

## **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)

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

zwischen denen ein (zumeist a priori unbekannter) funktionaler Zusammenhang

$$(1) \quad Y = f(X_1, \dots, X_p) + \varepsilon$$

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

$$f(X_1, \dots, X_p) = E(Y | X_1, \dots, X_p)$$

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  $Y, X_1, \dots, X_p, \varepsilon$  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

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

so dass der Zusammenhang

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

gilt. Die Konstanten  $\beta_0, \beta_1, \dots, \beta_p$  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

$$\text{Konsum} = \beta_0 + \beta_1 \cdot \text{Einkommen}$$

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

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

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  $\varepsilon$  überlagert wird:

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

## Skalenniveau der Variablen

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

Im Falle metrischer Variablen lassen wir auch Transformationen

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

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  $1, \dots, k$  werden durch  $m = k - 1$  sogenannte *Dummy-Variablen*

$$X_i^{(1)}, \dots, X_i^{(m)}$$

kodiert.

*Kodierung kategorialer Variablen  $V_i$* (a) *Binär-Kodierung* (binäre Dummies)

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

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

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

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

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

### 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  $\log(\text{Miete})$  ausüben, dann können wir dies z.B. wie folgt berücksichtigen:

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

bzw.

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

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



## Daten

Zum Zwecke der Modellschätzung beobachten wir die Variablen

$$Y, X_1, \dots, X_p$$

$n$ -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:

$$y_v, x_{v1}, \dots, x_{vp} \quad (v = 1, \dots, n) .$$

Im Falle von Querschnittsdaten kennzeichnet der Index  $v$  die statistischen Einheiten, an denen die Werte gemessen wurden. Liegen Zeitreihen vor, dann ist  $v$  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  $y_v, x_{v1}, \dots, x_{vp}$  ( $v = 1, \dots, n$ ) erfüllen die  $n$  Gleichungen

$$(3) \quad y_v = \beta_0 + \beta_1 x_{v1} + \dots + \beta_p x_{vp} + \varepsilon_v \quad (v = 1, \dots, n),$$

wobei die Störungen  $\varepsilon_v$  nicht direkt beobachtbar sind. Mit

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

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

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

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  $x_{v1}, \dots, x_{vp}$  *als deterministische Größen* betrachten. Die  $(n, p+1)$ -Regressormatrix  $\mathbf{X}$  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  $rg(\mathbf{X}) = p+1$  der Regressormatrix.

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

$$E(\varepsilon_v) = 0 ,$$

$$Var(\varepsilon_v) = E(\{\varepsilon_v - E(\varepsilon_v)\}^2) = E(\varepsilon_v^2) = \sigma^2 = \text{const.} \quad (\text{Homoskedastizität}) ,$$

$$Cov(\varepsilon_v, \varepsilon_\kappa) = E(\{\varepsilon_v - E(\varepsilon_v)\}\{\varepsilon_\kappa - E(\varepsilon_\kappa)\}) = E(\varepsilon_v \varepsilon_\kappa) = 0 \quad (\text{Unkorreliertheit}) .$$

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  *$n$ -Erwartungswertvektor*

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

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

$$\begin{aligned} \text{Cov}(\boldsymbol{\varepsilon}) &= E\left\{(\boldsymbol{\varepsilon} - E(\boldsymbol{\varepsilon}))\{\boldsymbol{\varepsilon} - E(\boldsymbol{\varepsilon})\}'\right\} = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = E \begin{pmatrix} \varepsilon_1\varepsilon_1 & \varepsilon_1\varepsilon_2 & \cdots & \varepsilon_1\varepsilon_n \\ \varepsilon_2\varepsilon_1 & \varepsilon_2\varepsilon_2 & \cdots & \varepsilon_2\varepsilon_n \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_n\varepsilon_1 & \varepsilon_n\varepsilon_2 & \cdots & \varepsilon_n\varepsilon_n \end{pmatrix} \\ &= \begin{pmatrix} E(\varepsilon_1^2) & E(\varepsilon_1\varepsilon_2) & \cdots & E(\varepsilon_1\varepsilon_n) \\ E(\varepsilon_2\varepsilon_1) & E(\varepsilon_2^2) & \cdots & E(\varepsilon_2\varepsilon_n) \\ \vdots & \vdots & \ddots & \vdots \\ E(\varepsilon_n\varepsilon_1) & E(\varepsilon_n\varepsilon_2) & \cdots & E(\varepsilon_n^2) \end{pmatrix} = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} = \sigma^2 \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} = \sigma^2 \mathbf{I} \end{aligned}$$

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

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

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

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

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

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

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

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)  $y = X\beta + \varepsilon$  (Linearität).

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

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

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

(A4) Zusatzannahme (optional)

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

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

$$konsum_v = \beta_0 + \beta_1 \cdot einkommen_v + \varepsilon_v \quad (v = 1, \dots, n).$$

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

$$E(Konsum | Einkommen = einkommen_v) = \beta_0 + \beta_1 \cdot einkommen_v .$$

Die Störungen  $\varepsilon_v$  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 Variabilität des Konsums konstant:

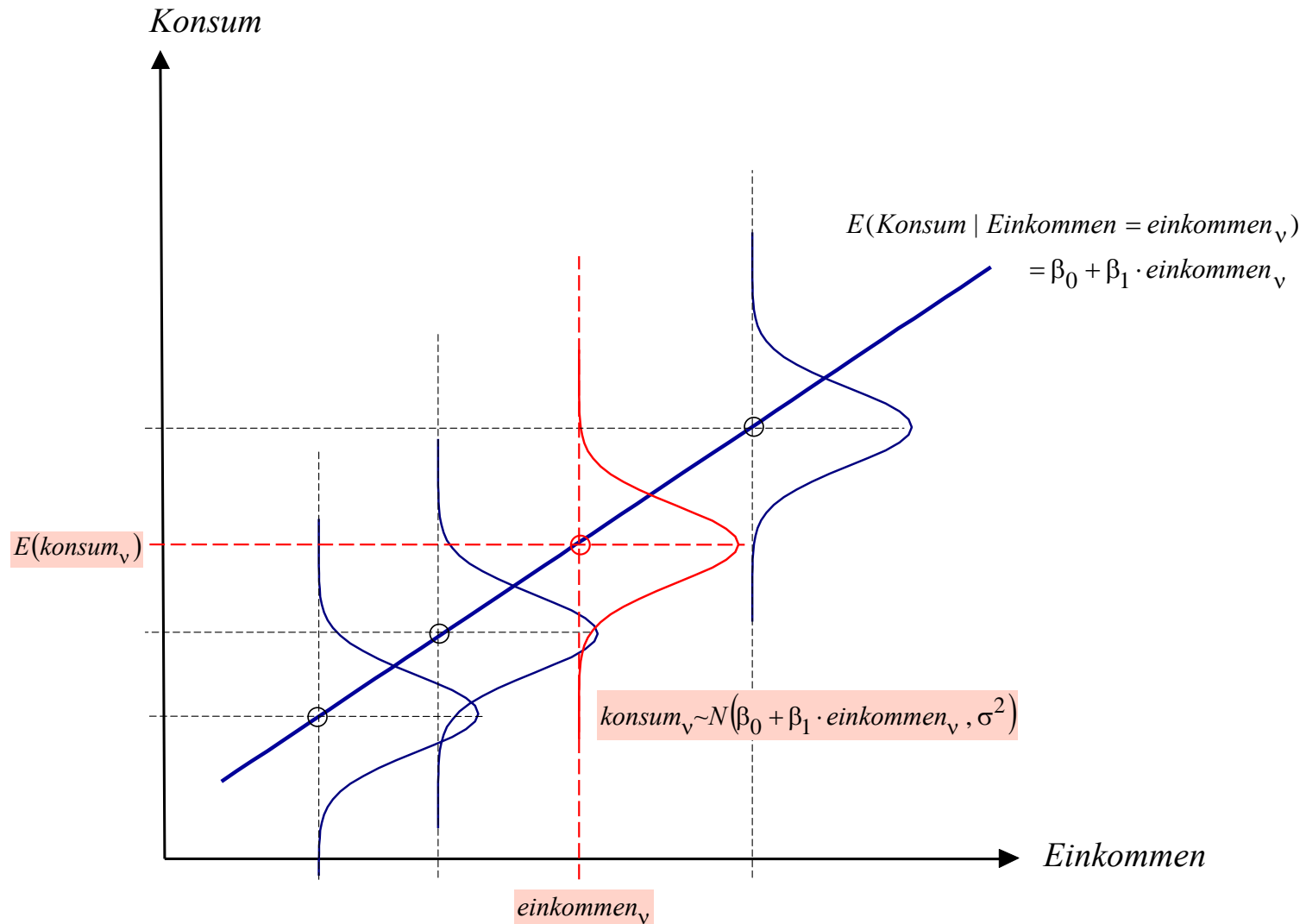
$$\text{Var}(\textit{konsum}_v) = \text{Var}(\beta_0 + \beta_1 \cdot \textit{einkommen}_v + \varepsilon_v) = \text{Var}(\varepsilon_v) = \sigma^2 \quad (v = 1, \dots, n).$$

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  $\textit{konsum}_v$  nicht nur vom Periodeneinkommen  $\textit{einkommen}_v$ , sondern auch vom Konsumniveau der Vorperiode abhängt, d. h.

$$\textit{konsum}_v = \beta_0 + \beta_1 \cdot \textit{einkommen}_v + \beta_2 \cdot \textit{konsum}_{v-1} + \varepsilon_v \quad (v = 1, \dots, n),$$

dann ist der *Regressor*  $\textit{konsum}_{v-1}$  *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 unbekanntem Modellparameter, die es im Folgenden auf der Datenbasis  $y_v, x_{v1}, \dots, x_{vp}$  ( $v=1, \dots, n$ ) 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

$$(5a) \quad Q(\boldsymbol{\beta}) = \sum_{v=1}^n \varepsilon_v^2 = \sum_{v=1}^n (y_v - \beta_0 - \beta_1 x_{v1} - \dots - \beta_p x_{vp})^2 = \sum_{v=1}^n (y_v - \mathbf{x}'_v \boldsymbol{\beta})^2,$$

wobei  $\mathbf{x}'_v = (x_{v1}, \dots, x_{vp})$  die  $v$ -te Zeile der Regressormatrix  $\mathbf{X}$  bezeichnet. In Matrixschreibweise notieren wir die Quadratsummenfunktion gemäß

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

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

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

$$Q(\boldsymbol{\beta}) = \sum_{v=1}^n (y_v - \beta_0 - \beta_1 x_{v1} - \dots - \beta_p x_{vp})^2$$

lauten

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_0} &= \sum_{v=1}^n 2(y_v - \beta_0 - \beta_1 x_{v1} - \dots - \beta_p x_{vp}) \cdot (-1) = -2 \sum_{v=1}^n (y_v - \beta_0 - \beta_1 x_{v1} - \dots - \beta_p x_{vp}) \\ \frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_1} &= \sum_{v=1}^n 2(y_v - \beta_0 - \beta_1 x_{v1} - \dots - \beta_p x_{vp}) \cdot (-x_{v1}) = -2 \sum_{v=1}^n (y_v x_{v1} - \beta_0 x_{v1} - \beta_1 x_{v1} x_{v1} - \dots - \beta_p x_{v1} x_{vp}) \\ &\vdots \\ \frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_p} &= \sum_{v=1}^n 2(y_v - \beta_0 - \beta_1 x_{v1} - \dots - \beta_p x_{vp}) \cdot (-x_{vp}) = -2 \sum_{v=1}^n (y_v x_{vp} - \beta_0 x_{vp} - \beta_1 x_{v1} x_{vp} - \dots - \beta_p x_{vp} x_{vp}). \end{aligned}$$

Wir vernachlässigen den Faktor 2 und setzen die Ableitungen gleich Null

$$\begin{aligned} \sum_{v=1}^n (-y_v + \hat{\beta}_0 + \hat{\beta}_1 x_{v1} + \dots + \hat{\beta}_p x_{vp}) & \stackrel{!}{=} 0 \\ \sum_{v=1}^n (-y_v x_{v1} + \hat{\beta}_0 x_{v1} + \hat{\beta}_1 x_{v1} x_{v1} + \dots + \hat{\beta}_p x_{v1} x_{vp}) & \stackrel{!}{=} 0 \\ & \vdots \\ \sum_{v=1}^n (-y_v x_{vp} + \hat{\beta}_0 x_{vp} + \hat{\beta}_1 x_{v1} x_{vp} + \dots + \hat{\beta}_p x_{vp} x_{vp}) & \stackrel{!}{=} 0 . \end{aligned}$$

