Einfache lineare Regression: Übung 2

Simulationsexperiment mit künstlich generierten Stichproben

Wahres Modell (datengenerierender Prozess):

$y_t = \alpha + \beta x_t + u_t$ mit $u_t \sim IN(0, \sigma^2)$

$\alpha = 4,\ \beta = 2$ und $\sigma = 1.5$

Mit diesem Modell werden 1000 künstliche Stichproben mit je 100 Beobachtungswerten (t = 1 bis 100) generiert. Dann werden für jede Stichprobe die Parameter α, β und σ mit der Methode der kleinsten Quadrate geschätzt. Dies ergibt 1000 Schätzwerte für die drei Parameter, die man mit den wahren Werten des datengenerierenden Prozesses vergleichen kann.

Zurest wird die Exogene x nach folgendem Schema generiert:

$(x_t - \mu_x) = 0.7071\,(x_{t-1} - \mu_x) + v_t$ mit $\mu_x = 3$ und Startwert $x_0 = \mu_x$

$v_t \sim IN(0,1)$

Die resultierende Variable x schwankt mit einer Varianz von 2 um einen Mittelwert von 3.

Dann wird die endogene Variable y als lineare Funktion von x bestimmt, wobei dieser Zusammenhang von einem Zufallseinfluss u überlagert wird:

$y_t = 4 + 2x_t + u_t,$ $u_t \sim IN(0, \sigma^2)$

Dieser Schritt wird 1000 mal wiederholt, indem mit einem Zufallszahlen-Generator immer wieder neue 100 Werte aus $u_t \sim IN(0, \sigma^2)$ gezogen werden. Dies ergibt 1000 Stichproben mit je 100 Beobachtungswerten für x und y. Für jede Stichprobe wird mit der Methode der kleinsten Quadrate eine Regressionsgerade geschätzt. So erhält man 1000 Schätzwerte für α, β und σ. Deren Verteilung entspricht der sogenannten Stichprobenverteilung.

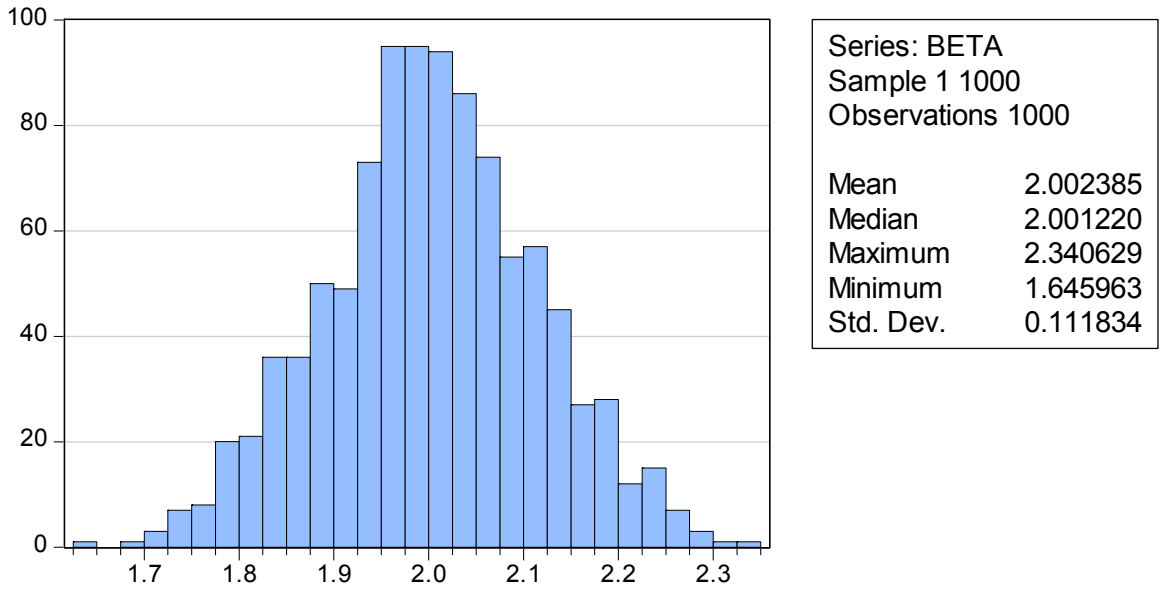
Für die Stichprobe Nummer 15 ergibt sich zum Beispiel:

Method: Least Squares
Included observations: 100 after adjustments
Y = C(1)+C(2)*X

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	3.756086	0.374231	10.03681	0.0000
C(2)	1.979577	0.113718	17.40783	0.0000

R-squared	0.755631	Mean dependent var	9.717834
Adjusted R-squared	0.753137	S.D. dependent var	3.036370
S.E. of regression	1.508629	Akaike info criterion	3.680077
Sum squared resid	223.0442	Schwarz criterion	3.732180
Log likelihood	-182.0038	Hannan-Quinn criter.	3.701164
F-statistic	303.0325	Durbin-Watson stat	2.360395

