Berner Fachhochschule
Haute école spécialisée bernoise
Bern University of Applied Sciences

CAS Datenanalyse

Modul Regression

Dozent: Prof. Dr. Raúl Gimeno

Übersicht Modul Regression

Übersicht

Kapitel 1: Das lineare Regressionsmodell
Kapitel 2: Statistische Bewertungen von Regressionen
Kapitel 3: Modellspezifikation
Kapitel 4: Dummy-Variablen
Kapitel 5: Beurteilung der Prognosequalität

Literatur

Ludwig von Auer, Ökonometrie, SpringerGabler, 7. Auflage
Hill/ Griffiths / Lim, Principles of Econometrics, John Wiley 4th edition

Berner Fachhochschule
Haute école spécialisée bernoise
Bern University of Applied Sciences

CAS Datenanalyse

Kapitel 1: Das lineare Regressionsmodell

Prof. Dr. Raúl Gimeno
FRM,CAIA, PRM

Inhalt

- Korrelationskoeffizient
- Regressionsmodell in Matrix-Notation
- Schätzen der Regressionskoeffizienten: OLS-Methode
- OLS-Schätzer, Rechenbeispiel
- Interpretation der Koeffizienten
- Standardfehler
- Eigenschaften von Schätzern (Erwartungstreue, Effizienz, Konsistenz)
- Regressionsannahmen
- Streuungszerlegung
- Eigenschaften der Residuen
- Zentrale Momente
- Schiefe und Kurtosis einer Verteilung
- Testen auf Normalität der Residuen

Was ist Ökonometrie?

Ökonometrie befasst sich mit der empirischen Analyse ökonomischer Zusammenhänge und der Überprüfung ökonomischer Theorien mithilfe statistisch/ mathematischer Verfahren.

Idealtypischer Ablauf ökonometrischer Untersuchungen

1. Relevanter Bereich der ökonomischen Theorie (je nach Fragestellung)
2. Vermuteter Zusammenhang zwischen ökonomischen Variablen (Hypothese)
3. Spezifikation des Zusammenhangs als mathematisches Modell
4. Annahmen bezüglich der Störeinflüsse und Messfehler (Stochastik)

ökonometrisches Modell

5. Konfrontation mit Daten (Schätzung der Modellparameter, Berechnung statistischer Testmasse)
6. Spezifikationstests
7. Hypothesentests
8. Erstellen von Prognosen, Ableitung von Politikmassnahmen
9. Erfolgskontrolle, Vergleich konkurrierender Modelle

Korrelationskoeffizient: Definition

Für zwei Zufallsvariablen X und Y: $\rho_{xy} = \dfrac{\operatorname{cov}(x,y)}{\sigma_x \sigma_y}$

Für eine Stichprobe (x_i, y_i) $i = 1,\dots,N$ zweier statistischer Variablen X und Y → empirischer Korrelationskoeffizient $= r_{xy} = \dfrac{S_{xy}}{s_x s_y}$

Vorteil: Dimensionsloses Mass für den Grad des **linearen Zusammenhangs** zwischen zwei Variablen X und Y.

Nachteil: Der Korrelationskoeffizient kann keine Angabe über die Kausalität eines Zusammenhanges liefern.

Im Gegensatz zur Kovarianz ist r_{XY} invariant gegenüber linearen Transformationen: $b \cdot d > 0 \rightarrow \rho(a+br_1, c+dr_2) = \rho(r_1, r_2)$

falls $b \cdot d < 0 \rightarrow \rho(a+br_1, c+dr_2) = -\rho(r_1, r_2)$

Die Korrelation ist massstabsunabhängig.

Es spielt z.B. keine Rolle, ob man die Zeit in Minuten, Stunden oder Tagen misst.

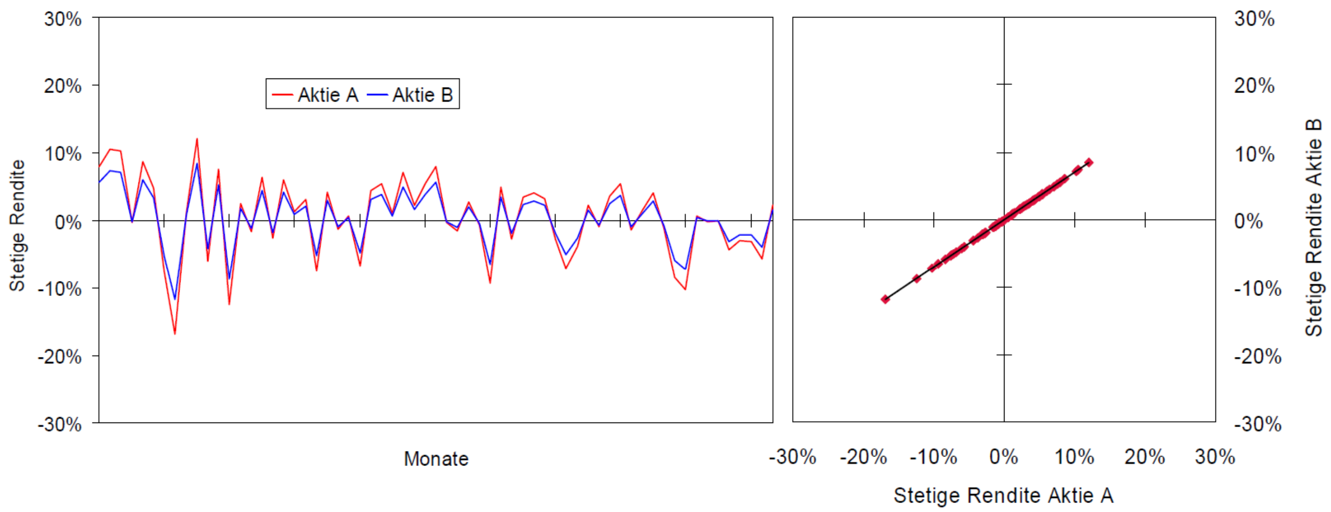
Korrelationskoeffizient: Wertebereich

Wertebereich: $r_{xy} \in [0, 1]$

- $r_{xy} = 0$: Keine Korrelation vorhanden oder beide Variablen hängen überhaupt nicht linear voneinander ab!
- $r_{xy} = 1$: Perfekte Korrelation, X und Y korrelieren vollständig miteinander → alle Messwerte liegen auf einer Gerade mit positiver Steigung.
- $r_{xy} = -1$: Perfekte Antikorrelation, X und Y korrelieren vollständig negativ miteinander → alle Messwerte liegen auf einer Gerade mit negativer Steigung.

