu^2 - n\bar{x}^2} = \frac{\sum_{\nu=1}^{n} (x_\nu - \bar{x})(y_\nu - \bar{y})}{\sum_{\nu=1}^{n} (x_\nu - \bar{x})^2}$$

B–2.1 *Makroökonomische Konsumfunktion.*

Wir wollen nun eine Keynesianische Konsumfunktion

$$konsum_\nu = \beta_0 + \beta_1 \cdot einkommen_\nu + \varepsilon_\nu \quad (\nu = 1,\ldots,n)$$

für die BRD (alte Bundesländer) im Zeitraum 1974 bis 1992 schätzen. Als Daten liegen jährliche Werte des realen privaten Verbrauchs und des realen verfügbaren Einkommens der privaten Haushalte in Preisen von 1985 zugrunde (siehe Tabelle 2.1, Quelle: Eckey et al. 2004, S. 31). Wir bezeichnen den Konsum und das Einkommen abkürzend mit Y bzw. X.

Die OLS-Schätzung für die Koeffizienten der Konsumfunktion ermitteln wir zunächst über Gleichung (8). Es ist:

$$\boldsymbol{y} = \begin{pmatrix} 837.60 \\ 863.82 \\ 897.32 \\ \vdots \\ 1274.63 \\ 1287.11 \end{pmatrix}, \qquad \boldsymbol{X} = \begin{pmatrix} 1 & 962.72 \\ 1 & 989.56 \\ 1 & 1017.68 \\ \vdots & \vdots \\ 1 & 1417.17 \\ 1 & 1421.38 \end{pmatrix} \qquad \text{und} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}.$$

Jahr	Einkommen x_ν in Mrd. DM	Konsum y_ν in Mrd. DM	x_ν^2	$x_\nu \cdot y_\nu$
1974	962.72	837.60	926829.8	806374.3
1975	989.56	863.82	979229.0	854801.7
1976	1017.68	897.32	1035672.6	913184.6
1977	1046.38	937.80	1094911.1	981295.2
1978	1093.39	971.48	1195501.7	1062206.5
1979	1130.98	1003.06	1279115.8	1134440.8
1980	1118.86	1015.57	1251847.7	1136280.7
1981	1096.80	1007.92	1202970.2	1105486.7
1982	1078.10	992.55	1162299.6	1070068.2
1983	1086.06	1005.92	1179526.3	1092489.5
1984	1100.16	1021.68	1210352.0	1124011.5
1985	1119.93	1036.53	1254243.2	1160841.0
1986	1205.44	1072.01	1453085.6	1292243.7
1987	1239.32	1106.88	1535914.1	1371778.5
1988	1299.72	1137.00	1689272.1	1477781.6
1989	1323.60	1167.37	1751917.0	1545130.9
1990	1420.53	1230.68	2017905.5	1748217.9
1991	1417.17	1274.63	2008370.8	1806367.4
1992	1421.38	1287.11	2020321.1	1829472.4
Summe	22167.78	19866.93	26249285.12	23512473.03

Tab. 2.1: Realer privater Verbrauch und reales verfügbare Einkommen in Preisen von 1985

Hiermit folgt

$$\boldsymbol{X'X} = \begin{pmatrix} n & \sum_{\nu=1}^{n} x_\nu \\ \sum_{\nu=1}^{n} x_\nu & \sum_{\nu=1}^{n} x_\nu^2 \end{pmatrix} = \begin{pmatrix} 19 & 22167.78 \\ 22167.78 & 26249285.12 \end{pmatrix}$$

und

$$\boldsymbol{X'y} = \begin{pmatrix} \sum_{\nu=1}^{n} y_\nu \\ \sum_{\nu=1}^{n} x_\nu y_\nu \end{pmatrix} = \begin{pmatrix} 19866.93 \\ 23512473.03 \end{pmatrix}.$$

Die Inverse der Kreuzproduktmatrix berechnen wir mit Hilfe ihrer Adjungierten (Anhang IV):

$$(\boldsymbol{X'X})^{-1} = \frac{1}{det(\boldsymbol{X'X})} adj(\boldsymbol{X'X}) = \frac{1}{19 \cdot 26249285.12 - 22167.78^2} \begin{pmatrix} 26249285.12 & -22167.78 \\ -22167.78 & 19 \end{pmatrix}$$

$$= \frac{1}{7325947} \begin{pmatrix} 26249285.12 & -22167.78 \\ -22167.78 & 19 \end{pmatrix} = \begin{pmatrix} 3.58306 & -0.00303 \\ -0.00303 & 2.59352\text{E}-06 \end{pmatrix}.$$

Einsetzen in (8) liefert

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \begin{pmatrix} 3.58306 & -0.00303 \\ -0.00303 & 2.59352\mathrm{E}-06 \end{pmatrix} \cdot \begin{pmatrix} 19866.93 \\ 23512473.03 \end{pmatrix} = \begin{pmatrix} 37.31676 \\ 0.86422 \end{pmatrix} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}.$$

Alternativ können wir auch die in Beispiel 2.3 hergeleiteten Formeln nutzen:

$$\hat{\beta}_1 = \frac{\sum_{\nu=1}^{n} x_\nu y_\nu - n\bar{x}\bar{y}}{\sum_{\nu=1}^{n} x_\nu^2 - n\bar{x}^2} = \frac{23512473.03 - \dfrac{22167.78 \cdot 19866.93}{19}}{26249285.12 - 19 \cdot \left(\dfrac{22167.78}{19}\right)^2} = \frac{333223.9}{385576.2} = 0.86422\ ,$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{19866.93}{19} - 0.86422 \cdot \frac{22167.78}{19} = 37.31676\ .$$

Der Schätzwert 0.864 für die marginale Konsumquote besagt, dass jede zusätzliche Milliarde DM Einkommen im Untersuchungszeitraum den Konsum um durchschnittlich 864 Millionen DM erhöhte. Der Schätzwert 37.317 für den autonomen Konsum ist mit Vorsicht zu interpretieren, da der Wert $x = 0$ weit außerhalb des Beobachtungsbereichs des Regressors X liegt.

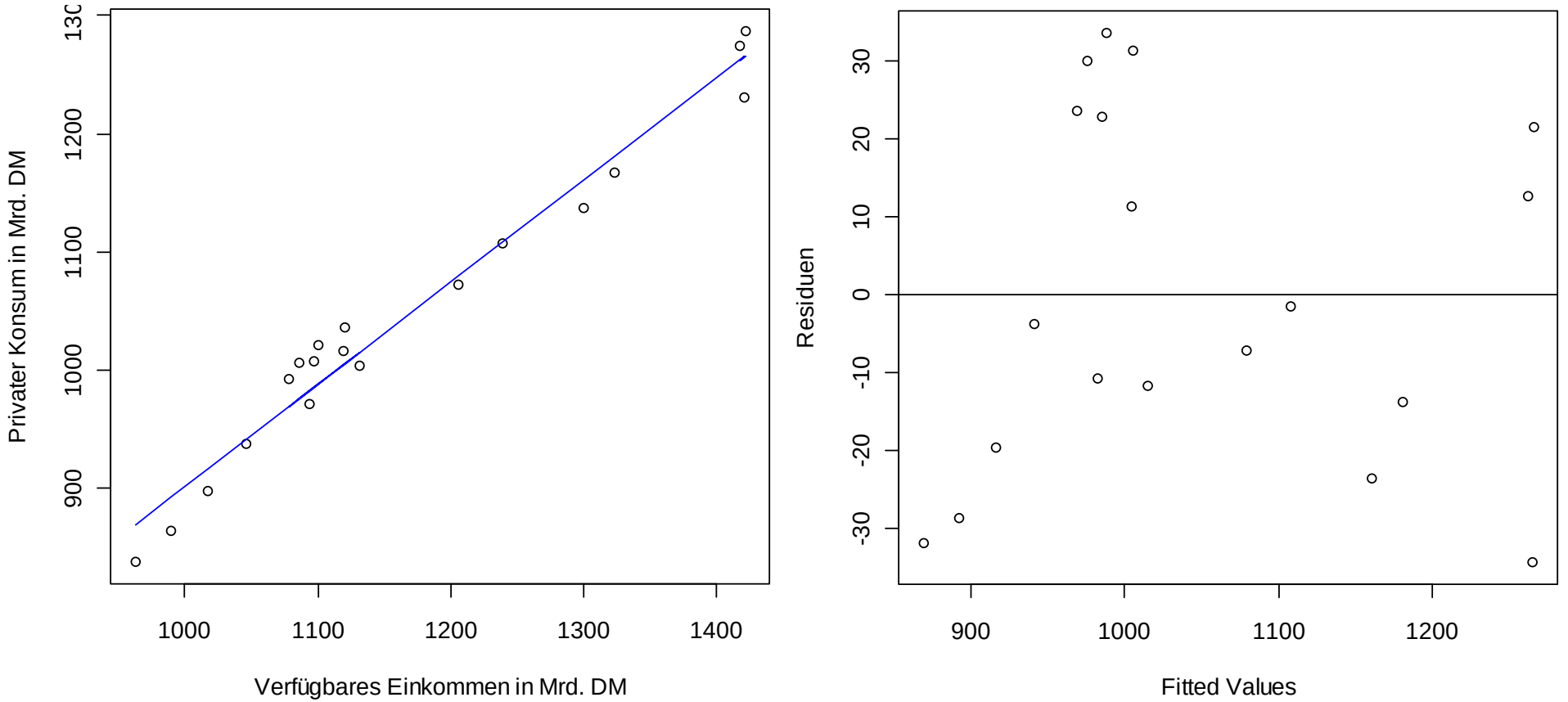


Abb. 2.2: Streudiagramm der Daten mit geschätzter Regressionsfunktion $\hat{y} = 37.317 + 0.864 \cdot x$ und Residuendiagramm ☒

2.3 Geschätzte Werte, Residuen, Streuungszerlegung

Wir kennzeichnen den OLS-Schätzer $\hat{\boldsymbol{\beta}}_{OLS}$ vereinfachend durch $\hat{\boldsymbol{\beta}}$. Setzen wir

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

in

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

ein, dann können wir die *fitted values* $\hat{\boldsymbol{y}}$ auch

(12) $$\hat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y} \quad \text{mit}$$

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

schreiben. Die (n,n)-Matrix $\boldsymbol{H}$ heißt *Prädiktionsmatrix* oder *„hat-matrix“* („Dach-Matrix“). Für die Residuen des geschätzten Modells folgt:

(13) $$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{H}\boldsymbol{y} = \boldsymbol{I}\boldsymbol{y} - \boldsymbol{H}\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} \ .$$

Hierbei ist $\boldsymbol{I}$ wieder die (n,n)-Einheitsmatrix.

S–2.2 Eigenschaften der Hat-Matrix

(*a*) $\boldsymbol{H}$ ist *symmetrisch*, d.h. $\boldsymbol{H}' = \boldsymbol{H}$;

(*b*) $\boldsymbol{H}$ ist *idempotent*, d.h. $\boldsymbol{H}^2 = \boldsymbol{H}$;

(*c*) $rg(\boldsymbol{H}) = \sum_{\nu=1}^{n} h_{\nu\nu} = spur(\boldsymbol{H}) = p+1$;

(*d*) $\frac{1}{n} \le h_{\nu\nu} \le 1 \quad (\nu = 1,\ldots,n)$;

(*e*) $(\boldsymbol{I}-\boldsymbol{H})' = \boldsymbol{I}-\boldsymbol{H}$, $(\boldsymbol{I}-\boldsymbol{H})^2 = \boldsymbol{I}-\boldsymbol{H}$, $rg(\boldsymbol{I}-\boldsymbol{H}) = rg(\boldsymbol{I}) - rg(\boldsymbol{H}) = n-(p+1) = n-p-1$.

Beweis

Wir beweisen lediglich (*a*) und (*b*):

(*a*) $\boldsymbol{H}' = [\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']' = \boldsymbol{X}''[(\boldsymbol{X}'\boldsymbol{X})^{-1}]'\boldsymbol{X}' = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \boldsymbol{H}$

(*b*) $\boldsymbol{H}^2 = \boldsymbol{H}\boldsymbol{H} = \boldsymbol{X}\underbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}}_{=\boldsymbol{I}}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \boldsymbol{H}$ *q.e.d.*

S–2.3 Geometrische Eigenschaften der OLS-Schätzung

Sei $\boldsymbol{x}^j$ die j-te Spalte von $\boldsymbol{X}$, d.h. $\boldsymbol{X} = (\boldsymbol{x}^1 \mid \boldsymbol{x}^2 \mid \ldots \mid \boldsymbol{x}^{p+1})$.

(*a*) $\hat{\boldsymbol{y}}'\hat{\boldsymbol{\varepsilon}} = 0$, d.h. $\hat{\boldsymbol{y}}$ und $\hat{\boldsymbol{\varepsilon}}$ sind *orthogonal*;

(*b*) $(\boldsymbol{x}^j)'\hat{\boldsymbol{\varepsilon}} = 0$, d.h. $\boldsymbol{x}^j$ und $\hat{\boldsymbol{\varepsilon}}$ sind *orthogonal*.

Beweis

(*a*) $$\hat{\boldsymbol{y}}'\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y}'\boldsymbol{H}'(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y} = \boldsymbol{y}'\boldsymbol{H}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{H}'\boldsymbol{H}\boldsymbol{y} = \boldsymbol{y}'\boldsymbol{H}\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{H}\boldsymbol{H}\boldsymbol{y} = \boldsymbol{y}'\boldsymbol{H}\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{H}\boldsymbol{y} = 0$$

(*b*) $$\boldsymbol{X}'\hat{\boldsymbol{\varepsilon}} = \boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y} = \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{H}\boldsymbol{y} = \boldsymbol{X}'\boldsymbol{y} - \underbrace{\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}}_{=\boldsymbol{I}}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{y} = \boldsymbol{0}$$

$\Rightarrow$ $(\boldsymbol{x}^j)'\hat{\boldsymbol{\varepsilon}} = 0$ für alle Spalten $j = 1,\ldots,p+1$ von $\boldsymbol{X}$ *q.e.d.*

Folgerungen aus S–2.3 (Geometrische Eigenschaften der OLS-Schätzung)

(*i*) $(\boldsymbol{x}^1)'\hat{\boldsymbol{\varepsilon}} = \mathbf{1}'\hat{\boldsymbol{\varepsilon}} = \sum_{\nu=1}^{n}\hat{\varepsilon}_\nu = 0\ , \quad \bar{\hat{\varepsilon}} = \frac{1}{n}\sum_{\nu=1}^{n}\hat{\varepsilon}_\nu = 0$ mit dem n-Einsvektor $\mathbf{1} = (1,1,\ldots,1)'$

Die Residuen sind im Mittel Null.

(*ii*) $\bar{y} = \frac{1}{n}\sum_{\nu=1}^{n} y_\nu = \frac{1}{n}\sum_{\nu=1}^{n}(\hat{y}_\nu + \hat{\varepsilon}_\nu) = \frac{1}{n}\sum_{\nu=1}^{n}\hat{y}_\nu + \frac{1}{n}\sum_{\nu=1}^{n}\hat{\varepsilon}_\nu = \frac{1}{n}\sum_{\nu=1}^{n}\hat{y}_\nu = \bar{\hat{y}}$

Der Mittelwert der geschätzten Werte ist gleich dem Mittelwert der beobachteten Werte.

(*iii*) $\bar{y} = \frac{1}{n}\sum_{\nu=1}^{n}\hat{y}_\nu = \frac{1}{n}\sum_{\nu=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_{\nu 1} + \ldots + \hat{\beta}_p x_{\nu p}) = \hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \ldots + \hat{\beta}_p\bar{x}_p$

Die Regressionsfunktion (Hyperebene) geht durch den Schwerpunkt $(\bar{y}, \bar{x}_1, \ldots, \bar{x}_p)$ der Daten.

(*iv*) $s_{\hat{Y},\hat{\varepsilon}} = \frac{1}{n}\hat{\boldsymbol{y}}'\hat{\boldsymbol{\varepsilon}} = 0, \quad r_{\hat{Y},\hat{\varepsilon}} = 0\ , \qquad s_{X_j,\hat{\varepsilon}} = \frac{1}{n}(\boldsymbol{x}^j)'\hat{\boldsymbol{\varepsilon}} = 0, \quad r_{X_j,\hat{\varepsilon}} = 0$

Die Residuen korrelieren weder mit den geschätzten Werten noch mit den Regressorwerten.

S–2.4 Streuungszerlegung

Es gilt

(14a) $$\sum_{\nu=1}^{n}(y_\nu-\bar{y})^2=\sum_{\nu=1}^{n}(\hat{y}_\nu-\bar{y})^2+\sum_{\nu=1}^{n}(y_\nu-\hat{y}_\nu)^2=\sum_{\nu=1}^{n}(\hat{y}_\nu-\bar{y})^2+\sum_{\nu=1}^{n}\hat{\varepsilon}_\nu^2\,.$$

Symbolisch schreiben wir hierfür

(14b) $$TSS=ESS+RSS$$

mit

$TSS=\sum_{\nu=1}^{n}(y_\nu-\bar{y})^2$ (*Total sum of squares*, *Streuung des Regressanden*)

$ESS=\sum_{\nu=1}^{n}(\hat{y}_\nu-\bar{y})^2$ (*Explained sum of squares*, *erklärte Streuung*)

$RSS=\sum_{\nu=1}^{n}\hat{\varepsilon}_\nu^2$ (*Residual sum of squares*, *nicht erklärte Streuung*)

Beweis

$$\sum_{\nu=1}^{n}(y_\nu-\bar{y})^2=\sum_{\nu=1}^{n}(y_\nu+[-\hat{y}_\nu+\hat{y}_\nu]-\bar{y})^2=\sum_{\nu=1}^{n}([y_\nu-\hat{y}_\nu]+[\hat{y}_\nu-\bar{y}])^2$$

$$=\sum_{\nu=1}^{n}(y_\nu-\hat{y}_\nu)^2+2\sum_{\nu=1}^{n}(y_\nu-\hat{y}_\nu)(\hat{y}_\nu-\bar{y})+\sum_{\nu=1}^{n}(\hat{y}_\nu-\bar{y})^2$$

$$=\sum_{\nu=1}^{n}(\hat{y}_\nu-\bar{y})^2+\sum_{\nu=1}^{n}(y_\nu-\hat{y}_\nu)^2\,,$$

da wegen Satz 2.3 gilt

$$\sum_{\nu=1}^{n}(y_\nu-\hat{y}_\nu)(\hat{y}_\nu-\bar{y})=\sum_{\nu=1}^{n}\hat{\varepsilon}_\nu(\hat{y}_\nu-\bar{y})=\sum_{\nu=1}^{n}\hat{\varepsilon}_\nu\hat{y}_\nu-\bar{y}\sum_{\nu=1}^{n}\hat{\varepsilon}_\nu=0$$ *q.e.d.*

Anwendung der Streuungszerlegung

Zur Beurteilung der Anpassungsgüte eines Regressionsmodells an die Daten dient das *Bestimmtheitsmaß*

(15) $$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \qquad \text{mit} \quad 0 \le R^2 \le 1 \,.$$

Es misst den Anteil der von der Regressionsfunktion erklärten Streuung an der Gesamtstreuung des Regressanden.