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

$$(6) \quad \begin{array}{ccccccc} \hat{\beta}_0 \sum_{v=1}^n 1 & + & \hat{\beta}_1 \sum_{v=1}^n x_{v1} & + \dots + & \hat{\beta}_p \sum_{v=1}^n x_{vp} & = & \sum_{v=1}^n y_v \\ \hat{\beta}_0 \sum_{v=1}^n x_{v1} & + & \hat{\beta}_1 \sum_{v=1}^n x_{v1} x_{v1} & + \dots + & \hat{\beta}_p \sum_{v=1}^n x_{v1} x_{vp} & = & \sum_{v=1}^n y_v x_{v1} \\ & & \vdots & & \vdots & & \vdots \\ \hat{\beta}_0 \sum_{v=1}^n x_{vp} & + & \hat{\beta}_1 \sum_{v=1}^n x_{v1} x_{vp} & + \dots + & \hat{\beta}_p \sum_{v=1}^n x_{vp} x_{vp} & = & \sum_{v=1}^n y_v x_{vp} . \end{array}$$

Die Normalgleichungen

$$(6) \quad \begin{array}{ccccccc} \hat{\beta}_0 \sum_{v=1}^n 1 & + & \hat{\beta}_1 \sum_{v=1}^n x_{v1} & + \dots + & \hat{\beta}_p \sum_{v=1}^n x_{vp} & = & \sum_{v=1}^n y_v \\ \hat{\beta}_0 \sum_{v=1}^n x_{v1} & + & \hat{\beta}_1 \sum_{v=1}^n x_{v1}x_{v1} & + \dots + & \hat{\beta}_p \sum_{v=1}^n x_{v1}x_{vp} & = & \sum_{v=1}^n y_v x_{v1} \\ & \vdots & & \vdots & & & \vdots \\ \hat{\beta}_0 \sum_{v=1}^n x_{vp} & + & \hat{\beta}_1 \sum_{v=1}^n x_{v1}x_{vp} & + \dots + & \hat{\beta}_p \sum_{v=1}^n x_{vp}x_{vp} & = & \sum_{v=1}^n y_v x_{vp} \end{array}$$

können wir kompakt als eine Matrixgleichung notieren:

$$\begin{pmatrix} \sum_{v=1}^n 1 & \sum_{v=1}^n x_{v1} & \cdots & \sum_{v=1}^n x_{vp} \\ \sum_{v=1}^n x_{v1} & \sum_{v=1}^n x_{v1}x_{v1} & \cdots & \sum_{v=1}^n x_{v1}x_{vp} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{v=1}^n x_{vp} & \sum_{v=1}^n x_{v1}x_{vp} & \cdots & \sum_{v=1}^n x_{vp}x_{vp} \end{pmatrix} \cdot \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} = \begin{pmatrix} \sum_{v=1}^n y_v \\ \sum_{v=1}^n y_v x_{v1} \\ \vdots \\ \sum_{v=1}^n y_v x_{vp} \end{pmatrix}.$$

Mit

$$\begin{pmatrix} \sum_{v=1}^n 1 & \sum_{v=1}^n x_{v1} & \cdots & \sum_{v=1}^n x_{vp} \\ \sum_{v=1}^n x_{v1} & \sum_{v=1}^n x_{v1}x_{v1} & \cdots & \sum_{v=1}^n x_{v1}x_{vp} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{v=1}^n x_{vp} & \sum_{v=1}^n x_{v1}x_{vp} & \cdots & \sum_{v=1}^n x_{vp}x_{vp} \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{pmatrix} \cdot \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} = \mathbf{X}'\mathbf{X}$$

$$\begin{pmatrix} \sum_{v=1}^n y_v \\ \sum_{v=1}^n y_v x_{v1} \\ \vdots \\ \sum_{v=1}^n y_v x_{vp} \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \mathbf{X}'\mathbf{y} \quad \text{und} \quad \hat{\boldsymbol{\beta}}_{OLS} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}$$



folgt

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

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

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

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

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

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

ist

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

Setzen wir die Ableitungen gleich Null

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

folgt die Matrix-Normalgleichung

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

mit der Lösung

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

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

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

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

$$\mathbf{c}'\mathbf{X}'\mathbf{X}\mathbf{c} > 0 \quad \text{für alle Vektoren } \mathbf{c} \in \mathbb{R}^{p+1} \text{ mit } \mathbf{c} \neq \mathbf{0} .$$

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

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

wegen

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

Ist nun

$$c'X'Xc > 0 \quad \text{für alle } c = \hat{\beta} - \beta \in \mathbb{R}^{p+1} \text{ mit } c \neq \mathbf{0},$$

folgt stets

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

### S-2.1 Eigenschaften der Kreuzproduktmatrix

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

(b)  $rg(X) = p + 1 \Rightarrow rg(X'X) = p + 1$

$\Rightarrow X'X$  regulär

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

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

(ohne Beweis)

### Folgerung

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

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

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

Die Funktionswerte der geschätzten Regressionsfunktion

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

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

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

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

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

Die Summe nennen wir im Folgenden *Residualquadratsumme*.

### B-2.3 Lineare Einfachregression.

Beispielhaft betrachten wir den Spezialfall der Einfachregression

$$y_v = \beta_0 + \beta_1 x_v + \varepsilon_v \quad (v = 1, \dots, n)$$

bzw.

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

Hier gilt

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \cdot \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n & \sum_{v=1}^n x_v \\ \sum_{v=1}^n x_v & \sum_{v=1}^n x_v^2 \end{pmatrix}$$

und

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_{v=1}^n y_v \\ \sum_{v=1}^n x_v y_v \end{pmatrix}.$$

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

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

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

$$\hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y} \quad \Rightarrow \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{mit} \quad \bar{x} = \frac{1}{n} \sum_{v=1}^n x_v \quad \text{und} \quad \bar{y} = \frac{1}{n} \sum_{v=1}^n y_v.$$

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

$$(\bar{y} - \hat{\beta}_1 \bar{x}) \sum_{v=1}^n x_v + \hat{\beta}_1 \sum_{v=1}^n x_v^2 = \sum_{v=1}^n y_v x_v \quad \Rightarrow \quad \hat{\beta}_1 = \frac{\sum_{v=1}^n x_v y_v - n \bar{x} \bar{y}}{\sum_{v=1}^n x_v^2 - n \bar{x}^2} = \frac{\sum_{v=1}^n (x_v - \bar{x})(y_v - \bar{y})}{\sum_{v=1}^n (x_v - \bar{x})^2} \quad \boxtimes$$

### B-2.1 Makroökonomische Konsumfunktion.

Wir wollen nun eine Keynesianische Konsumfunktion

$$\textit{konsum}_v = \beta_0 + \beta_1 \cdot \textit{einkommen}_v + \varepsilon_v \quad (v = 1, \dots, n)$$

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:

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



Jahr	Einkommen $x_v$ in Mrd. DM	Konsum $y_v$ in Mrd. DM	$x_v^2$	$x_v \cdot y_v$
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

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum_{v=1}^n x_v \\ \sum_{v=1}^n x_v & \sum_{v=1}^n x_v^2 \end{pmatrix} = \begin{pmatrix} 19 & 22167.78 \\ 22167.78 & 26249285.12 \end{pmatrix}$$

und

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} \sum_{v=1}^n y_v \\ \sum_{v=1}^n x_v y_v \end{pmatrix} = \begin{pmatrix} 19866.93 \\ 23512473.03 \end{pmatrix}.$$

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

$$\begin{aligned} (\mathbf{X}'\mathbf{X})^{-1} &= \frac{1}{\det(\mathbf{X}'\mathbf{X})} \text{adj}(\mathbf{X}'\mathbf{X}) = \frac{1}{19 \cdot 26249285.12 - 22167.78^2} \begin{pmatrix} 26249285.12 & -22167.78 \\ -22167.78 & 19 \end{pmatrix} \\ &= \frac{1}{7325947} \begin{pmatrix} 26249285.12 & -22167.78 \\ -22167.78 & 19 \end{pmatrix} = \begin{pmatrix} 3.58306 & -0.00303 \\ -0.00303 & 2.59352\text{E}-06 \end{pmatrix}. \end{aligned}$$

Einsetzen in (8) liefert

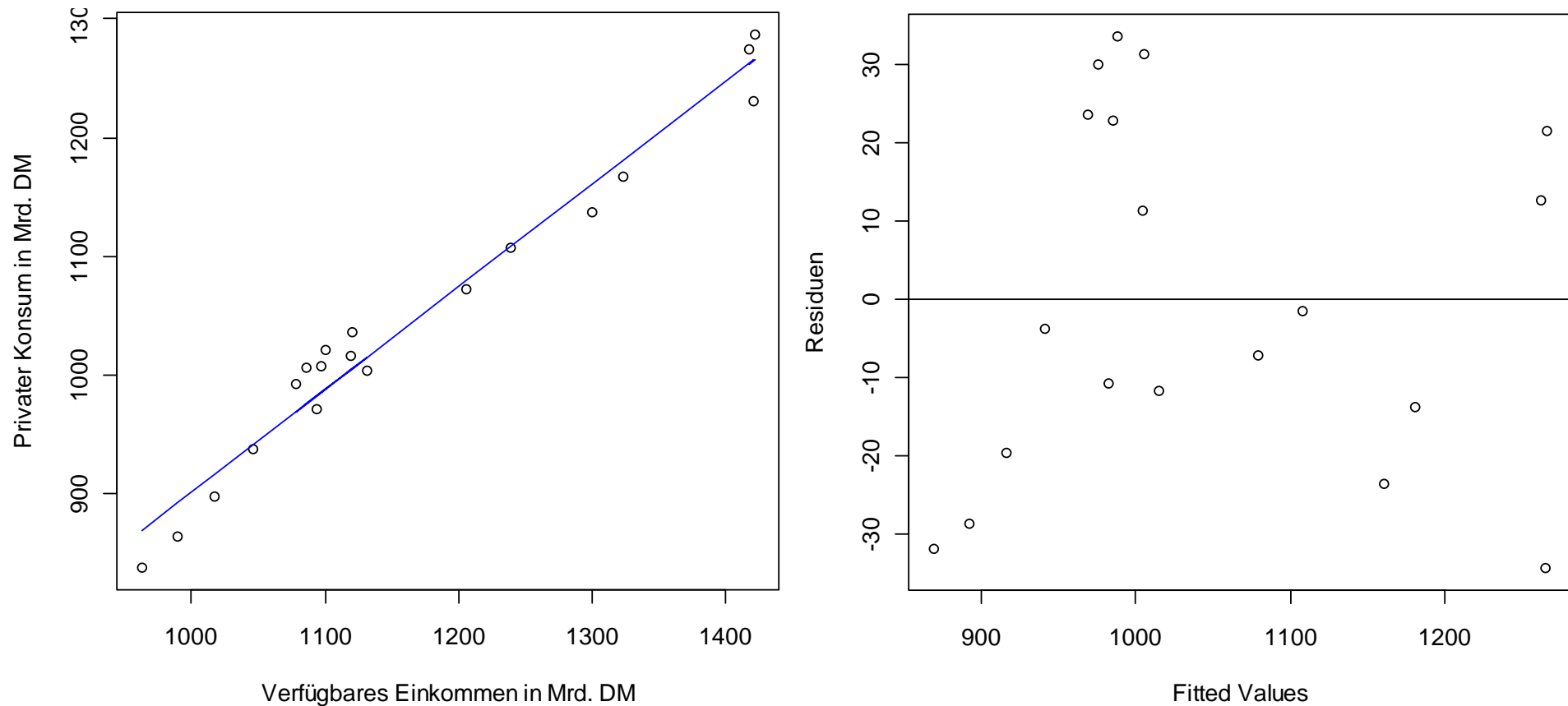
$$\hat{\beta} = (X'X)^{-1} X'y = \begin{pmatrix} 3.58306 & -0.00303 \\ -0.00303 & 2.59352E-06 \end{pmatrix} \cdot \begin{pmatrix} 19866.93 \\ 23512473.03 \end{pmatrix} = \begin{pmatrix} 37.31676 \\ 0.86422 \end{pmatrix} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}.$$

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

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

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

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 Residuen-  
diagramm ☒

## 2.3 Geschätzte Werte, Residuen, Streuungszerlegung

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

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

in

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

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

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

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

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

$$(13) \quad \hat{\varepsilon} = y - \hat{y} = y - Hy = Iy - Hy = (I - H)y .$$

Hierbei ist  $I$  wieder die  $(n,n)$ -Einheitsmatrix.

