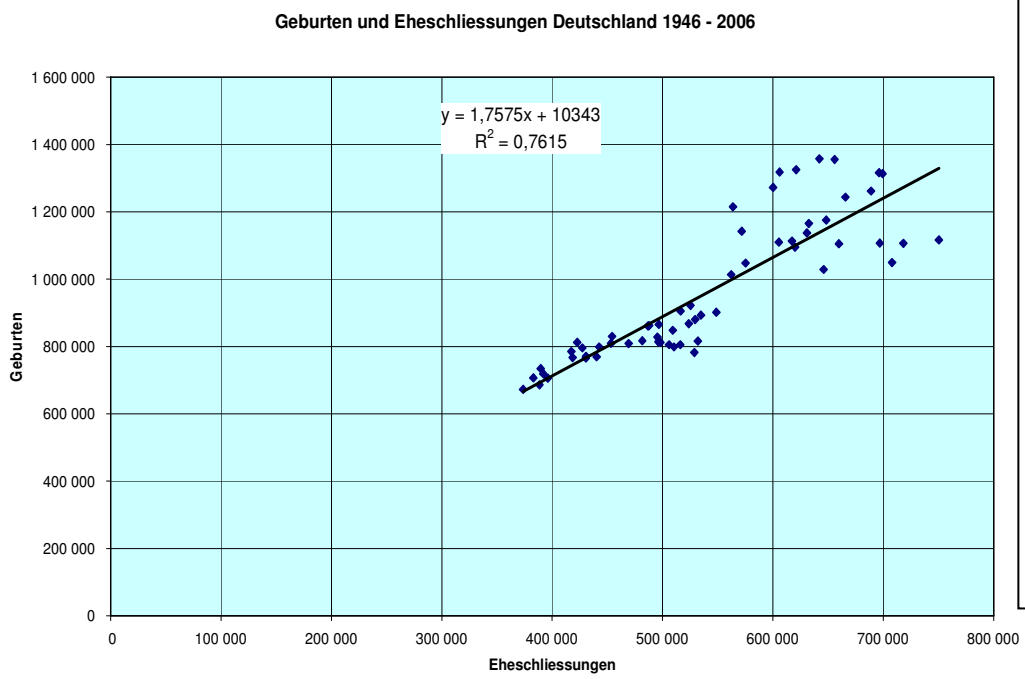
Prof. Dr. Peter von der Lippe (Übungsbeispiel F)

Übungsaufgaben zur Entwicklung der Geburten in Deutschland (Excel, EViews)

(auch Hinweise zur **Konfidenzellipse** und den "**Diagnostic Tests**", d.h. den **Annahmen B2 – B4** [Homoskedastizität, keine Autokorrelation, Normalverteiltheit] sowie zu (den Schwierigkeiten) der **Spezifikation** [Seite 7]).

1. Einfache Regression: Geburten in Abhängigkeit von der Anzahl der Eheschließungen

Daten der amtlichen Statistik (online service "GENESIS") für Deutschland insgesamt von 1946 bis 2006: Streuungsdiagramm