Verteilung der 1000 geschätzten BETA's



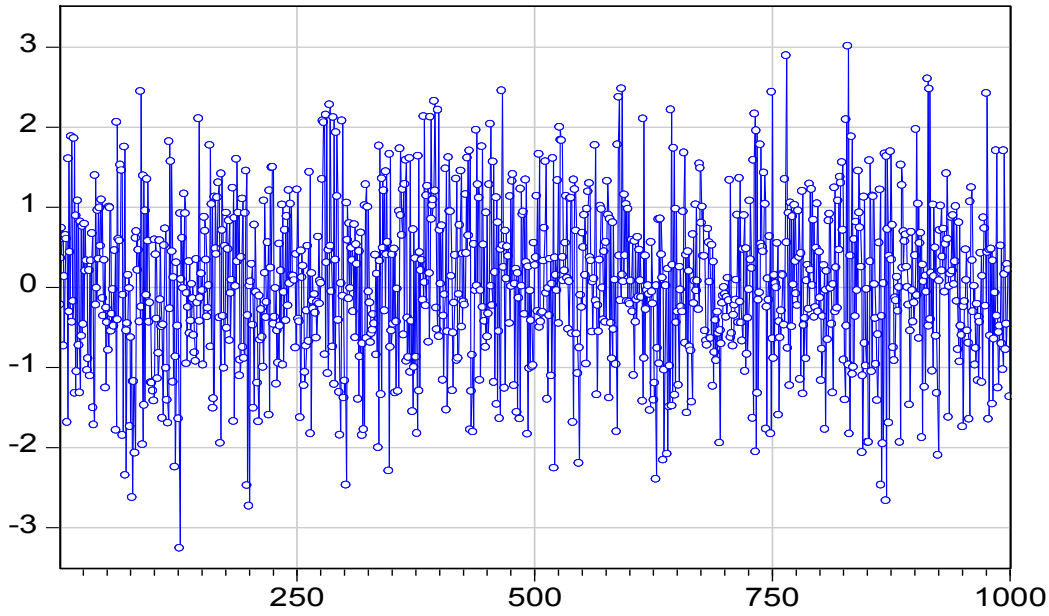
Die 1000 geschätzten BETA's sind symmetrisch um den wahren Wert von 2 verteilt. Ihr Mittelwert stimmt praktisch mit dem wahren Wert von 2 überein. Die Standardabweichung beträgt 0.1118. Dies entspricht ungefähr dem in der Stichprobe Nr. 15 geschätzten "Standard Error" von BETA (0.1137).

Testet man die richtige Nullhypothese BETA = 2 (richtig, weil die Stichproben mit diesem Parameterwert generiert wurden) mit einem *t*-Test,

$$t = \frac{\hat{\beta} - 2}{SE(\hat{\beta})} \ ,$$

so ergibt sich in 48 der 1000 Stichproben ein t-Wert, der absolut grösser ist als der kritische Wert $t^* = 1.9845$ (5%-Niveau). Die richtige Nullhypothese wird also fälschlicherweise in 48 der 1000 Stichproben verworfen, was aufgrund des gewählten Signifikanzniveaus von 5% in etwa zu erwarten war.

t-Werte für die Hypothese BETA = 2 (Stichproben 1 bis 1000)



Verworfen wird die Hypothese BETA = 2 z.B. in der Stichprobe Nr. 77:

Method: Least Squares
Included observations: 100 after adjustments
Y = C(1)+C(2)*X

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	4.941022	0.387209	12.76060	0.0000
C(2)	1.691168	0.117661	14.37318	0.0000

R-squared	0.678254	Mean dependent var	10.03419
Adjusted R-squared	0.674971	S.D. dependent var	2.737963
S.E. of regression	1.560949	Akaike info criterion	3.748262
Sum squared resid	238.7829	Schwarz criterion	3.800365
Log likelihood	-185.4131	Hannan-Quinn criter.	3.769349
F-statistic	206.5882	Durbin-Watson stat	2.182571

Die Schätzung für BETA weicht hier statistisch signifikant vom Wert 2 ab:

$$t = \frac{\hat{\beta} - 2}{SE(\hat{\beta})} = \frac{1.6912 - 2}{0.1177} = -2.62$$