## S-2.2 Eigenschaften der Hat-Matrix

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

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

(c)  $rg(\mathbf{H}) = \sum_{v=1}^n h_{vv} = spur(\mathbf{H}) = p + 1$ ;

(d)  $\frac{1}{n} \leq h_{vv} \leq 1 \quad (v = 1, \dots, n)$ ;

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

### Beweis

Wir beweisen lediglich (a) und (b):

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

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

### S–2.3 Geometrische Eigenschaften der OLS-Schätzung

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

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

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

*Beweis*

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

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

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

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

$$(i) \quad (\mathbf{x}^1)' \hat{\boldsymbol{\varepsilon}} = \mathbf{1}' \hat{\boldsymbol{\varepsilon}} = \sum_{v=1}^n \hat{\boldsymbol{\varepsilon}}_v = 0, \quad \bar{\boldsymbol{\varepsilon}} = \frac{1}{n} \sum_{v=1}^n \hat{\boldsymbol{\varepsilon}}_v = 0 \quad \text{mit dem } n\text{-Einsvektor } \mathbf{1} = (1, 1, \dots, 1)'$$

Die Residuen sind im Mittel Null.

$$(ii) \quad \bar{y} = \frac{1}{n} \sum_{v=1}^n y_v = \frac{1}{n} \sum_{v=1}^n (\hat{y}_v + \hat{\boldsymbol{\varepsilon}}_v) = \frac{1}{n} \sum_{v=1}^n \hat{y}_v + \frac{1}{n} \sum_{v=1}^n \hat{\boldsymbol{\varepsilon}}_v = \frac{1}{n} \sum_{v=1}^n \hat{y}_v = \bar{\hat{y}}$$

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

$$(iii) \quad \bar{y} = \frac{1}{n} \sum_{v=1}^n \hat{y}_v = \frac{1}{n} \sum_{v=1}^n (\hat{\beta}_0 + \hat{\beta}_1 x_{v1} + \dots + \hat{\beta}_p x_{vp}) = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \dots + \hat{\beta}_p \bar{x}_p$$

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

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

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



## S-2.4 Streuungszerlegung

Es gilt

$$(14a) \quad \sum_{v=1}^n (y_v - \bar{y})^2 = \sum_{v=1}^n (\hat{y}_v - \bar{y})^2 + \sum_{v=1}^n (y_v - \hat{y}_v)^2 = \sum_{v=1}^n (\hat{y}_v - \bar{y})^2 + \sum_{v=1}^n \hat{\varepsilon}_v^2.$$

Symbolisch schreiben wir hierfür

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

mit

$$TSS = \sum_{v=1}^n (y_v - \bar{y})^2 \quad (\textit{Total sum of squares, Streuung des Regressanden})$$

$$ESS = \sum_{v=1}^n (\hat{y}_v - \bar{y})^2 \quad (\textit{Explained sum of squares, erklärte Streuung})$$

$$RSS = \sum_{v=1}^n \hat{\varepsilon}_v^2 \quad (\textit{Residual sum of squares, nicht erklärte Streuung})$$

*Beweis*

$$\begin{aligned}
 \sum_{v=1}^n (y_v - \bar{y})^2 &= \sum_{v=1}^n (y_v + [-\hat{y}_v + \hat{y}_v] - \bar{y})^2 = \sum_{v=1}^n ([y_v - \hat{y}_v] + [\hat{y}_v - \bar{y}])^2 \\
 &= \sum_{v=1}^n (y_v - \hat{y}_v)^2 + 2 \sum_{v=1}^n (y_v - \hat{y}_v)(\hat{y}_v - \bar{y}) + \sum_{v=1}^n (\hat{y}_v - \bar{y})^2 \\
 &= \sum_{v=1}^n (\hat{y}_v - \bar{y})^2 + \sum_{v=1}^n (y_v - \hat{y}_v)^2,
 \end{aligned}$$

da wegen Satz 2.3 gilt

$$\sum_{v=1}^n (y_v - \hat{y}_v)(\hat{y}_v - \bar{y}) = \sum_{v=1}^n \hat{\varepsilon}_v (\hat{y}_v - \bar{y}) = \sum_{v=1}^n \hat{\varepsilon}_v \hat{y}_v - \bar{y} \sum_{v=1}^n \hat{\varepsilon}_v = 0 \quad \text{q.e.d.}$$

## Anwendung der Streuungszersetzung

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

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

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*

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

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

### B-2.3 *Lineare Einfachregression.*

Im Spezialfall der linearen Einfachregression

$$y_v = \beta_0 + \beta_1 x_v + \varepsilon_v \quad (v = 1, \dots, n)$$

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  $y_v, x_v$  ( $v = 1, \dots, n$ ) 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  $\hat{y} = \bar{y}$ .

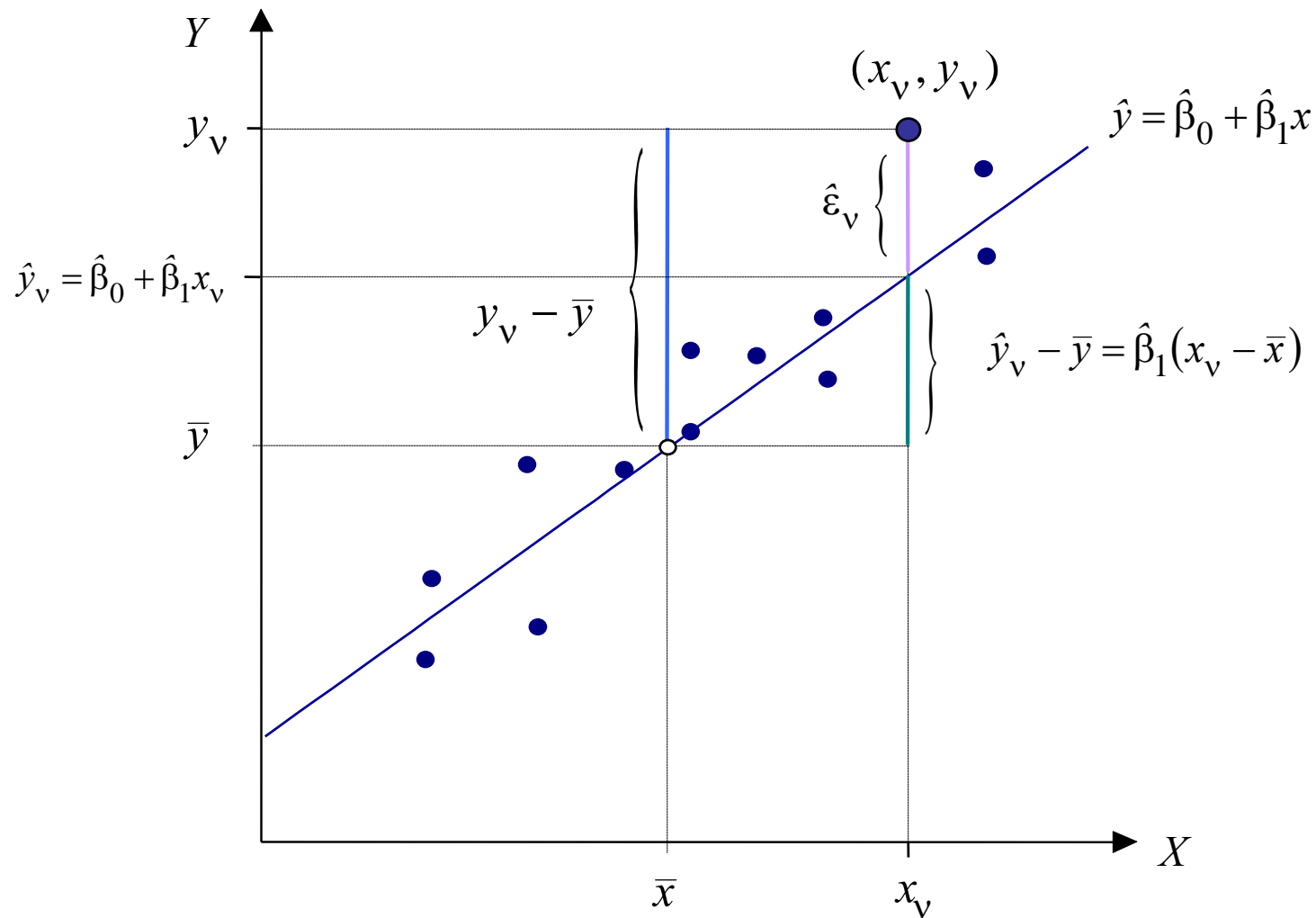


Abb. 2.3: Streudiagramm bivariater Daten, Regressionsgerade und Streuungszersetzung

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

$$\begin{aligned}
 R^2 &= \frac{\sum_{v=1}^n (\hat{y}_v - \bar{y})^2}{\sum_{v=1}^n (y_v - \bar{y})^2} = \frac{\sum_{v=1}^n (\hat{\beta}_1 (x_v - \bar{x}))^2}{\sum_{v=1}^n (y_v - \bar{y})^2} = \hat{\beta}_1^2 \cdot \frac{\sum_{v=1}^n (x_v - \bar{x})^2}{\sum_{v=1}^n (y_v - \bar{y})^2} \\
 &= \left( \frac{\sum_{v=1}^n (x_v - \bar{x})(y_v - \bar{y})}{\sum_{v=1}^n (x_v - \bar{x})^2} \right)^2 \cdot \left( \frac{\sqrt{\sum_{v=1}^n (x_v - \bar{x})^2}}{\sqrt{\sum_{v=1}^n (y_v - \bar{y})^2}} \right)^2 \\
 &= \left( \frac{\sum_{v=1}^n (x_v - \bar{x})(y_v - \bar{y})}{\sqrt{\sum_{v=1}^n (x_v - \bar{x})^2} \cdot \sqrt{\sum_{v=1}^n (y_v - \bar{y})^2}} \right)^2 = r_{XY}^2 .
 \end{aligned}$$

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

$$r_{XY} = \frac{\sum_{v=1}^n (x_v - \bar{x})(y_v - \bar{y})}{\sqrt{\sum_{v=1}^n (x_v - \bar{x})^2} \cdot \sqrt{\sum_{v=1}^n (y_v - \bar{y})^2}} = 0.98428 .$$

Folglich ist der von der geschätzten Konsumfunktion

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

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*  $\hat{\beta}$  und unterstellen dabei die Gültigkeit der Annahmen A1 – A3 aus Definition 2.1.