Die von Excel errechnete Regressionsgerade ergibt

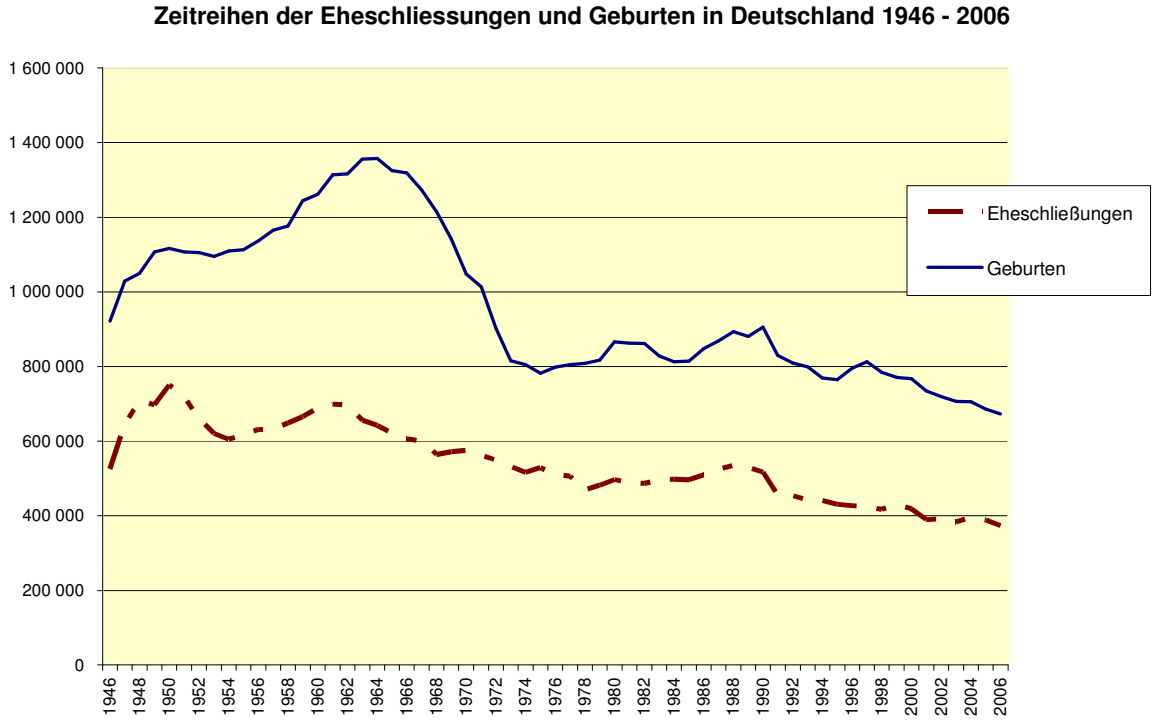
$\hat{\alpha} = 10343$

$\hat{\beta} = 1{,}7575$

und eine Bestimmtheit von $r^2 = 0{,}7615$.

Die Regressionsgerade ist als "Trend" in das Streuungsdiagramm eingezeichnet. Das Diagramm legt des Gedanken nahe, dass die Störgröße heteroskedastisch ist. Dies wird auch durch den White Test (siehe S. 3 unten) bestätigt.

Die mit Excel erstellte Graphik der beiden Zeitreihen, lässt vermuten, dass ab Mitte der 60er bis Anfang der 70er Jahre ein Strukturbruch bei den Geburten vorlag (der sog. "Pillenknick").



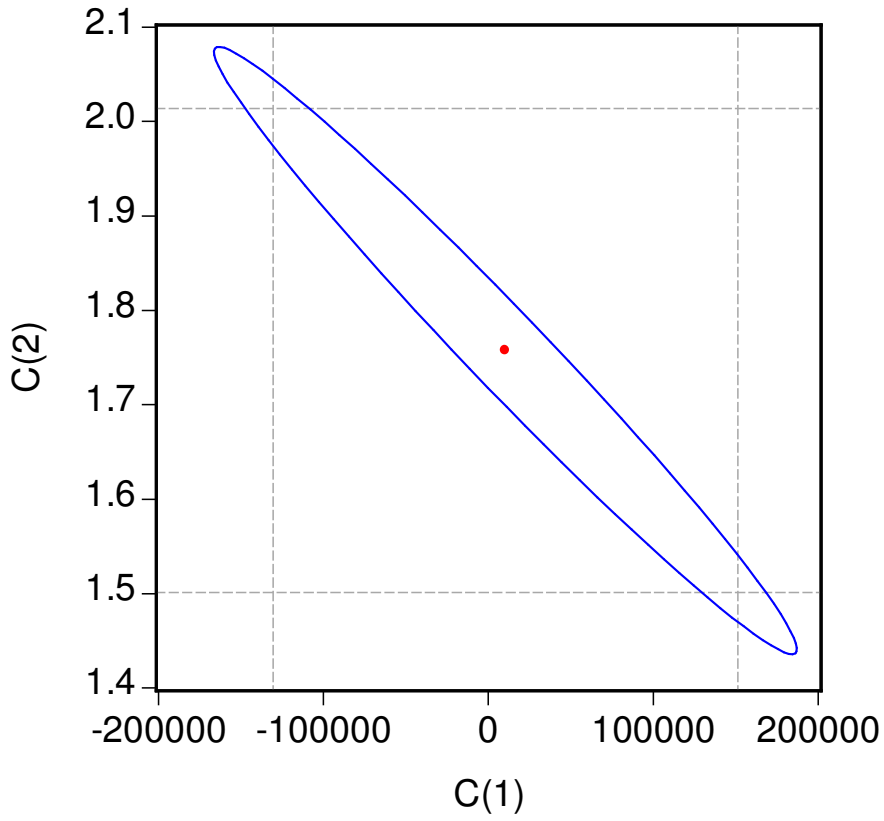
Die mit EViews[1] berechnete **Regressionsgleichung** lautet