Positive Korrelation

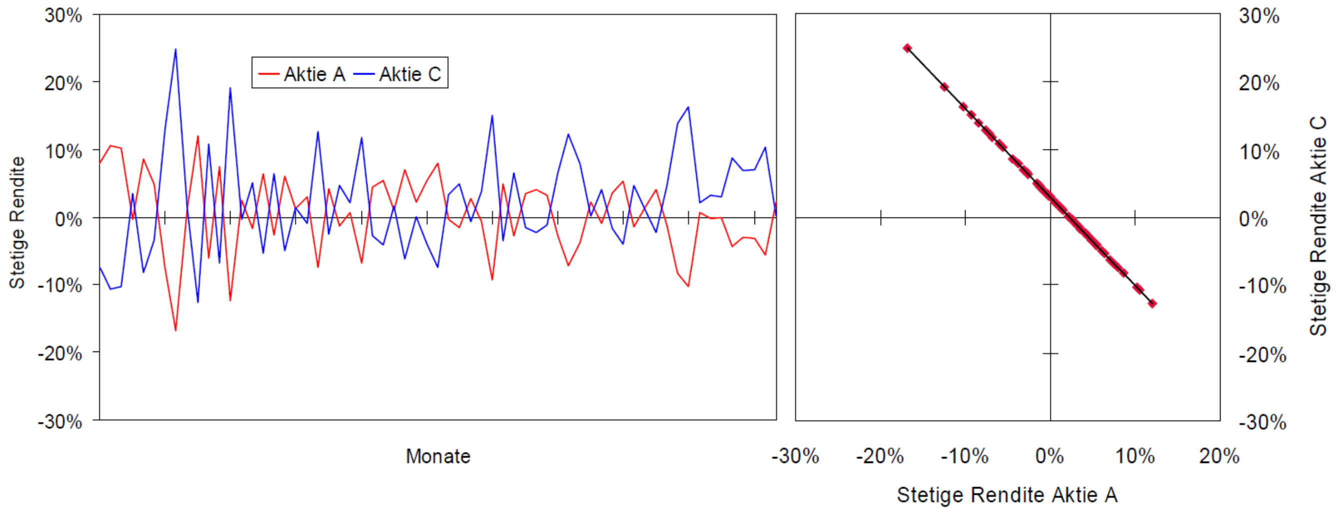
Perfekt positive Korrelation ($\rho_{AB} = 1$)



Allgemein: $r_B = a + br_A$

Negative Korrelation

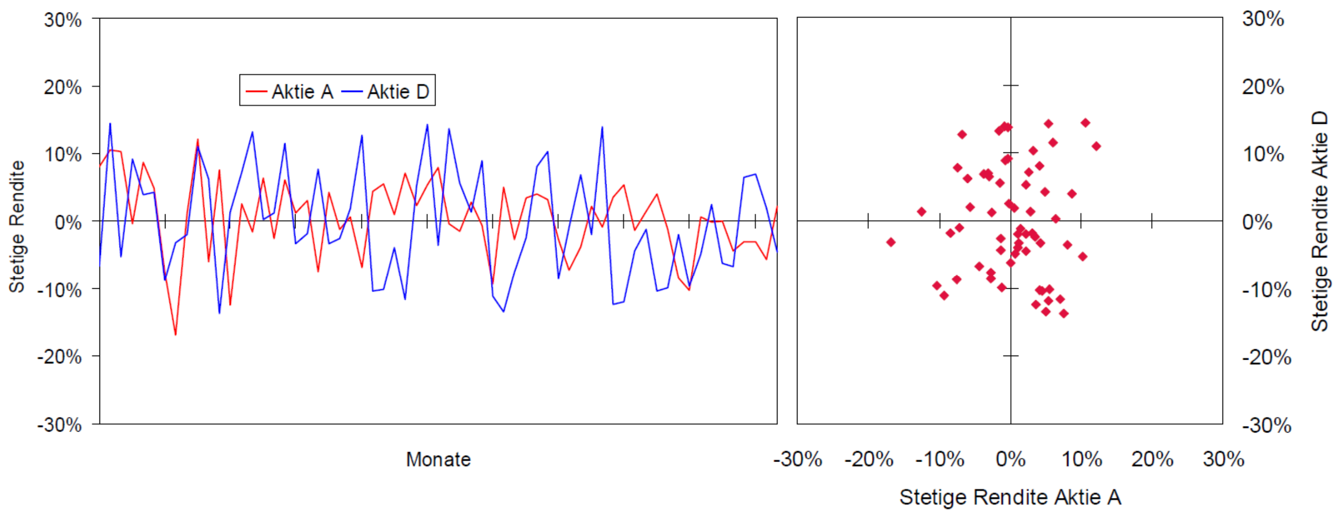
Perfekt negative Korrelation (ρ_{AC} = -1)



Allgemein: $r_c = a - br_A$

Keine Korrelation

Keine Korrelation (ρ_{AD} = 0)



Matrixschreibweise

Populationsmodell mit k Regressoren $x_1, \ldots, x_k$

N Beobachtungen: $\begin{cases} (y_1, x_{21}, \ldots, x_{k1}) \longrightarrow \text{erste Beobachtung} \\ \ldots \\ (y_N, x_{2N}, \ldots, x_{kN}) \longrightarrow \text{letzte Beobachtung} \end{cases}$

Modell: $y_t = \mathbf{x_t}'\beta + u_t, \qquad t = 1, \ldots, N$

Matrixform: $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$ **X: Matrix** → fettgedruckter Grossbuchstabe

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix} = \begin{pmatrix} 1 & X_{21} & \ldots & X_{k1} \\ 1 & X_{22} & \ldots & X_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{2N} & \ldots & X_{kN} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{pmatrix} \begin{matrix} \longrightarrow \text{erste Beobachtung} \\ \\ \\ \longrightarrow \text{letzte Beobachtung} \end{matrix}$$

Nx1 NxK kx1 Nx1

y, β und **u**: Spaltenvektoren

Beispiel

Zusammenhang zwischen der Höhe des Lohnes und seinen Bestimmungsgrössen (= Regressoren)

t	y_t	x_{2t}	x_{3t}	x_{4t}
1	1250	1	28	12
2	1950	9	34	8
3	2300	11	55	25
…	…	…	…	…
18	2600	7	58	30
19	1400	2	35	17
20	1550	2	41	6

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{20} \end{pmatrix} = \begin{pmatrix} 1 & X_{21} & X_{3,1} & X_{41} \\ 1 & X_{22} & X_{3,20} & X_{42} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{2,20} & X_{3,20} & X_{4,20} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{20} \end{pmatrix}$$

y_t: Lohnhöhe

x_{2t}: Ausbildungsjahre

x_{3t}: Alter

x_{4t}: Firmenzugehörigkeit in Jahren

Regressionsmodell: $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$

20x1 (20x4)(4x1)

Bezeichnungen

Alternative Bezeichnungen für y und x in der Funktion $y = b_1 + b_2x$

y	x
abhängige Variable (dependent variable)	unabhängige Variable (independent variable)
erklärte Variable (explained variable)	erklärende Variable (explanatory variable)
Regressand (regressand)	Regressor (regressor)
endogene Variable	exogene Variable

OLS: Ordinary Least Squares

KQ: Methode der kleinsten Quadrate

Steigung	slope
Achsenabschnitt/Interzept	intercept

Regressionsanalyse

- Statistische Methode um Zusammenhänge zwischen einer abhängigen und einer oder mehreren unabhängigen Variablen zu untersuchen.
- Alle einbezogenen Variablen müssen **metrisch** skaliert sein.
- Die Entscheidung, welche Variable als abhängige (y) und welche als unabhängige Variable (x) in die Analysen einbezogen werden, muss vorab aus einem theoretischen Bezugsrahmen abgeleitet werden → es werden kausale Beziehungen unterstellt.
- Beispiel: Preis eines Gebrauchtautos und Kilometerstand oder Alter

- Ziel: Mit Hilfe eines Modellansatzes die Punktwolke beschreiben

Störterm / Störgrösse

Regressionsmodell: $y = \beta_1 + \beta_2 x + u$

Es gibt viele weitere Faktoren, die einen Effekt auf y haben können, welche im Regressionsmodell nicht berücksichtigt wurden.

Der Störterm u und die erklärende Variable x dürfen nicht miteinander in Beziehung stehen, wobei hier die Unkorreliertheit bei der Betrachtung des Korrelationskoeffizienten (d.h. der linearen Korrelation) nicht ausreichend ist.