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

$$\begin{aligned}\hat{\beta} - \beta &= \hat{\beta} - (E(\hat{\beta}) - E(\hat{\beta})) - \beta \\ &= (\hat{\beta} - E(\hat{\beta})) + (E(\hat{\beta}) - \beta).\end{aligned}$$

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



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

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1} X'y = (X'X)^{-1} X'(X\beta + \varepsilon) = \underbrace{(X'X)^{-1} X'X}_{=I} \beta + (X'X)^{-1} X'\varepsilon \\ &= \beta + (X'X)^{-1} X'\varepsilon.\end{aligned}$$

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

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

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

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

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

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

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

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

$$\begin{aligned}
 (18) \quad \text{Cov}(\hat{\boldsymbol{\beta}}) &= E\left(\{\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\}\{\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\}'\right) = E\left(\{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\}\{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\}'\right) \quad \text{wegen } E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} \\
 &= E\left(\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\}\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\}'\right) \quad \text{wegen } \hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \\
 &= E\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}[(\mathbf{X}'\mathbf{X})^{-1}]'\right) \\
 &= E\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
 &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\underbrace{\mathbf{X}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}}_{=\mathbf{I}} \\
 &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}.
 \end{aligned}$$

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

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

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

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

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

Die Quadratwurzeln der Varianzen

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

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

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

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

des Vektors  $\mathbf{y}$  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

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

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

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

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

erwartungstreu sein soll, muss

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

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

$$\tilde{\beta} = \boldsymbol{\beta} + \mathbf{B}\boldsymbol{\varepsilon}$$

folgt. Hiermit erhalten wir

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

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

$$\forall_{\mathbf{c} \in \mathbb{R}^{p+1}} \mathbf{c}'(\boldsymbol{\Sigma}_{\tilde{\beta}} - \boldsymbol{\Sigma}_{\hat{\beta}})\mathbf{c} \geq 0 \quad \Rightarrow \quad \text{Var}(\tilde{\beta}_i) - \text{Var}(\hat{\beta}_i) \geq 0 \text{ für } i = 0, 1, \dots, p.$$

Es ist

$$\Sigma_{\tilde{\beta}} - \Sigma_{\hat{\beta}} = \sigma^2 \mathbf{B}\mathbf{B}' - \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 (\mathbf{B}\mathbf{B}' - (\mathbf{X}'\mathbf{X})^{-1}).$$

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

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

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

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

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

für alle  $\mathbf{c} \in \mathbb{R}^{p+1}$ . Der OLS-Schätzer  $\hat{\beta}$  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  $e > 0$

$$\lim_{n \rightarrow \infty} P(|\hat{\beta}_i - \beta_i| < e) = 1 \quad (i = 0, 1, \dots, p)$$

erfüllt ist. Man schreibt hierfür abkürzend

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

und sagt: „ $\hat{\beta}$  konvergiert mit wachsenden Stichprobenumfang  $n$  in Wahrscheinlichkeit gegen den wahren Parametervektor  $\beta$ “. 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:

$$\lim_{n \rightarrow \infty} E(\hat{\beta}_i) = \beta_i \quad \text{und} \quad \lim_{n \rightarrow \infty} \text{Var}(\hat{\beta}_i) = 0 \quad (i = 0, 1, \dots, p).$$



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:

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

Wir fordern, dass eine endliche Grenzwertmatrix

$$\lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{X}'\mathbf{X} = \mathbf{M}$$

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

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

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

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

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

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

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

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

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

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

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

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

wobei bei hinreichend großen Stichproben  $\frac{1}{n} \mathbf{M}^{-1}$  durch  $(\mathbf{X}'\mathbf{X})^{-1}$  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

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

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

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

für den linearen Schätzer

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

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

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

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

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

folgt:

- (a)  $E(\hat{\boldsymbol{\beta}}_{OLS}) = \boldsymbol{\beta}$  (Erwartungstreue)
- (b)  $\text{Cov}(\hat{\boldsymbol{\beta}}_{OLS}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$  (Varianz)
- (c)  $\text{plim} \hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{\beta}$  (Konsistenz)
- (d)  $\hat{\boldsymbol{\beta}}_{OLS} \overset{\text{asymptotisch}}{\sim} N_{p+1}(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1})$  (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 unbekannt Parameter  $\beta$  und  $\sigma^2$  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 unbekannt Parameter eines stochastischen Modells von der gemeinsamen Dichtefunktion der Stichprobenvariablen aus. Da die Parameter die gemeinsame Dichte beeinflussen, fasst man die Dichte als eine 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

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

Die einzelnen Störungen des Modells  $\varepsilon_v = y_v - \mathbf{x}'_v \boldsymbol{\beta}$  ( $v = 1, \dots, n$ ) 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  $y_1, \dots, y_n$  des Regressanden:

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

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

$$\begin{aligned}
 (21) \quad L(\boldsymbol{\beta}, \sigma^2) &= \prod_{v=1}^n f(\varepsilon_v) \\
 &= \prod_{v=1}^n \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{\varepsilon_v^2}{2\sigma^2}\right\} = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{\sum_{v=1}^n \varepsilon_v^2}{2\sigma^2}\right\} \\
 &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{2\sigma^2}\right\} = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\right\}.
 \end{aligned}$$

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

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

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

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

mit  $Q(\boldsymbol{\beta}) \equiv \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$  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

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

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



Nullsetzen der Ableitungen liefert die sogenannten *Likelihoodgleichungen*

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

und

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

Aus den Likelihoodgleichungen erhalten wir die Maximumstellen der Funktion:

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

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

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

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

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

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  $\beta$  sind identisch. Beide teilen somit auch die statistischen Eigenschaften (siehe Kapitel 2.4). Noch nicht untersucht haben wir die Eigenschaften des ML- Schätzers

$$\hat{\sigma}_{ML}^2 = \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{n}$$

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

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

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

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

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

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

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

folgt

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

Die Quadratsumme

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

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

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

folgt hieraus

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

mit

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

Der Erwartungswert des ML-Schätzers ist

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

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

Die Varianz des Schätzers ist

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

Aus

$$\lim_{n \rightarrow \infty} E(\hat{\sigma}_{ML}^2) = \sigma^2 \quad \text{und} \quad \lim_{n \rightarrow \infty} \text{Var}(\hat{\sigma}_{ML}^2) = 0$$

folgt die *Konsistenz* des ML-Schätzers:

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

Da wir den Bias des ML-Schätzers

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

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

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

mit

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

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

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

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

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

erwartungstreu schätzen:

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

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

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

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

### B-2.1 Makroökonomische Konsumfunktion.

Wir betrachten weiterhin die geschätzte Konsumfunktion

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

Die zugehörige Schätzung der Varianz der Störungen ist

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

mit

$$\hat{\varepsilon}_v = y_v - \hat{y}_v = y_v - 37.317 - 0.864 \cdot x_v \quad (v = 1, \dots, n).$$

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

$$\begin{aligned} \hat{\Sigma}_{\hat{\beta}} &= \hat{\sigma}^2 (\mathbf{X}'\mathbf{X})^{-1} = 545.5995 \cdot \begin{pmatrix} 3.58306 & -0.00303 \\ -0.00303 & 2.59352\text{E}-06 \end{pmatrix} \\ &= \begin{pmatrix} 1954.914 & -1.65094 \\ -1.65094 & 0.00142 \end{pmatrix}. \end{aligned}$$



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

$$\hat{\Sigma}_{\hat{\beta}} = \hat{\sigma}^2 (\mathbf{X}\mathbf{X})^{-1} = \hat{\sigma}^2 \begin{pmatrix} n & \sum_{v=1}^n x_v \\ \sum_{v=1}^n x_v & \sum_{v=1}^n x_v^2 \end{pmatrix}^{-1} = \frac{\hat{\sigma}^2}{n \sum x_v^2 - (\sum x_v)^2} \cdot \begin{pmatrix} \sum_{v=1}^n x_v^2 & -\sum_{v=1}^n x_v \\ -\sum_{v=1}^n x_v & n \end{pmatrix}$$

gilt

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

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

Die Quadratwurzeln der Varianzen sind die geschätzten Standardfehler

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

## 2.7 Hypothesentests

### 2.7.1 Konfidenzintervalle und Tests für einzelne Regressionskoeffizienten

Die Modellannahmen A1 – A4 des klassischen linearen Modells aus Definition D–2.1 seien erfüllt. Dann besitzt die Statistik

$$(31) \quad T_1 = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_{\hat{\beta}_i}}$$

eine  $t$ -Verteilung mit  $n - p - 1$  Freiheitsgraden. Wir schreiben abkürzend  $T_1 \sim t(n - p - 1)$ . Hierbei ist

$$\hat{\sigma}_{\hat{\beta}_i} = \hat{\sigma} \cdot \sqrt{a_{jj}}$$

der Standardfehler (30) des Schätzers  $\hat{\beta}_i = \hat{\beta}_{i,ML} = \hat{\beta}_{i,OLS}$  für den Regressionsparameter  $\beta_i$  und  $a_{jj}$  kennzeichnet wieder das  $j$ -te Diagonalelement von  $(\mathbf{X}'\mathbf{X})^{-1}$ .

Die  $t$ -Verteilung der Statistik (31) resultiert aus folgenden Überlegungen. Wir wissen bereits, dass

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2 a_{jj}) \Leftrightarrow \frac{\hat{\beta}_i - \beta_i}{\sigma \sqrt{a_{jj}}} \sim N(0,1)$$

und

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

gilt. Die Statistik (31) erhalten wir gemäß:

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

Entsprechend der Definition D-1.32 ist diese Zufallsvariable  $t$ -verteilt mit  $n - p - 1$  Freiheitsgraden (vgl. Kapitel 1.16).

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

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

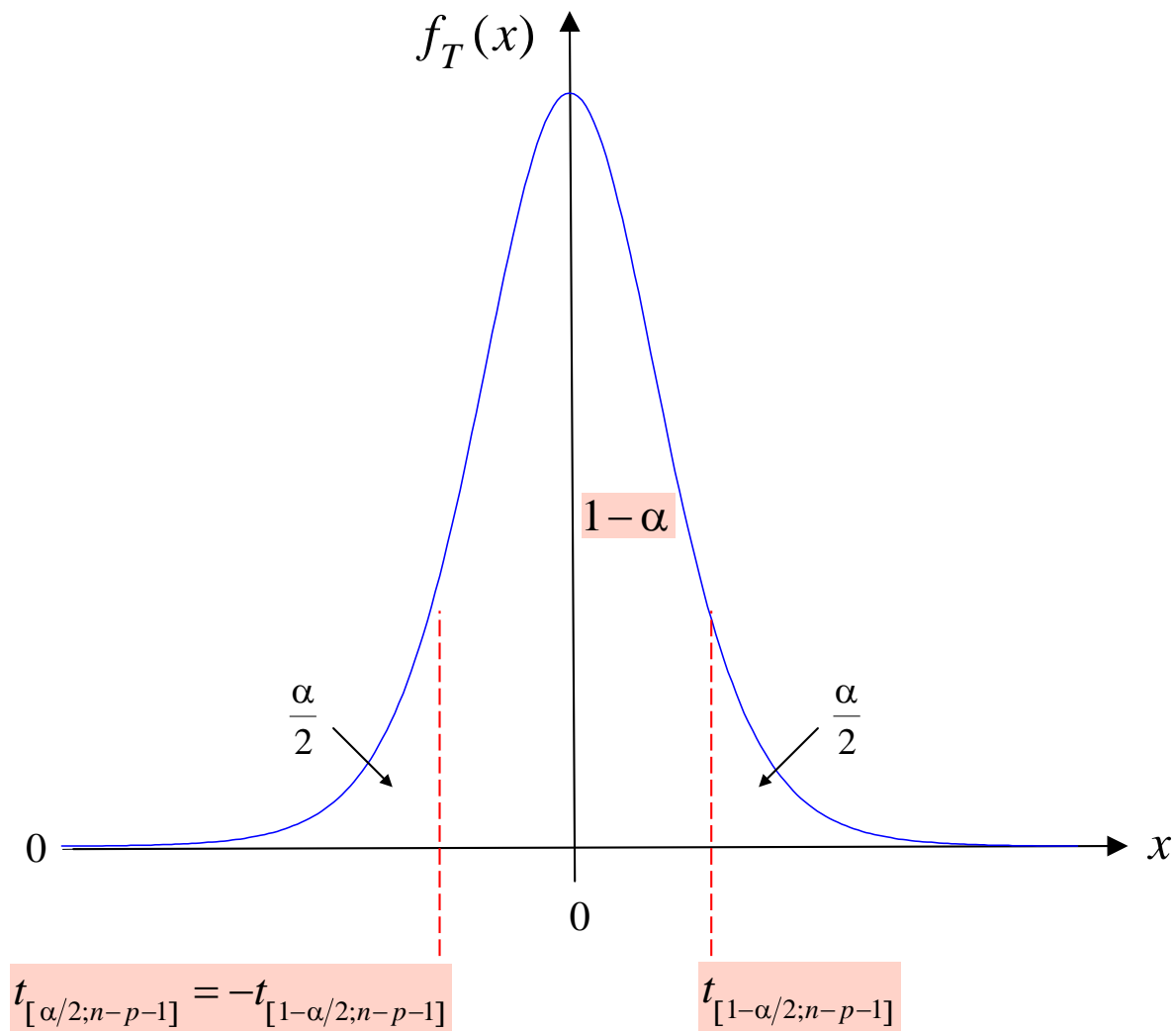
$$P\left(-t_{[1-\alpha/2;n-p-1]} < \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_{\hat{\beta}_i}} < +t_{[1-\alpha/2;n-p-1]}\right) = 1 - \alpha$$

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 sogenannte  *$(1-\alpha)$ -Konfidenzintervall* (Vertrauensintervall) für  $\beta_i$  ( $i = 0, 1, \dots, p$ ):

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



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

Das Konfidenzintervall (32)

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

überdeckt den wahren (aber unbekannt) Regressionsparameter  $\beta_i$  des linearen Regressionsmodells mit einer vorgegebenen Wahrscheinlichkeit  $1-\alpha$ . Die Zahl  $1-\alpha$  wird als *Konfidenzniveau* (*Vertrauenswahrscheinlichkeit*) bezeichnet. Die Zahl  $\alpha$  misst die Wahrscheinlichkeit, dass das Intervall den unbekannt Parameter nicht überdeckt und heißt deshalb die *Irrtumswahrscheinlichkeit*.