Das Simulationsexperiment wurde mit dem Programm simpreg.prg in *EViews Version 6* durchgeführt. Je nach EViews-Version können sich wegen eines anderen Zufallszahlengenerators numerisch leicht abweichende Resultate ergeben. Sie finden das Programm auf der Internetseite zur Vorlesung im Anhang unter *Simulationsprogramm 2*.

```
' SIMPREG:
' Schätzung einer Gleichung  y = alfa + beta*x + u
' in 1000 künstlich generierten Stichproben mit je 100 Beobachtungswerten

' Öffnen eines Workfiles und Festlegen des Stichprobenumfangs
        workfile simpreg U  1 1000

' Generieren einer erklärenden Variablen x mit Erwartungswert 3
' und Standardabweichung 2
         smpl 1 100                ' oder: smpl 1 50
        rndseed(type=kn) 76543      '513
        genr v = nrnd              ' oder: genr v = 2*nrnd
        genr x = 3
        smpl 2 100
        genr x = 3+0.7071*(x(-1)-3)+v
        smpl 1 100

' Initialisieren der Vektoren für Ablage der Schätzresultate
        series(1000) alfa
        series(1000) beta
        series(1000) sdbeta
        series(1000) tbeta2
        series(1000) seofeq
        series(1000) rquadrat
        smpl 1 1000

' Generieren von y mit einem "wahren Modell" mit den Parametern
' alfa=4,  beta=2  und sigma=1.5
' mit 1000 Replikationen

FOR !j = 1 to 1000
        scalar xseed = 8501+!j
        rndseed(type=kn) xseed
        genr u = 1.5*nrnd                  ' oder: genr u = 3*nrnd
        genr y = 4 + 2*x + u

        ' Schätzung der Gleichung
        equation test!j.ls y = c(1)+c(2)*x
        alfa(!j) = c(1)
        beta(!j) = c(2)
        sdbeta(!j) = sqr(@covariance(2,2))
        tbeta2(!j) = (beta(!j)-2)/sdbeta(!j)
        seofeq (!j) = @se
        rquadrat(!j) = @r2
NEXT
```

Erläuterungen zum EViews-Programm:

Der Parameter *BETA* wird aus 1000 Stichproben mit je 100 Beobachtungen geschätzt.

Der wahre Wert im datengenerierenden Prozess ist *BETA* = 2.

Die 1000 Schätzgleichungen werden vom Programm im Workfile *simpreg* abgelegt: *test1* bis *test1000.*

Die 1000 Schätzungen für *BETA* werden in einen Vektor *BETA* geschrieben.

Die Standardfehler dieser Schätzungen werden in einen Vektor *SDBEAT* geschrieben.

Die Standardfehler der Gleichung werden in einen Vektor *SEOFEQ* geschrieben.

Für jede Schätzung von *BETA* wird mit einem t-Test die richtige Hypothese *BETA* = 2 getestet. Die t-Werte dieser Tests werden in einen Vektor *TBETA2* geschrieben.

Beantworten Sie die folgenden Fragen:

1. In jeder einzelnen Gleichung wird der Standardfehler der Gleichung auf Basis der berechneten Residuen geschätzt. Dies ist eine Schätzung für die Standardabweichung des Störterms im datengenerierenden Prozess, die mit σ = 1.5 vorgegeben ist. Wie gut stimmen die 1000 Schätzungen mit diesem Wert überein?

2. Die "Präzision" bzw. "Unsicherheit" der Schätzung von *BETA* wird in jeder einzelnen Gleichung vom Standardfehler von *BETA* gemessen. (Je grösser dieser Standardfehler, desto weiter wird z.B. ein 95%-Vertrauensbereich für *BETA*.) Im den hier durchgeführten Simulationen lässt sich die Schätzunsicherheit auch experimentell bestimmen, indem man sich die Streuung der 1000 geschätzten *BETA's* anschaut. (Bei praktischen Fragestellung ist dies nicht möglich, da man nur über eine Stichprobe und somit nur einen Schätzwert für *BETA* verfügt.) Wie gut stimmen die aus den einzelnen Stichproben abgeleiteten Standardfehler mit der Streuung der *BETA's* in den 1000 Schätzungen überein?

3. Beim gegebenen Stichprobenumfang von 100 Beobachtungen ist der kritische t-Wert für einen Test auf dem 5%-Signifikanzniveau 1.9845. In wie vielen der 1000 Stichproben wird dieser kritische t-Wert (positiv oder negativ) überschritten und die richtige Nullhypothese *BETA* = 2 somit fälschlicherweise abgelehnt?

4. Modifizieren Sie den datengenerierenden Prozess und schauen Sie, wie sich dadurch die Präzision der Schätzung von *BETA* (gemessen an den einzelnen Standardfehlern bzw. an der Streuung der 1000 geschätzen *BETA's*) verändert.
   a) Vergrössern Sie die Varianz der erklärenden Variablen (z.B. genr v = 2*nrnd).
   b) Variieren Sie den Stichprobenumfang (z.B. smpl 1 50 oder smpl 1 500).
   c) Vergrössern Sie die Varianz des Störterms (z.B. genr u = 3*nrnd).
   d) Die Stichprobenverteilungen von ALFA und BETA sind nicht unabhängig voneinander, sondern aufgrund der im Simulationsexperiment getroffenen Annahmen negativ korreliert, d.h. Überschätzungen von BETA gehen typischerweise mit Unterschätzungen von ALFA einher. Zeigen Sie dies anhand eines Scatterplots.