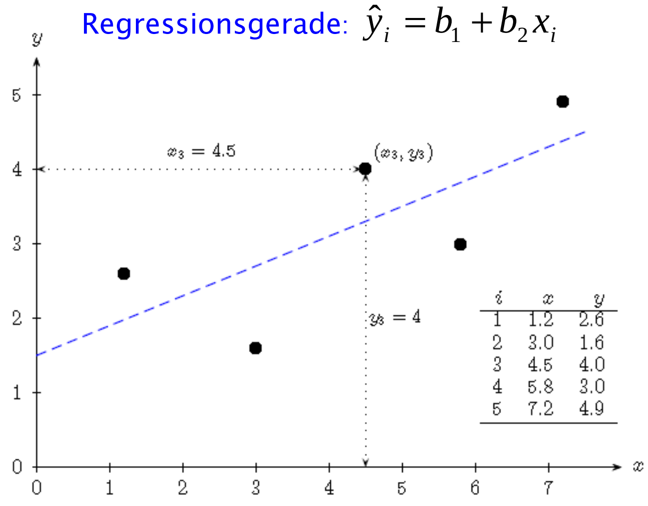
1. Annahme: $E(u|x) = E(u)$

→ der Störterm u ist im Erwartungswert von der erklärenden Variablen x unabhängig (mean independent), d.h. der bedingte Erwartungswert von u, gegeben ein beliebiger Wert von x, ist gleich dem unbedingten Erwartungswert von u und damit konstant.

2. Annahme: $E(u) = 0$ ⇔ Erwartungswert von u (bezogen auf die Grundgesamtheit) ist gleich null → nicht einschränkende Annahme, falls das Interzept (Konstante) β_1 in das einfache lineare Regressionsmodell einbezogen wird. Aus der Kombination beider Annahmen ergibt sich dann: $E(u|x) = 0$

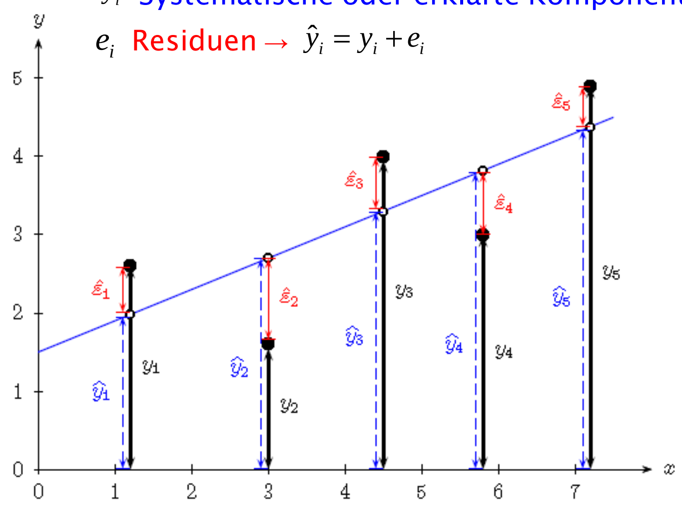
Regressionsgerade

Regressionsgerade: $\hat{y}_i = b_1 + b_2 x_i$



$\hat{y}_i$ Systematische oder erklärte Komponente

e_i Residuen → $\hat{y}_i = y_i + e_i$

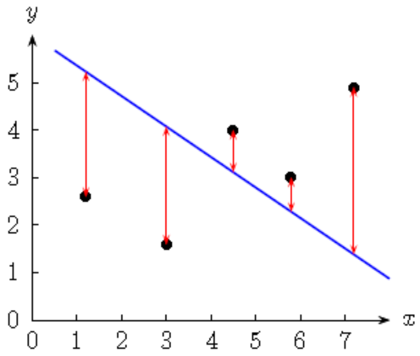


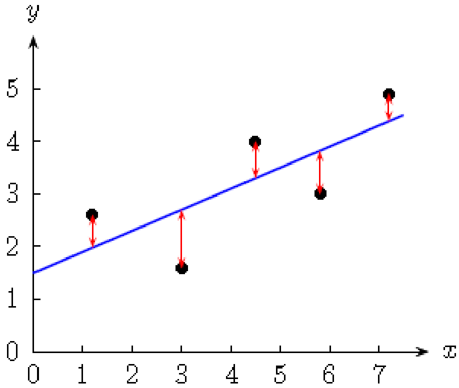
Eine gute Regressionsgerade sollte zwei Bedingungen erfüllen:

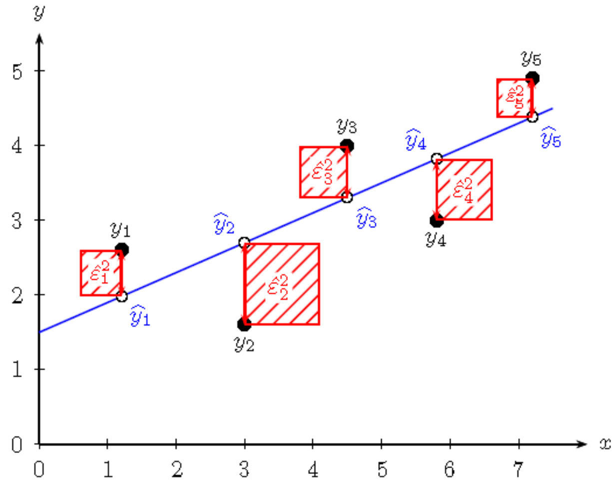
- ✓ Anteil der systematischen, bzw. erklärten Komponente sollte möglichst gross sein → die Residuen (e) sollten möglichst klein sein;
- ✓ die Korrelation ($r_{x\hat{y}}$) zwischen systematischer Komponente ($\hat{y}$) und Residuen sollte möglichst klein sein.

Residuen einer Regression

Die Summe der Abweichungen ($\hat{y}_i$ -y_i) hat in beiden Abbildungen den gleichen Wert!







Nach der OLS-Methode werden b_1 und b_2 derart gewählt, dass die Summe der quadrierten Abweichungen möglichst klein wird, d.h., die Gesamtfläche der schraffierten Quadrate wird minimiert.

Residuen

Residuen (geschätzte Störterme): Differenz zwischen den tatsächlich beobachteten Werten der abhängigen Variable (y_t) und der OLS-Regressionswerte ($\hat{y}_t$)

Residuum positiv: $e_t = y_t - \hat{y}_t > 0$

Tatsächlich beobachtete abhängige Variable y ist grösser als der entsprechende Regressionswert und wird somit unterschätzt: $y > \hat{y}$

Residuum negativ: $e_t = y_t - \hat{y}_t < 0$

Tatsächlich beobachtete abhängige Variable y ist kleiner als der entsprechende Regressionswert und wird somit überschätzt: $y < \hat{y}$

Wichtige Eigenschaften:

(1) $\sum_t e_t = 0$

(2) $\sum_t y_t = \sum_t (\hat{y}_t + e_t) = \sum_t \hat{y}_t \quad \Rightarrow \quad \bar{y} = \bar{\hat{y}}$

(3) $\sum_t x_{jt} e_t = 0 \qquad j = 1,...,k$

Die Kovarianz zwischen jeder erklärenden Variablen und den Residuen ist null. Die unabhängigen Variablen x_j und die Residuen sind orthogonal → : sie stehen geometrisch senkrecht aufeinander → unkorreliert!