Das Konfidenzniveau sagt grob gesprochen aus, dass bei wiederholten Stichprobenentnahmen aus einer Grundgesamtheit der Parameter  $\beta_i$  in  $(1-\alpha) \cdot 100\%$  der Fälle von den Konfidenzintervallen überdeckt wird. Der Fehleranteil liegt bei  $\alpha \cdot 100\%$ .

Das *Konfidenzniveau* eines Intervalls ist frei *wählbar*. Man wählt i.d.R. hohe Vertrauenswahrscheinlichkeiten  $1-\alpha$  (und damit niedrige Irrtumswahrscheinlichkeiten  $\alpha$ ), z.B.  $1-\alpha = 0.9, 0.95, 0.99$  oder dergleichen. Für den *Informationsgewinn* durch die sogenannte *Intervallschätzung* ist zusammen mit dem Konfidenzniveau die Länge des Intervalls entscheidend. Ein Intervall ist bei einem vorgegebenen Niveau  $1-\alpha$  umso informativer, je enger es die unbekannt Parameter umschließt.

### B-2.1 Makroökonomische Konsumfunktion.

Unsere bisherigen Ergebnisse fassen wir wie folgt zusammen:

$$\hat{k}onsum = 37.317 + 0.864 \cdot einkommen, \quad R^2 = 0.9688.$$

(44.2144) (0.0376)

Mittels der Standardfehler können wir die Grenzen von Konfidenzintervallen für die beiden Modellparameter  $\beta_0$  und  $\beta_1$  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  $\alpha = 0.05$ ) ☒

Häufig möchte man bestimmte *Vermutungen* oder *Hypothesen* über einen oder mehrere Parameter eines Regressionsmodells auf Stichprobenbasis überprüfen. Ein statistisches Verfahren zur Prüfung von Hypothesen heißt *statistischer Test*.

Wir betrachten zunächst Hypothesen der Form

$$(33) \quad H_0 : \beta_i = b \quad \text{versus} \quad H_1 : \beta_i \neq b .$$

Die Zahl  $b$  ist hierbei ein vorgegebener hypothetischer Parameterwert. Die Aussage  $H_0$  bezeichnet man als *Arbeits- oder Nullhypothese* über  $\beta_i$ .  $H_0$  kann wahr oder falsch sein. Ist  $H_0$  falsch, dann muss die Negation von  $H_0$  - wir schreiben hierfür  $H_1$  - wahr sein.  $H_1$  heißt *Gegenhypothese* (zu  $H_0$ ).

Von besonderem Interesse ist das Hypothesenpaar

$$H_0 : \beta_i = 0 \quad \text{versus} \quad H_1 : \beta_i \neq 0 \quad (\text{mit } b = 0) .$$

Falls die Nullhypothese wahr ist, liefert der  $i$ -te Regressor  $X_i$  keinen Beitrag zur Erklärung des Regressanden  $Y$  und ist verzichtbar.



Ein statistischer Test prüft, ob eine

Nullhypothese im *signifikanten* (deutlichen) *Widerspruch* zum Stichprobenbefund

steht oder ob eine

Nullhypothese mit der vorliegenden Stichprobe verträglich

ist. Im ersten Falle wird *H<sub>0</sub> verworfen*, andernfalls wird *H<sub>0</sub> nicht verworfen*.

Ein Test auf *Stichprobenbasis* kann eine Nullhypothese *niemals sicher* als wahr bzw. als falsch erkennen. Es sind auch Fehlentscheidungen möglich:

Entscheidung	Unbekannte Realität	
	$H_0$ ist wahr	$H_1$ ist wahr
$H_0$ nicht verworfen	richtige Entscheidung	Fehler (2. Art)
$H_0$ verworfen	Fehler (1. Art)	richtige Entscheidung

Zum Testen der Hypothesen

$$H_0 : \beta_i = b \quad \text{versus} \quad H_1 : \beta_i \neq b$$

können wir das *(1- $\alpha$ )-Konfidenzintervall* für  $\beta_i$

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

nutzen. Es überdeckt den wahren Regressionsparameter  $\beta_i$  mit der Wahrscheinlichkeit  $1-\alpha$ .

Fällt der hypothetische Wert  $b$  nicht in obiges Intervall hinein, so werten wir dies als signifikanten Widerspruch zur Nullhypothese und **verwerfen  $H_0$** . Die Entscheidung kann richtig sein oder zu einem Fehler 1. Art führen. Die Wahrscheinlichkeit für den Fehler 1. Art ist mit der Irrtumswahrscheinlichkeiten  $\alpha$  des Konfidenzintervalls identisch. Sie wird in diesem Zusammenhang auch als *Signifikanzniveau des Tests* bezeichnet.

Fällt der hypothetische Wert  $b$  in das Intervall hinein, so werten wir dies als Unterstützung der Nullhypothese und **verwerfen  $H_0$  nicht**. Die Entscheidung kann richtig sein oder zu einem Fehler 2. Art führen. Die Wahrscheinlichkeit für den Fehler 2. Art ist *unbekannt*.

Äquivalent zu obigem Vorgehen ist der sogenannte *t-Tests* der Hypothesen (33). Er benutzt die *t*-Statistik (31)

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

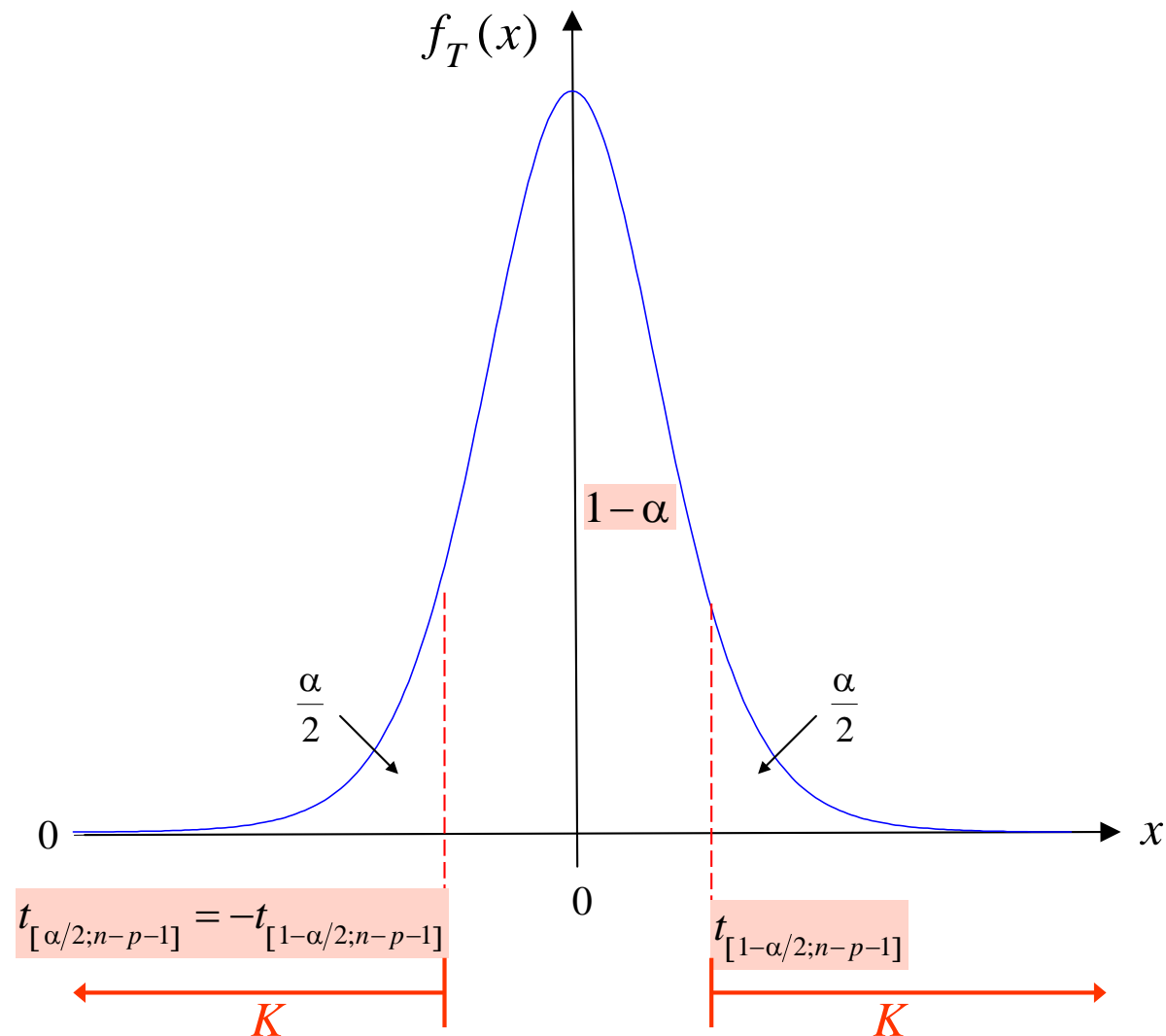
als eine sogenannte *Teststatistik*. Unter Gültigkeit der Modellannahmen A1 – A4 und unter der Bedingung  $\beta_i = b$  gilt:

$$(34) \quad T_1 = \frac{\hat{\beta}_i - b}{\hat{\sigma}_{\hat{\beta}_i}} \sim t(n - p - 1).$$

$H_0 : \beta_i = b$  wird zum *Signifikanzniveau*  $\alpha$  (= Irrtumswahrscheinlichkeit) verworfen, wenn der realisierte Wert  $t$  der Statistik  $T_1$  in den sogenannten kritischen Bereich der Teststatistik

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

fällt, d.h. falls  $|t| > t_{[1-\alpha/2; n-p-1]}$ . Andernfalls wird  $H_0 : \beta_i = b$  nicht verworfen.