Für den *Vergleich der Güte von alternativen Regressionsmodellen* mit unterschiedlicher Anzahl von Regressoren verwendet man das *adjustierte Bestimmtheitsmaß*

(16) $$\overline{R}^2 = 1 - \frac{\frac{RSS}{n-p-1}}{\frac{TSS}{n-1}} = 1 - \frac{n-1}{n-p-1} \cdot \frac{RSS}{TSS} = 1 - \frac{n-1}{n-p-1} \cdot \left(1 - R^2\right).$$

Liegen alternative Modelle mit identischer Anpassungsgüte vor, dann bevorzugt $\overline{R}^2$ das Modell mit der geringsten Regressorenanzahl (*Prinzip der Sparsamkeit*).

B–2.3 *Lineare Einfachregression.*

Im Spezialfall der linearen Einfachregression

$$y_\nu = \beta_0 + \beta_1 x_\nu + \varepsilon_\nu \qquad (\nu = 1,\ldots,n)$$

stellt die geschätzte Regressionsfunktion $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ geometrisch eine Gerade in der *XY*-Ebene dar, die stets durch den Schwerpunkt $(\bar{x}, \bar{y})$ des bivariaten Datensatzes verläuft. Die Abbildung 2.3 zeigt ein Streudiagramm bivariater Daten y_ν, x_ν $(\nu = 1,\ldots,n)$ und die zugehörige Regressionsgerade.

Setzt man $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ in $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ ein, so folgt

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

und weiter

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

Für $x = \bar{x}$ ist also $\hat{y} = \bar{y}$.

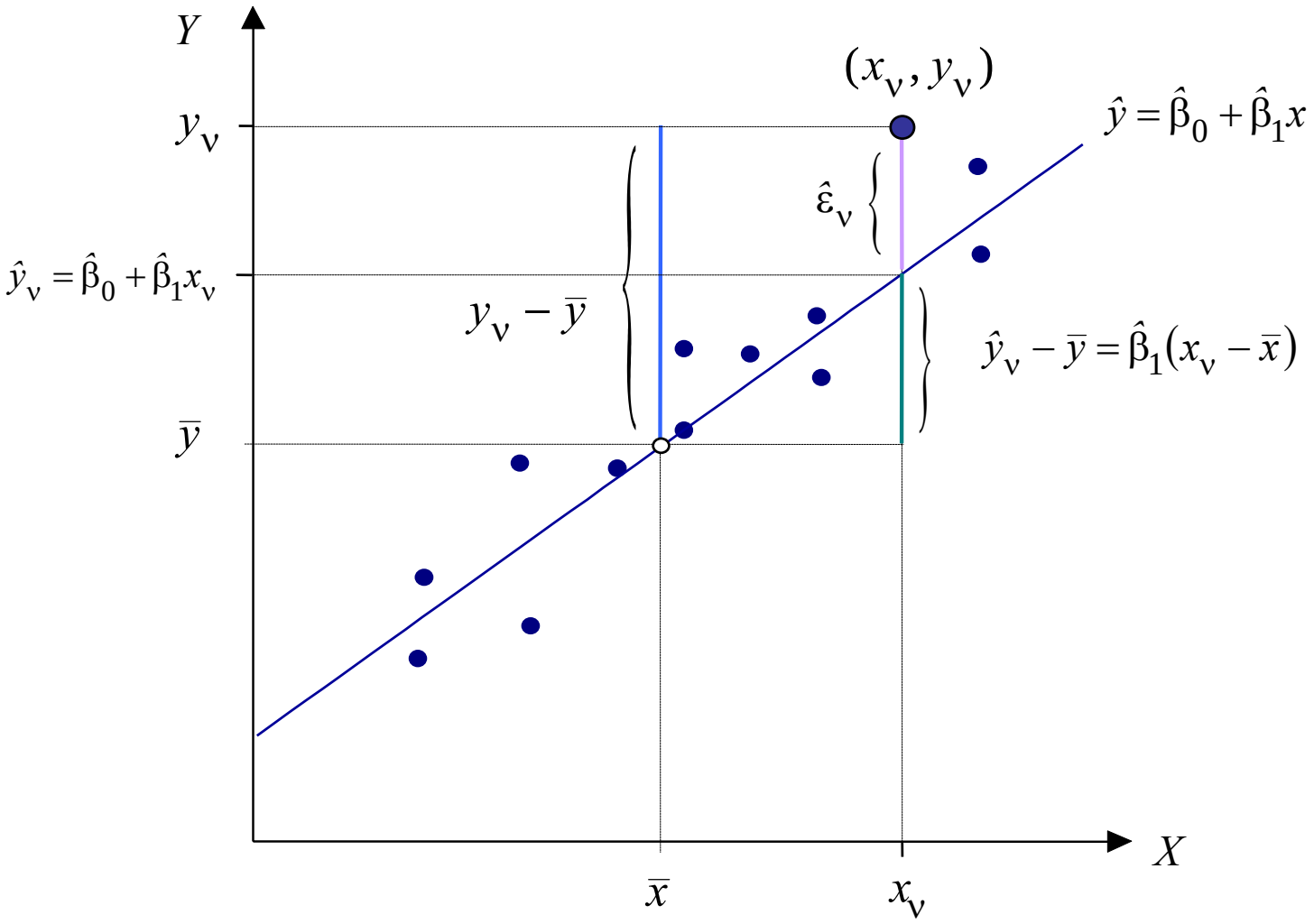


Abb. 2.3: Streudiagramm bivariater Daten, Regressionsgerade und Streuungszerlegung

Zwischen dem Bestimmtheitsmaß der Regressionsgeraden und dem *Bravais-Pearson-Korrelationskoeffizienten* r_{XY} (siehe Anhang I)) besteht ein unmittelbarer Zusammenhang:

$$R^2 = \frac{\sum_{\nu=1}^{n}(\hat{y}_\nu - \bar{y})^2}{\sum_{\nu=1}^{n}(y_\nu - \bar{y})^2} = \frac{\sum_{\nu=1}^{n}\left(\hat{\beta}_1(x_\nu - \bar{x})\right)^2}{\sum_{\nu=1}^{n}(y_\nu - \bar{y})^2} = \hat{\beta}_1^2 \cdot \frac{\sum_{\nu=1}^{n}(x_\nu - \bar{x})^2}{\sum_{\nu=1}^{n}(y_\nu - \bar{y})^2}$$

$$= \left(\frac{\sum_{\nu=1}^{n}(x_\nu - \bar{x})(y_\nu - \bar{y})}{\sum_{\nu=1}^{n}(x_\nu - \bar{x})^2}\right)^2 \cdot \left(\frac{\sqrt{\sum_{\nu=1}^{n}(x_\nu - \bar{x})^2}}{\sqrt{\sum_{\nu=1}^{n}(y_\nu - \bar{y})^2}}\right)^2$$

$$= \left(\frac{\sum_{\nu=1}^{n}(x_\nu - \bar{x})(y_\nu - \bar{y})}{\sqrt{\sum_{\nu=1}^{n}(x_\nu - \bar{x})^2} \cdot \sqrt{\sum_{\nu=1}^{n}(y_\nu - \bar{y})^2}}\right)^2 = r_{XY}^2 \quad .$$

Der Wertebereich der Kenngröße R^2 ist das Intervall [0, 1]. Gilt $R^2 = 1$, liegt eine perfekte Anpassung der Regressionsgeraden vor. Die Regressionsgerade verläuft exakt durch alle Punkte des Streudiagramms. Je kleiner R^2 ist, umso schwächer werden die Werte von *Y durch* die Werte von *X bestimmt*. Sind X und Y unkorreliert, dann ist $R^2 = 0$. Es gilt jetzt $\hat{\beta}_1 = 0$ und die Regressionsgerade $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = \bar{y}$ liefert nicht mehr Information bzgl. Y als das arithmetische Mittel $\bar{y}$ ☒

B–2.1 *Makroökonomische Konsumfunktion.*

Der reale private Verbrauch und das reale verfügbare Einkommen der privaten Haushalte in der Bundesrepublik in den Jahren 1974 bis 1992 sind sehr stark positiv korreliert. Der Bravais-Pearson-Korrelationskoeffizient beträgt

$$r_{XY} = \frac{\sum_{\nu=1}^{n}(x_\nu - \bar{x})(y_\nu - \bar{y})}{\sqrt{\sum_{\nu=1}^{n}(x_\nu - \bar{x})^2} \cdot \sqrt{\sum_{\nu=1}^{n}(y_\nu - \bar{y})^2}} = 0.98428 \ .$$

Folglich ist der von der geschätzten Konsumfunktion

$$\hat{y} = 37.317 + 0.864 \cdot x \qquad \text{bzw.} \qquad \hat{konsum} = 37.317 + 0.864 \cdot einkommen$$

erklärte Anteil der Streuung des privaten Konsums mit 96.88% sehr hoch:

$$R^2 = \frac{ESS}{TSS} = \frac{287979.9}{297255.0} = 1 - \frac{RSS}{TSS} = 1 - \frac{9275.191}{297255.0} = r_{XY}^2 = 0.98428^2 = 0.96880 \ .$$

Die Datenpunkte weichen im Streudiagramm nur geringfügig von der Regressionsgeraden ab (siehe Abbildung 2.2) ☒

2.4 Statistische Eigenschaften des OLS-Schätzers

Wir untersuchen nun die statistischen Eigenschaften des *OLS-Schätzers* $\hat{\boldsymbol{\beta}}$ und unterstellen dabei die Gültigkeit der Annahmen A1 – A3 aus Definition 2.1.

Den „wahren“ Regressionskoeffizientenvektor $\boldsymbol{\beta}$ kann man auf Stichprobenbasis im Allgemeinen nicht exakt schätzen. Der *Schätzfehler* $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$, mit dem wir rechnen müssen, lässt sich in zwei Summanden zerlegen:

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \hat{\boldsymbol{\beta}} - \left(E(\hat{\boldsymbol{\beta}}) - E(\hat{\boldsymbol{\beta}})\right) - \boldsymbol{\beta}$$

$$= \left(\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\right) + \left(E(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta}\right).$$

Der erste Summand $\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})$ ist eine zentrierte Zufallsgröße und heißt *Zufallsfehler*. Der zweite Summand $E(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta}$ ist eine Konstante und heißt *systematischer Fehler*, *Verzerrung* oder *Bias des Schätzers*.

Den OLS-Schätzer können wir in der folgenden Form schreiben:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}) = \underbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}}_{=\boldsymbol{I}}\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}$$

$$= \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\,.$$

Der Erwartungswert- oder *Mittelwertvektor* von $\hat{\boldsymbol{\beta}}$ ist

(17) $$E(\hat{\boldsymbol{\beta}}) = E\left(\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\right) = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{0} = \boldsymbol{\beta}\,,$$

$\hat{\boldsymbol{\beta}}$ ist ein so genannter *unverzerrter* oder *erwartungstreuer* Schätzer für $\boldsymbol{\beta}$.

Könnten wir OLS-Schätzwerte $\hat{\boldsymbol{\beta}}$ über alle denkbaren Stichproben gleichen Umfangs berechnen und mitteln, dann wäre der Mittelwert der $\hat{\boldsymbol{\beta}}$ mit dem wahren Regressionskoeffizientenvektor $\boldsymbol{\beta}$ identisch. Der OLS-Schätzer weist keine Verzerrung in Form eines systematischen Schätzfehlers (Bias) auf.

Auch wenn OLS-Schätzer frei von systematischen Fehlern $E(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta}$ sind, treten Schätzfehler in Form von Zufallsfehlern $\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})$ auf.

Das Ausmaß des Zufallsfehlers $\hat{\beta}_i - E(\hat{\beta}_i)$ eines einzelnen geschätzten Regressionskoeffizienten misst dessen Varianz

$$Var(\hat{\beta}_i) = E\left(\left[\hat{\beta}_i - E(\hat{\beta}_i) \right]^2 \right).$$

Je kleiner die Varianz ist, umso präziser oder effizienter ist die Schätzung.

Einen erwartungstreuen Schätzer, der im Vergleich zu allen alternativen erwartungstreuen Schätzfunktionen die kleinste Varianz bzw. den kleinsten Standardfehler aufweist, bezeichnet man als den *besten erwartungstreuen* oder den *effizientesten erwartungstreuen Schätzer*.

Die *Varianz-Kovarianz-Matrix* des OLS-Schätzers besitzt die Form

(18)
$$Cov(\hat{\boldsymbol{\beta}}) = E\left(\{\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\}\{\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\}'\right) = E\left(\{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\}\{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\}'\right) \quad \text{wegen} \quad E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

$$= E\left(\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\}\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\}'\right) \quad \text{wegen} \quad \hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}$$

$$= E\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{X}[(\boldsymbol{X}'\boldsymbol{X})^{-1}]'\right)$$

$$= E\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\right) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

$$= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\left(\sigma^2\boldsymbol{I}\right)\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\underbrace{\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}}_{=\boldsymbol{I}}$$

$$= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} = \boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}} \; .$$

Bei der Herleitung wurde ausgenutzt, dass die Kreuzproduktmatrix $\boldsymbol{X}'\boldsymbol{X}$ eine symmetrische Matrix ist, d.h. $(\boldsymbol{X}'\boldsymbol{X})' = \boldsymbol{X}'\boldsymbol{X}$. Damit ist auch ihre Inverse symmetrisch, d.h. $[(\boldsymbol{X}'\boldsymbol{X})^{-1}]' = (\boldsymbol{X}'\boldsymbol{X})^{-1}$.

Auf der Hauptdiagonalen der Matrix stehen die Varianzen der geschätzten Regressionskoeffizienten:

$$Cov(\hat{\boldsymbol{\beta}}) = \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{pmatrix} E(\hat{\beta}_0 - \beta_0)^2 & E(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1) & \cdots & E(\hat{\beta}_0 - \beta_0) \\ E(\hat{\beta}_1 - \beta_1)(\hat{\beta}_0 - \beta_0) & E(\hat{\beta}_1 - \beta_1)^2 & \cdots & E(\hat{\beta}_1 - \beta_1)(\hat{\beta}_p - \beta_p) \\ \vdots & \vdots & \ddots & \vdots \\ E(\hat{\beta}_p - \beta_p)(\hat{\beta}_0 - \beta_0) & E(\hat{\beta}_p - \beta_p)(\hat{\beta}_1 - \beta_1) & \cdots & E(\hat{\beta}_p - \beta_p)^2 \end{pmatrix} = \boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}} .$$

Es kennzeichne a_{jj} das j-te Diagonalelement von $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Hiermit erhalten wir

(19) $$Var(\hat{\beta}_i) = E\left(\left[\hat{\beta}_i - \beta_i\right]^2\right) = \sigma^2 \cdot a_{jj} = \sigma^2_{\hat{\beta}_i} \quad \text{mit} \quad j = i+1 \ \text{ und } \ i = 0,1,\ldots,p .$$

Die Quadratwurzeln der Varianzen

(20) $$\sigma_{\hat{\beta}_i} = \sigma \cdot \sqrt{a_{jj}} \quad \text{mit} \quad j = i+1 \ \text{ und } \ i = 0,1,\ldots,p$$

bezeichnet man als *Standardfehler* der OLS-Schätzungen $\hat{\beta}_i$ für die Regressionskoeffizienten β_i.

Es lässt sich zeigen, dass der OLS-Schätzer der beste erwartungstreue Schätzer innerhalb der Klasse der linearen Schätzfunktionen ist. Diese Aussage ist als *Gauss-Markov-Theorem* bekannt. Der OLS-Schätzer ist der beste (effizienteste) erwartungstreue lineare Schätzer:

Best Linear Unbiased Estimate, kurz *BLUE*.

Der OLS-Schätzer heißt linearer Schätzer, da er eine Linearkombination

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{A}\boldsymbol{y} \qquad \text{mit} \quad \boldsymbol{A} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

des Vektors $\boldsymbol{y}$ ist. Das Gauss-Markov-Theorem vergleicht den OLS-Schätzer mit der umfangreichen Klasse aller denkbaren linearen erwartungstreuen Schätzern für $\boldsymbol{\beta}$, die wir hier gemäß

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{B}\boldsymbol{y}$$

schreiben. $\boldsymbol{B}$ ist eine geeignete, aber ansonsten beliebige $(p+1,n)$-Matrix.