Einfachregression: Schätzen der Koeffizienten

β_1, β_2: wahre Regressionskoeffizienten der Population (Grundgesamtheit)

b_1, b_2: geschätzte Regressionskoeffizienten aus der Stichprobe

Regressionsfunktion: $y_t = \beta_1 + \beta_2 x_t + u_t$ $\quad$ $k = 2$

β_1, β_2: Parameter der Grundgesamtheit → feste aber unbekannte Zahlen

Geschätztes Modell: $\hat{y}_t = b_1 + b_2 x_t$

b_1, b_2: Geschätzte Koeffizienten für die wahren Parameter β_1, β_2 → unterscheiden sich von Stichprobe zu Stichprobe.

b_1, b_2: Interpretation

- Schätzfunktion /Schätzer (estimator) → für jede Stichprobe einen Zahlenwert → Zufallsvariable → Stichproben(kennwert)verteilung (sampling distribution)
- Schätzung (estimate) für eine konkrete Stichprobe → Realisation einer Zufallsvariable nach der Stichprobenziehung

Problem: Nur 1 Stichprobe zur Verfügung → Verteilung nicht beobachtbar

Einfachregression: Schätzen der Koeffizienten

$\hat{y}_t$ = prognostizierter Wert / gefitteter Wert:

Die gefitteten Werte liegen auf der Regressionsgerade!

Störgrössen: $u_t = y_t - (\beta_1 + \beta_2 x_t)$ $\quad$ u = Zufallsvariable

Regressionsmodell mit Schätzer b: $y_t = b_1 + b_2 x_t + e_t$

Residuen: $e_t = y_t - \hat{y}_t = y_t - (b_1 + b_2 x_t)$ $\quad$ e_t = realisierter Wert

Interpretation der Koeffizienten

$y_t = \beta_1 + \beta_2 x_t + u_t$

Erwartungswertfunktion: $E(y|x) = \beta_1 + \beta_2 x$ da $E(u_t) = 0$

Interzept: $E(y|x = 0) = \beta_1$

In praktischen Fällen ist das Interzept selten von Bedeutung

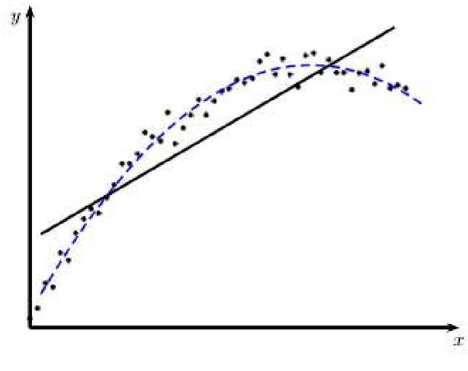
β_2 : Steigungskoeffizient → quantitativer Zusammenhang zwischen erklärender und abhängiger Variable: $dE(y|x)/dx = \beta_2$

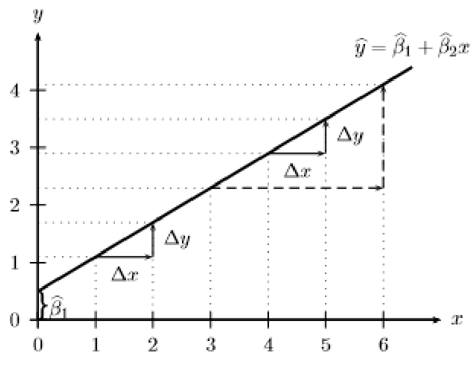
Interpretation: Um wie viele Einheiten ändert sich der erwartete (mittlere) Wert von y, wenn die Variable x um eine Einheit zunimmt → misst den marginalen Effekt.

Notwendig für die Interpretation: In welchen Einheiten wurden x und y gemessen?

Die Steigung einer linearen Funktion ist konstant → der marginale Effekt hat für alle Werte von x den gleichen Wert β_2.

Nichtlinearer Zusammenhang





Lineare Funktion liefert sehr schlechten Fit (Anpassung), wenn der tatsächliche Zusammenhang nichtlinear ist!

Vor jeder Datenanalyse Streudiagramm der Daten ansehen!

Ausdruck: Ceteris-paribus (c.p.)

= unter der Voraussetzung, dass alle anderen Variablen ausser die betrachtete Variable gleich bleiben

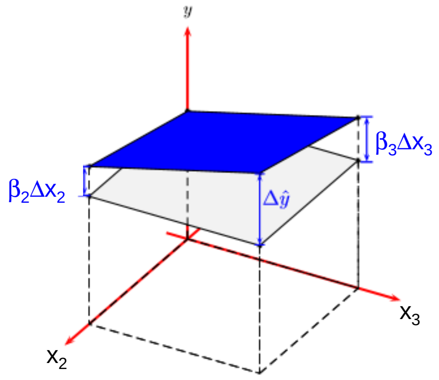
Regressionsebene

Regressionsgleichung mit zwei Regressoren x_2 und x_3:

$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + u_t$

Die gefitteten Werte $\hat{y}$ liegen auf einer Regressionsebene

1 erklärende Variable: $\hat{y}$ liegen auf eine Regressionsgerade



b_1: Interzept = Achsenabschnitt des Schnittpunktes mit der y-Achse

b_2, b_3: messen die Steigungen in Richtung der beiden x-Achsen
→ partielle Ableitungen der Regressionsgleichung
→ als marginale Effekte interpretierbar

$$b_2 = \left.\frac{d\hat{y}}{dx_2}\right|_{dx_3=0} = \frac{\partial\hat{y}}{\partial x_2} \qquad b_3 = \left.\frac{d\hat{y}}{dx_3}\right|_{dx_2=0} = \frac{\partial\hat{y}}{\partial x_3}$$

Ceteris paribus (c.p.) Interpretation von b_2: $\hat{y} = b_1 + b_2 x_{2t} + b_3 x_{3t}$

Um wie viele Einheiten verändert sich y, wenn x_2 um eine Einheit zunimmt und x_3 unverändert bleibt → partielle Ableitung

Skalierung

Eine erklärende Variable x wird mit einer Konstanten b multipliziert

Der neue Koeffizient misst um wie viele Einheiten sich y ändert, wenn x um «eine neue Einheit» = «b alte Einheiten» zunimmt.

Der ursprüngliche Koeffizient wird durch b dividiert

$$\hat{y} = b_1 + \underbrace{(b_2 \frac{1}{b})}_{b_2^*}(bx)$$

b_2^* → Koeffizient der skalierten Gleichung

Änderung der Skalierung der abhängigen Variable y → y wird mit einer Konstante c multipliziert:

$y^* = cy = cb_1 + cb_2 x$

$$y^* = b_1^* + b_1^* x + e^*$$

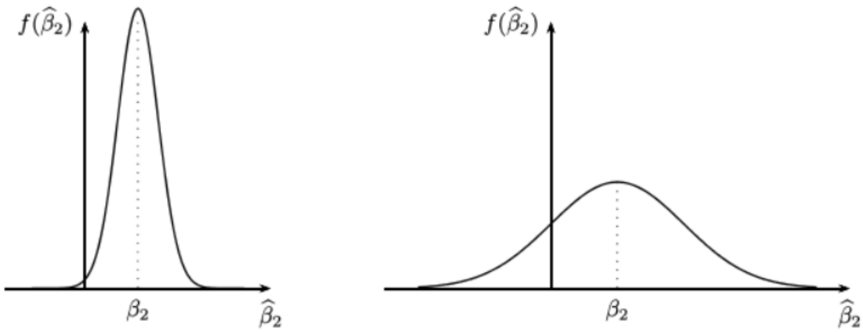
Bivariates Modell: $y^* = cy \quad \Rightarrow \quad b_1^* = cb_1$ und $b_2^* = cb_2$

Standardfehler

Standardfehler (standard error): Standardabweichung einer Stichproben-kennwertverteilung

Abbildung: Stichprobenkennwertverteilung

Realisationen liegen näher beim wahren Wert b_2, Streuung ist geringer → grössere Verlässlichkeit



Standardfehler: Masszahl für die Genauigkeit einer Schätzung

Darstellungsform: Werden in Klammern unter den Koeffizienten angegeben

$$\widehat{\text{Preis}} = \underset{(377{,}44)}{23183{,}6} - \underset{(217{,}99)}{2202{,}77}\,\text{Alter} - \underset{(0{,}0070489)}{0{,}0215039}\,\text{KM}$$

Standardfehler

Allgemein: Streuungsmass einer Schätzfunktion $\hat{\theta}$ für einen unbekannten Parameter θ der Grundgesamtheit.