Dependent Variable: GEB
Method: Least Squares
Date: 02/18/08 Time: 15:00
Sample: 1946 2006
Included observations: 61

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	10343.32	70325.26	0.147078	0.8836
EHEN	1.757469	0.128065	13.72329	0.0000

R-squared	0.761451	Mean dependent var	959278.8
Adjusted R-squared	0.757408	S.D. dependent var	203209.3
S.E. of regression	100088.0	Akaike info criterion	25.89772
Sum squared resid	5.91E+11	Schwarz criterion	25.96693
Log likelihood	-787.8806	F-statistic	188.3287
Durbin-Watson stat	0.203681	Prob(F-statistic)	0.000000

Konfidenzellipse 95% gemeinsames Konfidenzintervall (-gebiet)



Die Anzahl der Eheschließungen als Regressor heißt "EHEN". Wie man sieht, ist die Konstante C (also $\hat{\alpha}$) nicht signifikant verschieden von Null, wohl aber die Steigung β (das ist am t- und auch am F-Test zu erkennen).

Das bestätigt sich auch an dem Bild des gemeinsamen Konfidenzbereichs (= Konfidenzellipse 95%) von EViews, das unten abgebildet ist

C(1) ist hier $\hat{\alpha}$ und C(2) ist $\hat{\beta}$. Der rote Punkt stellt die Punktschätzung dar (ein Begriff, der hier sehr anschaulich wird).

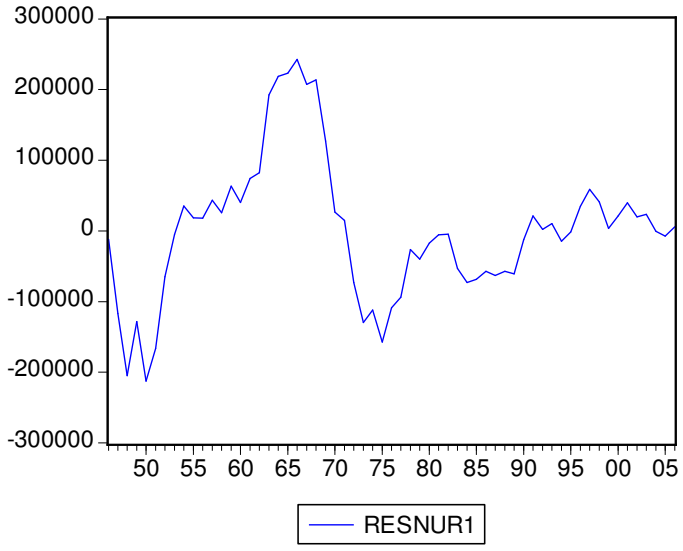
Alle Punkte innerhalb der Ellipse t bedeuten, dass α *und* β *nicht* signifikant ist (H_0: $\alpha = \beta = 0$ angenommen wird.

Die waagrechten und senkrechten Linien markieren die beiden eindimensionalen Konfidenzintervalle (isolierte Konfidenzintervalle für jeweis nur einen Regressionskoeffizienten).

Die waagrechten Linien kann man leicht nachrechnen: der t-Wert bei 95% zweiseitig und T-2 = 59 Freiheitsgraden ist 2,0. $\hat{\beta} = 1{,}757469$ und $\hat{\sigma}_{\beta} = 0{,}128065$. Für die Grenzen $\hat{\beta} \pm 2\,\hat{\sigma}_{\beta}$ (waagrechten Linien) erhält man 1,5013 und 2,0128, wie man das auch auf dem Bild eingezeichnet sieht. Die senkrechten Linien erhält man bei der entsprechenden Berechnung für $\hat{\alpha}$.

[1] Hinweis zum **Datenimport**: Sämtliche Daten sind aus verschiedenen Excel-Tabellen des Statistischen Bundesamts zu einem Excel Tabellenblatt (oder man bildet damit eine neue Mappe) zusammengefasst worden. In E-Views zunächst workfile definieren und damit auch Art der Daten und auf welchen Zeitraum sie sich beziehen. Dann lautet der Befehl Proc (procedure) → Import → Read Text-Lotus-Excel dann Tabelle suchen und nennen, die Reihen benennen und die Bereiche angeben derin EViews kopiert werden soll. EViews erkennt dann leicht, welche Zahlen der Tabelle der betreffende Variable zuzuordnen sind.