Für den Beweis benötigt man die Varianz-Kovarianz-Matrix von $\widetilde{\boldsymbol{\beta}}$. Da der lineare Schätzer

$$\widetilde{\boldsymbol{\beta}} = \boldsymbol{B}\boldsymbol{y} = \boldsymbol{B}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{B}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{B}\boldsymbol{\varepsilon}$$

erwartungstreu sein soll, muss

$$E(\widetilde{\boldsymbol{\beta}}) = E(\boldsymbol{B}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{B}\boldsymbol{\varepsilon}) = \boldsymbol{B}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{B}E(\boldsymbol{\varepsilon}) = \boldsymbol{B}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

gelten, woraus $\boldsymbol{B}\boldsymbol{X} = \boldsymbol{I}$ und

$$\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + \boldsymbol{B}\boldsymbol{\varepsilon}$$

folgt. Hiermit erhalten wir

$$\begin{aligned} Cov(\widetilde{\boldsymbol{\beta}}) &= E\left(\{\widetilde{\boldsymbol{\beta}} - E(\widetilde{\boldsymbol{\beta}})\}\{\widetilde{\boldsymbol{\beta}} - E(\widetilde{\boldsymbol{\beta}})\}' \right) = E\left(\{\boldsymbol{B}\boldsymbol{\varepsilon}\}\{\boldsymbol{B}\boldsymbol{\varepsilon}\}'\right) \\ &= E\left(\boldsymbol{B}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{B}'\right) = \boldsymbol{B}E\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right)\boldsymbol{B}' = \boldsymbol{B}\left(\sigma^2\boldsymbol{I}\right)\boldsymbol{B}' = \sigma^2\boldsymbol{B}\boldsymbol{B}' = \boldsymbol{\Sigma}_{\widetilde{\boldsymbol{\beta}}} \,. \end{aligned}$$

Der OLS-Schätzer $\hat{\boldsymbol{\beta}}$ wäre der effizienteste lineare erwartungstreue Schätzer, wenn wir beweisen könnten, dass die Differenz der Kovarianz-Matrizen $\boldsymbol{\Sigma}_{\widetilde{\boldsymbol{\beta}}} - \boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}$ eine positiv semidefinite Matrix ist, denn es gilt die Implikation:

$$\forall_{\boldsymbol{c} \in IR^{p+1}} \; \boldsymbol{c}'(\boldsymbol{\Sigma}_{\widetilde{\boldsymbol{\beta}}} - \boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}})\boldsymbol{c} \geq 0 \quad \Rightarrow \quad Var(\widetilde{\beta}_i) - Var(\hat{\beta}_i) \geq 0 \text{ für } i = 0,1,\ldots,p \,.$$

Es ist

$$\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}} - \boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}} = \sigma^2 \boldsymbol{B}\boldsymbol{B}' - \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2 \left(\boldsymbol{B}\boldsymbol{B}' - (\boldsymbol{X}'\boldsymbol{X})^{-1} \right).$$

Weil $\boldsymbol{B}\boldsymbol{X} = \boldsymbol{I} = (\boldsymbol{B}\boldsymbol{X})' = \boldsymbol{X}'\boldsymbol{B}'$ ist, gilt

$$\begin{aligned} \boldsymbol{B}\boldsymbol{B}' - (\boldsymbol{X}'\boldsymbol{X})^{-1} &= \boldsymbol{B}\boldsymbol{B}' - \boldsymbol{B}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{B}' = \boldsymbol{B}\left(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \right)\boldsymbol{B}' \\ &= \boldsymbol{B}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{B}' = \boldsymbol{B}\boldsymbol{Q}\boldsymbol{B}'. \end{aligned}$$

Wegen der Idempotenz und Symmetrie von $\boldsymbol{Q} = (\boldsymbol{I} - \boldsymbol{H})$ sowie $\sigma^2 > 0$ folgt

$$\boldsymbol{c}'(\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}} - \boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}})\boldsymbol{c} = \sigma^2 \boldsymbol{c}'\boldsymbol{B}\boldsymbol{Q}\boldsymbol{B}'\boldsymbol{c} = \sigma^2 \boldsymbol{c}'\boldsymbol{B}\boldsymbol{Q}\boldsymbol{Q}\boldsymbol{B}'\boldsymbol{c} = \sigma^2 \boldsymbol{d}'\boldsymbol{d} = \sigma^2 \sum_{i=1}^{p+1} d_i^2 \geq 0$$

$$\text{mit} \quad \boldsymbol{d} = \boldsymbol{Q}\boldsymbol{B}'\boldsymbol{c} = (d_1, \ldots, d_{p+1})'$$

für alle $\boldsymbol{c} \in IR^{p+1}$. Der OLS-Schätzer $\hat{\boldsymbol{\beta}}$ ist somit der beste erwartungstreue lineare Schätzer (*BLUE*). Das Gauss-Markov-Theorem ist bewiesen.

Abschließend untersuchen wir die *asymptotischen Eigenschaften* des OLS-Schätzers. Unterliegt eine Schätzfunktion dem (schwachen) *Gesetz der Großen Zahl*, dann bezeichnet man sie als (schwach) *konsistenten Schätzer*. Im vorliegenden Fall hieße dies, dass für beliebig kleine reelle Zahlen $e > 0$

$$\lim_{n\to\infty} P\left(\left|\hat{\beta}_i - \beta_i\right| < e\right) = 1 \qquad (i = 0,1,\ldots,p)$$

erfüllt ist. Man schreibt hierfür abkürzend

$$p\lim \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$$

und sagt: „$\hat{\boldsymbol{\beta}}$ konvergiert mit wachsenden Stichprobenumfang n in Wahrscheinlichkeit gegen den wahren Parametervektor $\boldsymbol{\beta}$“. Die Präzision einer konsistenten Schätzfunktion verbessert sich mit dem Stichprobenumfang.

Hinreichende Bedingung für die Konsistenz ist, dass mit wachsendem Stichprobenumfang systematische Fehler und Zufallsfehler der Schätzung verschwinden:

$$\lim_{n\to\infty} E\left(\hat{\beta}_i\right) = \beta_i \qquad \text{und} \qquad \lim_{n\to\infty} Var\left(\hat{\beta}_i\right) = 0 \qquad (i = 0,1,\ldots,p)\,.$$

Wir haben weiter oben gezeigt, dass $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ bereits für endlich große Stichproben gilt. Um die Konsistenz des OLS-Schätzers sicherzustellen, bedarf es einer Zusatzannahme. Wir können die Varianz-Kovarianz-Matrix wie folgt umschreiben:

$$Cov(\hat{\boldsymbol{\beta}}) = \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1} = \frac{\sigma^2}{n}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.$$

Wir fordern, dass eine endliche Grenzwertmatrix

$$\lim_{n\to\infty} \frac{1}{n}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{M}$$

existiert und diese regulär (invertierbar) ist. Dann gilt:

$$\lim_{n\to\infty} Cov(\hat{\boldsymbol{\beta}}) = \lim_{n\to\infty} \frac{\sigma^2}{n}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1} = \left(\lim_{n\to\infty} \frac{\sigma^2}{n}\right) \cdot \lim_{n\to\infty}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1} = 0 \cdot \boldsymbol{M}^{-1} = \mathbf{O}.$$

Unter Gültigkeit der kaum restriktiven Zusatzforderung konvergiert $Cov(\hat{\boldsymbol{\beta}})$ für $n \to \infty$ gegen eine $(p+1, p+1)$-Nullmatrix **O**. Es folgt die Konsistenz des OLS-Schätzers:

$$p\lim \hat{\boldsymbol{\beta}} = \boldsymbol{\beta} \ .$$

Aus der Gleichung $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}$ erhalten wir durch Umformung

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \sqrt{n}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon} = n\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\frac{1}{\sqrt{n}}\boldsymbol{X}'\boldsymbol{\varepsilon} = \left(\frac{\boldsymbol{X}'\boldsymbol{X}}{n}\right)^{-1}\frac{1}{\sqrt{n}}\boldsymbol{X}'\boldsymbol{\varepsilon} \ .$$

Unterstellen wir nicht nur unkorrelierte, sondern *unabhängige* Störungen und existiert

$$\lim_{n\to\infty}\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{M} \ , \quad \boldsymbol{M} \text{ ist regulär,}$$

dann ist der Vektor $\frac{1}{\sqrt{n}}\boldsymbol{X}'\boldsymbol{\varepsilon}$ der gewogenen Summen der Störungen gemäß dem zentralen Grenzwertsatz von Lindeberg-Lévy *asymptotisch normalverteilt*:

$$\frac{1}{\sqrt{n}}\boldsymbol{X}'\boldsymbol{\varepsilon} \overset{asymptotisch}{\sim} N_{p+1}\left(\mathbf{0}, \sigma^2\boldsymbol{M}\right).$$

Einen formalen Beweis liefert z.B. Green 2008, S. 65 ff. Hieraus folgt die asymptotische Verteilung von $\hat{\boldsymbol{\beta}}$

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \overset{asymptotisch}{\sim} N_{p+1}\left(\mathbf{0}, \sigma^2\boldsymbol{M}^{-1}\right) \quad \text{bzw.} \quad \hat{\boldsymbol{\beta}} \overset{asymptotisch}{\sim} N_{p+1}\left(\boldsymbol{\beta}, \frac{\sigma^2}{n}\boldsymbol{M}^{-1}\right),$$

wobei bei hinreichend großen Stichproben $\frac{1}{n}\boldsymbol{M}^{-1}$ durch $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ approximiert werden kann.

Ist die Zusatzannahme A4 aus Definition 2.1 erfüllt, dann besitzt der OLS-Schätzer für endlich große Stichproben exakt eine Normalverteilung. Es gilt nun

$$\boldsymbol{y} \sim N_n\left(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}\right) .$$

Da Linearkombinationen multivariat normalverteilter Zufallsgrößen ebenfalls normalverteilt sind (vgl. Satz 1.22), folgt

$$\hat{\boldsymbol{\beta}} \sim N_{p+1}\left(\boldsymbol{\beta}, \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}\right)$$

für den linearen Schätzer

$$\hat{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{y} \quad \text{mit} \quad \boldsymbol{A} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' .$$

Für die einzelnen Komponenten $\hat{\beta}_i$ von $\hat{\boldsymbol{\beta}}$ gilt:

$$\hat{\beta}_i \sim N\left(\beta_i, \sigma^2 a_{jj}\right) \quad \text{mit} \quad j = i+1 \text{ und } i = 0,1,\ldots,p .$$

Die Normalverteilung des Schätzers nutzen wir später zur Konstruktion von Konfidenzintervallen und Tests.

Wir fassen die Eigenschaften des OLS-Schätzers in dem folgenden Satz zusammen.

S–2.5 Statistische Eigenschaften der OLS-Schätzer

Unter Gültigkeit der Annahmen A1 – A3 sowie den Zusatzannahmen unabhängiger Störungen und

$$\lim_{n\to\infty} \frac{1}{n}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{M} \quad , \quad \boldsymbol{M} \text{ ist regulär}$$

folgt:

(*a*)	$E(\hat{\boldsymbol{\beta}}_{OLS}) = \boldsymbol{\beta}$	(Erwartungstreue)
(*b*)	$Cov(\hat{\boldsymbol{\beta}}_{OLS}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$	(Varianz)
(*c*)	$plim\,\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{\beta}$	(Konsistenz)
(*d*)	$\hat{\boldsymbol{\beta}}_{OLS} \overset{asymptotisch}{\sim} N_{p+1}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$	(asymptotische Normalverteilung)
(*e*)	$\hat{\boldsymbol{\beta}}_{OLS}$ ist BLUE.	(Gauss-Markov-Theorem)

2.5 Parameterschätzung: Maximum-Likelihood-(ML)-Methode

Wir nutzen nun die ML-Methode zur Schätzung der unbekannten Parameter β und σ^2 des klassischen linearen Modells. Diese setzt im Gegensatz zur OLS-Methode die Kenntnis des Typs der Wahrscheinlichkeitsverteilung der Zufallsgrößen voraus.

Die ML-Methode geht bei der Schätzung der unbekannten Parameter eines stochastischen Modells von der gemeinsamen Dichtefunktion der Stichprobenvariablen aus. Da die Parameter die gemeinsame Dichte beeinflussen, fasst man die Dichte als ein Funktion der Parameter auf (während wir die Werte der Stichprobenvariablen als gegeben betrachtet werden). Die gemeinsame Dichtefunktion nennt man in diesem Zusammenhang *Likelihoodfunktion*. Likelihood ist das umgangssprachliche englische Wort für Wahrscheinlichkeit oder Plausibilität. Die ML-Schätzer für die Modellparameter erhält man als Maximumstellen der Likelihoodfunktion. D.h., die ML-Schätzer werden so konstruiert, dass der tatsächlich vorliegenden Stichprobe eine maximale Wahrscheinlichkeitsdichte zugeordnet wird (siehe Anhang V).

Unter den Annahmen A1 – A3 des klassischen linearen Modells und der zusätzlichen Normalverteilungsannahme A4 gilt

$$\boldsymbol{\varepsilon} = \boldsymbol{y} - \boldsymbol{X\beta} \sim N_n\left(\mathbf{0}, \sigma^2 \boldsymbol{I}\right).$$

Die einzelnen Störungen des Modells $\varepsilon_\nu = y_\nu - \boldsymbol{x}_\nu' \boldsymbol{\beta}$ $(\nu = 1,\ldots,n)$ sind *stochastisch unabhängig* und identisch $N(0,\sigma^2)$-verteilt mit der Dichte

$$f(\varepsilon) = \frac{1}{(2\pi\sigma^2)^{1/2}} \cdot exp\left\{-\frac{\varepsilon^2}{2\sigma^2}\right\}.$$

Es folgen unabhängige und normalverteilte Stichprobenvariable $y_1,\ldots,y_n$ des Regressanden:

$$\boldsymbol{y} \sim N_n\left(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}\right).$$

Die *Likelihoodfunktion* für das lineare Regressionsmodell in D–2.1 besitzt die Form:

(21)
$$L(\boldsymbol{\beta},\sigma^2) = L(\boldsymbol{\beta},\sigma^2 \mid \varepsilon_1,\ldots,\varepsilon_n) = \prod_{\nu=1}^{n} f(\varepsilon_\nu)$$

$$= \prod_{\nu=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} exp\left\{-\frac{\varepsilon_\nu^2}{2\sigma^2}\right\} = \frac{1}{(2\pi\sigma^2)^{n/2}} exp\left\{-\frac{\sum_{\nu=1}^{n}\varepsilon_\nu^2}{2\sigma^2}\right\}$$

$$= (2\pi\sigma^2)^{-n/2} exp\left\{-\frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{2\sigma^2}\right\} = (2\pi\sigma^2)^{-n/2} exp\left\{-\frac{(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})}{2\sigma^2}\right\} .$$

Die ML-Schätzer $\hat{\boldsymbol{\beta}}_{ML}, \hat{\sigma}^2_{ML}$ für die Parameter $\boldsymbol{\beta}$ und σ^2 erfüllen die Bedingung:

(22)
$$L(\hat{\boldsymbol{\beta}}_{ML}, \hat{\sigma}^2_{ML}) = \max_{\boldsymbol{\beta},\sigma^2} L(\boldsymbol{\beta},\sigma^2) .$$

Da die Extremwerte einer Funktion durch monotone Transformationen nicht beeinflusst werden, können die ML-Schätzer auch durch Maximieren der logarithmierten Likelihoodfunktion, die sog. *Loglikelihoodfunktion*, ermittelt werden.

Die *Loglikelihoodfunktion* besitzt hier die Form

(23)
$$l(\boldsymbol{\beta},\sigma^2) := ln\, L(\boldsymbol{\beta},\sigma^2) = ln\left((2\pi\sigma^2)^{-n/2}\, exp\left\{ -\frac{(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})}{2\sigma^2} \right\} \right)$$

$$= ln(2\pi\sigma^2)^{-n/2} - \frac{(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})}{2\sigma^2}$$

$$= -\frac{n}{2} ln\, 2\pi - \frac{n}{2} ln\, \sigma^2 - \frac{Q(\boldsymbol{\beta})}{2\sigma^2}$$

mit $Q(\boldsymbol{\beta}) \equiv \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})$ gemäß Gleichung 5b. Notwendige Bedingung für ein Maximum der Funktion ist, dass ihre partiellen Ableitungen erster Ordnung verschwinden.

Die partiellen Ableitungen erster Ordnung der Loglikelihoodfunktion lauten

(24)
$$\frac{\partial l(\boldsymbol{\beta},\sigma^2)}{\partial \boldsymbol{\beta}} = 0 - 0 - \frac{1}{2\sigma^2} \cdot \frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\frac{1}{\sigma^2}(-\boldsymbol{X'y} + \boldsymbol{X'X\beta}) \qquad \text{(vgl. Formel 6)}$$

(25)
$$\frac{\partial l(\boldsymbol{\beta},\sigma^2)}{\partial \sigma^2} = 0 - \frac{n}{2\sigma^2} + \frac{Q(\boldsymbol{\beta})}{2\sigma^4}.$$

Nullsetzen der Ableitungen liefert die sogenannten *Likelihoodgleichungen*

$$\frac{\partial l(\hat{\boldsymbol{\beta}}_{ML}, \hat{\sigma}^2_{ML})}{\partial \boldsymbol{\beta}} = -\frac{1}{\hat{\sigma}^2_{ML}}\left(-\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}_{ML}\right) \stackrel{!}{=} \mathbf{0}$$

und

$$\frac{\partial l(\hat{\boldsymbol{\beta}}_{ML}, \hat{\sigma}^2_{ML})}{\partial \sigma^2} = -\frac{n}{2\hat{\sigma}^2_{ML}} + \frac{Q(\hat{\boldsymbol{\beta}}_{ML})}{2\hat{\sigma}^4_{ML}} \stackrel{!}{=} 0 \ .$$

Aus den Likelihoodgleichungen erhalten wir die Maximumstellen der Funktion:

(26) $$-\frac{1}{\hat{\sigma}^2_{ML}}\left(-\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}_{ML}\right) = \mathbf{0} \quad \Rightarrow \quad \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}_{ML} = \boldsymbol{X}'\boldsymbol{y} \quad \Rightarrow \quad \hat{\boldsymbol{\beta}}_{ML} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \ ,$$

(27) $$-\frac{n}{2\hat{\sigma}^2_{ML}} + \frac{Q(\hat{\boldsymbol{\beta}}_{ML})}{2\hat{\sigma}^4_{ML}} = 0 \quad \Rightarrow \quad \frac{Q(\hat{\boldsymbol{\beta}}_{ML})}{2\hat{\sigma}^4_{ML}} = \frac{n}{2\hat{\sigma}^2_{ML}} \quad \Rightarrow \quad \hat{\sigma}^2_{ML} = \frac{Q(\hat{\boldsymbol{\beta}}_{ML})}{n} = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n}$$

mit $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{ML}$.

