3. Einfache lineare Regression (Maddala Kapitel 3)

Regressionsrechnung: Empirische Analyse der Abhängigkeit einer Variablen (y) von einem oder mehreren Bestimmungsfaktoren (x_1, x_2, .. x_k).

einfache Regression   multiple Regression

Terminologie:

y:	*x:*
erklärte Variable	erklärende Variable
abhängige Variable	unabhängige Variable
endogene Variable	exogene Variable
Regressand	Regressor

z.B.	Konsumausgaben	Einkommen, Zinssatz
	Investitionen	Kapazitätsauslastung, Zinssatz
	Exporte	Wirtschaftsentwicklung im Ausland, Wechselkurs des Frankens
	Umsatz	Werbeaufwand

Wahl der Funktionsform:

allgemeine Formulierung: $y = f(x) + u$

$y_i = f(x_i) + u_i \qquad i = 1, 2, \ldots n$

$y_t = f(x_t) + u_t \qquad t = 1, 2, \ldots T$ (für Zeitreihen)

einfache lineare Regression: $y_i = \alpha + \beta x_i + u_i$

Ansatz mit konstanter Elastizität: $y_i = e^{\alpha} x_i^{\beta} e^{u_i}$

$\ln(y_i) = \alpha + \beta \ln(x_i) + u_i$

Quadratische Funktion: $y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + u_i$

Im Folgenden betrachten wir den Fall der einfachen linearen Regression.

Datengenerierender Prozess ("wahres" Modell):

$$\underbrace{y_i = \alpha + \beta x_i}_{\text{deterministischer Teil}} + u_i \qquad u_i \sim IN(0, \sigma^2)$$

Störterm: unabhängig normalverteilt mit Mittelwert 0 und Varianz σ^2

Regressionsparameter α und β (Achsenabschnitt und Steigung der Regressionsgeraden)

Das Modell enthält somit drei Parameter: α, β, σ^2

Stichprobe: $x_1 \ x_2 \ x_3 \ .... \ x_n$ exogen
$y_1 \ y_2 \ y_3 \ .... \ y_n$ endogen

=> Schätzwerte $\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2$

Die geschätzte Regressionsgleichung lautet somit:

$$y_i = \hat{\alpha} + \hat{\beta} x_i + \hat{u}_i$$

Residuum

Annahmen zum Störterm der Gleichung:

1) $E(u_i) = 0$
2) $\text{var}(u_i) = E(u_i^2) = \sigma^2$ konstant
3) u_i und u_j sind unabhängig für alle i und j $(i \neq j)$
4) u_i und x_j sind unabhängig für alle i und j, d.h. x ist exogen
5) u_i ist *normalverteilt*

Schätzung der Regressionsparameter nach der Methode der "kleinsten Quadrate" (Ordinary Least Squares, OLS)

In der geschätzten Regressionsgleichung

$$y_i = \hat{\alpha} + \hat{\beta} x_i + \hat{u}_i$$

werden die Parameter $\hat{\alpha}$ und $\hat{\beta}$ so festgelegt, dass die Summe der quadrierten Residuen minimal wird.

Summe der quadrierten Residuen (residual sum of squares):

$$RSS = \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{\alpha} - \hat{\beta} x_i \right)^2 \Rightarrow \min$$

Bei gegebener Stichprobe (Werte für x_i und y_i) ist *RSS* eine Funktion von $\hat{\alpha}$ und $\hat{\beta}$. Man sucht also das Minimum der Funktion $RSS(\hat{\alpha}, \hat{\beta})$.

Diese Minimierung führt zu folgenden Schätzformeln (vgl. Maddala):

$$\boxed{\hat{\beta} = S_{xy} / S_{xx}} \quad \text{und} \quad \boxed{\hat{\alpha} = \overline{y} - \hat{\beta}\, \overline{x}}, \quad \text{wobei}$$

$$S_{xx} = \sum \left(x_i - \overline{x} \right)^2 \qquad \overline{x} = \frac{\sum x_i}{n}$$

$$S_{yy} = \sum \left(y_i - \overline{y} \right)^2 \qquad \overline{y} = \frac{\sum y_i}{n}$$

$$S_{xy} = \sum \left(x_i - \overline{x} \right)\left(y_i - \overline{y} \right)$$

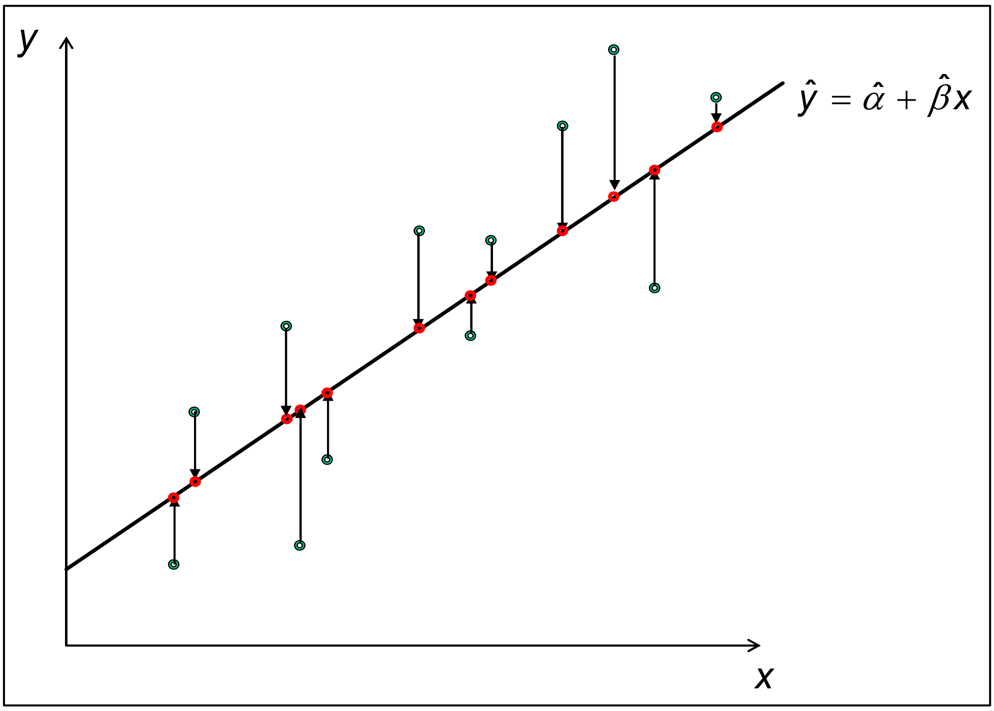
Die Schätzformel für $\hat{\alpha}$ impliziert, dass die Regressionsgerade genau durch den Punkt $(\overline{x}, \overline{y})$ läuft (Stichprobenmittelwert).