Standardfehler = Standardabweichung der Schätzfunktion $se(\hat{\theta}) = \sqrt{\mathrm{var}(\hat{\theta})}$

Je grösser die Stichprobe, desto geringer der Standardfehler und desto höher die Zuverlässigkeit der Schätzergebnisse und deren Interpretation.

Unverzerrter Schätzer für die Störgrössenvarianz (s^2): $s_e^2 = \frac{1}{N-k}\sum_i e_i^2 = \frac{\mathbf{e}'\mathbf{e}}{N-k}$

Zu unterscheiden:

1. Standardfehler der Schätzung oder Regression: $s_e = \sqrt{\frac{\mathbf{e}'\mathbf{e}}{N-k}}$
2. Standardfehler der Regressionskoeffizienten: $se(\mathbf{b}) = \sqrt{s_e^2(\mathbf{X}'\mathbf{X})^{-1}}$

Der Begriff Standardfehler bezieht sich auf eine Schätzfunktion und Standardabweichung auf eine Variable.

Standardabweichung der endogenen Variable y: $s_y = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2}$

Auswertung des gretl-Output

	Koeffizient	Std.-fehler	t-Quotient	p-Wert	
const	23183,6	377,445	61,42	1,76e-054	***
Alter	-2202,77	217,994	-10,10	2,11e-014	***
KM	-0,0215039	0,00704890	-3,051	0,0034	***

Mittel d. abh. Var.	16140,16	Stdabw. d. abh. Var.	4029,835
Summe d. quad. Res.	95049375	Stdfehler d. Regress.	1280,149

Abhängige Variable: y = Preis
Geschätztes Modell: Preis = 23’183.6 – 2’202.77Alter – 0.0215KM

- Mittelwert: $\frac{1}{N}\sum_{i=1}^{N} y_i$ = 16’140.16
- Standardabweichung: $s_y = \sqrt[2]{\frac{\sum_i (y_i-\bar{y})^2}{N-1}}$ = 4029.835
- Summe der quadrierten Residuen: $S_{ee} = \sum_i (y_i - \hat{y}_i)^2 = 95'049'375$
- Standardfehler der Regression: $s_e = \sqrt[2]{\frac{\sum_i (y_i-\hat{y}_i)^2}{N-k}}$ = 1280.149 mit k = 3

Auswertung des gretl-Output

	Koeffizient	Std.-fehler	t-Quotient	p-Wert	
const	23183,6	377,445	61,42	1,76e-054	***
Alter	-2202,77	217,994	-10,10	2,11e-014	***
KM	-0,0215039	0,00704890	-3,051	0,0034	***

Mittel d. abh. Var.	16140,16	Stdabw. d. abh. Var.	4029,835
Summe d. quad. Res.	95049375	Stdfehler d. Regress.	1280,149

Abhängige Variable: y = Preis
Geschätztes Modell: Preis = 23’183.6 – 2’202.77Alter – 0.0215KM

$$\mathrm{var}(\mathbf{b}) = s_e^2(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 142'464 & -58'755 & 0.78068 \\ -58'755 & 47'521 & -1.2823 \\ 0.7807 & -1.2824 & 0.0000497 \end{pmatrix}$$

$$se(b_1) = \sqrt{142'464} = 377.445$$
$$se(b_2) = \sqrt{47'521} = 217.994$$
$$se(b_3) = \sqrt{0.0000497} = 0.00704$$

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} 23'183.6 \\ -2'202.77 \\ -0.0215 \end{pmatrix}$$

OLS-Schätzer

Liste der Annahmen

A1	lineare funktionale Form des Modells
A2	$r(\boldsymbol{X}) = k$
A3	$\lim \boldsymbol{X}_n{}' \boldsymbol{X}_n/n = Q$ hat vollen Rang
A4	X_i unabhängig von u für alle i (Exogenität)
A5	$E(u) = 0$
A6	$var(u) = \sigma^2 I$
$A6_1$	$var(u_t) = \sigma^2$ für alle t Homoskedastizität
$A6_2$	$cov(u_t, u_s) = 0$ für alle t und s mit $t \neq s$
A7	u_t normalverteilt für alle t

A1: Linearität des Regressionsmodells

Die Beobachtung y_t ist eine lineare Funktion

$$y_t = \mathbf{x}_t'\boldsymbol{\beta} + u_t$$

der Beobachtungen der erklärenden Variablen X_{ti} und der Störgrösse u_t.

Die Linearität bezieht sich auf die unbekannten Regressionsparameter β_1, β_2, β_3,…, nicht jedoch auf die Struktur der erklärenden und abhängigen Variablen. Somit können auch Nichtlinearitäten in lineare Regressions-modelle einbezogen werden.

Beispiele linearer Modelle

$y = \beta_1 + \beta_2 x^2 + u$

$y = \beta_1 + \beta_2 \ln(x) + u$

$y = \beta_1 + \beta_2 / x + u$

$\ln y = \beta_1 + \beta_2 x_1 + \beta_3 x_1^2 + \beta_4 \ln x_2 + u \quad (1)$

$y^* = \ln y \quad x_2 = x_1^2 \qquad x_3 = \ln x_2$

$(1) \Leftrightarrow \; y^* = \beta_1 + \beta_2 x_1 + \beta_3 x_2 + \beta_4 x_3 + u$

Annahme A2

Y = **Xb** + **u** → OLS-Schätzer: $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$

A2: Die (nxk)–Matrix **X** hat vollen Rang: r(X) = k

✓ Anzahl Beobachtungen (N) ≥ Anzahl Regressoren (k)

✓ Zwischen den k-Spaltenvektoren von **X** besteht keine lineare Beziehung

rg(**X**) = k ⇒ **X´X** invertierbar ⇒ Schätzer existiert

Annahme A2 = keine perfekte Kollinearität

In der Stichprobe (und daher auch in der Grundgesamtheit) ist keine der erklärenden Variablen x_i konstant und es besteht keine exakte lineare Beziehung zwischen den erklärenden Variablen x_i

Verletzung von A2

✓ eine Variable ist ein Vielfaches der anderen Variablen.

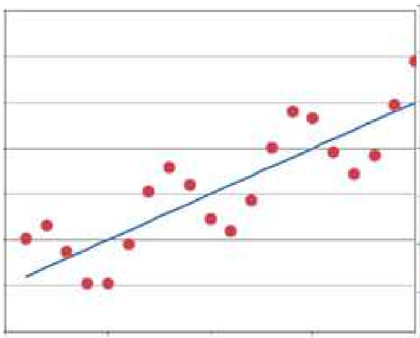
Beispiel: Sowohl Einkommen in Euro als auch Einkommen in Dollar (bzw. Einkommen in 1000 Euro) wurden als erklärende Variablen einbezogen.