Unter *Gültigkeit der Normalverteilungsannahme* A4 (und nur dann!) sind somit der ML- und der OLS-Schätzer für die Regressionskoeffizienten identisch:

$$\hat{\boldsymbol{\beta}}_{ML} \equiv \hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \ .$$

Die ML-Methode liefert auf diesem Wege eine *statistische Rechtfertigung der OLS-Methode*, die von Gauss als ein mathematisches Verfahren der Funktionsapproximation begründet wurde.

2.6 Schätzung der Störvarianz

Der OLS-Schätzer und der ML-Schätzer für die Regressionskoeffizienten $\boldsymbol{\beta}$ sind identisch. Beide teilen somit auch die statistischen Eigenschaften (siehe Kapitel 2.4). Noch nicht untersucht haben wir die Eigenschaften des ML- Schätzers

$$\hat{\sigma}^2_{ML} = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n}$$

für die Varianz σ^2 der Störungen. Wir setzen hierbei die Gültigkeit der Annahmen A1 – A4 des klassischen linearen Modells voraus.

Wegen $\boldsymbol{\varepsilon} \sim N_n\left(\mathbf{0}, \sigma^2 \boldsymbol{I}\right)$ ist die Quadratsumme der standardisierten Störungen

$$\sum_{\nu=1}^{n}\left(\frac{\varepsilon_\nu}{\sigma}\right)^2 = \frac{1}{\sigma^2}\sum_{\nu=1}^{n}\varepsilon_\nu^2 = \frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{\sigma^2} \sim \chi^2(n)$$

eine Summe quadrierter unabhängiger standardnormalverteilter Zufallsvariablen und besitzt somit eine χ^2-Verteilung mit n Freiheitsgraden (vgl. Anhang A.5). Wir betrachten nun folgende Quadratsumme der Residuen $\hat{\varepsilon} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$:

$$\frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\sigma^2} = \frac{1}{\sigma^2}\sum_{\nu=1}^{n}\hat{\varepsilon}_\nu^2 \ .$$

Wir setzen $\boldsymbol{Q} = \boldsymbol{I} - \boldsymbol{H}$ mit $\boldsymbol{Q}' = \boldsymbol{Q}$, $\boldsymbol{QQ} = \boldsymbol{Q}$ und $rg(\boldsymbol{Q}) = n - p - 1$ (siehe Satz 2.2). Wegen

$$\boldsymbol{QX} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{X}\underbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}}_{=\boldsymbol{I}} = \boldsymbol{X} - \boldsymbol{X} = \mathbf{0}$$

folgt

$$\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y}'\boldsymbol{Q}'\boldsymbol{Q}\boldsymbol{y} = (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})'\boldsymbol{Q}'\boldsymbol{Q}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = (\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{Q}' + \boldsymbol{\varepsilon}'\boldsymbol{Q}')(\boldsymbol{QX}\boldsymbol{\beta} + \boldsymbol{Q}\boldsymbol{\varepsilon}) = \boldsymbol{\varepsilon}'\boldsymbol{Q}'\boldsymbol{Q}\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}'\boldsymbol{Q}\boldsymbol{\varepsilon} \ .$$

Die Quadratsumme

$$\frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\sigma^2} = \frac{\boldsymbol{\varepsilon}'\boldsymbol{Q}\boldsymbol{\varepsilon}}{\sigma^2} \sim \chi^2(n-p-1)$$

mit $\boldsymbol{Q} = \boldsymbol{I} - \boldsymbol{H}$ besitzt ebenfalls eine χ^2-Verteilung, die allerdings wegen $rg(\boldsymbol{Q}) = n - p - 1$ nur $n - p - 1$ Freiheitsgrade aufweist. Für den ML-Schätzer

$$\hat{\sigma}^2_{ML} = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n}$$

folgt hieraus

$$\frac{n\hat{\sigma}^2_{ML}}{\sigma^2} \sim \chi^2(n-p-1)$$

mit

$$E\left(\frac{n\hat{\sigma}^2_{ML}}{\sigma^2}\right) = \frac{nE\left(\hat{\sigma}^2_{ML}\right)}{\sigma^2} = n-p-1 \quad \text{und} \quad Var\left(\frac{n\hat{\sigma}^2_{ML}}{\sigma^2}\right) = \frac{n^2 Var\left(\hat{\sigma}^2_{ML}\right)}{\sigma^4} = 2(n-p-1)\ .$$

Der Erwartungswert des ML-Schätzers ist

$$E\left(\hat{\sigma}^2_{ML}\right) = \frac{n-p-1}{n}\sigma^2 \ .$$

Da $\lim_{n\to\infty} \frac{n-p-1}{n} = 1$ gilt, ist $\hat{\sigma}^2_{ML}$ ein *asymptotisch erwartungstreuer* Schätzer der Störvarianz σ^2. Für endlich Stichprobenumfänge wird die Varianz systematisch unterschätzt.

Die Varianz des Schätzers ist

$$Var\left(\hat{\sigma}^2_{ML}\right) = \frac{2(n-p-1)}{n^2}\sigma^4 \ .$$

Aus

$$\lim_{n\to\infty} E\left(\hat{\sigma}^2_{ML}\right) = \sigma^2 \qquad \text{und} \qquad \lim_{n\to\infty} Var\left(\hat{\sigma}^2_{ML}\right) = 0$$

folgt die *Konsistenz* des ML-Schätzers:

$$p\lim \hat{\sigma}^2_{ML} = \sigma^2 \ .$$

Da wir den Bias des ML-Schätzers

$$E\left(\hat{\sigma}^2_{ML}\right) - \sigma^2 = \frac{n-p-1}{n}\cdot\sigma^2 - \sigma^2 = \frac{-(1+p)}{n}\cdot\sigma^2 \ .$$

kennen, können wir die Verzerrung problemlos ausschalten und einen *erwartungstreuen Schätzer* konstruieren:

$$\hat{\sigma}^2 = \frac{n}{n-p-1}\hat{\sigma}^2_{ML} = \frac{n}{n-p-1}\cdot\frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n} = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-p-1}$$

mit

$$E\left(\hat{\sigma}^2\right) = \frac{n}{n-p-1}\cdot E\left(\hat{\sigma}^2_{ML}\right) = \frac{n}{n-p-1}\cdot\frac{n-p-1}{n}\cdot\sigma^2 = \sigma^2 \ .$$

Das Ergebnis besitzt auch dann Gültigkeit, wenn die Normalverteilungsannahme A4 nicht erfüllt ist. Auf einen Beweis der Behauptung wird an dieser Stelle verzichtet.

Wir ziehen im Folgenden stets den erwartungstreuen und konsistenten Schätzer

(28) $$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-p-1}$$

dem Schätzer $\hat{\sigma}^2_{ML}$ vor. Mit $\hat{\sigma}^2$ können wir die Varianz-Kovarianz-Matrix

$$Cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} = \boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}$$

erwartungstreu schätzen:

(29) $$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} = \hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

Es kennzeichne a_{jj} wieder das j-te Diagonalelement von $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Dann erhalten wir gemäß

(30) $$\hat{\sigma}_{\hat{\beta}_i} = \hat{\sigma} \cdot \sqrt{a_{jj}} \quad \text{mit} \quad j = i+1 \ \text{ und } \ i = 0,1,\ldots,p$$

die *geschätzten Standardfehler* der OLS- bzw. ML-Schätzungen $\hat{\beta}_i$ für die Regressionskoeffizienten β_i. Mit Hilfe von (30) messen wir die Präzision der Schätzer $\hat{\beta}_i$.

B–2.1 *Makroökonomische Konsumfunktion.*

Wir betrachten weiterhin die geschätzte Konsumfunktion

$$\hat{y} = 37.317 + 0.864 \cdot x \quad \text{bzw.} \quad \hat{konsum} = 37.317 + 0.864 \cdot einkommen \ .$$

Die zugehörige Schätzung der Varianz der Störungen ist

$$\hat{\sigma}^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-p-1} = \frac{\sum_{\nu=1}^{n} \varepsilon_\nu^2}{n-p-1} = \frac{RSS}{n-p-1} = \frac{9275.191}{19-1-1} = 545.5995$$

mit

$$\hat{\varepsilon}_\nu = y_\nu - \hat{y}_\nu = y_\nu - 37.317 - 0.864 \cdot x_\nu \quad (\nu = 1,\ldots,n).$$

Hiermit erhalten wir folgende Schätzung der Varianz-Kovarianz-Matrix des OLS-Schätzers $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} = \hat{\sigma}^2 (\boldsymbol{X}'\boldsymbol{X})^{-1} = 545.5995 \cdot \begin{pmatrix} 3.58306 & -0.00303 \\ -0.00303 & 2.59352\text{E}-06 \end{pmatrix}$$

$$= \begin{pmatrix} 1954.914 & -1.65094 \\ -1.65094 & 0.00142 \end{pmatrix}.$$

Auf der Hauptdiagonalen von $\hat{\Sigma}_{\hat{\beta}}$ stehen die geschätzten Varianzen der Regressionskoeffizienten. Wegen

$$\hat{\Sigma}_{\hat{\beta}} = \hat{\sigma}^2 (\boldsymbol{X}'\boldsymbol{X})^{-1} = \hat{\sigma}^2 \begin{pmatrix} n & \sum_{\nu=1}^{n} x_\nu \\ \sum_{\nu=1}^{n} x_\nu & \sum_{\nu=1}^{n} x_\nu^2 \end{pmatrix}^{-1} = \frac{\hat{\sigma}^2}{n\sum x_\nu^2 - (\sum x_\nu)^2} \cdot \begin{pmatrix} \sum_{\nu=1}^{n} x_\nu^2 & -\sum_{\nu=1}^{n} x_\nu \\ -\sum_{\nu=1}^{n} x_\nu & n \end{pmatrix}$$

gilt

$$\hat{\sigma}^2_{\hat{\beta}_0} = \frac{\hat{\sigma}^2 \cdot \sum x_\nu^2}{n\sum x_\nu^2 - (\sum x_\nu)^2} = \frac{\hat{\sigma}^2 \cdot \frac{1}{n}\sum x_\nu^2}{\sum x_\nu^2 - n\bar{x}^2} = \frac{\hat{\sigma}^2 \cdot \frac{1}{n}\sum x_\nu^2}{\sum (x_\nu - \bar{x})^2} = 1954.914 \,,$$

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{\hat{\sigma}^2 \cdot n}{n\sum x_\nu^2 - (\sum x_\nu)^2} = \frac{\hat{\sigma}^2}{\sum x_\nu^2 - n\bar{x}^2} = \frac{\hat{\sigma}^2}{\sum (x_\nu - \bar{x})^2} = 0.00142 \,.$$

Die Quadratwurzeln der Varianzen sind die geschätzten Standardfehler

$$\hat{\sigma}_{\hat{\beta}_0} = \sqrt{1954.914} = 44.214 \qquad \text{und} \qquad \hat{\sigma}_{\hat{\beta}_1} = \sqrt{0.00142} = 0.0376 \quad \boxtimes$$

2.7 Konfidenz- und Prognoseintervalle

Die Modellannahmen A1 – A4 seien erfüllt. a_{jj} sei wieder das j-te Diagonalelement von $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Wir wissen bereits

$$\hat{\beta}_i \sim N\left(\beta_i, \sigma^2 a_{jj}\right) \Leftrightarrow \frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{a_{jj}}} \sim N(0,1) \qquad \text{mit} \qquad \hat{\beta}_i = \hat{\beta}_{i,ML} = \hat{\beta}_{i,OLS}$$

sowie

$$\frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\sigma^2} = \frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p-1)\ .$$

Hiermit und mit $\hat{\sigma}_{\hat{\beta}_i} = \hat{\sigma}\cdot\sqrt{a_{jj}}$ erhalten wir die folgende t-Statistik (vgl. Anhang V):

(31)
$$T_1 = \frac{\frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{a_{jj}}}}{\sqrt{\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2(n-p-1)}}} = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}\cdot\sqrt{a_{jj}}} = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_{\hat{\beta}_i}} \sim t(n-p-1).$$

Es bezeichne $t_{[1-\alpha/2;n-p-1]}$ das $(1-\alpha/2)$-Quantil der t-Verteilung mit $n-p-1$ Freiheitsgraden (vgl. Abbildung 2.4). Aus

$$P\left(-t_{[1-\alpha/2;n-p-1]} < T_1 < +t_{[1-\alpha/2;n-p-1]}\right) =$$

$$P\left(-t_{[1-\alpha/2;n-p-1]} < \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_{\hat{\beta}_i}} < +t_{[1-\alpha/2;n-p-1]}\right) = 1-\alpha$$

folgt

$$P\left(\hat{\beta}_i - t_{[1-\alpha/2;n-p-1]} \cdot \hat{\sigma}_{\hat{\beta}_i} < \beta_i < \hat{\beta}_i + t_{[1-\alpha/2;n-p-1]} \cdot \hat{\sigma}_{\hat{\beta}_i}\right) = 1-\alpha$$

und hiermit das $(1-\alpha)$-*Konfidenzintervall für* β_i $(i = 0,1,\ldots,p)$:

(32) $$\left[\hat{\beta}_i - t_{[1-\alpha/2;n-p-1]} \cdot \hat{\sigma}_{\hat{\beta}_i} \;\;,\;\; \hat{\beta}_i + t_{[1-\alpha/2;n-p-1]} \cdot \hat{\sigma}_{\hat{\beta}_i}\right].$$

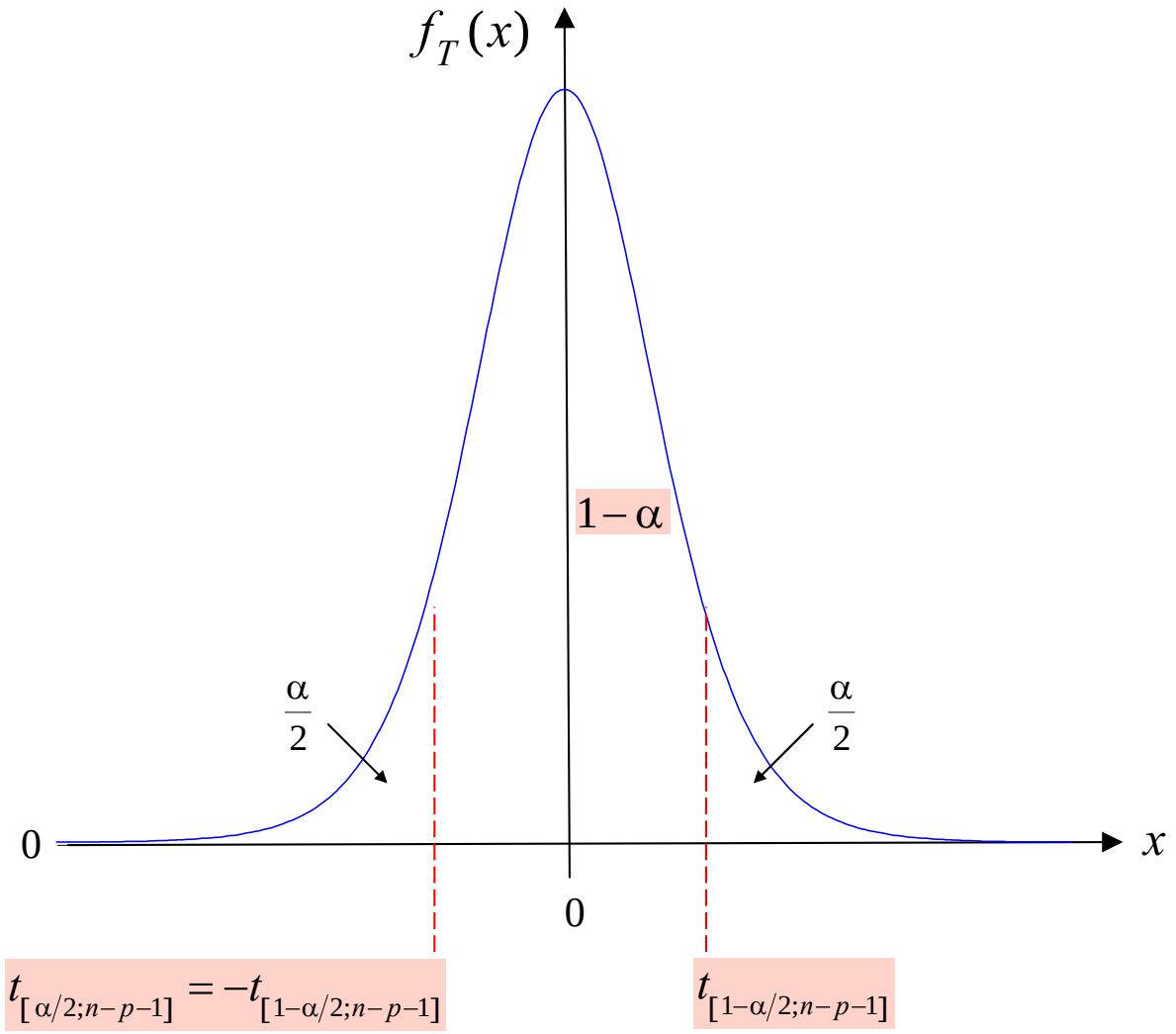


Abb. 2.4: Dichtefunktion sowie das $(\alpha/2)$- und das $(1-\alpha/2)$-Quantil der $t(n-p-1)$-Verteilung