Schätzung der Störtermvarianz: $\boxed{\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2}}$

Man dividiert durch $n - 2$, weil der Berechnung der Residuen $\hat{u}$ zwei geschätzte Parameter zugrunde liegen ($\hat{\alpha}$ und $\hat{\beta}$) und die $\hat{u}$ deshalb tendenziell kleiner sind (absolut) als die unbeobachtbaren Störterme *u*.

Varianzanalyse

Auf Basis der geschätzten Regressionsgeraden kann man die Varianz der abhängigen Variablen *y* aufteilen in einen durch *x* erklärten Teil (Punkte auf der Regressionsgeraden) und eine unerklärte Restvarianz (Abweichungen der Beobachtungspunkte von der Regressionsgeraden).



Für diese Aufteilung verwendet man anstelle der Varianzen bequemer die folgenden Quadratsummen:

y_i: $TSS = S_{yy} = \sum (y_i - \bar{y})^2$ (total sum of squares)

$\hat{y}_i$: $ESS = \sum (\hat{y}_i - \bar{y})^2$ (explained sum of squares)

$\hat{u}_i$: $RSS = \sum \hat{u}_i^2$ (residual sum of squares)

Es gilt: $TSS = ESS + RSS$ (Gesamtvarianz in *y* gleich erklärte Varianz plus unerklärte Restvarianz)

Das Bestimmtheitsmass R^2 gibt an, welcher Anteil der Varianz von *y* durch *x* erklärt wird:

$$R^2 = ESS/TSS = (TSS - RSS)/TSS = 1 - RSS/TSS$$

Die Wurzel aus R^2 entspricht dem (positiven oder negativen) Korrelationskoeffizienten.

ESS kann wie folgt berechnet werden:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i \qquad \text{Punkte auf der Regressionsgeraden}$$

$$\hat{y}_i - \bar{y} = \hat{\alpha} + \hat{\beta} x_i - \bar{y}$$

$$\begin{aligned} \hat{y}_i - \bar{y} &= \hat{\beta} x_i - \hat{\beta} \bar{x} \qquad \text{da } \hat{\alpha} - \bar{y} = -\hat{\beta} \bar{x} \\ &= \hat{\beta}\left(x_i - \bar{x}\right) \end{aligned}$$

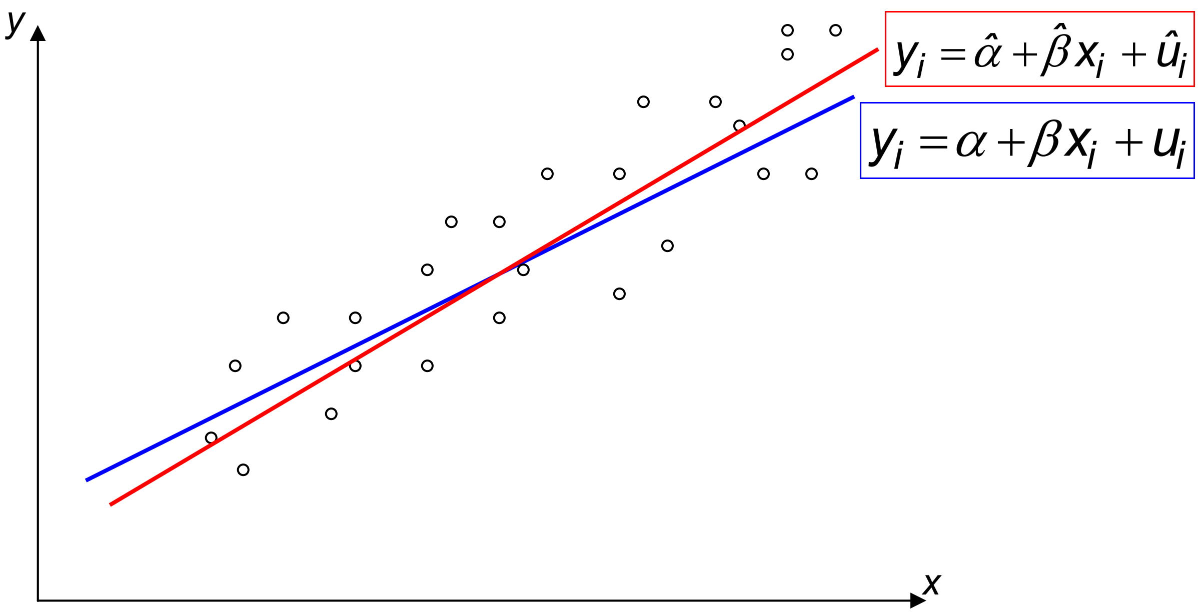
$$\begin{aligned} ESS = \sum (\hat{y}_i - \bar{y})^2 &= \hat{\beta}^2 \sum \left(x_i - \bar{x}\right)^2 \\ &= \hat{\beta}^2 S_{xx} \\ &= \hat{\beta} S_{xy} \qquad \text{da } \hat{\beta} S_{xx} = S_{xy} \end{aligned}$$

Statistische Aussagen

Welche statistischen Rückschlüsse lassen sich von den geschätzten Parametern auf die Parameter des datengenerierenden Prozesses ziehen?