✓ eine erklärende Variable kann als lineare Funktion mehrerer anderer erklärender Variablen formuliert werden.

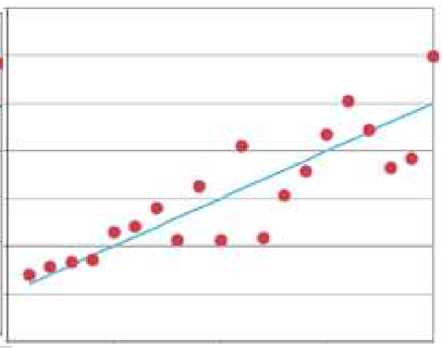
✓ bei einem zu kleinen Stichprobenumfang, d.h. wenn N < k

Annahme A6

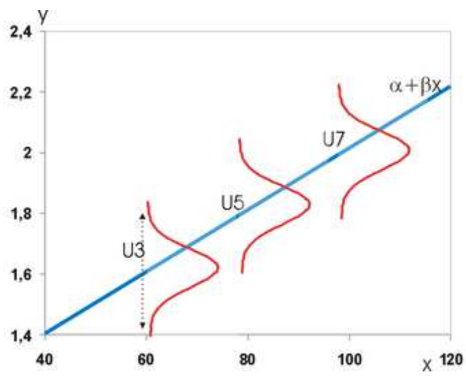
- Keinerlei Systematik in den Residuen: → nur rein zufällig streuen ⇔ die Zielvariable y soll durch x vollständig erklärt werden.
- Systematik in den Residuen: Spezifikationsfehler im Regressionsmodell
- Überprüfung dieser Modellvoraussetzungen mittels (x;y)-Streudiagramm → vermittelt einen optischen Eindruck von der Verteilung der Störgrössen.



Korrelierte Residuen: Schwingungskomponente ist vorhanden



Verschiedene Varianz der Residuen: Vermutlich zwei verschiedene Populationen



Homoskedastizität der Residuen: Varianz bleibt konstant

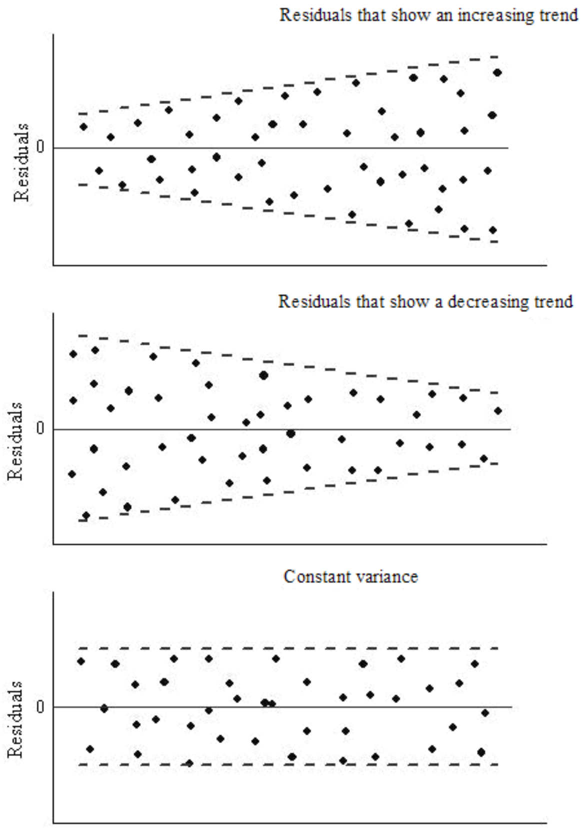
Residuendiagramm

Ansteigender Trend: Fehlervarianz steigt mit der unabhängigen Variable an.

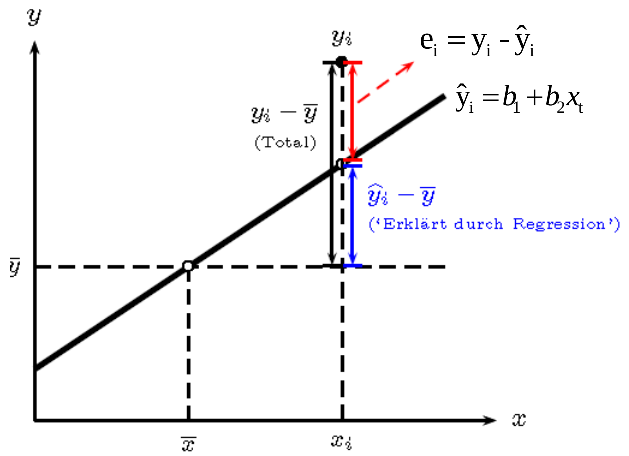
Abfallender Trend: Fehlervarianz fällt mit der unabhängigen Variable ab.

Keine dieser Verteilungen besitzt ein konstantes Varianzmuster→ Annahme einer konstanten Varianz trifft nicht zu → schlechte Güte der Regression.

Ein horizontales Muster deutet auf eine konstante Varianz der Residuen hin.



Streuungszerlegung



Streuung = Summe der quadrierten Abweichungen $\sum_i (y_i - \bar{y})^2$

$$(y_i - \bar{y})^2 = [(\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)]^2$$

$$\underbrace{\sum_i (y_i - \bar{y})^2}_{\text{TSS}} = \underbrace{\sum_i (\hat{y}_i - \bar{y})^2}_{\text{ESS}} + \underbrace{\sum_i (y_i - \hat{y}_i)^2}_{\text{RSS}} + \underbrace{2\sum_i (\hat{y}_i - \bar{y}_i)(y_i - \hat{y}_i)}_{= 0 \text{ wenn Regression Interzept hat}}$$

Erklärte Streuung

Unerklärte Streuung

Streuungszerlegung

ESS: Explained Sum Squares= Erklärte Abweichungsquadratsumme

- Streuung der gefitteten Werte $\hat{y}_i$ um den Mittelwert $\bar{y}$
- durch die Regression erklärte Variation

$$ESS = S_{\hat{y}\hat{y}} = \sum_i (\hat{y}_i - \bar{y})^2 = \mathbf{b}'\mathbf{X}'\mathbf{y} - N\bar{y}^2$$

RSS: Residual Sum of Squares = Residualabweichungsquadratsumme

- Streuung der y_i um die Regressionsgerade
- residuale (nicht erklärte) Variation (Reststreuung)

$$RSS = S_{ee} = \sum_i (y_i - \hat{y}_i)^2$$

TSS: Total Sum of Squares

- Gesamte Abweichungsquadratsumme
- Gesamtvariation = gesamte Streuung der y_i um den Mittelwert $\bar{y}$.

$$TSS = S_{yy} = \sum_i (y_i - \bar{y})^2 = \mathbf{y}'\mathbf{y} - N\bar{y}^2$$

TSS = RSS + ESS

Residuen: Eigenschaften

$$\mathbf{e} = \mathbf{y} - \mathbf{Xb} = \mathbf{y} - \hat{\mathbf{y}}$$

Eigenschaften (inhomogene Regression):

- $(1/N)\Sigma_i e_i = 0$
- $\bar{Y} = \bar{\hat{Y}} \qquad \sum_t y_t = \sum_t (\hat{y}_t + e_t) = \sum_t \hat{y}_t$
- $\bar{Y} = b_1 + b_2\bar{X}_2 + \ldots + b_k\bar{X}_k$
- $\mathbf{y}'\mathbf{y} = \hat{\mathbf{y}}'\hat{\mathbf{y}} + \mathbf{e}'\mathbf{e}$