Für die Beurteilung der einfachen Regression (mit dem einzigen Regressor "Anzahl der Ehen") ist die Betrachtung der Residuen der Regressionsgleichung wichtig:



Der nebenstehende **Graph der Residuen** scheint für sich genommen nicht sehr aufschlussreich zu sein.

Es wird daher im Folgenden mit entsprechenden Tests untersucht, ob für die Residuen erfüllt ist

B2 Homoskedastizität (**White** Test)

B3 keine Autokorrelation (autocorrelation = serial correlation) (**Box-Pierce** Test und **Breusch-Godfrey** Test)

B4 Normalverteiltheit (**Jarque-Bera** Test [JB Test]).

Homoskedastitzität (B2) erfüllt?

White Heteroskedasticity Test: (wählen mit ***View/Residual Tests/White Heteroskedasticity***)

F-statistic	9.510683	Prob. F (Freiheitsgrade 2, 58)	0.000268
Obs*R-squared	15.06469	Prob. Chi-Square (Freiheitsgrade 2)	0.000535

Dependent Variable: RESID^2

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1.78E+10	5.36E+10	-0.331321	0.7416
EHEN	24044.04	199165.9	0.120724	0.9043
EHEN^2	0.047979	0.180369	0.266002	0.7912

R-squared	0.246962	Durbin-Watson stat	9.69E+09
Adjusted R-squared	0.220995	Prob(F-statistic)	0.000268

Die maßgebliche Größe für den Test ist TR^2 (oben: "Obs*R-squared"), die χ^2 verteilt ist mit 2 Freiheitsgraden signifikant ist (der Tabellenwert bei 1% Signifikanzniveau ist 9,21). E-Views berechnet die Hilfsregression $\hat{u}_t^2 = C + 24044 \cdot x_t + 0{,}047979 \cdot x_t^2 + v_t$. Es zeigt sich dass die Residuen nicht systematisch vom einzigen Regressor x = EHEN und von x^2 abhängen und R^2 für sich genommen klein ist. Gleichwohl ist wegen TR^2 die H_0 (= Annahme B2 erfüllt, Homoskedastizität) abzulehnen. *Mehr zum White Test auf **Seite 9** unten.*

Keine Autokorrelation der Residuen (B3)?

Der kleine Wert der DW Statistik (0,203681 auf S. 2) zeigt, dass die Residuen offenbar positiv autokorreliert sind, was hier auch mit dem – in EViews verfügbaren - Test "residual test – **correlogram Q statistic**" (dieser Test ist in der Vorlesung eingeführt unter dem Namen **Box-Pierce Test**) gezeigt wird. Man erhält ihn, wenn die Reihe der Residuen geöffnet ist und man ***View/Residual Tests/Correlogram-Q-statistics*** auf der equation tool bar wählt (anklickt).[2]

Die dann mit EViews geschätzten Autokorrelationskoeffizienten (AC) erster und zweiter Ordnung sind 0,898 und 0,762. Das Original der Graphik sieht (auf dem Bildschirm) wesentlich besser aus als hier im Druck[3] (man erhält farbige Säulen statt *****)

[2] Man erhält jetzt die Autokorrelationsfunktion (AC), partielle Autokorrelationsfunktion (PAC) und die Ljung-Box Q-Statistik und deren prob-values (wenn diese klein sind < 0,01, wird H_0 auf dem 1% Niveau verworfen).

[3] Bild ab Lag 5 ausgeblendet.

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
. |******* |	. |******* |	1	0.898	0.898	51.651	0.000
. |****** |	**| . |	2	0.762	-0.230	89.456	0.000
. |**** |	**| . |	3	0.587	-0.259	112.32	0.000
. |*** |	**| . |	4	0.384	-0.233	122.26	0.000

Die Q Statistik ist signifikant bei allen Lags (auch bei > 4), was zeigt, das Autokorreliertheit der Störgröße gegeben ist (also B3 verletzt ist). **Autokorrelation der Residuen ist praktisch das Hauptproblem bei allen Versuchen, die hier unternommen werden, GEB mit einer Regressionsfunktion zu "erklären".** Das wird auch bestätigt durch den folgenden Breusch Godfrey Test, der mit EViews leicht durchzuführen ist[4] (auch hier ist der Ergebnis-Ausdruck graphisch etwas verändert).