Datengenerierender Prozess (wahres Modell) $\rightarrow$ Stichprobe

α, β, σ^2 $\leftarrow$ $\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2$



Da die Lage der Beobachtungspunkte einer vorliegenden Stichprobe von den Zufallsrealisationen des Störtems u_i abhängt, weicht die geschätzte Regressionsgerade mehr oder weniger stark vom wahren Zusammenhang im datengenerierenden Prozess ab. Die Schätzwerte für die Parameter haben mit anderen Worten den Charakter von Zufallsvariablen, die einer bestimmten ***Stichprobenverteilung*** unterliegen. Diese leitet sich aus der unterstellten Normalverteilung des Störterms u_i ab.

Stichprobenverteilung von $\hat{\alpha}$, $\hat{\beta}$ und $\hat{\sigma}^2$

$\Downarrow$

Vertrauensbereiche und Hypothesentests für α, β und σ^2

Stichprobenverteilung von $\hat{\alpha}$, $\hat{\beta}$ *und* $\hat{\sigma}^2$:

Die mit der Methode der kleinsten Quadrate geschätzten Parameter $\hat{\alpha}$, $\hat{\beta}$ und $\hat{\sigma}^2$ sind in Relation zu den wahren Parameterwerten wie folgt verteilt:

$$\hat{\alpha} \sim N\left(\alpha,\ \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$$

$$\hat{\beta} \sim N\left(\beta,\ \frac{\sigma^2}{S_{xx}}\right)$$

=> Die Schätzungen $\hat{\alpha}$, $\hat{\beta}$ sind normalverteilt, unverzerrt und konsistent.

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} = \frac{RSS}{\sigma^2} \sim \chi^2_{n-2}$$

=> Die Schätzung $\hat{\sigma}^2$ ist Chi-Quadrat-verteilt, unverzerrt und konsistent.

Die Verteilung von $\hat{\beta}$ wird in eine Standard-Normalverteilung übergeführt:

$$\frac{\hat{\beta} - \beta}{\sqrt{\sigma^2 / S_{xx}}} \sim N(0, 1)$$

Die Störterm-Varianz σ^2 ist unbekannt und muss durch die Schätzung $\hat{\sigma}^2$ ersetzt werden, welche Chi-Quadrat-verteilt ist. Der resultierende Ausdruck ist das Verhältnis einer standard-normalverteilten und der Wurzel aus einer gemittelten Chi-Quadrat-verteilten Variablen und ist somit *t*-verteilt:

$$\frac{\dfrac{\hat{\beta} - \beta}{\sqrt{\sigma^2 / S_{xx}}}}{\sqrt{\dfrac{(n-2)\hat{\sigma}^2}{(n-2)\sigma^2}}} = \frac{\hat{\beta} - \beta}{\sqrt{\hat{\sigma}^2 / S_{xx}}} = \frac{\hat{\beta} - \beta}{SE(\hat{\beta})} \sim t_{n-2}$$

Bei $\sqrt{\hat{\sigma}^2 / S_{xx}}$ handelt es sich um den geschätzten Standardfehler von $\hat{\beta}$, wofür kurz auch $SE(\hat{\beta})$ geschrieben wird.

Prognose auf Basis einer einfachen linearen Regression

Let x_0 be the given value of x. Then we predict the corresponding value y_0 of y by

$$\hat{y}_0 = \hat{\alpha} + \hat{\beta} x_0 \tag{3.11}$$

The true value y_0 is given by

$$y_0 = \alpha + \beta x_0 + u_0$$

where u_0 is the error term.

Hence the prediction error is

$$\hat{y}_0 - y_0 = (\hat{\alpha} - \alpha) + (\hat{\beta} - \beta) x_0 - u_0$$

Since $E(\hat{\alpha} - \alpha) = 0$, $E(\hat{\beta} - \beta) = 0$, and $E(u_0) = 0$ we have

$$E(\hat{y}_0 - y_0) = 0$$

This equation shows that the predictant given by equation (3.11) is unbiased. Note that the predictant is unbiased in the sense that $E(\hat{y}_0) = E(y_0)$ since both $\hat{y}_0$ and y_0 are random variables.

The variance of the prediction error is

$$\begin{aligned} V(\hat{y}_0 - y_0) &= V(\hat{\alpha} - \alpha) + x_0^2 V(\hat{\beta} - \beta) + 2x_0 \operatorname{cov}(\hat{\alpha} - \alpha, \hat{\beta} - \beta) + V(u_0) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) + \sigma^2 \frac{x_0^2}{S_{xx}} - 2x_0 \sigma^2 \frac{\bar{x}}{S_{xx}} + \sigma^2 \\ &= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right] \end{aligned}$$

Thus the variance increases the farther away the value of x_0 is from $\bar{x}$, the mean of the observations on the basis of which $\hat{\alpha}$ and $\hat{\beta}$ have been computed. This is illustrated in Figure 3.5, which shows the confidence bands for y.