Streuungszerlegung: $\boxed{s_y^2 = s_{\hat{y}}^2 + s_e^2}$ $\qquad s_{\hat{y}}^2 = \frac{1}{N-1}\sum_t (\hat{y}_t - \bar{\hat{y}})^2 = \frac{1}{N-1}\sum_t (\hat{y}_t - \bar{y})^2$

Stichproben-Varianz

$$s_y^2 = \frac{1}{N-1}\sum_t (y_t - \bar{y})^2$$

Prognose-Varianz

erklärter Teil

Varianz der Residuen

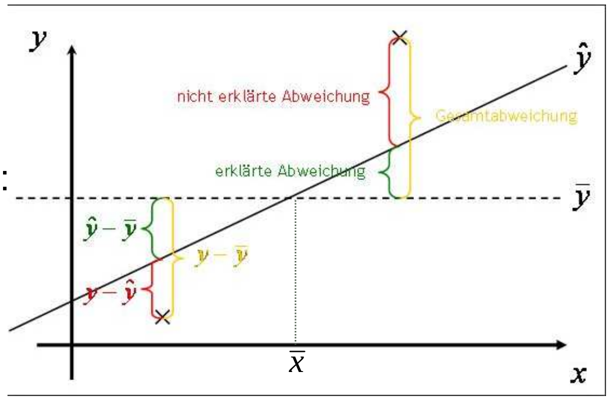
unerklärter Teil

Bestimmtheitsmass R²: Definition

Nicht erklärte Streuung: $RSS = \sum_i (y_i - \hat{y}_i)^2 = S_{ee}$

Erklärte Streuung: $ESS = \sum_i (\hat{y}_i - \bar{y})^2 = S_{\hat{y}\hat{y}}$

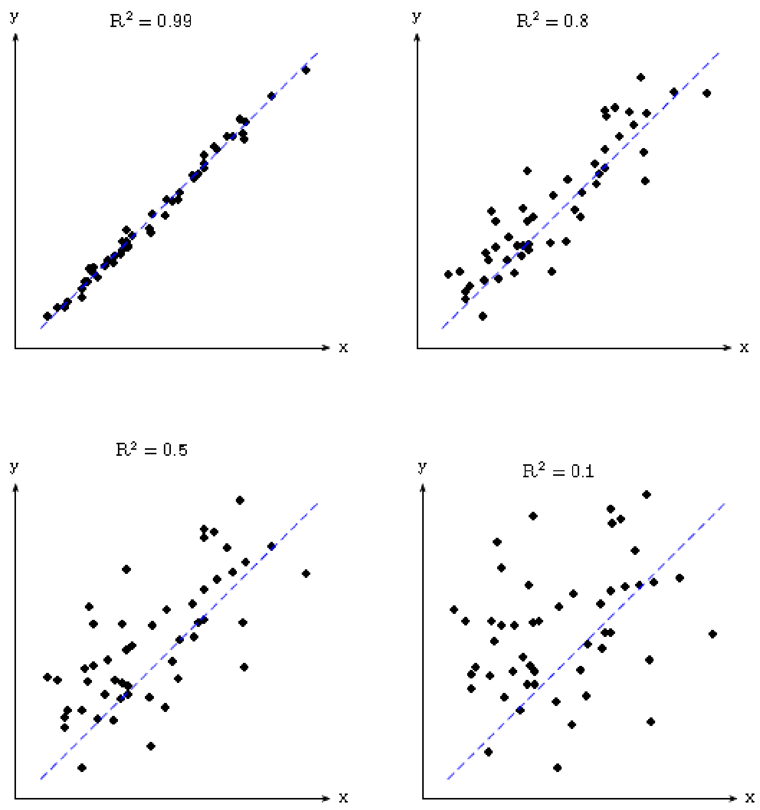
Gesamtstreuung: $TSS = \sum_i (y_i - \bar{y})^2 = S_{yy}$



Erklärungskraft einer Regressionsgerade: desto höher, je höher der Anteil der erklärten Variation an der gesamten Streuung (Variation) von y.

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{N} (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y}_i)^2} = \frac{S_{\hat{y}\hat{y}}}{S_{yy}} = \frac{\text{erklärte Streuung}}{\text{Gesamtstreuung}}$$

R²: Güte der Regressionsfunktion



R^2 : Erklärungen

- Deskriptives Mass zur Beurteilung der Güte der Anpassung der Regressionsgerade an die Beobachtungspunkte.
- Anteil der durch die Regressionsgerade erklärten Streuung ESS an der gesamten Streuung TSS

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{S_{ee}}{S_{yy}}$$

- Anteil der Varianz der abhängigen Variable, die durch das Modell erklärt wird.
- Wertebereich: $0 \leq R^2 \leq 1$
- Je besser die Güte der Anpassung (Fit) der Regressionsgerade ist, desto näher liegt das Bestimmtheitsmass bei 1.

R^2: Einfach- versus Mehrfachregression

- Achtung: Wenn eine Regressionsgleichung fehlspezifiziert ist, kann sie ein sehr hohes R^2 aufweisen, obwohl sie unbrauchbar ist!
- Mehrfachregression: R^2 = Quadrat des Korrelationskoeffizienten zwischen den Beobachtungen y_i und den Prognosen $\hat{y}_i$: $R^2 = r_{y\hat{y}}^2$
- Einfachregression: R^2 = Quadrat des Korrelationskoeffizienten zwischen den Beobachtungen y_i und dem Regressor.

$$r_{xy} = \frac{\operatorname{cov}(x,y)}{s_x s_y} = \frac{S_{xy}/(N-1)}{\sqrt{S_{xx}/(N-1)}\sqrt{S_{yy}/(N-1)}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \pm \sqrt{\frac{S_{xy}^2}{S_{xx}S_{yy}}} = \pm\sqrt{R^2}$$

$$R^2 = \frac{S_{\hat{y}\hat{y}}}{S_{yy}} = \frac{b^2 S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx}S_{yy}} = r_{xy}^2$$

Bestimmtheitsmass: Eigenschaften

$$R^2 = \frac{S_{\hat{y}\hat{y}}}{S_{yy}} = 1 - \frac{S_{ee}}{S_{yy}} \qquad 0 \le R^2 \le 1$$

- $R^2 = 1 \Rightarrow S_{\hat{y}\hat{y}} = \sum_i (\hat{y}_i - \bar{y})^2 = S_{yy} = \sum_i (y_i - \bar{y})^2 \Leftrightarrow S_{ee} = \sum_i (y_i - \hat{y}_i)^2 = 0$
  $\Rightarrow$ alle Residuen sind null $\Leftrightarrow$ vollständige Korrelation
- $R^2 = 0 \Rightarrow S_{\hat{y}\hat{y}} = 0 \Leftrightarrow S_{ee} = S_{yy}$
  alle Regressionskoeffizienten haben den Wert null!
  $\Leftrightarrow$ keine Korrelation
- R^2 sollte einen hohen Wert $\rightarrow$ nah von 1
- Bestimmtheitsmass und Korrelationskoeffizient: $R^2 = r_{xy}^2 = r_{y\hat{y}}^2$
- R^2 kann durch Hinzufügen eines Regressors nicht kleiner werden $\rightarrow$ R^2 ist ein schlechtes Mass zur Beurteilung der Anpassungsgüte eines linearen Regressionsmodells.