**Abb. 2.5:** Dichtefunktion sowie das  $(\alpha/2)$ - und das  $(1-\alpha/2)$ -Quantil der  $t(n-p-1)$ -Verteilung

Alternativ kann anstelle der  $t$ -Statistik auch die  $F$ -Statistik

$$(35) \quad F_1 = T_1^2 \sim F(1, n - p - 1)$$

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

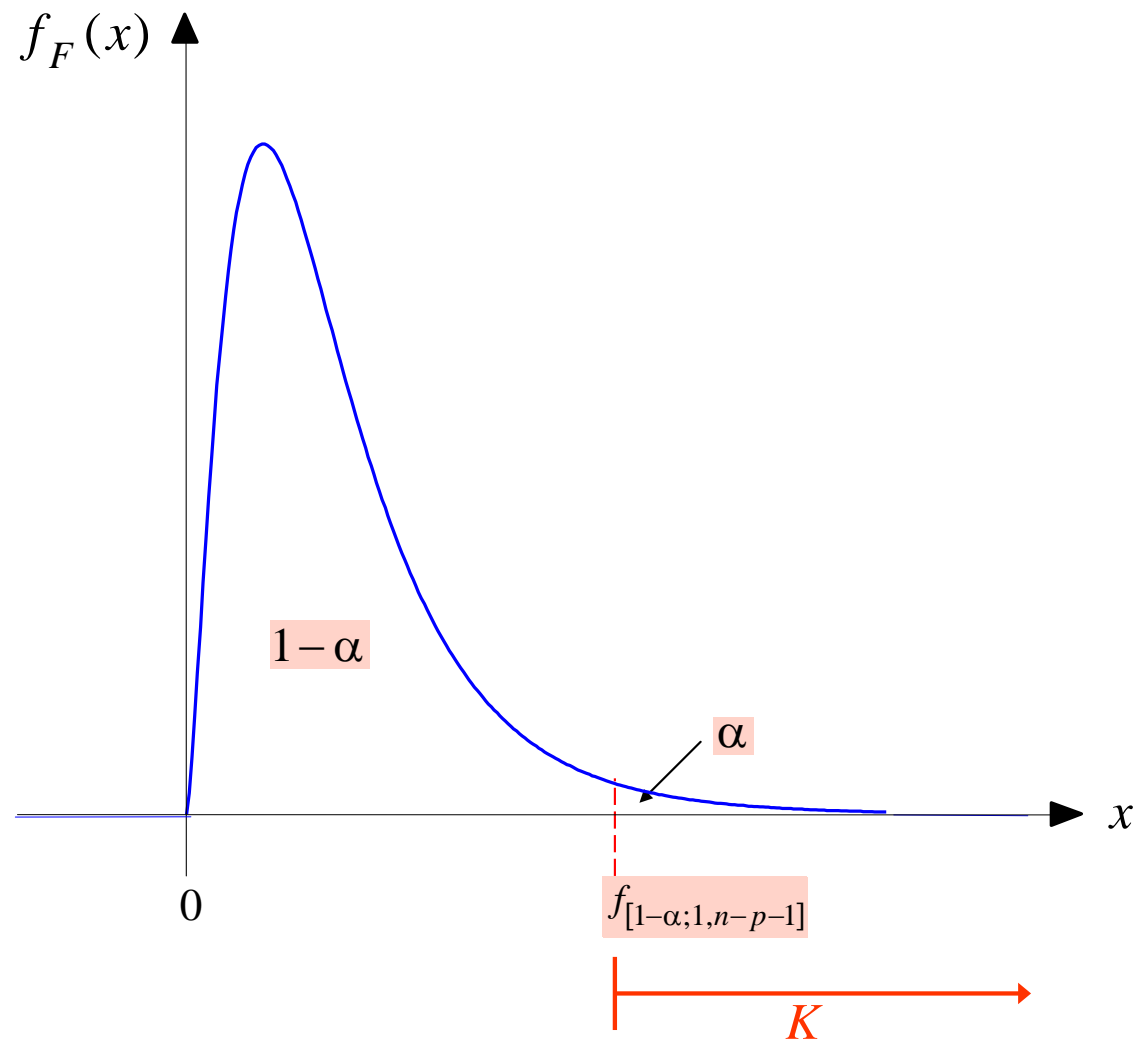
$$K = (f_{[1-\alpha; 1, n-p-1]}, \infty)$$

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*. Er führt zu identischen Entscheidungen wie der  $t$ -Test.

Das *Signifikanzniveau* eines Tests ist frei *wählbar*. Man wählt i.d.R. niedrige Irrtumswahrscheinlichkeiten  $\alpha$ , z.B.  $\alpha = 0.1, 0.05, 0.01$  oder dergleichen.



**Abb. 2.6:** Dichtefunktion sowie das  $(1-\alpha)$ -Quantil der  $F$ -Verteilung mit 1 und  $n-p-1$  Freiheitsgraden

### 2.7.2 F-Test für alle Regressionskoeffizienten

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

$$(36) \quad H_0 : \beta_1 = \dots = \beta_p = 0 \quad \text{versus} \quad H_1 : \beta_i \neq 0 \quad \text{für mindestens ein } i .$$

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

$$(37) \quad F_2 = \frac{n-p-1}{p} \cdot \frac{ESS}{RSS} = \frac{n-p-1}{p} \cdot \frac{R^2}{1-R^2} \sim F(p, n-p-1).$$

Hierbei ist  $R^2$  das Bestimmtheitsmaß (15). Große Werte  $f$  der Statistik  $F_2$  führen zur Ablehnung der Nullhypothese.  $H_0$  wird zum Signifikanzniveau  $\alpha$  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  $\beta_1$  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

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

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

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



Die beim overall  $F$ -Test benötigten Quadratsummen werden oft in Form einer ANOVA-Tabelle (*analysis 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

**Exkurs:  $p$ -Wert oder empirisches Signifikanzniveau**

Sei  $s$  der realisierte Wert der Teststatistik  $S$  des Tests der Hypothesen  $H_0$  versus  $H_1$ . Der  $p$ -Wert ist die Wahrscheinlichkeit  $p^*$ , mit der  $S$  einen Wert annimmt, der unter Gültigkeit von  $H_0$  genauso oder noch plausibler ist als der realisierte Wert  $s$ .

Es gilt also

$$p^* = P(S \geq s | H_0)$$

bei einem Test mit einem einseitigen kritischen Bereich (z. B. einem  $F$ -Test) und

$$p^* = P(|S| \geq s | H_0) = 1 - P(-s < S < s | H_0)$$

bei einem Test mit einem zweiseitigen kritischen Bereich (z. B. einem  $t$ -Test mit zweiseitiger Gegenhypothese).

Je kleiner der  $p$ -Wert ist, umso eher sollte  $H_0$  verworfen werden. Konkret wird  $H_0$  verworfen, wenn  $p^*$  kleiner als ein vorgegebenes Signifikanzniveau  $\alpha$  ist.

## B-2.1 Makroökonomische Konsumfunktion.

Wir wählen  $\alpha = 0.05$ . Der  $p$ -Wert für den  $t$ -Test der Nullhypothese  $H_0 : \beta_0 = 0$  versus  $H_0 : \beta_0 \neq 0$  ist

$$\begin{aligned} p^* &= P(|T_1| \geq 0.844 | H_0) \\ &= 1 - P(-0.844 < T_1 < 0.844 | H_0) \\ &= 0.41. \end{aligned}$$

Wegen

$$p^* = 0.41 > 0.05 = \alpha$$

kann  $H_0$  nicht verworfen werden.

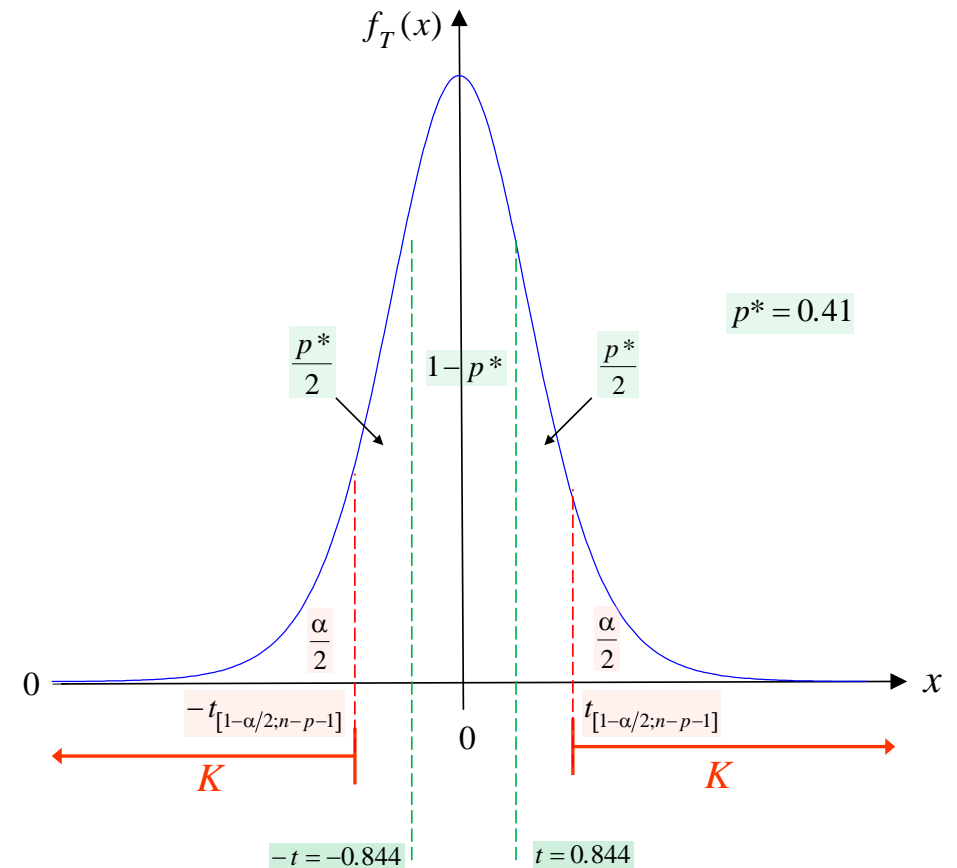


Abb. 2.7: Dichte der  $t(17)$ -Verteilung (blau),  $p$ -Wert (grün) und kritischer Bereich (rot)

Der  $p$ -Wert für den  $t$ -Test der Nullhypothese  $H_0 : \beta_1 = 0$  versus  $H_0 : \beta_1 \neq 0$  ist

$$p^* = P(|T_1| \geq 22.974 | H_0) = 1 - P(-22.974 < T_1 < 22.974 | H_0) = 3.07 \cdot 10^{-14}.$$

Wegen  $p^* = 3.07 \cdot 10^{-14} < 0.05 = \alpha$  wird  $H_0$  deutlich verworfen. Das verfügbare Einkommen liefert einen signifikanten Beitrag zur Erklärung des privaten Konsums.

Der overall  $F$ -Tests für  $H_0 : \beta_1 = 0$  liefert ein identisches Ergebnis. Es gilt:

$$p^* = P(F_2 \geq 527.82 | H_0) = 3.07 \cdot 10^{-14}.$$

Wieder ist  $p^* = 3.07 \cdot 10^{-14} < 0.05 = \alpha$  ☒

### 2.7.3 F-Test der allgemeinen linearen Hypothese

Die Testprobleme (33) und (36) lassen sich als Spezialfälle des folgenden allgemeinen Testproblems auffassen:

$$(38) \quad H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{b} \quad \text{versus} \quad H_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{b} ,$$

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

$$\begin{array}{ll}
 H_0 : \beta_i = b & \mathbf{C} = (0, \dots, 0, 1, 0, \dots, 0), \quad \mathbf{b} = b \quad \text{mit } s = 1 \\
 \\
 H_0 : \beta_1 = \dots = \beta_p = 0 & \mathbf{C} = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{mit } s = p \\
 \\
 H_0 : \beta_2 = \beta_3 = 0 & \mathbf{C} = \begin{pmatrix} 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{mit } s = 2 \\
 \\
 H_0 : \beta_1 = \beta_4 \Leftrightarrow \beta_1 - \beta_4 = 0 & \mathbf{C} = (0, 1, 0, 0, -1, 0, \dots, 0), \quad \mathbf{b} = 0 \quad \text{mit } s = 1 \\
 \\
 H_0 : \beta_2 + \beta_3 + \beta_4 = 1 & \mathbf{C} = (0, 0, 1, 1, 1, 0, \dots, 0), \quad \mathbf{b} = 1 \quad \text{mit } s = 1
 \end{array}$$

Die  $F$ -Teststatistik der Hypothesen (38) ist analog zu (37) aufgebaut:

$$(39) \quad F_3 = \frac{\frac{RSS(H_0) - RSS}{s \cdot \sigma^2}}{\frac{RSS}{(n-p-1) \cdot \sigma^2}} = \frac{n-p-1}{s} \cdot \frac{RSS(H_0) - RSS}{RSS} \sim F(s, n-p-1).$$

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

$RSS$  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  $\alpha$  verworfen, falls

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

Man beachte, dass  $RSS(H_0) = TSS$  im Falle des Testproblems (36) 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:

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

mit

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

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)

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



schreiben wir um gemäß

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

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 (39) nimmt den Wert

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

an. Bei allen üblichen Signifikanzniveaus  $\alpha$  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  $\beta_2, \beta_3$  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  $\beta_2 = \beta_3$ 
```

```
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  $H_0: \beta_2 = \beta_3$ 
```

```
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	☒

## 2.8 Konfidenzintervalle für Regressionsfunktionen und Prognoseintervalle

Sei  $\mathbf{x}_0 = (1, x_{01}, \dots, x_{0p})'$  ein Beobachtungsvektor der Regressoren und  $y_0$  eine Zufallsvariable, die für die potenziellen Werte des Regressanden  $Y$  gegeben die Regressorwerte  $\mathbf{x}_0$  steht.

Der *Wert der Regressionsfunktion* an der Stelle  $\mathbf{x}_0$  entspricht dem Erwartungswert von  $Y$  unter der Bedingung  $\mathbf{x}_0$ :

$$E(y_0) = \beta_0 + \beta_1 x_{01} + \dots + \beta_p x_{0p} = \mathbf{x}_0' \boldsymbol{\beta} .$$

Wegen

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

gilt

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

(vgl. Satz 1.22).