Sei $\boldsymbol{x}_0 = (1, x_{01}, \ldots, x_{0p})'$ ein Beobachtungsvektor der Regressoren und y_0 eine Zufallsvariable, die für die potenziellen Werte des Regressanden Y gegeben die Regressorwerte $\boldsymbol{x}_0$ steht. Der *Wert der Regressionsfunktion* an der Stelle $\boldsymbol{x}_0$

$$E(y_0) = \boldsymbol{x}_0'\boldsymbol{\beta}$$

entspricht dem Erwartungswert von Y unter der Bedingung $\boldsymbol{x}_0$. Wegen $\hat{\boldsymbol{\beta}} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$ gilt

$$\boldsymbol{x}_0'\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{x}_0'\boldsymbol{\beta}, \sigma^2\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0\right)$$

(vgl. Satz 1.22). Durch Standardisieren des *Werts der geschätzten Regressionsfunktion*

$$\frac{\boldsymbol{x}_0'\hat{\boldsymbol{\beta}} - \boldsymbol{x}_0'\boldsymbol{\beta}}{\sqrt{\sigma^2\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}} \sim N(0,1)$$

und Schätzen von σ^2 erhalten wir die t-Statistik

(33)
$$T_2 = \frac{\dfrac{\boldsymbol{x}_0'\hat{\boldsymbol{\beta}} - \boldsymbol{x}_0'\boldsymbol{\beta}}{\sqrt{\sigma^2\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}}}{\sqrt{\dfrac{(n-p-1)\hat{\sigma}^2}{\sigma^2(n-p-1)}}} = \frac{\boldsymbol{x}_0'\hat{\boldsymbol{\beta}} - \boldsymbol{x}_0'\boldsymbol{\beta}}{\sqrt{\hat{\sigma}^2\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}} \sim t(n-p-1).$$

Aus

$$P\left(-t_{[1-\alpha/2;n-p-1]} < T_2 < t_{[1-\alpha/2;n-p-1]}\right) =$$

$$P\left(-t_{[1-\alpha/2;n-p-1]} < \frac{\boldsymbol{x}_0'\hat{\boldsymbol{\beta}} - \boldsymbol{x}_0'\boldsymbol{\beta}}{\sqrt{\hat{\sigma}^2 \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}} < t_{[1-\alpha/2;n-p-1]}\right) = 1-\alpha$$

und

$$P\left(\boldsymbol{x}_0'\hat{\boldsymbol{\beta}} - t_{[1-\alpha/2;n-p-1]} \cdot \sqrt{\hat{\sigma}^2 \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0} < \boldsymbol{x}_0'\boldsymbol{\beta} < \boldsymbol{x}_0'\hat{\boldsymbol{\beta}} + t_{[1-\alpha/2;n-p-1]} \cdot \sqrt{\hat{\sigma}^2 \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}\right) = 1-\alpha$$

resultiert das (1–α)-*Konfidenzintervall für* $E(y_0)$ oder das (1–α)-*Konfidenzintervall der Regressionsfunktion an der Stelle* $\boldsymbol{x}_0$:

(34)
$$\left[\boldsymbol{x}_0'\hat{\boldsymbol{\beta}} - t_{[1-\alpha/2;n-p-1]} \cdot \sqrt{\hat{\sigma}^2 \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0} \;,\; \boldsymbol{x}_0'\hat{\boldsymbol{\beta}} + t_{[1-\alpha/2;n-p-1]} \cdot \sqrt{\hat{\sigma}^2 \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}\right].$$

Will man darüber hinaus ein Intervall konstruieren, in dem y_0 mit vorgegebener Wahrscheinlichkeit Werte annimmt, so muss man neben der Streuung von $\hat{\boldsymbol{\beta}}$ auch die Streuung von y_0 berücksichtigen. Wir betrachten die Prognose $\hat{y}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}$ für y_0 mit dem Prognosefehler $\hat{\varepsilon}_0 = y_0 - \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}$. Aus

$$y_0 - \boldsymbol{x}_0'\hat{\boldsymbol{\beta}} \sim N\left(0\,,\, \sigma^2 + \sigma^2 \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0\right)$$

gemäß Satz 1.22 folgt durch Standardisieren und Einsetzen von $\hat{\sigma}^2$ für σ^2

(35)
$$T_3 = \frac{y_0 - \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}}{\sqrt{\hat{\sigma}^2 + \hat{\sigma}^2 \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}} \sim t(n-p-1)$$

und damit

$$P\left(-t_{[1-\alpha/2;n-p-1]} < \frac{y_0 - \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}}{\sqrt{\hat{\sigma}^2 + \hat{\sigma}^2 \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}} < +t_{[1-\alpha/2;n-p-1]}\right) = 1-\alpha$$

Es resultiert das (1–α)-*Prognoseintervall für* y_0:

(36)
$$\left[\boldsymbol{x}_0'\hat{\boldsymbol{\beta}} - t_{[1-\alpha/2;n-p-1]} \cdot \sqrt{\hat{\sigma}^2 + \hat{\sigma}^2 \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}\ ,\ \boldsymbol{x}_0'\hat{\boldsymbol{\beta}} + t_{[1-\alpha/2;n-p-1]} \cdot \sqrt{\hat{\sigma}^2 + \hat{\sigma}^2 \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}\right].$$

Anmerkung

Ist die Normalverteilungsannahme A4 verletzt, dann gelten aufgrund des asymptotischen Normalverteilung des Schätzers $\hat{\boldsymbol{\beta}}$ die Konfidenzintervalle (32) und (34) für große Stichprobenumfänge n weiterhin approximativ. Das Prognoseintervall (36) ist hingegen nicht mehr valide, da y_0 nun keine Normalverteilung besitzt.

B–2.1 *Makroökonomische Konsumfunktion.*

Unsere bisherigen Ergebnisse fassen wir wie folgt zusammen:

$$\underset{}{\hat{konsum}} = \underset{(44.2144)}{37.317} + \underset{(0.0376)}{0.864} \cdot einkommen\ , \qquad R^2 = 0.9688.$$

Mittels der Standardfehler können wir die Grenzen von Konfidenzintervallen für die beiden Modellparameter β_0 und β_1 berechnen. Zum Konfidenzniveau $1-\alpha = 0.95$ erhalten wir mit dem 0.975-

Quantil $t_{[0.975;17]} = 2.1098$ der t-Verteilung mit $n - p - 1 = 17$ Freiheitsgraden die konkreten Konfidenzintervalle

$$\left[\hat{\beta}_0 - t_{[0.975;17]} \cdot \hat{\sigma}_{\hat{\beta}_0} \;,\; \hat{\beta}_0 + t_{[0.975;17]} \cdot \hat{\sigma}_{\hat{\beta}_0}\right] = [-55.967 \;,\; 130.601] \,,$$

$$\left[\hat{\beta}_1 - t_{[0.975;17]} \cdot \hat{\sigma}_{\hat{\beta}_1} \;,\; \hat{\beta}_1 + t_{[0.975;17]} \cdot \hat{\sigma}_{\hat{\beta}_1}\right] = [0.785 \;,\; 0.944] \,.$$

Man beachte, dass das erste Intervall die Zahl Null einschließt, so dass der Schätzwert 37.317 nicht signifikant von Null verschieden ist (Signifikanzniveau $\alpha = 0.05$).

Auf die numerische Angabe von Konfidenzintervallen für die Regressionsfunktion und von Prognoseintervallen wird hier verzichtet. Stattdessen zeigt die Abbildung 2.5 graphisch das Konfidenzband zum Vertrauensniveau von 0.95 für die Regressionsfunktion (rote Linien) und das 0.95-Konfidenzband für Prognosen (grüne Linien).

Hinweis: Im Falle einer linearen Einfachregression hängt die Breite der Konfidenzbänder von der Varianz σ^2 der Störungen, der Varianz des Regressors, der Anzahl n der zur Parameterschätzung genutzten Daten sowie dem Abstand $|x_0 - \bar{x}|$ ab. Je größer der Abstand $|x_0 - \bar{x}|$ desto breiter werden die Intervalle. Die Bänder haben deshalb eine hyperbolische Form.

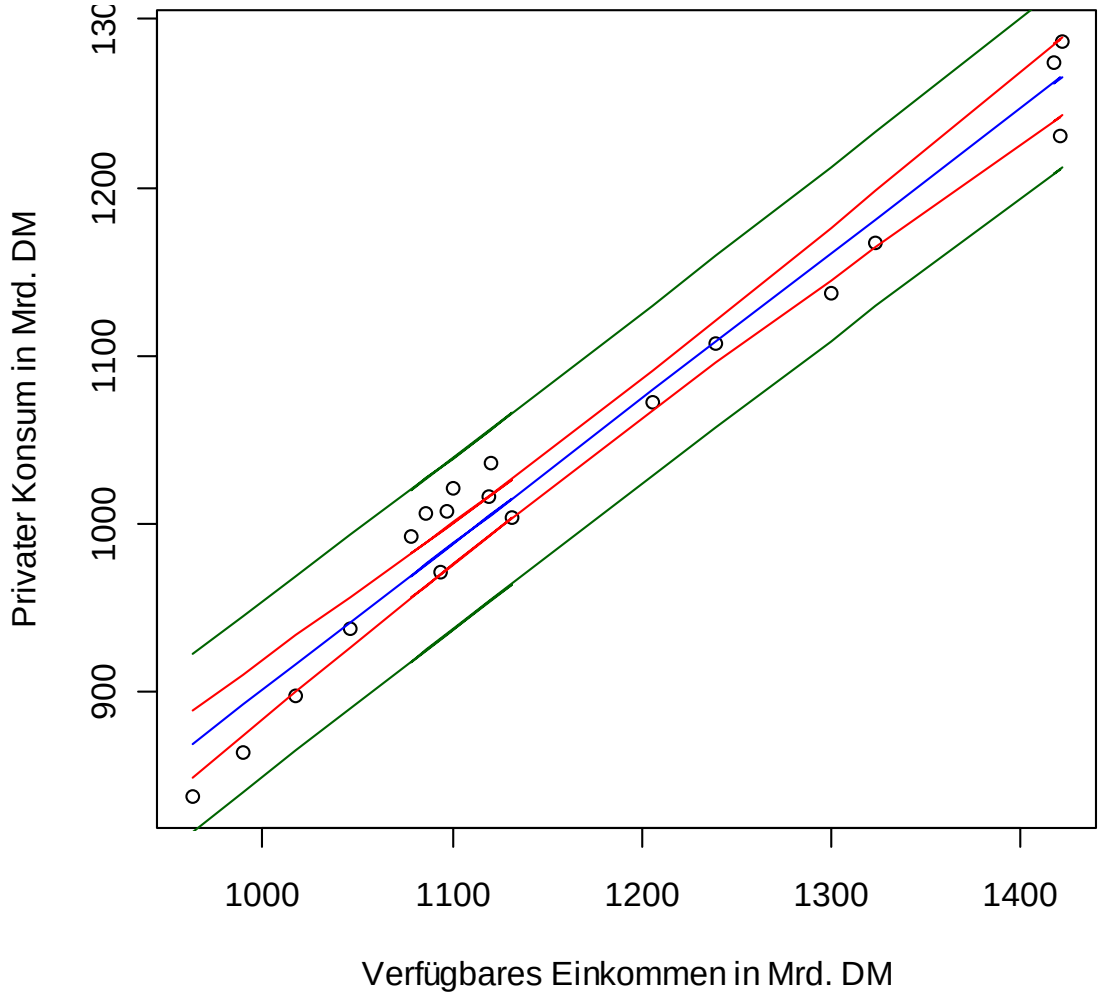


Abb. 2.5: Streudiagramm der Daten mit geschätzter Regressionsfunktion (blau) und 95%-Konfidenzbänder für die Regressionsfunktion (rot) und die Prognose (grün) ☒

2.8 Hypothesentests

Tests für einen einzelnen Regressionskoeffizienten

Die Modellannahmen A1 – A4 seien erfüllt. Unter der Bedingung $\beta_i = b$ gilt dann gemäß (31)

(37) $$T_1 = \frac{\hat{\beta}_i - b}{\hat{\sigma}_{\hat{\beta}_i}} \sim t(n-p-1) \ .$$

Die Statistik T_1 kann als Teststatistik des *t-Tests* der Hypothesen

(38) $$H_0 : \beta_i = b \qquad \text{versus} \qquad H_1 : \beta_i \neq b$$

verwendet werden. H_0 wird zum Signifikanzniveau α (= Irrtumswahrscheinlichkeit) verworfen, wenn der realisierte Wert t der Statistik T_1 in den kritischen Bereich

$$K = \left(-\infty\,, -t_{[1-\alpha/2;n-p-1]}\right) \cup \left(+t_{[1-\alpha/2;n-p-1]}\,, +\infty\right)$$

fällt, d.h. falls

$$|t| > t_{[1-\alpha/2;n-p-1]} \ .$$

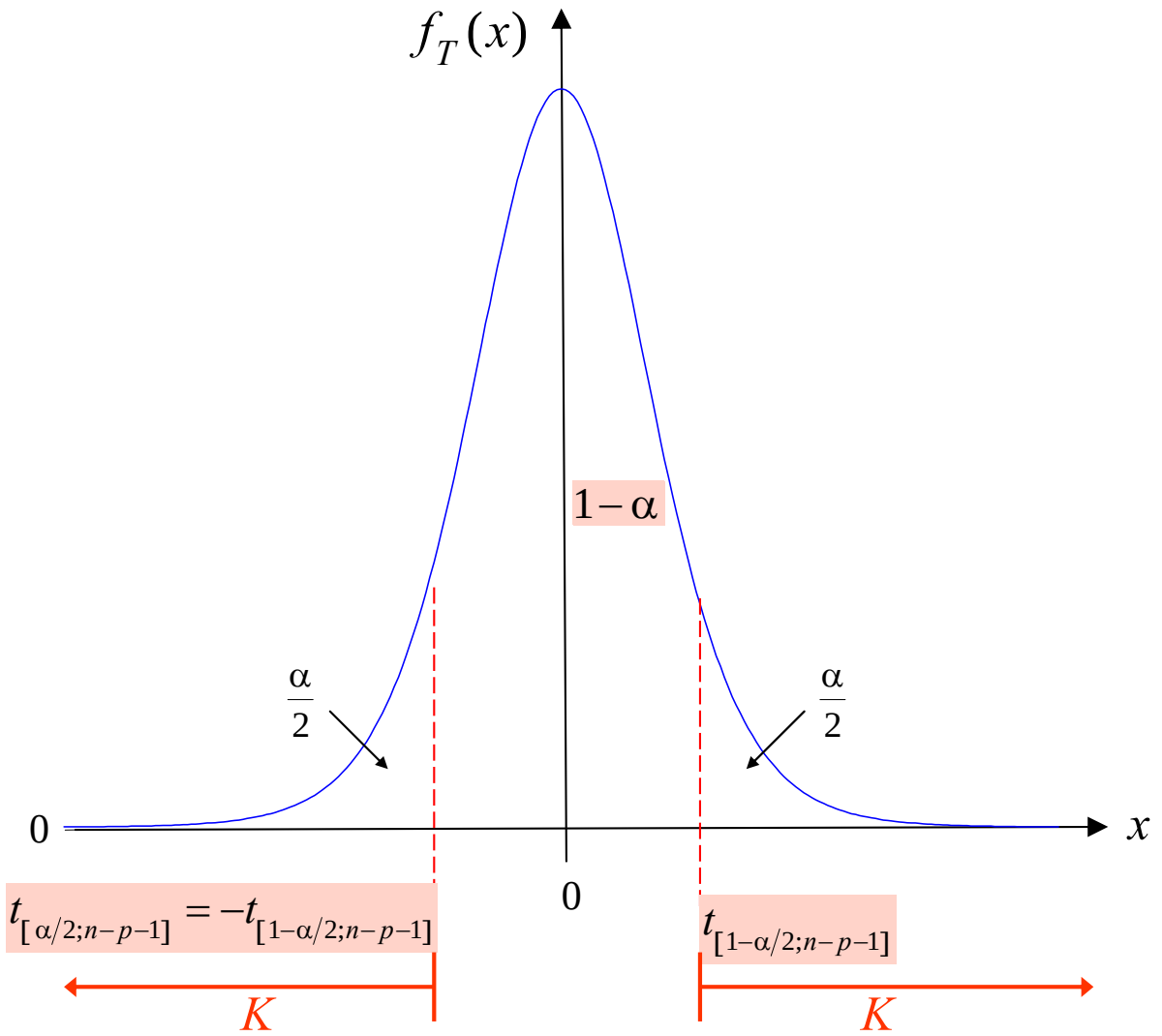


Abb. 2.6: Dichtefunktion sowie das $(\alpha/2)$- und das $(1-\alpha/2)$-Quantil der $t(n-p-1)$-Verteilung

Von besonderem Interesse ist oft der Test der Hypothesen

$$H_0 : \beta_i = 0 \qquad \text{versus} \qquad H_1 : \beta_i \neq 0 \ ,$$

der prüft, ob der *i*-te Regressor X_i einen statistisch signifikanten Beitrag zur Erklärung des Regressanden Y liefert.

Äquivalent kann anstelle der *t*-Statistik auch die *F*-Statistik

(39) $$F_1 = T_1^2 \sim F(1, n-p-1)$$

verwendet werden. H_0 wird zum Signifikanzniveau α (= Irrtumswahrscheinlichkeit) verworfen, wenn der realisierte Wert f der Statistik F_1 in den kritischen Bereich

$$K = \left(f_{[1-\alpha;1,n-p-1]} \ , \ \infty \right)$$

fällt, d.h. falls

$$f > f_{[1-\alpha;1,n-p-1]} \ .$$

$f_{[1-\alpha;1,n-p-1]}$ ist das $(1-\alpha)$-Quantil der *F*-Verteilung mit 1 und $n-p-1$ Freiheitsgraden. Man spricht hier von einem *partiellen F-Test*.