If x_0 lies within the range of the sample observations on x, we can call it within-sample prediction, and if x_0 lies outside the range of the sample observations, we call the prediction out-of-sample prediction. As an illustration, consider a consumption function estimated on the basis of 12 observations. The equation is (this is just a hypothetical example)

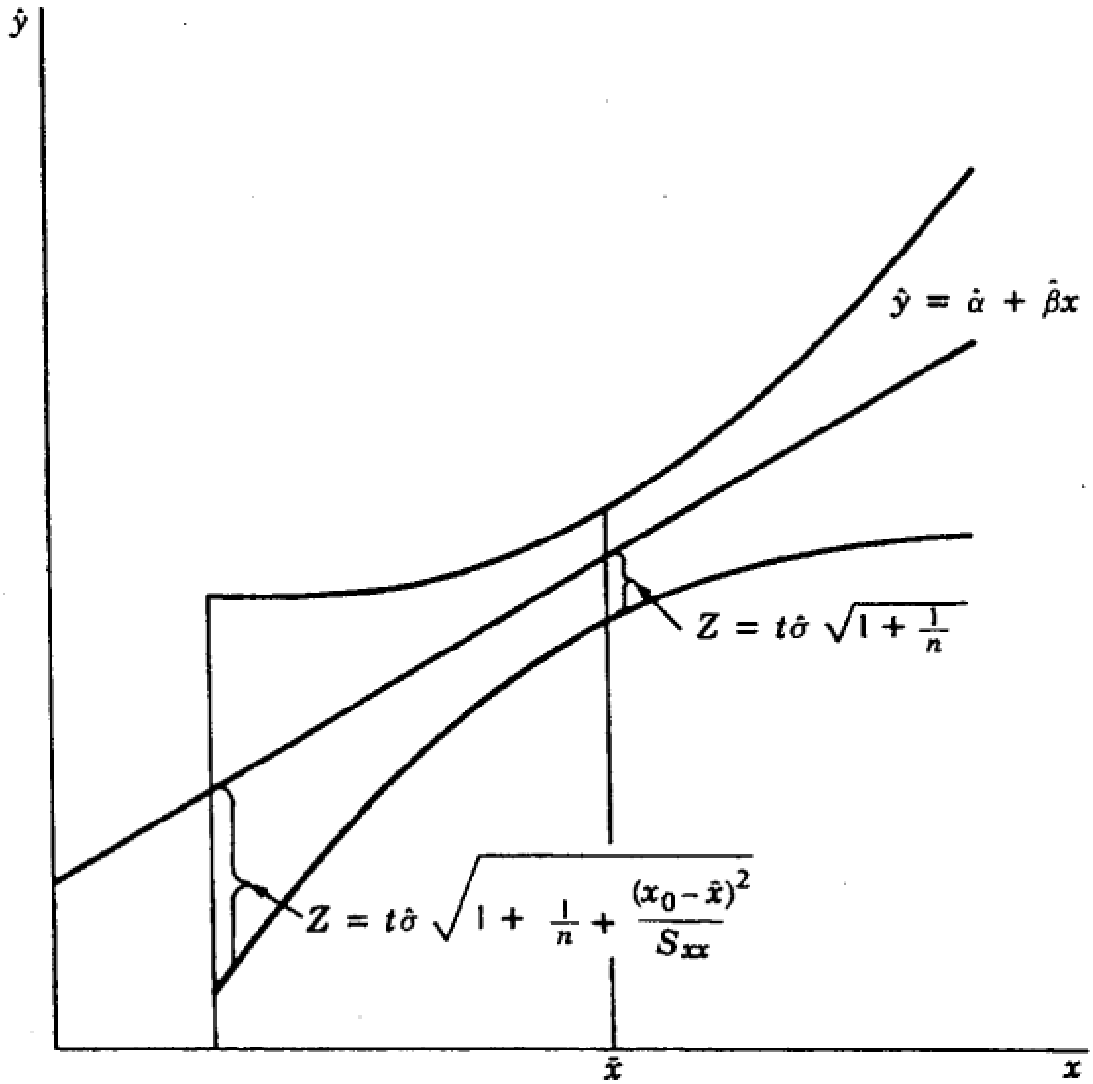
$$y = 10.0 + 0.90x$$

where y = consumer expenditures
x = disposable income

We are given $\hat{\sigma}^2 = 0.01, \bar{x} = 200$, and $S_{xx} = 4000$. Given $x_0 = 250$, our prediction of y_0 is

$$\hat{y}_0 = 10.0 + 0.9(250) = 235$$

$$\mathrm{SE}(\hat{y}_0) = \sqrt{0.01 \left(1 + \frac{1}{12} + \frac{2500}{4000} \right)} = 0.131$$



Der in der Grafik dargestellte 95%-Vertrauensbereich für *y* ist beim Stichprobenmittel von *x* am engsten und wird mit zunehmender Entfernung von *x* gegenüber dem Stichprobenmittel immer grösser, weil Fehlschätzungen von *β* mit zunehmendem Abstand vom Stichprobenmittelwert immer stärker ins Gewicht fallen.