Durch Standardisieren des *Werts der geschätzten Regressionsfunktion*

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

und Schätzen von  $\sigma^2$  erhalten wir die  $t$ -Statistik

$$(40) \quad T_2 = \frac{\mathbf{x}'_0 \hat{\boldsymbol{\beta}} - \mathbf{x}'_0 \boldsymbol{\beta}}{\sqrt{\sigma^2 \mathbf{x}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0}} = \frac{\mathbf{x}'_0 \hat{\boldsymbol{\beta}} - \mathbf{x}'_0 \boldsymbol{\beta}}{\sqrt{\hat{\sigma}^2 \mathbf{x}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0}} \sim t(n-p-1).$$

$$\sqrt{\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2(n-p-1)}}$$

Aus

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

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

und

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

resultiert das  $(1-\alpha)$ -Konfidenzintervall für  $E(y_0)$  oder das  $(1-\alpha)$ -Konfidenzintervall der Regressionsfunktion an der Stelle  $\mathbf{x}_0$ :

$$(41) \quad \left[ \mathbf{x}'_0 \hat{\boldsymbol{\beta}} - t_{[1-\alpha/2;n-p-1]} \cdot \sqrt{\hat{\sigma}^2 \mathbf{x}'_0 (\mathbf{X}\mathbf{X})^{-1} \mathbf{x}_0}, \mathbf{x}'_0 \hat{\boldsymbol{\beta}} + t_{[1-\alpha/2;n-p-1]} \cdot \sqrt{\hat{\sigma}^2 \mathbf{x}'_0 (\mathbf{X}\mathbf{X})^{-1} \mathbf{x}_0} \right].$$

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

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

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

$$(42) \quad T_3 = \frac{y_0 - \mathbf{x}'_0 \hat{\beta}}{\sqrt{\hat{\sigma}^2 + \hat{\sigma}^2 \mathbf{x}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0}} \sim t(n-p-1)$$

und damit

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

Es resultiert das *(1- $\alpha$ )-Prognoseintervall für  $y_0$* :

$$(43) \quad \left[ \mathbf{x}'_0 \hat{\beta} - t_{[1-\alpha/2; n-p-1]} \cdot \sqrt{\hat{\sigma}^2 + \hat{\sigma}^2 \mathbf{x}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0}, \mathbf{x}'_0 \hat{\beta} + t_{[1-\alpha/2; n-p-1]} \cdot \sqrt{\hat{\sigma}^2 + \hat{\sigma}^2 \mathbf{x}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0} \right].$$



## Anmerkungen

(a) Im Falle einer *linearen Einfachregression* gilt:

$$\hat{y}_0 = \mathbf{x}'_0 \hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 x_0 \quad \text{und} \quad \mathbf{x}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0 = \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{v=1}^n (x_v - \bar{x})^2} .$$

Somit folgt

$$\begin{aligned} \hat{y}_0 \pm t_{[1-\alpha/2; n-2]} \cdot \sqrt{\hat{\sigma}^2 + \hat{\sigma}^2 \mathbf{x}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0} \\ = \hat{y}_0 \pm t_{[1-\alpha/2; n-2]} \cdot \sqrt{\hat{\sigma}^2 \cdot \left( 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{v=1}^n (x_v - \bar{x})^2} \right)} . \end{aligned}$$

(b) Ist die Normalverteilungsannahme A4 verletzt, dann gelten aufgrund des asymptotischen Normalverteilung des Schätzers  $\hat{\boldsymbol{\beta}}$  die Konfidenzintervalle (32) und (41) für große Stichprobenumfänge  $n$  weiterhin approximativ. Das Prognoseintervall (43) ist hingegen nicht mehr valide, da  $y_0$  nun keine Normalverteilung besitzt.

### B-2.1 Makroökonomische Konsumfunktion.

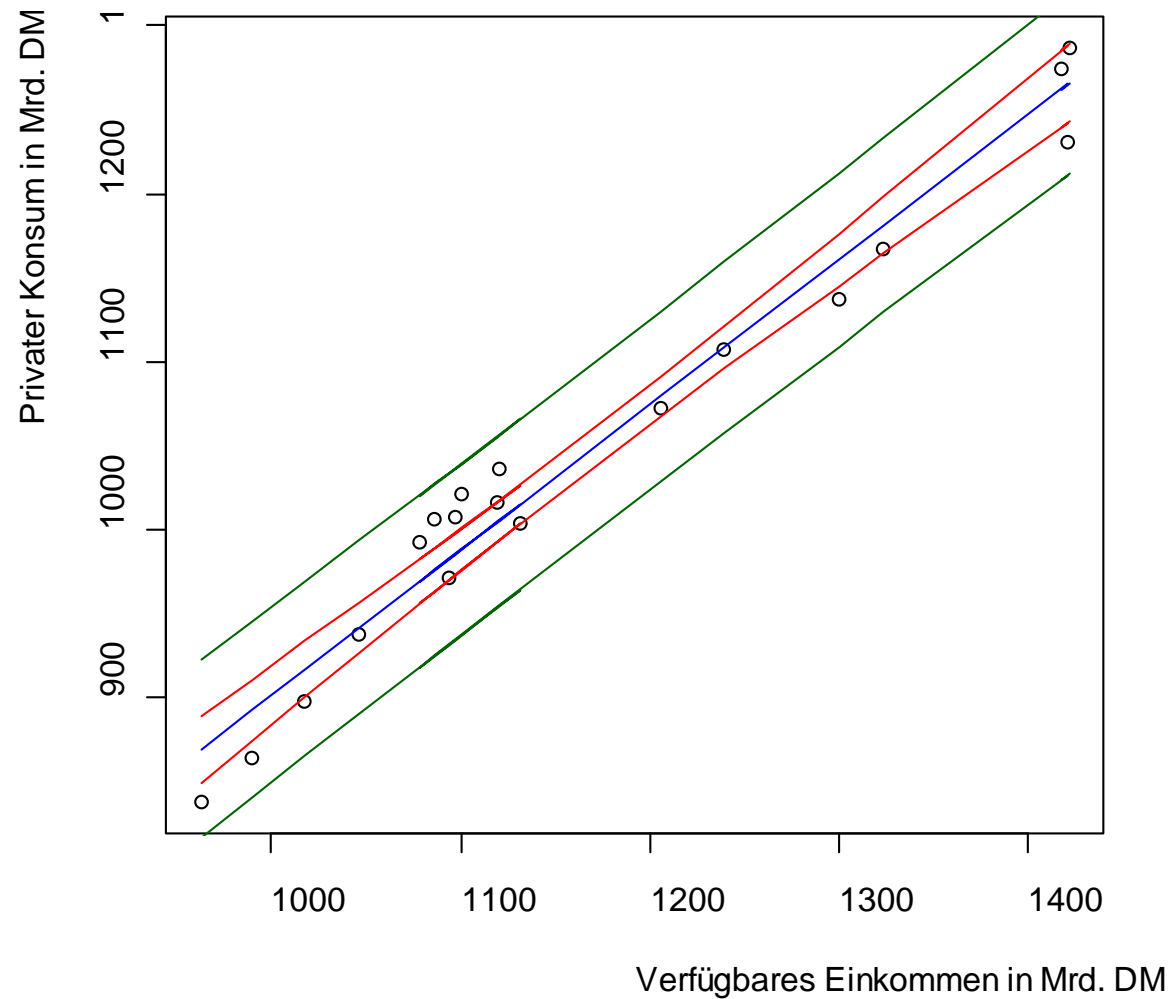
Auf die numerische Angabe von Konfidenzintervallen für die Regressionsfunktion

$$\hat{k}onsum = 37.317 + 0.864 \cdot einkommen, \quad R^2 = 0.9688.$$

(44.2144) (0.0376)

und von Prognoseintervallen wird hier verzichtet. Stattdessen zeigt die Abbildung 2.8 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  $\sigma^2$  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. 2.8:** Streudiagramm der Daten mit geschätzter Regressionsfunktion (blau) und 95%-Konfidenzbänder für die Regressionsfunktion (rot) und die Prognose (grün) ☒

## 2.9 Variablenauswahl

Die Festlegung, welche Regressoren  $X_1, \dots, X_p$  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

$$(44) \quad AIC(M) = -2 \cdot l(\hat{\boldsymbol{\beta}}_{ML}, \hat{\sigma}_{ML}^2) + 2 \cdot (p + 2)$$

$$\propto n \cdot \ln \hat{\sigma}_{ML}^2 + 2 \cdot (p + 2) = n \cdot \ln \left( \frac{RSS}{n} \right) + 2 \cdot (p + 2)$$

an, wobei hier  $\hat{\boldsymbol{\beta}}_{ML}$ ,  $\hat{\sigma}_{ML}^2$  die ML-Schätzer der Parameter  $\boldsymbol{\beta}$  und  $\sigma^2$  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

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

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

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

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

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):

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

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 (*forward 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  $X_1, \dots, X_p$ ,
- linear abhängige Regressoren  $X_1, \dots, X_p$  (*Multikollinearität*),
- *heterogene Varianzen* der Störungen  $\varepsilon_1, \dots, \varepsilon_n$  (*Heteroskedastizität*),
- *autokorrelative Beziehungen* zwischen den Störungen  $\varepsilon_1, \dots, \varepsilon_n$  bei Analysen von Zeitreihendaten,
- ggf. *nicht-normalverteilte* Störungen  $\varepsilon_1, \dots, \varepsilon_n$ ,

beeinflussen die statistischen Eigenschaften der OLS- und ML-Schätzer. Die Schätzer *können* nun substantiell verzerrt und ineffizient werden. Ebenso *können* Konfidenzintervalle und Tests 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 skaliertener Variabler. Ausgehend von  $n$  beobachteten Wertepaaren  $(x_v, y_v)$  ( $v = 1, \dots, n$ ) der Variablen  $X$  und  $Y$  ist die *empirische Kovarianz* definiert durch die Gleichung

$$s_{XY} = \frac{1}{n} \sum_{v=1}^n (x_v - \bar{x})(y_v - \bar{y}) = \frac{1}{n} \sum_{v=1}^n x_v \cdot y_v - \bar{x} \cdot \bar{y}$$

mit

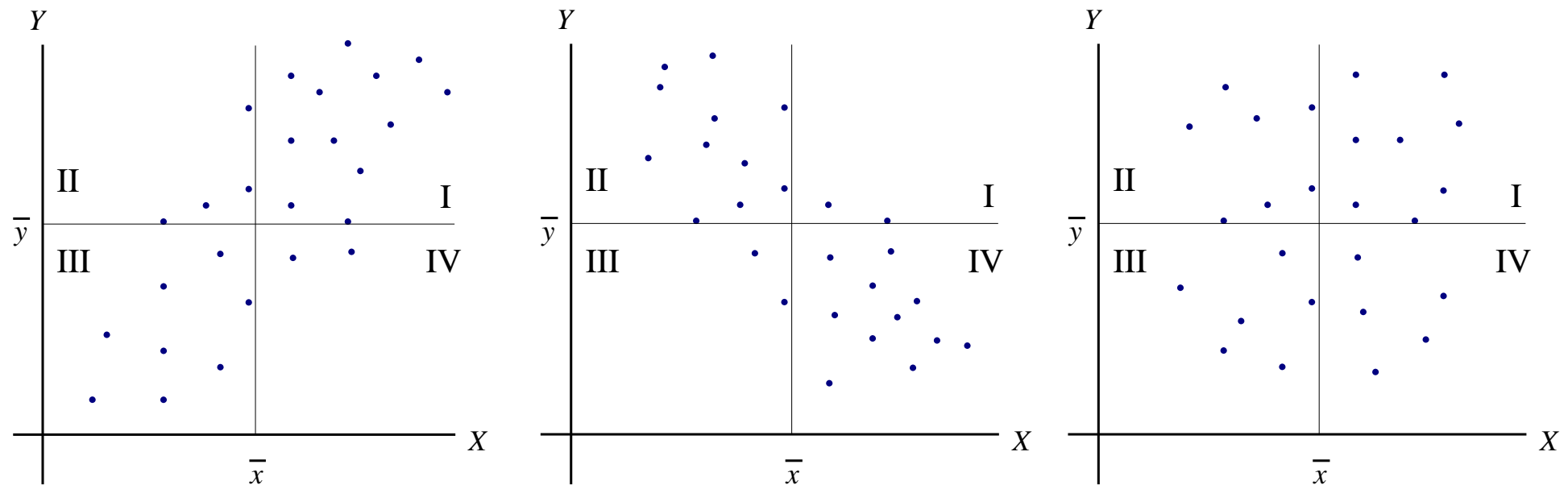
$$\bar{x} = \frac{1}{n} \sum_{v=1}^n x_v \quad \text{und} \quad \bar{y} = \frac{1}{n} \sum_{v=1}^n y_v.$$

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

$$s_X^2 = \frac{1}{n} \sum_{v=1}^n (x_v - \bar{x})^2 = \frac{1}{n} \sum_{v=1}^n x_v^2 - \bar{x}^2 \quad \text{und} \quad s_Y^2 = \frac{1}{n} \sum_{v=1}^n (y_v - \bar{y})^2 = \frac{1}{n} \sum_{v=1}^n y_v^2 - \bar{y}^2$$

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

$$(x_v - \bar{x})(y_v - \bar{y}) .$$

Liegt ein Datenpaar  $(x_v, y_v)$  in...

dann gilt...