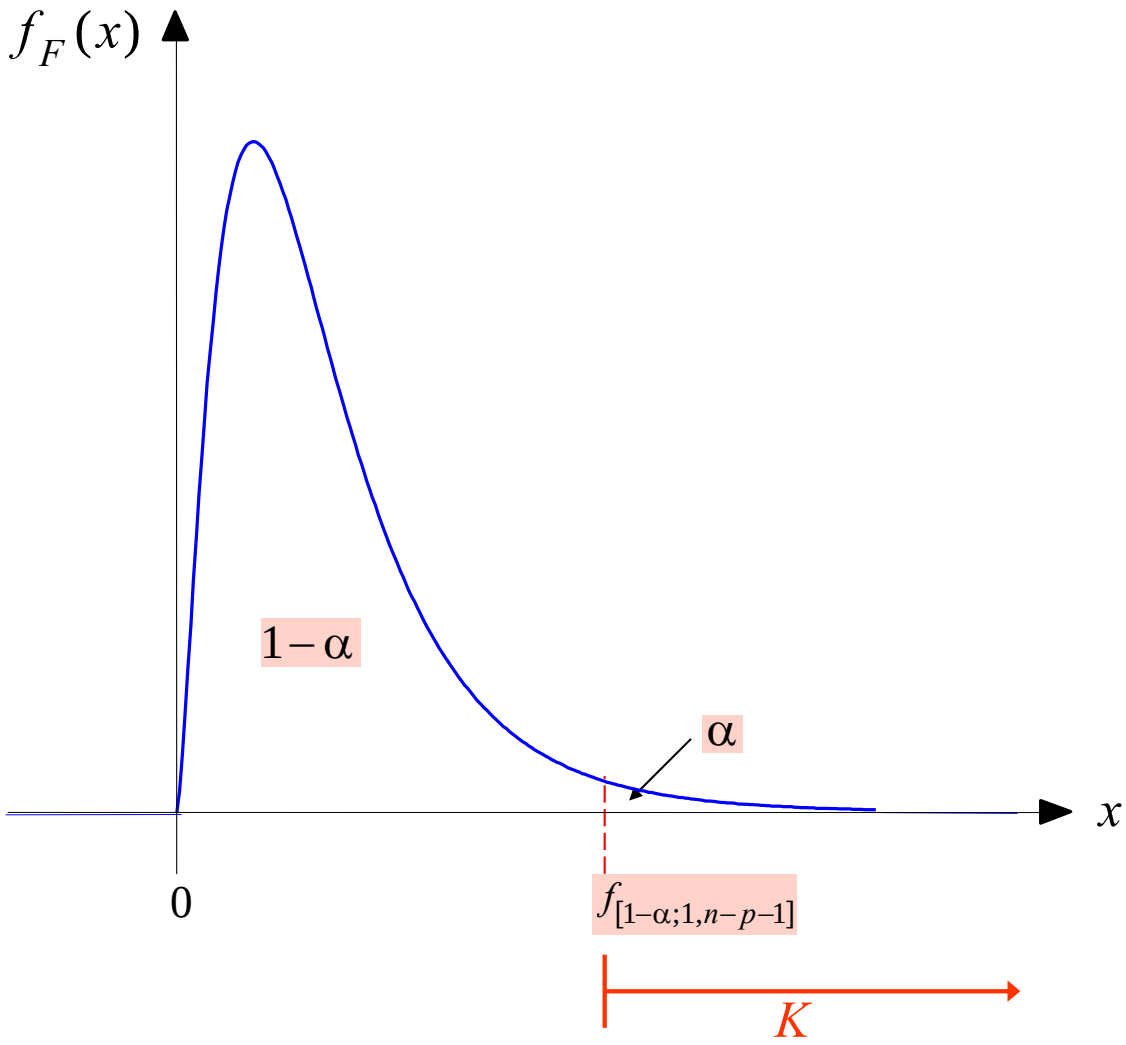


Abb. 2.7: Dichtefunktion sowie das $(1-\alpha)$-Quantil der F-Verteilung mit 1 und $n-p-1$ Freiheitsgraden

F-Test für alle Regressionskoeffizienten

Will man untersuchen, ob die Regressoren insgesamt zur Erklärung des Regressanden Y beitragen, so führt dies auf das Testproblem

(40) $$H_0 : \beta_1 = \ldots = \beta_p = 0 \quad \text{versus} \quad H_1 : \beta_i \neq 0 \ \text{ für mindestens ein } i \, .$$

Die Nullhypothese kann mittels der Statistik des sogenannten *overall F-Tests* geprüft werden:

(41) $$F_2 = \frac{n-p-1}{p} \cdot \frac{ESS}{RSS} = \frac{n-p-1}{p} \cdot \frac{R^2}{1-R^2} \sim F(p\,, n-p-1) \, .$$

Hierbei ist R^2 das Bestimmtheitsmaß (15). Große Werte f der Statistik F_2 führen zur Ablehnung der Nullhypothese. H_0 wird zum Signifikanzniveau α verworfen, falls

$$f > f_{[1-\alpha\,;\,p,n-p-1]} \, .$$

Im Falle einer linearen Einfachregression ist der overall F-Test natürlich mit dem partiellen F-Test für den Parameter β_1 identisch.

Hintergrund

Die Modellannahmen A1 – A4 seien erfüllt. Dann gilt

$$\frac{RSS}{\sigma^2} = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\sigma^2} \sim \chi^2(n-p-1) \ .$$

Ferner ist

$$\frac{Y_\nu - E(Y_\nu)}{\sigma} \overset{i.i.d}{\sim} N(0,1) \quad \text{und} \quad \sum_{\nu=1}^{n}\left(\frac{Y_\nu - \bar{Y}}{\sigma}\right)^2 = \frac{1}{\sigma^2}\sum_{\nu=1}^{n}(Y_\nu - \bar{Y})^2 = \frac{TSS}{\sigma^2} \sim \chi^2(n-1) \ .$$

Hieraus folgt zunächst

$$\frac{ESS}{\sigma^2} = \frac{TSS}{\sigma^2} - \frac{RSS}{\sigma^2} \sim \chi^2(p)$$

und wegen der Unabhängigkeit von *ESS* und *RSS* ist

$$F_2 = \frac{\dfrac{TSS - RSS}{p \cdot \sigma^2}}{\dfrac{RSS}{(n-p-1)\cdot\sigma^2}} = \frac{n-p-1}{p} \cdot \frac{ESS}{RSS} \sim F(p\,, n-p-1) \ .$$

Die beim overall F-Test benötigten Quadratsummen werden oft in Form einer ANOVA-Tabelle (***a-nalysis of variance***) dargestellt.

ANOVA-Tabelle

Source	Degree of Freedom (DF)	Sum of Squares	Mean Squares	F-Statistic
Regression	p	ESS	$\frac{ESS}{p}$	F_2
Residual	$n-p-1$	RSS	$\frac{RSS}{n-p-1}$	-
Total	$n-1$	TSS	-	-

B–2.1 *Makroökonomische Konsumfunktion.*

Das Programm *lm* (*linear model*) des *Statistik-Pakets R* berechnet standardmäßig *t-Tests* und den *overall F-Test* für Regressionsparameter:

```
kf.modell <- lm(konsum~einkommen, data=kf)
summary(kf.modell)

Coefficients:
              Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)   37.31676     44.21441     0.844       0.41
einkommen      0.86422      0.03762    22.974    3.07e-14 ***
---
Residual standard error: 23.36 on 17 degrees of freedom
Multiple R-squared: 0.9688,     Adjusted R-squared: 0.967
F-statistic: 527.8 on 1 and 17 DF,  p-value: 3.07e-14

anova(kf.modell)
Analysis of Variance Table
Response: konsum
             Df    Sum Sq    Mean Sq    F value      Pr(>F)
einkommen     1    287980     287980     527.82    3.07e-14 ***
Residuals    17      9275        546
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

F-Test der allgemeinen linearen Hypothese

Die Testprobleme (38) und (40) lassen sich als Spezialfälle des folgenden allgemeinen Testproblems auffassen:

(42) $$H_0 : \boldsymbol{C\beta} = \boldsymbol{b} \qquad \text{versus} \qquad H_1 : \boldsymbol{C\beta} \neq \boldsymbol{b} \ ,$$

wobei $\boldsymbol{C}$ eine geeignete $(s, p+1)$-Matrix mit $rg(\boldsymbol{C}) = s$ und $\boldsymbol{b}$ ein s-Vektor ist. Wir betrachten die Matrix $\boldsymbol{C}$ und den Vektor $\boldsymbol{b}$ in einigen speziellen Hypothesen:

$H_0 : \beta_i = b$	$\boldsymbol{C} = (0,\ldots,0,1,0,\ldots,0),\ \boldsymbol{b} = b$ mit $s = 1$
$H_0 : \beta_1 = \ldots = \beta_p = 0$	$\boldsymbol{C} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix},\ \boldsymbol{b} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ mit $s = p$
$H_0 : \beta_2 = \beta_3 = 0$	$\boldsymbol{C} = \begin{pmatrix} 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \end{pmatrix},\ \boldsymbol{b} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ mit $s = 2$
$H_0 : \beta_1 = \beta_4 \Leftrightarrow \beta_1 - \beta_4 = 0$	$\boldsymbol{C} = (0,1,0,0,-1,0,\ldots,0),\ \boldsymbol{b} = 0$ mit $s = 1$
$H_0 : \beta_2 + \beta_3 + \beta_4 = 1$	$\boldsymbol{C} = (0,0,1,1,1,0,\ldots,0),\ \boldsymbol{b} = 1$ mit $s = 1$

Die F-Teststatistik der Hypothesen (42) ist analog zu (41) aufgebaut:

$$(43) \qquad F_3 = \frac{\dfrac{RSS(H_0) - RSS}{s \cdot \sigma^2}}{\dfrac{RSS}{(n-p-1)\cdot\sigma^2}} = \frac{n-p-1}{s} \cdot \frac{RSS(H_0) - RSS}{RSS} \sim F(s\,, n-p-1)\,.$$

$RSS(H_0)$ steht hierbei für die Residualquadratsumme der Regression von Y bezüglich $X_1,...,X_p$ unter Beachtung der Nebenbedingung $\boldsymbol{C\beta} = \boldsymbol{b}$.

RSS ist die Residualquadratsumme der Regression ohne Beachtung der Nebenbedingung. Offensichtlich muss stets $RSS(H_0) \geq RSS$ erfüllt sein.

Große Werte f der Statistik führen zur Ablehnung der Nullhypothese. H_0 wird zum Signifikanzniveau α verworfen, falls

$$f > f_{[1-\alpha\,;\,s,n-p-1]}\,.$$

Man beachte, dass $RSS(H_0) = TSS$ im Falle des Testproblems (40) gilt.

B–2.2 *Mikroökonomisches Mietpreismodell* (*Binär-Kodierung kategorialer Variablen*).

Wir schätzen nun ein einfaches mikroökonomisches Mietpreismodell für Mietwohnungen in München. Die Modellgleichung besitzt die Form:

$$log(Miete) = \beta_0 + \beta_1 \cdot Fläche + \beta_2 \cdot Lage^{+} + \beta_3 \cdot Lage^{++} + \varepsilon$$

mit

$$Lage^{+} = \begin{cases} 1 & \text{falls Wohnung in guter Lage} \\ 0 & \text{sonst} \end{cases}, \quad Lage^{++} = \begin{cases} 1 & \text{falls Wohnung in sehr guter Lage} \\ 0 & \text{sonst.} \end{cases}$$

Als Daten stehen Angaben über die Miete (in DM), Wohnfläche und Wohnlage von 799 zufällig ausgewählte Wohnungen mit normalem Ausstattungsniveau und einem Alter von höchstens 30 Jahren zur Verfügung. Die Daten wurden im Jahr 1999 zur Erstellung eines Mietspiegels für München erhoben (Quelle: Fahrmeir et al. 2009, S. 5 f).

Das mittels OLS geschätzte Modell (siehe unten)

$$log(Miete) = 6.059 + 0.012 \cdot Fläche + 0.106 \cdot Lage^{+} + 0.172 \cdot Lage^{++} + \hat{\varepsilon}$$

schreiben wir um gemäß

$$\begin{aligned} Miete &= exp\left(6.059 + 0.012 \cdot Fläche + 0.106 \cdot Lage^{+} + 0.172 \cdot Lage^{++} + \hat{\varepsilon}\right) \\ &= exp(6.059) \cdot exp(0.012 \cdot Fläche) \cdot exp\left(0.106 \cdot Lage^{+}\right) \cdot exp\left(0.172 \cdot Lage^{++}\right) \cdot exp(\hat{\varepsilon}) . \end{aligned}$$

Die exponierten geschätzten Koeffizienten der Regressoren

$$exp(0.012) = 1.012 \ , \quad exp(0.106) = 1.112 \ , \quad exp(0.172) = 1.188$$

können als Änderungsfaktoren bezüglich des Mietpreises interpretiert werden. Sie entsprechen dem durchschnittlichen Änderungsfaktor der Miete, die aus der Erhöhung des Wertes des zugehörigen Regressors um die Zahl 1 resultiert (*marginaler Effekt*). Ein zusätzlicher Quadratmeter Wohnfläche erhöht also die Miete durchschnittlich um 1.2%. Eine gute Lage (bzw. sehr gute Lage) führt im Mittel zu einem Preisanstieg von 11.2% (bzw. 18.8%) gegenüber einer Wohnung mit normaler Lage.

Gute und sehr gute Lagen führen zu signifikant höheren Mieten als normale Lagen. Eine noch unbeantwortete Frage ist, ob eine sehr gute Lage zu signifikant höheren Mieten führt als eine gute Lage. Wir testen deshalb die Hypothese

$$H_0 : \beta_2 = \beta_3 \ .$$

Die zugehörige F-Statistik (43) nimmt den Wert

$$f = \frac{\frac{RSS(H_0) - RSS}{1}}{\frac{RSS}{(n-p-1)}} = \frac{\frac{57.545 - 57.453}{57.453}}{\frac{}{799-3-1}} = 1.2717$$

an. Bei allen üblichen Signifikanzniveaus α ist der Testwert zu klein, um die Nullhypothese verwerfen zu können. Es gilt z.B. $f < f_{[0.9;1,795]} = 2.712$. Bei der Interpretation der Schätzwerte von β_2 , β_3 ist also Vorsicht geboten.

```
# OLS-Schätzung des Mietpreismodells

msm.modell <- lm(log(miete)~flaeche+lage.+lage..)
summary(msm.modell)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.0584521  0.0305612 198.240  < 2e-16 ***
flaeche     0.0116935  0.0004455  26.246  < 2e-16 ***
lage.       0.1059089  0.0207395   5.107 4.11e-07 ***
lage..      0.1724369  0.0576373   2.992  0.00286 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.2688 on 795 degrees of freedom
Multiple R-squared: 0.4835,     Adjusted R-squared: 0.4816
F-statistic: 248.1 on 3 and 795 DF,  p-value: < 2.2e-16

sum(msm.modell$res^2)
[1] 57.45262


# OLS-Schätzung des restringierten Modells mit ß2 = ß3

msm.modell.r <- lm(log(miete)~flaeche+I(lage.+lage..))
summary(msm.modell.r)

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.0548497  0.0303990 199.179  < 2e-16 ***
flaeche           0.0117503  0.0004427  26.540  < 2e-16 ***
I(lage. + lage..) 0.1115329  0.0201344   5.539 4.13e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2689 on 796 degrees of freedom
Multiple R-squared: 0.4827,     Adjusted R-squared: 0.4814
F-statistic: 371.4 on 2 and 796 DF,  p-value: < 2.2e-16

sum(msm.modell.r$res^2)
[1] 57.54452
```

```
# F-Test der Hypothese H0: ß2 = ß3

anova(msm.modell.r, msm.modell)

Analysis of Variance Table
Model 1: log(miete) ~ flaeche + I(lage. + lage..)
Model 2: log(miete) ~ flaeche + lage. + lage..