Verschiedene Funktionsformen, die sich als lineare Regressionen schätzen lassen

a) Linearer Grundansatz:

$y_t = \alpha + \beta x_t + u_t$ datengenerierender Prozess

$y_t = \hat{\alpha} + \hat{\beta} x_t + \hat{u}_t$ Schätzgleichung

$\hat{y}_t = \hat{\alpha} + \hat{\beta} x_t$ theoretische Werte auf der Regressionsgeraden

$\hat{u}_t = y_t - \hat{y}_t$ Residuen

b) Logarithmischer Ansatz:

$$\ln y_t = \alpha + \beta \ln x_t + u_t$$

$y_t = A\, x_t^{\beta} e^{u_t}$ zugrundeliegender multiplikativer Ansatz mit Elastizität β

$$u_t = 0 \quad \Leftrightarrow \quad e^{u_t} = 1$$
$$u_t = 0.01 \quad \Leftrightarrow \quad e^{u_t} \cong 1.01$$
$$u_t = -0.01 \quad \Leftrightarrow \quad e^{u_t} \cong 0.99$$

$$\ln y_t = \hat{\alpha} + \hat{\beta} \ln x_t + \hat{u}_t$$
$$\ln \hat{y}_t = \hat{\alpha} + \hat{\beta} \ln x_t$$

$$\hat{u}_t = \ln y_t - \ln \hat{y}_t = \ln\left(y_t / \hat{y}_t\right) \cong y_t / \hat{y}_t - 1 = \frac{y_t - \hat{y}_t}{\hat{y}_t}$$

relative Abweichung

c) Gleichung in ersten Differenzen:

$$\Delta y_t = \gamma + \beta \Delta x_t + u_t$$ Differenzengleichung

$$y_t - y_{t-1} = \gamma + \beta(x_t - x_{t-1}) + u_t$$

$$y_t = y_{t-1} + \gamma + \beta(x_t - x_{t-1}) + u_t$$

$$y_t = \underbrace{(y_0 - \beta x_0)}_{\alpha} + \gamma t + \beta x_t + \underbrace{\sum_{i=1}^{t} u_i}_{v_t}$$ Niveaugleichung

$$y_t = \alpha + \gamma t + \beta x_t + v_t$$

Wenn u_t in der Differenzengleichung ein zeitlich unkorrelierte Zufalls-variable ist, dann ist v_t in der Niveaugleichung hochgradig autokorreliert, weil v_t. und v_{t-1} sich nur durch das zuletzt aufaddierte u_t unterscheiden. Umgekehrt folgt daraus, dass man eine ausgeprägte Autokorrelation der Residuen in einer Niveaugleichung als Resultat einer Kumulation von Störungen über die Zeit interpretieren kann. Geht man dann zu einer Differenzengleichung über, so wird diese Kumulation rückgängig gemacht, d.h. man erhält durch die Differenzenbildung eine Gleichung, in der die einzelnen Störungen pro Periode voneinander unabhängig sind.

d) Gleichung in ersten Differenzen logarithmierter Variablen:

$$\Delta \ln y_t = \gamma + \beta \Delta \ln x_t + u_t$$

$$\Delta \ln y_t = \ln y_t - \ln y_{t-1} = \ln(y_t / y_{t-1}) \cong y_t / y_{t-1} - 1 = \frac{y_t - y_{t-1}}{y_{t-1}}$$

Veränderungsrate

analog für $\Delta \ln x_t$.

Spezifikationstests

Datengenerierender Prozess: $y_i = \alpha + \beta x_i + u_i$

Annahmen bezüglich des Störterms:

1) $E(u_i) = 0$
2) $var(u_i) = E(u_i^2) = \sigma^2$ ist konstant
3) u_i und u_j sind unabhängig für alle $i \neq j$
4) u_i und x_j sind unabhängig für alle *i* und *j*
5) u_i ist normalverteilt

Geschätzte Regressionsgleichung: $y_i = \hat{\alpha} + \hat{\beta} x_i + \hat{u}_i$

Die Residuen einer geschätzten Regressionsgleichung können Aufschluss darüber geben, ob die der Schätzung zugrundeliegende Stichprobe mit dem unterstellten datengenerierenden Prozess vereinbar ist und die ***Spezifikation des Modells*** somit als ***adäquat*** zu betrachten ist:

- Ist es zulässig, von einem linearen Zusammenhang zwischen *x* und *y* auszugehen?
- Sind wichtige Einflussfaktoren vernachlässigt worden?
- Sind die Annahmen bezüglich des Störterms u_i - insbesondere konstante Varianz (2) und keine Autokorrelation (3) - haltbar?

⇨ ***Beurteilung anhand der berechneten Residuen*** $\hat{u}_i$ (vgl. die folgenden Beispiele). Wenn die Annahmen zum datengenerierenden Prozess verletzt sind, so sind die oben hergeleiteten Eigenschaften der Schätzungen von $\hat{\alpha}$, $\hat{\beta}$ *und* $\hat{\sigma}^2$ (z.B. die Eigenschaft, dass $\hat{\beta}$ *t*-verteilt ist) nicht gültig, denn sie basieren alle auf der Annahme, dass der Störterm der Gleichung unabhängig normalverteilt ist.

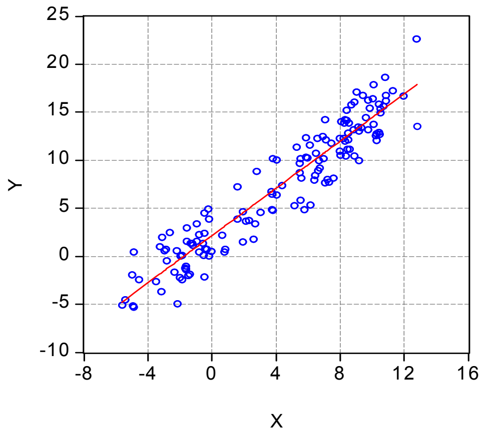
Residuen aus verschiedenen Regressionsgleichungen

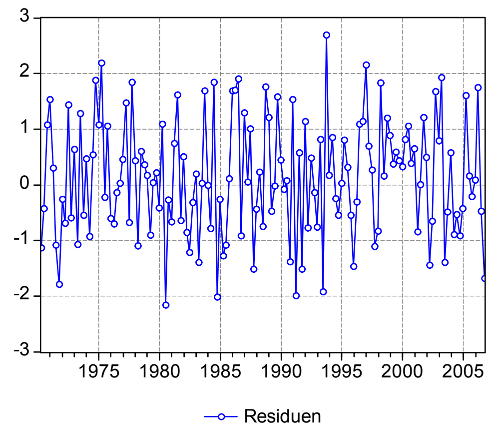
Nur im Fall 1 erscheint die Modellspezifikation adäquat.