Quadrant I ,	$x_v > \bar{x}, y_v > \bar{y}$	$\Rightarrow$	$(x_v - \bar{x})(y_v - \bar{y}) > 0;$
Quadrant III ,	$x_v < \bar{x}, y_v < \bar{y}$	$\Rightarrow$	$(x_v - \bar{x})(y_v - \bar{y}) > 0;$
Quadrant II ,	$x_v < \bar{x}, y_v > \bar{y}$	$\Rightarrow$	$(x_v - \bar{x})(y_v - \bar{y}) < 0;$
Quadrant IV ,	$x_v > \bar{x}, y_v < \bar{y}$	$\Rightarrow$	$(x_v - \bar{x})(y_v - \bar{y}) < 0.$

- 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:

$$(1) \quad -1 \leq r_{XY} \leq +1 \quad \text{oder} \quad |r_{XY}| \leq 1 \quad (\text{Normierung})$$

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

$$y_v = a + bx_v \quad (v = 1, \dots, n)$$

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 ...

$$r_{XY} = 0$$

$$0 < r_{XY} \leq 0.5 \quad (0 > r_{XY} \geq -0.5)$$

$$0.5 < r_{XY} \leq 0.8 \quad (-0.5 > r_{XY} \geq -0.8)$$

$$0.8 < r_{XY} < 1 \quad (-0.8 > r_{XY} > -1)$$

$$r_{XY} = 1 \quad (r_{XY} = -1)$$

ist, dann sind  $X$  und  $Y$  ...

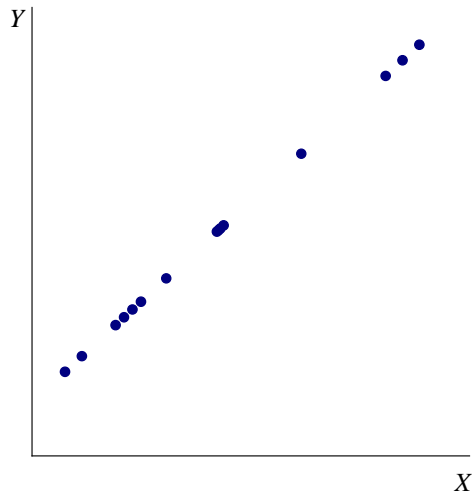
*unkorreliert;*

*schwach* positiv (negativ) korreliert;

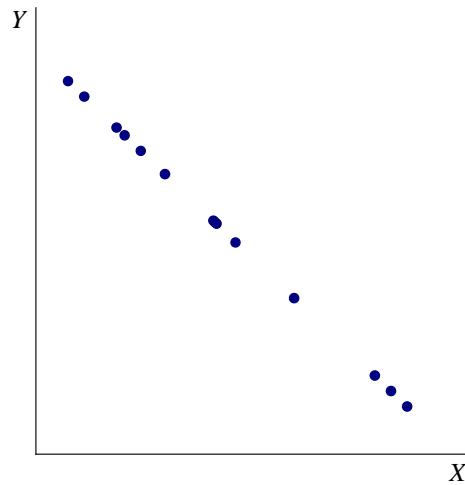
*mittelstark* positiv (negativ) korreliert;

*stark* positiv (negativ) korreliert;

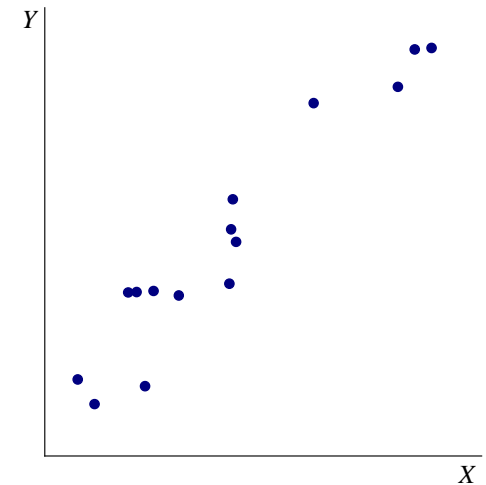
*perfekt* positiv (negativ) korreliert.



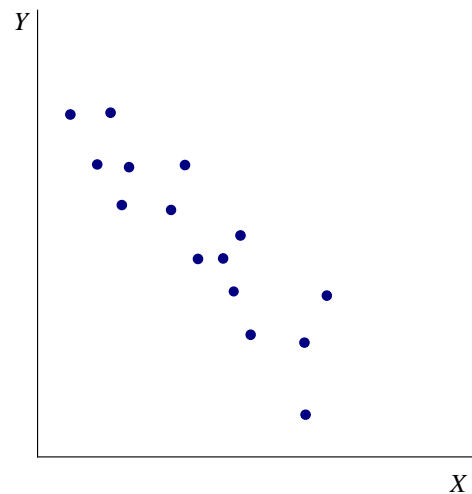
$r_{XY} = +1$



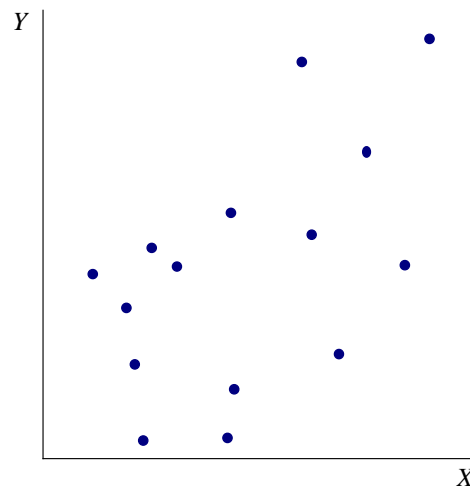
$r_{XY} = -1$



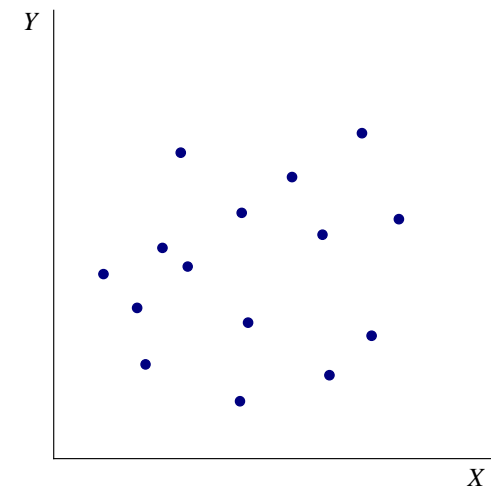
$r_{XY}$  stark positiv



$r_{XY}$  stark negativ



$r_{XY}$  schwach positiv



$r_{XY} = 0$



## II. Vektordifferentiation / Symbolische Differentiation

Betrachtet sei eine überall zweimal differenzierbare Funktion  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  der Veränderlichen  $x_1, x_2, \dots, x_n$  mit  $y = f(x_1, x_2, \dots, x_n)$ . Wir schreiben kurz

$$y = f(\mathbf{x}) \quad \text{mit} \quad \mathbf{x} = (x_1, x_2, \dots, x_n)' \in \mathbb{R}^n.$$

Die partiellen Ableitungen erster Ordnung

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

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

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

Die partiellen Ableitungen zweiter Ordnung

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

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

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

Diese Schreibvereinfachungen bietet arbeitsökonomische Vorteile.

## Beispiele

### Konstante Funktion

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

### Vektorschreibweise

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

### Lineare Funktion

$$f(x_1, \dots, x_n) = c_1 x_1 + \dots + c_n x_n, \quad \frac{\partial f(x_1, \dots, x_n)}{\partial x_i} = c_i \quad (i = 1, \dots, n), \quad \frac{\partial^2 f(x_1, \dots, x_n)}{\partial x_i \partial x_j} = 0 \quad (i, j = 1, \dots, n)$$

### Vektorschreibweise

$$f(\mathbf{x}) = \mathbf{c}'\mathbf{x}, \quad \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \mathbf{c}, \quad \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}'} = \mathbf{O} \quad \text{mit } \mathbf{c} = (c_1, \dots, c_n)' \in \mathbb{R}^n \text{ und der } (n, n)\text{-Nullmatrix } \mathbf{O}'$$

## Quadratische Funktion

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

### Vektorschreibweise

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

## Quadratische Funktion

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

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

### Vektorschreibweise

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

## Quadratische Form

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

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

## Vektorschreibweise

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

## III. Bestimmung lokaler Extrema

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

## Notwendige Bedingung für ein Extremum

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

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

## Hinreichende Bedingung für ein Minimum / Maximum

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

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

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

$$\mathbf{c}' \mathbf{H}(\mathbf{x}_0) \mathbf{c} > 0 \quad (\mathbf{c}' \mathbf{H}(\mathbf{x}_0) \mathbf{c} < 0) .$$

## Definite Matrizen

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

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

## IV. Inverse Matrix

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

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

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

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

Es sei  $A$  eine invertierbare  $(n,n)$ -Matrix. Dann gilt:

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

Dabei ist  $\text{adj}(A) = \mathbf{B} = (b_{ij})_{(n,n)}$  die Adjungierte von  $A$  mit den Komponenten  $b_{ij} = (-1)^{i+j} \cdot \det(A_{ij})$ .

$A_{ij}$  entsteht durch Streichen der  $i$ -ten Zeile und der  $j$ -ten Spalte aus  $A$ .

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

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

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



## V. Maximum-Likelihood-Methode

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

Aufgrund der Unabhängigkeit und identischen Verteilung der Stichprobenvariablen lautet ihre gemeinsame Wahrscheinlichkeits- oder Dichtefunktion einfach

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n / \theta) = f_X(x_1 / \theta) \cdot \dots \cdot f_X(x_n / \theta) = \prod_{v=1}^n f_X(x_v / \theta).$$

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

$$L(\theta / x_1, \dots, x_n) = \prod_{v=1}^n f_X(x_v / \theta) \quad \text{oder kurz} \quad L(\theta).$$

Die Funktion

$$L(\theta) = L(\theta / x_1, \dots, x_n) = \prod_{v=1}^n f_X(x_v / \theta)$$

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

$$L(\hat{\theta} / x_1, \dots, x_n) \geq L(\theta / x_1, \dots, x_n)$$

für alle möglichen Werte von  $\theta$ , so ist

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

der *Maximum-Likelihood-(ML)-Schätzwert* für  $\theta$ . Die zugehörigen zufälligen Stichprobenfunktionen

$$L(\theta / X_1, \dots, X_n) \quad \text{und} \quad \hat{\Theta} = T(X_1, \dots, X_n)$$

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

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

$$l(\theta / x_1, \dots, x_n) := \ln L(\theta / x_1, \dots, x_n) = \ln \prod_{v=1}^n f_X(x_v / \theta) = \sum_{v=1}^n \ln f_X(x_v / \theta)$$

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*

$$\frac{dl(\theta / x_1, \dots, x_n)}{d\theta} = \sum_{v=1}^n \frac{d \ln f_X(x_v / \theta)}{d\theta} = 0 .$$

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

$$\frac{\partial l(\theta / x_1, \dots, x_n)}{\partial \theta_i} = \sum_{v=1}^n \frac{\partial \ln f_X(x_v / \theta)}{\partial \theta_i} = 0 \quad i = 1, \dots, m .$$

## Rechnen mit Logarithmen

$$(1) \quad \ln 1 = 0$$

$$(2) \quad \ln e = 1$$

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

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

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

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

## 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, 6<sup>th</sup> ed.; Prentice-Hall 2008, Chapter 1-5 und 16

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

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