  Res.Df    RSS  Df Sum of Sq      F Pr(>F)
1    796 57.545
2    795 57.453   1     0.092 1.2717 0.2598
```

2.9 Variablenauswahl

Die Festlegung, welche Regressoren $X_1,\ldots,X_p$ zur Erklärung des Regressanden Y benötigt werden (*Variablenauswahl / Variablenselektion*), sollte möglichst aufgrund substanzwissenschaftlicher Erkenntnissen erfolgen. Liegen diese nicht vor oder führen sie zu unbefriedigenden Prognosen, dann bedarf es statistischer Auswahlverfahren basierend auf

- *Hypothesentests* (siehe Abschnitt 2.8) oder
- *Modellauswahlkriterien*.

Modellauswahlkriterien dienen dem Vergleich mehrerer konkurrierender Modelle basierend auf verschiedenen Regressorteilmengen aus einer vorgegebenen Grundmenge potentieller Regressoren. Sie setzen sich i. d. R. aus zwei gegenläufigen Komponenten zusammen:

- Komponente 1 belohnt/bestraft eine gute/schlechte Anpassung (*fit*) des Modells an die Daten,
- Komponente 2 belohnt Modelleinfachheit (*parsimony*) bzw. bestraft Modellkomplexität durch einen *Strafterm*.

Gesucht ist dann ein geeigneter Kompromiss aus *fit* und *parsimony*.

AIC (Information Criterion A oder Akaike Information Criterion)

Das AIC (Hirotugu Akaike 1974) ist ein im Zusammenhang mit der ML-Schätzung vielseitig einsetzbares Modellauswahlkriterium. Es ist allgemein definiert gemäß

$$AIC = -2 \cdot \text{Loglikelihood} + 2 \cdot (\text{Anzahl der zu schätzenden Modellparameter}) \ .$$

In dem hier interessierenden Kontext nimmt es die Form

(44)
$$AIC(M) = -2 \cdot l\left(\hat{\boldsymbol{\beta}}_{ML}, \hat{\sigma}^2_{ML}\right) + 2 \cdot (p+2)$$
$$\propto n \cdot ln\,\hat{\sigma}^2_{ML} + 2 \cdot (p+2) = n \cdot ln\left(\frac{RSS}{n}\right) + 2 \cdot (p+2)$$

an, wobei hier $\hat{\boldsymbol{\beta}}_{ML}$, $\hat{\sigma}^2_{ML}$ die ML-Schätzer der Parameter $\boldsymbol{\beta}$ und σ^2 eines Regressionsmodells M mit $p+1$ Regressionskoeffizienten sind.

Das Modell M wird dem konkurrierenden Modell M* vorgezogen, falls $AIC(M) < AIC(M\,^*)$ gilt.

BIC (Bayesian Information Criterion) / SIC (Schwartz Information Criterion)

Das BIC (Hirotugu Akaike 1974) ist in dem hier interessierenden Kontext definiert gemäß

(45)
$$BIC(M) = -2 \cdot l\left(\hat{\boldsymbol{\beta}}_{ML}, \hat{\sigma}^2_{ML}\right) + ln(n) \cdot (p+2)$$
$$\propto n \cdot ln\,\hat{\sigma}^2_{ML} + ln(n) \cdot (p+2) = n \cdot ln\left(\frac{RSS}{n}\right) + ln(n) \cdot (p+2)\ ,$$

wobei wieder $\hat{\boldsymbol{\beta}}_{ML}$, $\hat{\sigma}^2_{ML}$ die ML-Schätzer der Parameter $\boldsymbol{\beta}$ und σ^2 eines Regressionsmodells M mit $p+1$ Regressionskoeffizienten sind.

$BIC(M)/n$ ist auch als Schwartz Information Criterion (Gideon Schwartz 1978) bekannt:

(46)
$$SIC(M) = ln\,\hat{\sigma}^2_{ML} + \frac{ln(n) \cdot (p+2)}{n} = ln\left(\frac{RSS}{n}\right) + \frac{ln(n) \cdot (p+2)}{n}\ .$$

Das Modell M wird dem konkurrierenden Modell M^* vorgezogen, falls $BIC(M) < BIC(M^*)$ oder äquivalent $SIC(M) < SIC(M^*)$ gilt.

Adjustiertes (korrigiertes) Bestimmtheitsmaß

Das adjustierte Bestimmtheitsmaß besitzt die Form (vgl. 16):

$$(47) \qquad \bar{R}^2(M) = 1 - \frac{\frac{RSS}{n-p-1}}{\frac{TSS}{n-1}} = 1 - \frac{n-1}{n-p-1} \cdot \left(1 - R^2(M)\right).$$

Das Modell M wird dem konkurrierenden Modell M^* vorgezogen, falls $\bar{R}^2(M) > \bar{R}^2(M^*)$ gilt.

Algorithmen der Variablenauswahl

- Alle möglichen Teilmodelle (*all subset regressions*)
- Vorwärtsauswahl (*foreward selection*)
- Rückwärtselimination (*backward selection*)
- Schrittweise Auswahl (*stepwise selection, stepwise regression*)

2.10 Ausblick

Die Ergebnisse in Kapitel 2 setzen die Gültigkeit der Modellannahmen in Definition 2.1 voraus. Verletzungen der Annahmen, insbesondere

- ein *nichtlinearer Zusammenhang* zwischen dem Regressanden Y und den Regressoren $X_1,...,X_p$,
- linear abhängige Regressoren $X_1,...,X_p$ (*Multikollinearität*),
- *heterogene Varianzen* der Störungen $\varepsilon_1,...,\varepsilon_n$ (*Heteroskedastizität*),
- *autokorrelative Beziehungen* zwischen den Störungen $\varepsilon_1,...,\varepsilon_n$ bei Analysen von Zeitreihendaten,
- ggf. *nicht-normalverteilte* Störungen $\varepsilon_1,...,\varepsilon_n$,

beeinflussen die statistischen Eigenschaften der OLS- und ML-Schätzer. Die Schätzer *können* nun substanziell verzerrt und ineffizient werden. Ebenso *können* Konfidenzintervalle (32) – (36) und Tests (37) – (43) invalide Ergebnisse liefern. An die Schätzung der Modellparameter muss sich daher eine gründliche *Diagnose des geschätzten Modells* anschließen, die nach Verletzungen der Modellannahmen sucht (siehe hierzu z.B. Fahrmeir et al. 2009, Kapitel 3).

Anhang

I. Empirische Kovarianzen und Korrelationen

Die empirische Kovarianz ist ein Maß für den statistischen Zusammenhang zweier metrisch skalierter Variabler. Ausgehend von n beobachteten Wertepaaren (x_ν, y_ν) ($\nu = 1,...,n$) der Variablen X und Y ist die *empirische Kovarianz* definiert durch die Gleichung

$$s_{XY} = \frac{1}{n}\sum_{\nu=1}^{n}(x_\nu - \bar{x})(y_\nu - \bar{y}) = \frac{1}{n}\sum_{\nu=1}^{n} x_\nu \cdot y_\nu - \bar{x}\cdot\bar{y}$$

mit

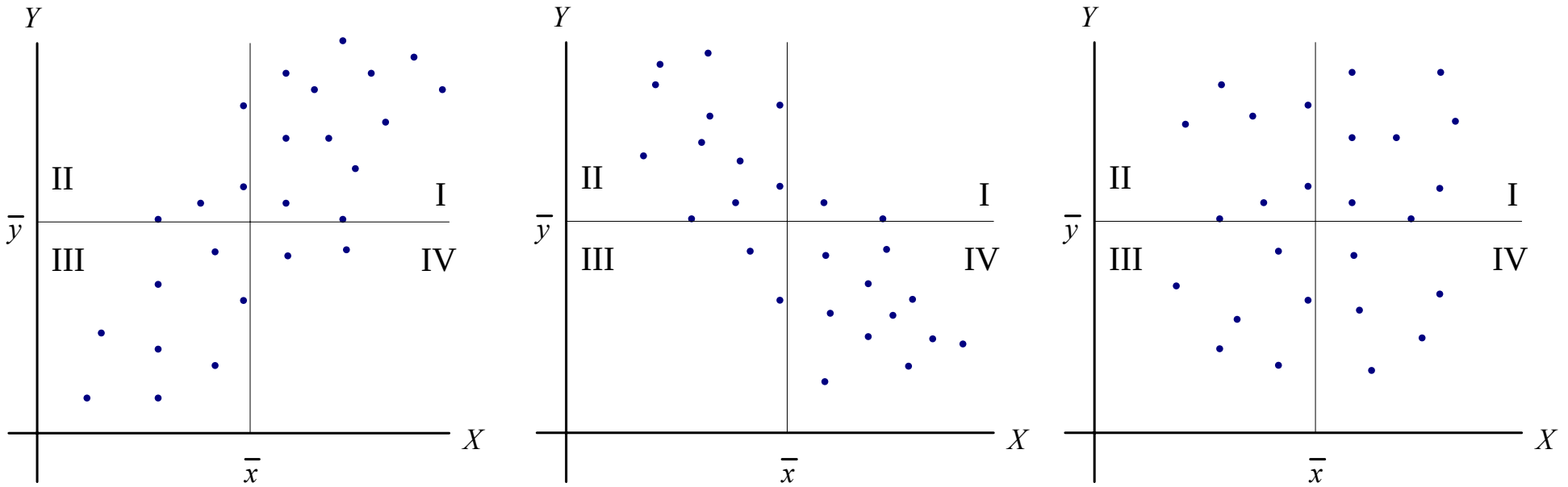
$$\bar{x} = \frac{1}{n}\sum_{\nu=1}^{n} x_\nu \qquad \text{und} \qquad \bar{y} = \frac{1}{n}\sum_{\nu=1}^{n} y_\nu .$$

Sie beschreibt die gemeinsame Variation oder Streuung der Variablen X und Y. Im Gegensatz zu den empirischen Varianzen

$$s_X^2 = \frac{1}{n}\sum_{\nu=1}^{n}(x_\nu - \bar{x})^2 = \frac{1}{n}\sum_{\nu=1}^{n} x_\nu^2 - \bar{x}^2 \qquad \text{und} \qquad s_Y^2 = \frac{1}{n}\sum_{\nu=1}^{n}(y_\nu - \bar{y})^2 = \frac{1}{n}\sum_{\nu=1}^{n} y_\nu^2 - \bar{y}^2$$

der einzelnen Variablen X und Y kann sie auch negative Werte annehmen.

Um diese Maßzahl zu motivieren, sollen drei typische *Streudiagramme* bivariater Datensätze betrachtet werden. Die Streudiagramme wurden jeweils durch ein Fadenkreuz ergänzt, dessen Mittelpunkt der Schwerpunkt oder Zentroid $(\bar{x}, \bar{y})$ des Datensatzes ist. Es entstehen so in jedem Diagramm 4 Quadranten I, II, III und IV.



Typische Streudiagramme bivariater Datensätze (mit Schwerpunkt)

Die ersten beiden Streudiagramme zeigen jeweils eine gemeinsame Tendenz der Beobachtungen von X und der von Y.

Im ersten Diagramm liegen die Punkte hauptsächlich in den Quadranten I und III. Große X-Werte korrespondieren mit großen Y-Werten und kleine X-Werte korrespondieren mit kleinen Y-Werten. Dies soll als *positiver Zusammenhang* bezeichnet werden.

Im zweiten Diagramm liegt ein *negativer Zusammenhang* vor. Die Punkte befinden sich hauptsächlich in den Quadranten II und IV. Große X-Werte korrespondieren mit kleinen Y-Werten und umgekehrt.

Im letzten Diagramm lässt sich *kein Zusammenhang* erkennen. Korrespondierend zu kleinen Werten der einen Variablen wurden sowohl kleine als auch große Werte der anderen Variablen beobachtet. Die Bezeichnungen „klein“ und „groß“ sind immer relativ zum jeweiligen Mittelwert gemeint.

Den Schlüssel zum Verständnis der Kovarianz liefern die Abweichungsprodukte

$$(x_\nu - \bar{x})(y_\nu - \bar{y}) .$$

Liegt ein Datenpaar (x_ν, y_ν) in...	dann gilt...
Quadrant I ,	$x_\nu > \bar{x}, y_\nu > \bar{y} \quad \Rightarrow \quad (x_\nu - \bar{x})(y_\nu - \bar{y}) > 0;$
Quadrant III ,	$x_\nu < \bar{x}, y_\nu < \bar{y} \quad \Rightarrow \quad (x_\nu - \bar{x})(y_\nu - \bar{y}) > 0;$
Quadrant II ,	$x_\nu < \bar{x}, y_\nu > \bar{y} \quad \Rightarrow \quad (x_\nu - \bar{x})(y_\nu - \bar{y}) < 0;$
Quadrant IV ,	$x_\nu > \bar{x}, y_\nu < \bar{y} \quad \Rightarrow \quad (x_\nu - \bar{x})(y_\nu - \bar{y}) < 0.$

- Liegen die Punkte hauptsächlich in den Quadranten I und III, so besteht ein positiver Zusammenhang. s_{XY} ist positiv.
- Liegen die Punkte hauptsächlich in den Quadranten II und IV, so besteht ein negativer Zusammenhang. s_{XY} ist negativ.
- Sind die Punkte gleichmäßig auf alle Quadranten verteilt, so besteht kein Zusammenhang. Positive und negative Produkte heben sich bei der Mittelung weitgehend auf, s_{XY} ist näherungsweise null.

Die Kovarianz s_{XY} zeigt in dem oben skizzierten Sinne den Zusammenhang der Variablen X und Y auf. Mit Hilfe der Kovarianz kann allerdings die *Stärke dieses Zusammenhanges* nur schwer beurteilt werden. Beispielsweise deutet ein sehr großer positiver Wert nicht zwangsläufig auf einen starken positiven Zusammenhang hin. Die Maßzahl ist dimensionsbehaftet. Allein durch die Änderung der Maßeinheiten der Variablen kann sie größer oder kleiner werden.

Dieses Problem kann gelöst werden, wenn man eine *normierte Kovarianz* als Kenngröße verwendet:

$$r_{XY} = \frac{s_{XY}}{s_X \cdot s_Y} \ .$$

r_{XY} heißt *Produkt-Moment-Korrelationskoeffizient* oder *Bravais-Pearson-Korrelationskoeffizient*. r_{XY} ist eine normierte und dimensionslose Maßzahl mit den Eigenschaften:

(1) $-1 \le r_{XY} \le +1$ oder $| r_{XY} | \le 1$ (Normierung)

(2) $r_{XY} = \pm 1$ gilt genau dann, wenn es zwei Konstanten a, b so gibt, dass

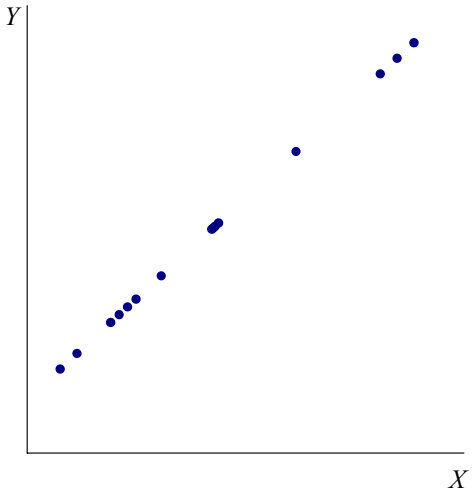
$$y_\nu = a + bx_\nu \ (\nu = 1,...,n)$$

erfüllt ist. Man sagt dann, zwischen X und Y besteht eine exakte lineare Beziehung.

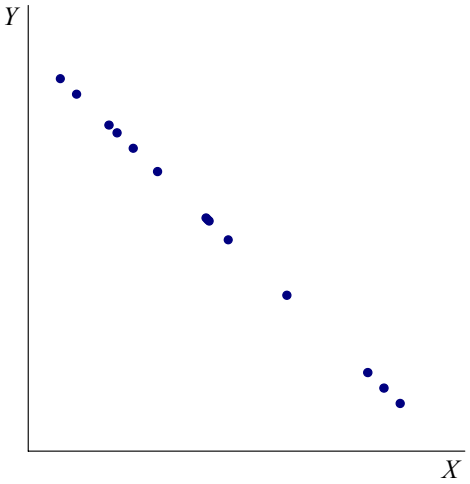
Folgerung

Aufgrund der genannten Eigenschaften des Korrelationskoeffizienten können wir folgern, dass der Korrelationskoeffizient r_{XY} (und damit natürlich auch die Kovarianz) ein *Maß für den linearen Zusammenhang* der Variablen X und Y ist. Gilt $|r_{XY}|=1$, besteht eine exakte lineare Beziehung. Der lineare Zusammenhang ist umso schwächer, je kleiner der Absolutbetrag $|r_{XY}|$ des Koeffizienten ist. Gilt $r_{XY} = 0$, besteht keine lineare Beziehung.

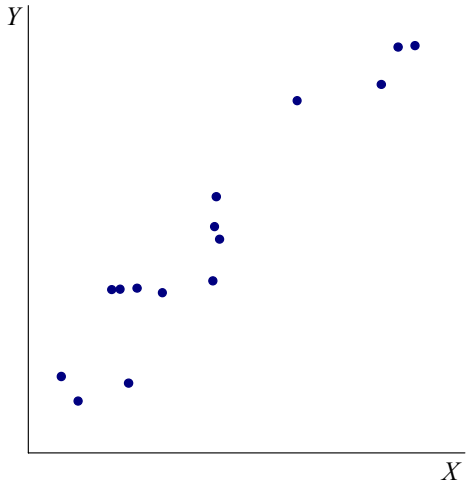
Man sagt, falls ...		ist, dann sind X und Y ...
$r_{XY} = 0$		*unkorreliert*;
$0 < r_{XY} \leq 0.5$	$(0 > r_{XY} \geq -0.5)$	*schwach* positiv (negativ) korreliert;
$0.5 < r_{XY} \leq 0.8$	$(-0.5 > r_{XY} \geq -0.8)$	*mittelstark* positiv (negativ) korreliert;
$0.8 < r_{XY} < 1$	$(-0.8 > r_{XY} > -1)$	*stark* positiv (negativ) korreliert;
$r_{XY} = 1$	$(r_{XY} = -1)$	*perfekt* positiv (negativ) korreliert.



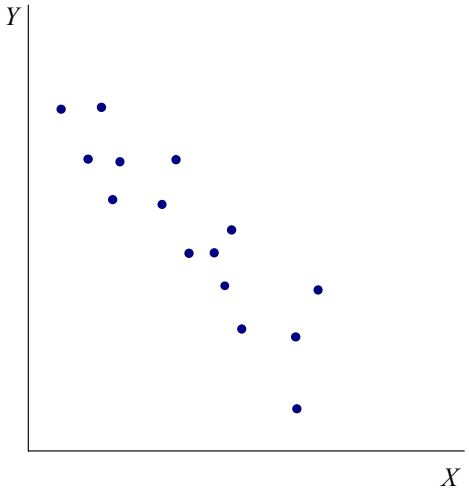
$r_{XY} = +1$



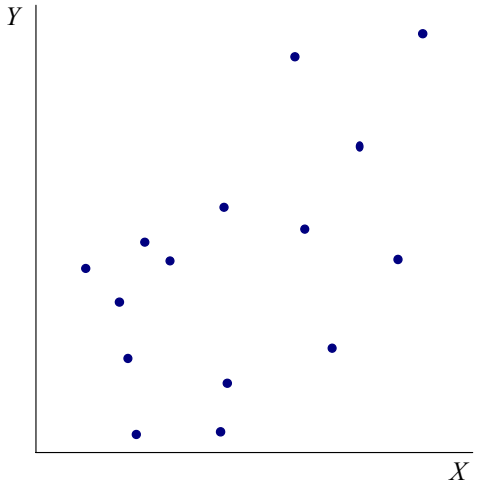
$r_{XY} = -1$



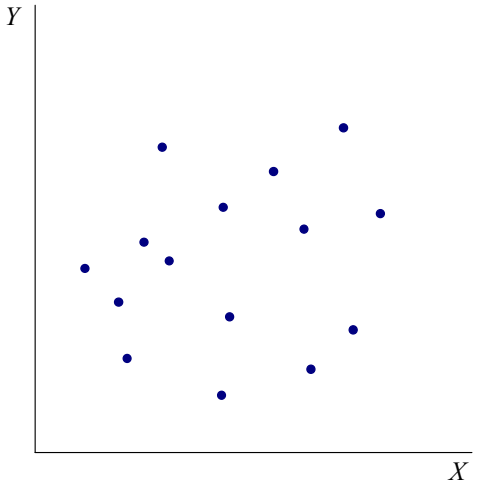
r_{XY} **stark positiv**



r_{XY} stark negativ



r_{XY} schwach positiv



$r_{XY} = 0$

II. Vektordifferentiation / Symbolische Differentiation

Betrachtet sei eine überall zweimal differenzierbare Funktion $f: IR^n \to IR$ der Veränderlichen $x_1, x_2, ..., x_n$ mit $y = f(x_1, x_2, ..., x_n)$. Wir schreiben kurz

$$y = f(\boldsymbol{x}) \quad \text{mit} \quad \boldsymbol{x} = (x_1, x_2, ..., x_n)' \in IR^n .$$

Die partiellen Ableitungen erster Ordnung

$$\frac{\partial f(x_1, x_2, ..., x_n)}{\partial x_i} \qquad (i = 1,2,...,n)$$

können wir in einem Vektor (*Gradient*) zusammenfassen. Den Gradienten schreiben wir symbolisch

$$g(\boldsymbol{x}) = \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = \begin{pmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} \\ \frac{\partial f(\boldsymbol{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\boldsymbol{x})}{\partial x_n} \end{pmatrix} .$$

Die partiellen Ableitungen zweiter Ordnung

$$\frac{\partial^2 f(x_1, x_2, \ldots, x_n)}{\partial x_i \partial x_j} \qquad (i, j = 1,2,\ldots,n)$$

können in einer Matrix zusammengefasst werden (*Hesse-Matrix*). Die Hesse-Matrix schreiben wir symbolisch