Gretl-Output

	Koeffizient	Std.-fehler	t-Quotient	p-Wert	
const	23183,6	377,445	61,42	1,76e-054	***
Alter	-2202,77	217,994	-10,10	2,11e-014	***
KM	-0,0215039	0,00704890	-3,051	0,0034	***

	Mittel d. abh. Var.	16140,16	Stdabw. d. abh. Var.	4029,835	S_y
S_{ee}	Summe d. quad. Res.	95049375	Stdfehler d. Regress.	1280,149
	R-Quadrat	0,902451	Korrigiertes R-Quadrat	0,899087
	F(2, 58)	268,2860	P-Wert(F)	4,87e-30
	Log-Likelihood	-521,4558	Akaike-Kriterium	1048,912
	Schwarz-Kriterium	1055,244	Hannan-Quinn-Kriterium	1051,393

Bestimmtheitsmass: $R^2 = 1 - \frac{S_{ee}}{S_{yy}} = 1 - \frac{95'049'375}{60 \cdot 4029.835^2} = 0.9024$

$$\mathrm{var}(y) = s_y^2 = \frac{1}{N-1}\sum_t (y_t - \bar{y})^2 = \frac{S_{yy}}{N-1}$$

Interpretation: 90% der Varianz des Autopreises wird durch das Regressionsmodell erklärt.

Zentrale Momente

Zur Kennzeichnung von Verteilungen können auch höhere Momente verwendet werden:

- Es sei X eine Zufallsvariable mit E(X) = μ
- Das zentrale Moment der Ordnung k entspricht dem Erwartungswert der k-ten Potenz der zentrierten Zufallsgrösse (X– μ):

- Zentrales Moment der Ordnung k von X: $m_k = E(X-\mu)^k$

$$m_k = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^k$$

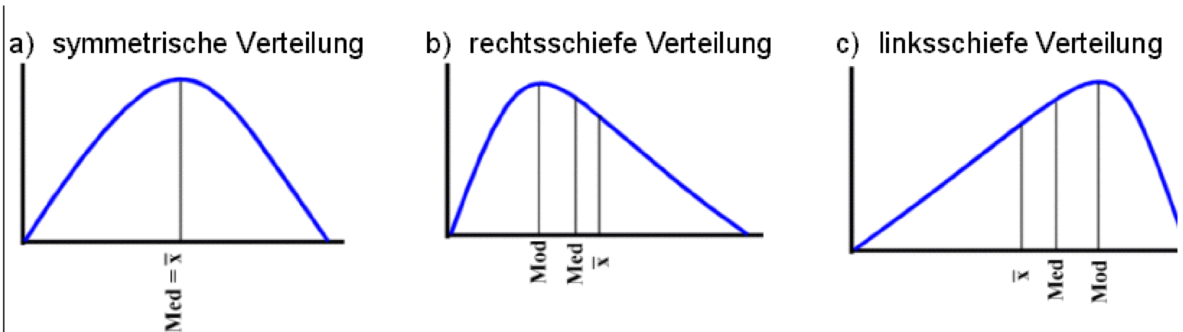
- Das zweite zentrale Moment ist die Varianz: $m_2 = E\left((X-\mu)^2\right)$

$$m_2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

Schiefe

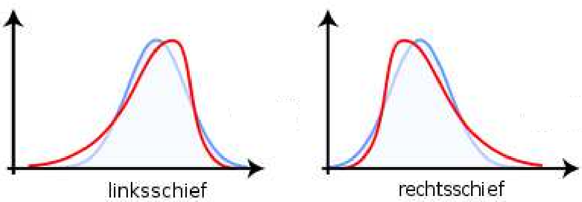
Normalverteilung: symmetrische Verteilung → Schiefe = 0

Das dritte zentrale Moment ist nach Normierung die Schiefe



$\bar{x}$ = med = Mod   Mod < Med < $\bar{x}$ $\bar{x}$ < Med < Mod

$$S = \frac{\mu_3}{\sigma^3} = \frac{\mu_3}{(\sigma^2)^{3/2}} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^3 / N}{\left(\sum_{i=1}^{N}(x_i - \bar{x})^2 / N\right)^{\frac{3}{2}}}$$



S < 0   S > 0

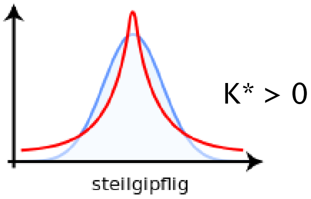
S(aX+b) = S(x): invariant unter linearer Transformation

Kurtosis

Mass für die Wölbung einer Verteilung. Bei Normalverteilung = 3 definitionsgemäss. Werte, die darüber liegen, zeigen an, dass die Verteilung fette Enden hat.

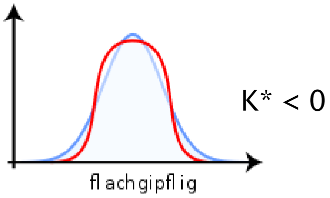
Kurtosis = das vierte zentrale Moment nach Normierung

$$K = \frac{\mu_4}{\sigma^4} = \frac{\mu_4}{\left(\sigma^2\right)^2} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^4 / N}{\left(\sum_{i=1}^{N}(x_i - \bar{x})^2 / N\right)^2}$$

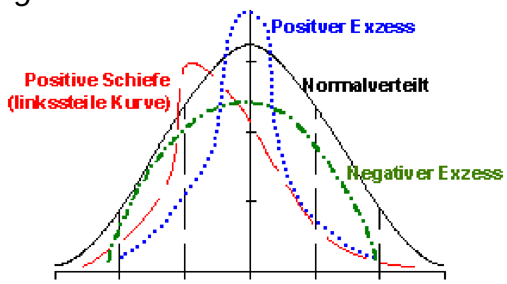


Vergleich mit der Wölbung der Normalverteilung (=3)
→ Wölbung der Normalverteilung wird auf 0 normiert (Subtraktion von 3); diese Grösse wird als Exzess (Kurtosis) bezeichnet.

$$K^* = \frac{m_4}{\left(s_x^2\right)^2} - 3$$



- K* = 0 normalgipflig (*mesokurtisch) →Normalverteilung*
- K* > 0 steilgipflig (*leptokurtisch)*. Im Vergleich zur Normalverteilung spitzere Verteilungen, d.h. Verteilungen mit starken Peaks.
- K* < 0 : flachgipflig (*platykurtisch)*. Im Vergleich zur Normalverteilung →abgeflachte Verteilung.



Testen auf Normalität der Daten

Jarque-Bera-Test: statistischer Test, der anhand der Schiefe und der Kurtosis in den Daten prüft, ob eine Normalverteilung vorliegt → Anpassungstest

Die Teststatistik *JB* des Jarque-Bera-Tests

$$JB = \frac{N}{6}\left(S^2 + \frac{(K-3)^2}{4}\right)$$

N: Anzahl der Beobachtungen
S: Schiefe K: Kurtosis

Grosse Werte für Schiefe und Werte über 3 für die Kurtosis führen zu grossen Werten für die Jacque-Bera Statistik.
Es gilt $JB \sim \chi_2^2$ → die Teststatistik JB ist asymptotisch Chi-Quadrat-verteilt mit zwei Freiheitsgraden.

Das Hypothesenpaar lautet:
H_0: Die Stichprobe ist normalverteilt.
H_1: Die Stichprobe ist nicht normalverteilt.

Bei einem Signifikanzniveau $\alpha = 0.10$ gilt: Für Werte der Teststatistik über 4.6 wird die Hypothese der Normalverteilung verworfen;

Für die Signifikanzniveaus $\alpha = 0.05$ und $\alpha = 0.02$ ergeben sich die Schranken 6 und 7.8.