Residuen als Abweichungen von von der Regressionsgeraden

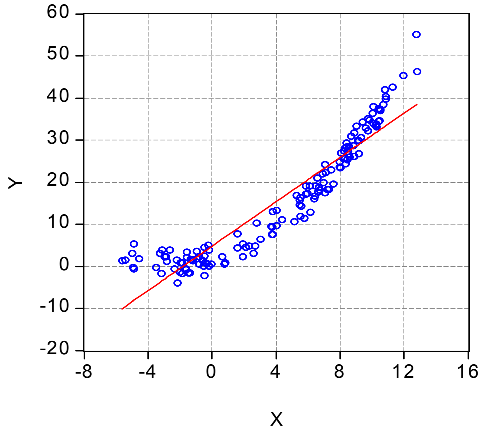
Residuen über die Zeit

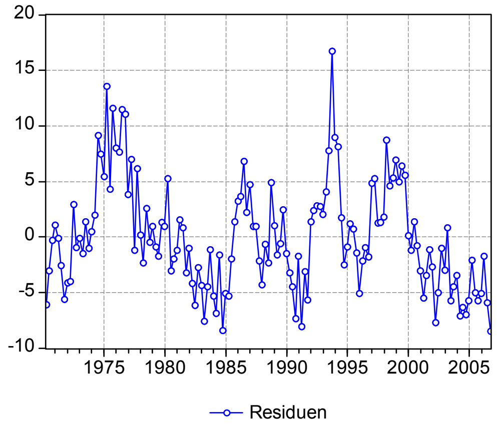
Fall 1:



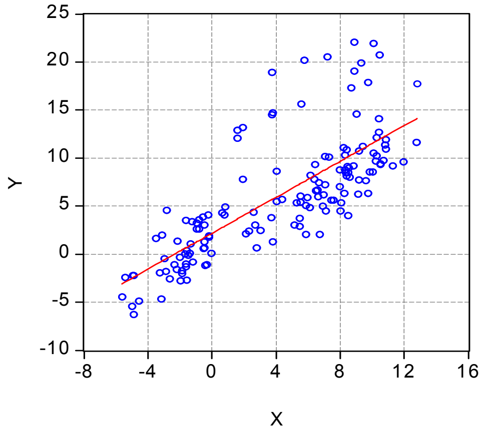


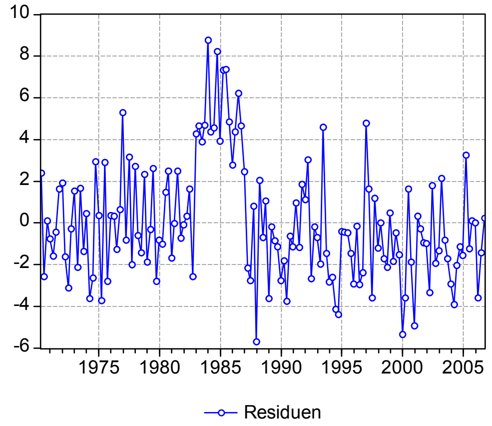
Fall 2:





Fall 3:

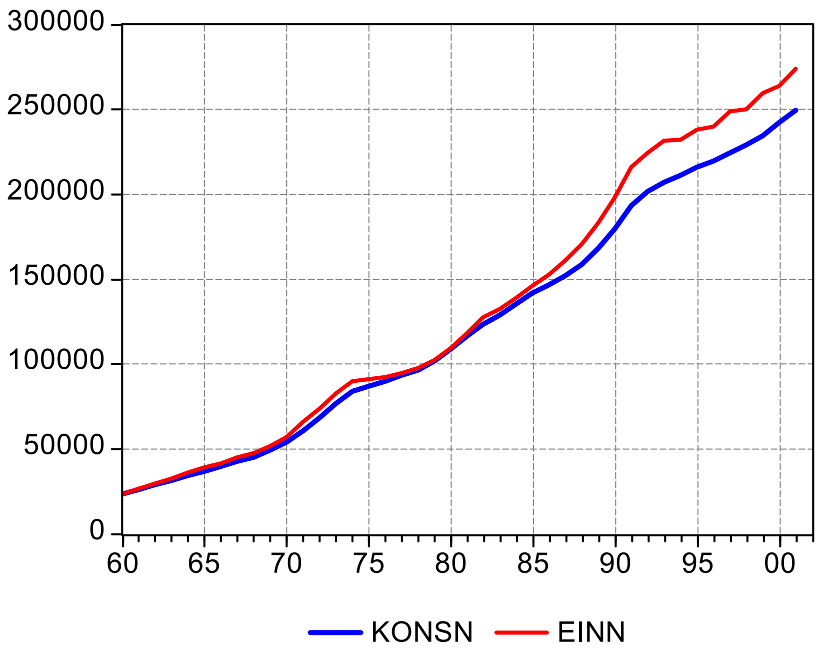




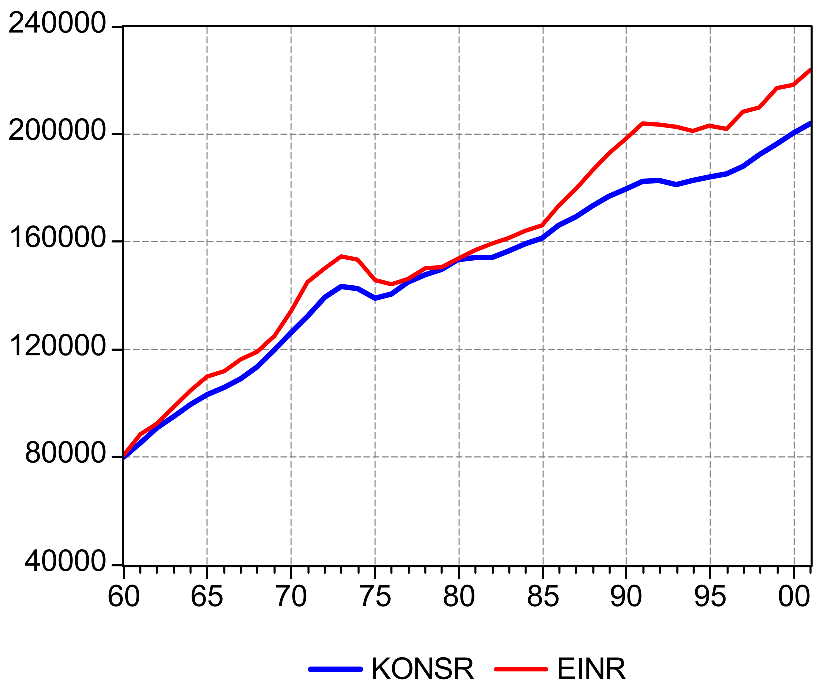
Beispiel: Schätzung einer Konsumfunktion für die Schweiz

Daten: Konsumausgaben und Haushaltseinkommen aus der Volkswirtschaftlichen Gesamtrechnung der Schweiz (Jahreswerte 1960 bis 2001)

Konsum und Einkommen
nominell, in Mio Franken

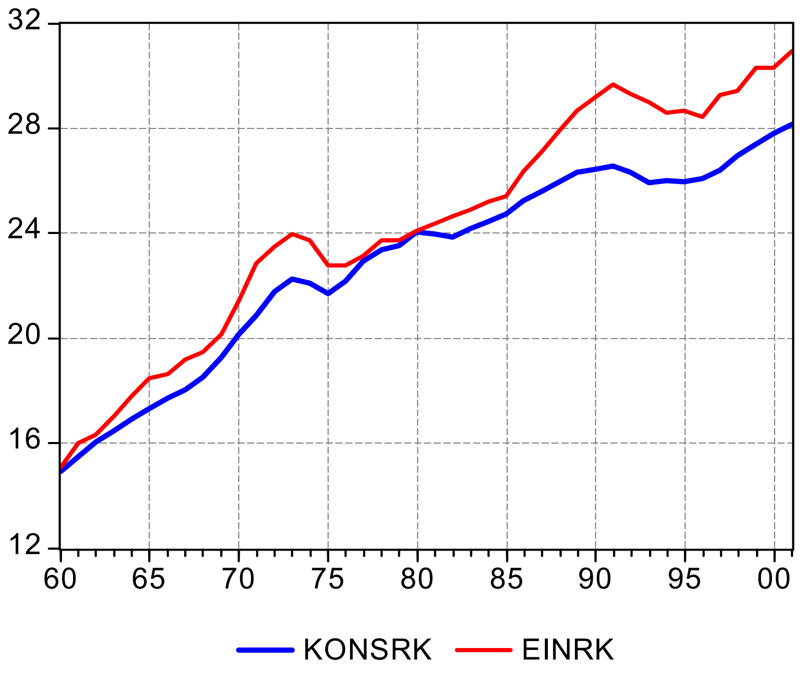


Konsum und Einkommen
real, in Mio Franken zu Preisen von 1990, 1960-2001



Konsum und Einkommen

real pro Kopf, in 1000 Franken zu Preisen von 1990



Schätzung einer linearen Regression: Zwei Darstellungsarten in EVIEWS

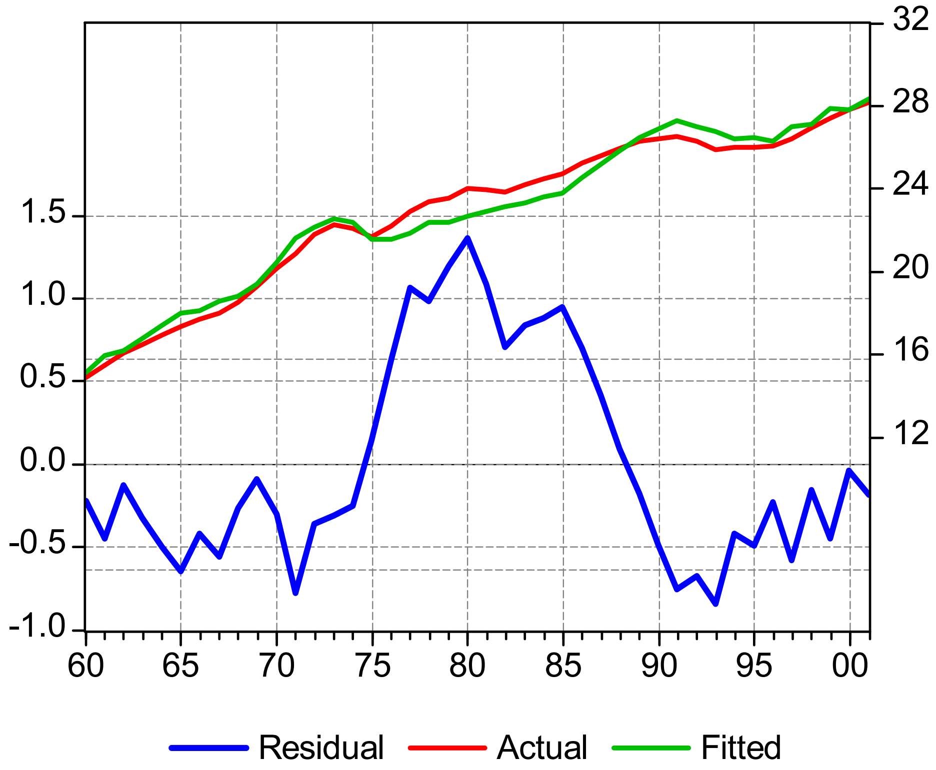
Sample: 1960 2001
Included observations: 42
KONSRK=C(1)+C(2)*EINRK