$$H(\boldsymbol{x}) = \frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x} \partial \boldsymbol{x}'} = \begin{pmatrix} \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_1} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n \partial x_1} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n \partial x_n} \end{pmatrix}.$$

Diese Schreibvereinfachungen bietet arbeitsökonomische Vorteile.

Beispiele

Konstante Funktion

$$f(x_1,\ldots,x_n)=c\ , \qquad \frac{\partial f(x_1,\ldots,x_n)}{\partial x_i}=0 \quad (i=1,\ldots,n)\ , \qquad \frac{\partial^2 f(x_1,\ldots,x_n)}{\partial x_i \partial x_j}=0 \quad (i,j=1,\ldots,n)$$

Vektorschreibweise

$$f(\boldsymbol{x})=c\ , \qquad \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}=\mathbf{0}\ , \qquad \frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x}\partial \boldsymbol{x}'}=\mathbf{O}$$ mit dem n-Nullvektor $\mathbf{0}$ und der (n,n)-Nullmatrix $\mathbf{O}=\mathbf{00'}$

Lineare Funktion

$$f(x_1,\ldots,x_n)=c_1x_1+\ldots+c_nx_n\ , \qquad \frac{\partial f(x_1,\ldots,x_n)}{\partial x_i}=c_i \quad (i=1,\ldots,n), \qquad \frac{\partial^2 f(x_1,\ldots,x_n)}{\partial x_i \partial x_j}=0 \quad (i,j=1,\ldots,n)$$

Vektorschreibweise

$$f(\boldsymbol{x})=\boldsymbol{c}'\boldsymbol{x}\ , \qquad \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}=\boldsymbol{c}\ , \qquad \frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x}\partial \boldsymbol{x}'}=\mathbf{O}$$ mit $\boldsymbol{c}=(c_1,\ldots,c_n)'\in IR^n$ und der (n,n)-Nullmatrix $\mathbf{O}'$

Quadratische Funktion

$$f(x_1,\ldots,x_n)=x_1^2+\ldots+x_n^2\ , \qquad \frac{\partial f(x_1,\ldots,x_n)}{\partial x_i}=2x_i \quad (i=1,\ldots,n)\ , \qquad \frac{\partial^2 f(x_1,\ldots,x_n)}{\partial x_i \partial x_j}=\begin{cases}2 & \text{falls } i=j\\ 0 & \text{falls } i\neq j\end{cases}$$

Vektorschreibweise

$$f(\boldsymbol{x})=\boldsymbol{x}'\boldsymbol{x}\ , \qquad \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}=2\boldsymbol{x}\ , \qquad \frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x}\partial \boldsymbol{x}'}=2\boldsymbol{I} \quad \text{mit der } (n,n)\text{-Einheitsmatrix } \boldsymbol{I}$$

Quadratische Funktion

$$f(x_1,\ldots,x_n)=\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}x_i x_j\ ,$$

$$\frac{\partial f(x_1,\ldots,x_n)}{\partial x_i}=\sum_{j=1}^{n}(a_{ij}+a_{ji})x_j \quad (i=1,\ldots,n)\ , \qquad \frac{\partial^2 f(x_1,\ldots,x_n)}{\partial x_i \partial x_j}=a_{ij}+a_{ji} \quad (i,j=1,\ldots,n)$$

Vektorschreibweise

$$f(\boldsymbol{x})=\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}\,, \qquad \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}=(\boldsymbol{A}+\boldsymbol{A}')\boldsymbol{x}\,, \qquad \frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x}\partial \boldsymbol{x}'}=\boldsymbol{A}+\boldsymbol{A}' \quad \text{mit der } (n,n)\text{-Matrix } \boldsymbol{A}=(a_{ij})$$

Quadratische Form

$$f(x_1, x_2, ..., x_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j \quad \text{mit} \quad a_{ij} = a_{ji} \ (i, j = 1,2,...,n)$$

$$\frac{\partial f(x_1, x_2, ..., x_n)}{\partial x_i} = 2 \sum_{j=1}^{n} a_{ij} x_j \quad (i = 1,2,...,n), \qquad \frac{\partial^2 f(x_1, ..., x_n)}{\partial x_i \partial x_j} = 2 a_{ij} \quad (i, j = 1, ..., n)$$

Vektorschreibweise

$$f(\boldsymbol{x}) = \boldsymbol{x}' \boldsymbol{A} \boldsymbol{x}, \qquad \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = 2 \boldsymbol{A} \boldsymbol{x}, \qquad \frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x} \partial \boldsymbol{x}'} = 2 \boldsymbol{A} \quad \text{mit} \quad \boldsymbol{A} = (a_{ij}) \text{ und } \boldsymbol{A} = \boldsymbol{A}'$$

III. Bestimmung lokaler Extrema

Betrachtet sei eine überall zweimal differenzierbare Funktion $f : IR^n \to IR$ der Veränderlichen $\boldsymbol{x} = (x_1, ..., x_n)' \in IR^n$ mit $y = f(\boldsymbol{x})$. Gesucht sind mögliche lokale (relative) Extremstellen $\boldsymbol{x}_0$ der Funktion ohne Nebenbedingung.

Notwendige Bedingung für ein Extremum

Besitzt die Funktion f in der Stelle $\boldsymbol{x}_0$ ein lokales Extremum, dann gilt

$$\boldsymbol{g}(\boldsymbol{x}_0) = \boldsymbol{0} \quad \text{mit} \quad \boldsymbol{g}(\boldsymbol{x}_0) = \frac{\partial f(\boldsymbol{x}_0)}{\partial \boldsymbol{x}} = \left.\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}=\boldsymbol{x}_0} .$$

Hinreichende Bedingung für ein Minimum / Maximum

Der Gradient $\boldsymbol{g}$ der Funktion f sei in der Stelle $\boldsymbol{x}_0$ ein Nullvektor, also $\boldsymbol{g}(\boldsymbol{x}_0) = \boldsymbol{0}$. Dann besitzt die Funktion in $\boldsymbol{x}_0$ ein lokales Minimum (Maximum), wenn die Hesse-Matrix in der Stelle $\boldsymbol{x}_0$

$$\boldsymbol{H}(\boldsymbol{x}_0) = \frac{\partial^2 f(\boldsymbol{x}_0)}{\partial \boldsymbol{x} \partial \boldsymbol{x}'} = \left.\frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x} \partial \boldsymbol{x}'}\right|_{\boldsymbol{x}=\boldsymbol{x}_0}$$

positiv definit (negative definit) ist. $\boldsymbol{H}(\boldsymbol{x}_0)$ heißt positiv definit (negative definit), falls die quadratische Form $\boldsymbol{c}'\boldsymbol{H}(\boldsymbol{x}_0)\boldsymbol{c}$ für alle Vektoren $\boldsymbol{c} \in IR^n$ mit $\boldsymbol{c} \neq \boldsymbol{0}$ positiv (negativ) ist:

$$\boldsymbol{c}'\boldsymbol{H}(\boldsymbol{x}_0)\boldsymbol{c} > 0 \qquad (\boldsymbol{c}'\boldsymbol{H}(\boldsymbol{x}_0)\boldsymbol{c} < 0) .$$

Definite Matrizen

Die quadratische Form $\boldsymbol{c}'\boldsymbol{A}\boldsymbol{c}$ und die symmetrische (n,n)-Matrix $\boldsymbol{A}$ heißen

1. *positiv definit*, falls $\boldsymbol{c}'\boldsymbol{A}\boldsymbol{c} > 0$ für alle Vektoren $\boldsymbol{c} \in IR^n$ mit $\boldsymbol{c} \neq \boldsymbol{0}$.
2. *positiv semidefinit*, falls $\boldsymbol{c}'\boldsymbol{A}\boldsymbol{c} \geq 0$ und $\boldsymbol{c}'\boldsymbol{A}\boldsymbol{c} = 0$ für mindestens ein $\boldsymbol{c} \neq \boldsymbol{0}$.
3. *nichtnegativ definit*, falls $\boldsymbol{c}'\boldsymbol{A}\boldsymbol{c}$ bzw. $\boldsymbol{A}$ entweder positiv oder positiv semidefinit ist.
4. *negativ definit*, wenn $-\boldsymbol{A}$ positiv definit ist.
5. *negativ semidefinit*, wenn $-\boldsymbol{A}$ positiv semidefinit ist.
6. *indefinit* in allen anderen Fällen.

IV. Inverse Matrix

Es sei $\boldsymbol{A}$ eine quadratische (n,n)-Matrix und $\boldsymbol{I}$ die (n,n)-Einheitsmatrix. Die Matrix $\boldsymbol{A}^{-1}$ heißt *inverse Matrix* von $\boldsymbol{A}$, wenn gilt:

$$\boldsymbol{A} \cdot \boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I} \quad .$$

Die Matrix $\boldsymbol{A}$ heißt invertierbar (*nichtsingulär, regulär*), wenn eine zu $\boldsymbol{A}$ inverse Matrix existiert.

Eine quadratische (*n,n*)-Matrix $\boldsymbol{A}$ ist invertierbar, wenn $rg(\boldsymbol{A}) = n \Leftrightarrow det(\boldsymbol{A}) \neq 0$ erfüllt ist.

Es sei $\boldsymbol{A}$ eine invertierbare (*n,n*)-Matrix. Dann gilt:

$$\boldsymbol{A}^{-1} = \frac{1}{det(\boldsymbol{A})} \cdot adj(\boldsymbol{A}) \ .$$

Dabei ist $adj(\boldsymbol{A}) = \boldsymbol{B} = \left(b_{ij}\right)_{(n,n)}$ die Adjungierte von $\boldsymbol{A}$ mit den Komponenten $b_{ij} = (-1)^{i+j} \cdot det\left(\boldsymbol{A}_{ij}\right)$.
$\boldsymbol{A}_{ij}$ entsteht durch Streichen der *i*-ten Zeile und der *j*-ten Spalte aus $\boldsymbol{A}$.

Im Falle von (2,2)- und (3,3)-Matrizen gilt:

$$(\boldsymbol{A})^{-1} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - cb} \cdot \begin{pmatrix} d & -b \\ -c & a \end{pmatrix},$$

$$(\boldsymbol{A})^{-1} = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}^{-1} = \frac{1}{a(ei - hf) + d(bi - hc) + g(bf - ec)} \cdot \begin{pmatrix} ei - fh & ch - bi & bf - ce \\ fg - di & ai - cg & cd - af \\ dh - eg & bg - ah & ae - bd \end{pmatrix}.$$

V. Maximum-Likelihood-Methode

Es sei $X_1,...,X_n$ eine einfache Zufallsstichprobe aus einer Grundgesamtheit X. Die Stichprobenvariablen $X_1,...,X_n$ sind stochastisch unabhängig und alle identisch wie X verteilt. Dabei kann X diskret oder stetig verteilt sein mit der Verteilungsfunktion $F_X(x \mid \theta)$ und der Wahrscheinlichkeits- bzw. Dichtefunktion $f_X(x \mid \theta)$. Wir nehmen an, dass die Verteilung von X mit Ausnahme eines Parameters θ bekannt ist, wobei θ auch ein Vektor sein darf.

Aufgrund der Unabhängigkeit und identischen Verteilung der Stichprobenvariablen lautet ihre gemeinsame Wahrscheinlichkeits- oder Dichtefunktion *e*infach

$$f_{X_1,...,X_n}(x_1,...,x_n \mid \theta) = f_X(x_1 \mid \theta) \cdot ... \cdot f_X(x_n \mid \theta) = \prod_{\nu=1}^{n} f_X(x_\nu \mid \theta) .$$

$x_1,...,x_n$ stehe jetzt für eine konkret beobachtete Stichprobe. Obige Funktion gibt dann die Wahrscheinlichkeit bzw. Dichte (”*likelihood*”) dieser Stichprobe für alle möglichen Werte von θ an. Die Funktion fassen wir als Funktion von θ auf und schreiben sie

$$L(\theta \mid x_1,...,x_n) = \prod_{\nu=1}^{n} f_X(x_\nu \mid \theta) \quad \text{oder kurz} \quad L(\theta) .$$

Die Funktion

$$L(\theta) = L(\theta \mid x_1,...,x_n) = \prod_{\nu=1}^{n} f_X(x_\nu \mid \theta)$$

wird *Likelihoodfunktion der konkreten Stichprobe* $x_1,...,x_n$ genannt. Gilt in der Stelle $\theta = \hat{\theta}$

$$L(\hat{\theta} \mid x_1,...,x_n) \geq L(\theta \mid x_1,...,x_n)$$

für alle möglichen Werte von θ, so ist

$$\hat{\theta} = T(x_1,\cdots,x_n)$$

der *Maximum-Likelihood-(ML)-Schätzwert* für θ. Die zugehörigen zufälligen Stichprobenfunktionen

$$L(\theta \mid X_1,...,X_n) \qquad \text{und} \qquad \hat{\Theta} = T(X_1,\cdots,X_n)$$

heißen *Likelihoodfunktion der einfachen Zufallsstichprobe* $X_1,...,X_n$ sowie *Maximum-Likelihood-(ML)-Schätzer* für θ.

Zur praktischen Bestimmung von ML-Schätzern bedient man sich häufig (wenn auch nicht immer) der Differentialrechnung. Hierbei kann man sich zunutze machen, dass $L(\theta \mid x_1,...,x_n)$ genau dort ein Maximum besitzt, wo auch die logarithmierte Likelihoodfunktion oder *Loglikelihoodfunktion*

$$l(\theta \mid x_1,...,x_n) := ln\, L(\theta \mid x_1,...,x_n) = ln \prod_{\nu=1}^{n} f_X(x_\nu \mid \theta) = \sum_{\nu=1}^{n} ln\, f_X(x_\nu \mid \theta)$$

ein Maximum besitzt. Eine Summe ist i.d.R. leichter zu differenzieren als ein Produkt. Notwendige Bedingung für eine Maximumstelle $\hat{\theta}$ ist, dass die erste Ableitung der Loglikelihoodfunktion verschwindet. Wir erhalten den ML-Schätzwert $\hat{\theta}$ als Lösung der *Likelihood-Gleichung*

$$\frac{dl(\theta \mid x_1,...,x_n)}{d\theta} = \sum_{\nu=1}^{n} \frac{d\, ln\, f_X(x_\nu \mid \theta)}{d\theta} = 0 \ .$$

Im Falle eines *m*-Parametervektors $\theta = (\theta_1,...,\theta_m)'$ ist der ML-Schätzwert $\hat{\theta}$ durch Lösen eines *Likelihood-Gleichungssystems* zu bestimmen:

$$\frac{\partial l(\theta \mid x_1,...,x_n)}{\partial \theta_i} = \sum_{\nu=1}^{n} \frac{\partial\, ln\, f_X(x_\nu \mid \theta)}{\partial \theta_i} = 0 \qquad\qquad i = 1,...,m \ .$$

Rechnen mit Logarithmen

(1) $ln\,1 = 0$

(2) $ln\,e = 1$

(3) $ln(x \cdot y) = ln\,x + ln\,y$

(4) $ln\,x^y = y \cdot ln\,x$

(5) $ln\,\frac{x}{y} = ln\,x - ln\,y$ speziell: $ln\,\frac{1}{y} = -ln\,y$ wegen $ln\,1 = 0$

(6) $f(x) = ln\,x \quad \Rightarrow \quad f'(x) = \frac{1}{x}$

VI. Literatur

Eckey, H.-F., Kosfeld, R. & C. Dreger: Ökonometrie, Grundlagen, Methoden, Beispiele, 3. Aufl.; Gabler 2004

Fahrmeir, L., Kneib, T. & S. Lang: Regression, 2. Aufl.; Springer 2009, Kapitel 2 und 3

Green, W.H.: Econometric analysis, 6th ed.; Prentice-Hall 2008, Chapter 1-5 und 16

Gujarati, D.N.: Basic econometrics, 4th ed.; McGraw-Hill 2003, Part I

Rinne, H.: Taschenbuch der Statistik, 4. Aufl.; Harri Deutsch 2008, Teil D1