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	2.580351	0.545681	4.728683	0.0000
C(2)	0.833541	0.022067	37.77306	0.0000
R-squared	0.972730	Mean dependent var		22.85573
Adjusted R-squared	0.972048	S.D. dependent var		3.807354
S.E. of regression	0.636545	Akaike info criterion		1.980925
Sum squared resid	16.20758	Schwarz criterion		2.063671
Log likelihood	-39.59942	Durbin-Watson stat		0.192419

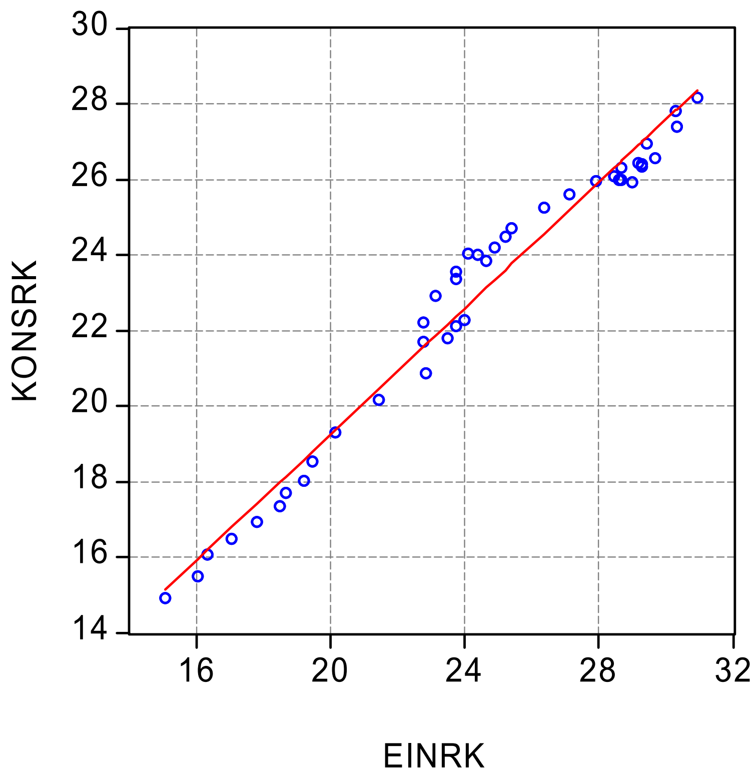
Dependent Variable: KONSRK
Sample: 1960 2001
Included observations: 42

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2.580351	0.545681	4.728683	0.0000
EINRK	0.833541	0.022067	37.77306	0.0000
R-squared	0.972730	Mean dependent var		22.85573
Adjusted R-squared	0.972048	S.D. dependent var		3.807354
S.E. of regression	0.636545	Akaike info criterion		1.980925
Sum squared resid	16.20758	Schwarz criterion		2.063671
Log likelihood	-39.59942	Durbin-Watson stat		0.192419

Tatsächliche Werte von KONSRK (y), theoretische Gleichungswerte ($\hat{y}$) und Residuen ($\hat{u} = y - \hat{y}$):



Residuen als Abweichungen der Beobachtungspunkte von der Regressionsgeraden



Die vorstehende Gleichung erscheint aufgrund der Verteilung der Residuen fehlspezifiziert. Ignoriert man diesen Umstand, so lässt sich die Berechnung von ***Vertrauensintervallen und Hypothesentests*** am Beispiel der marginalen Konsumquote β wie folgt demonstrieren:

Punktschätzung und geschätzter Standardfehler:

$$C(2) = \hat{\beta} = 0.833541, \quad SE(\hat{\beta}) = 0.022067$$

Stichprobenverteilung von $\hat{\beta}$:

$$\frac{\hat{\beta}-\beta}{SE(\hat{\beta})} \sim t_{40} \qquad t\text{-Verteilung mit 42-2 = 40 Freiheitsgraden}$$

a) 95%-Vertrauensintervall für β:

kritischer t-Wert: $t^* = 2.021$

$$PROB(-2.021 \leq \frac{\hat{\beta}-\beta}{SE(\hat{\beta})} \leq 2.021) = 0.95$$

$$PROB(0.833541 - 2.021 * 0.022067 \leq \beta \leq 0.83351 + 2.021 * 0.022067) = 0.95$$

$$PROB(0.788944 \leq \beta \leq 0.878138) = 0.95$$

b) Hypothesentest H_0: $\beta = 0$

Vergleich des berechneten t-Wertes mit dem kritischen t-Wert für einen zweiseitigen Test auf dem 5%-Niveau:

$$t_{berechnet} = \frac{\hat{\beta}-0}{SE(\hat{\beta})} = 37.77306 >> t^* = 2.021$$

Berechnung der beobachteten Signifikanzniveaus:

$$PROB(|t| \geq 37.77306) = 0.0000$$

$$\left.\begin{array}{l}\text{Der berechnete t-Wert ist viel grösser als der kritische t-Wert.}\\ \text{Das beobachtete Signifikanzniveau ist praktisch gleich Null.}\\ \text{Das 95\%-Vertrauensintervall schliesst den Wert } \beta = 0 \text{ nicht ein.}\end{array}\right\} H_0 \text{ wird verworfen}$$