Breusch-Godfrey Serial Correlation LM Test:

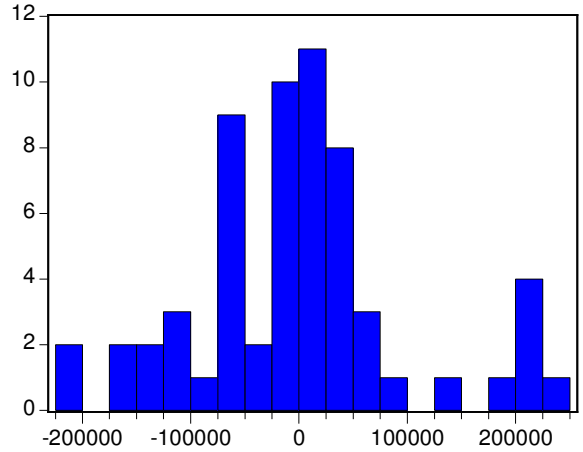
F-statistic	90.58283	Prob. F(3,56)	
Obs*R-squared	50.57737	Prob. Chi-Square(3)	

Dependent Variable: RESID — Mean dependent var -2.93E-11

Variable	Coefficient	t-Statistic	Prob.		
C	-4261.147	-0.142773	0.8870	R-squared	0.829137
EHEN	0.008086	0.148780	0.8823	Durbin-Watson stat	2.121909
RESID(-1)	1.045620	8.095507	0.0000	Prob(F-statistic)	0.000000
RESID(-2)	0.055914	0.293770	0.7700		
RESID(-3)	-0.259523	-2.008431	0.0494		

Die maximale Lag-Länge (hier 3) wird gewählt, nicht vom Programm bestimmt. H_0 bei dem Test ist: keine Autokorrelation. Sie wird hier nicht angenommen, sondern verworfen (also B3 *nicht* erfüllt). Die Prüfgröße TR^2 (oben: "Obs*R-squared") ist χ^2 verteilt mit 3 Freiheitsgraden [Tabellenwerte: bei 5% 7,81 und bei 1% 11,34, der Wert 50,577 ist also hochsignifikant]. Das hohe R^2 von 0,829 sowie die beiden signifikanten Regressoren sprechen für Autokorrelation. Der (orange markierte) DW Wert ist unbrauchbar, da lagged **dependent** variables auftreten.

Normalitätstest der Residuen (**Jarque Bera** Test)



Series: RES1	
Sample 1946 2006	
Observations 61	
Mean	-2.93e-11
Median	-665.1236
Maximum	242699.5
Minimum	-212538.7
Std. Dev.	99250.42
Skewness	0.464545
Kurtosis	3.557840
Jarque-Bera	2.984916
Probability	0.224819

Der nebenstehdne JB Test ist erreichbar mit ***View/Residual Tests/ Histogram-Normality***

Der JB Test zeigt, dass die Störgröße normalverteilt ist, da prob = 0,2248 > 0,01 (H_0 wird angenommen).

Es gibt grundsätzlich folgende Diagnostic Tests in EViews:

Coefficient Tests
Residual T. und
Stability Tests

Daneben aber auch Unit-Root- oder Granger-Causality Test usw.

[4] Zu wählen mit ***View/Residual Tests/Serial Correlation LM Test*** (LM = Lagrange Multiplier). Der EViews-Ausdruck ist auch hier hinsichtlich des Layouts verändert. Zum gleichen Test bei multipler Regression siehe unten Seite 8 und 9.

2. Multiple Regression (Teil 1):

Geburten in Abhängigkeit von mehreren Einflussfaktoren (EHEN, UEBER. NEGEB) Längerer Datensatz (1946 – 2006, wie bei der einfachen Regression)

Neben der Anzahl der Ehen (Variable EHEN) standen uns zunächst nur die folgenden Zeitreihen als zwei weitere mögliche Regressoren zur Verfügung

Geburtenüberschuss (UEBER) und
Anzahl der nichtehelichen Geburten (NEGEB)

Die abhängige Variable y ist stets die Anzahl der Geburten (GEB). Alle Daten beziehen sich auf Deutschland insgesamt 1946 bis 2006.

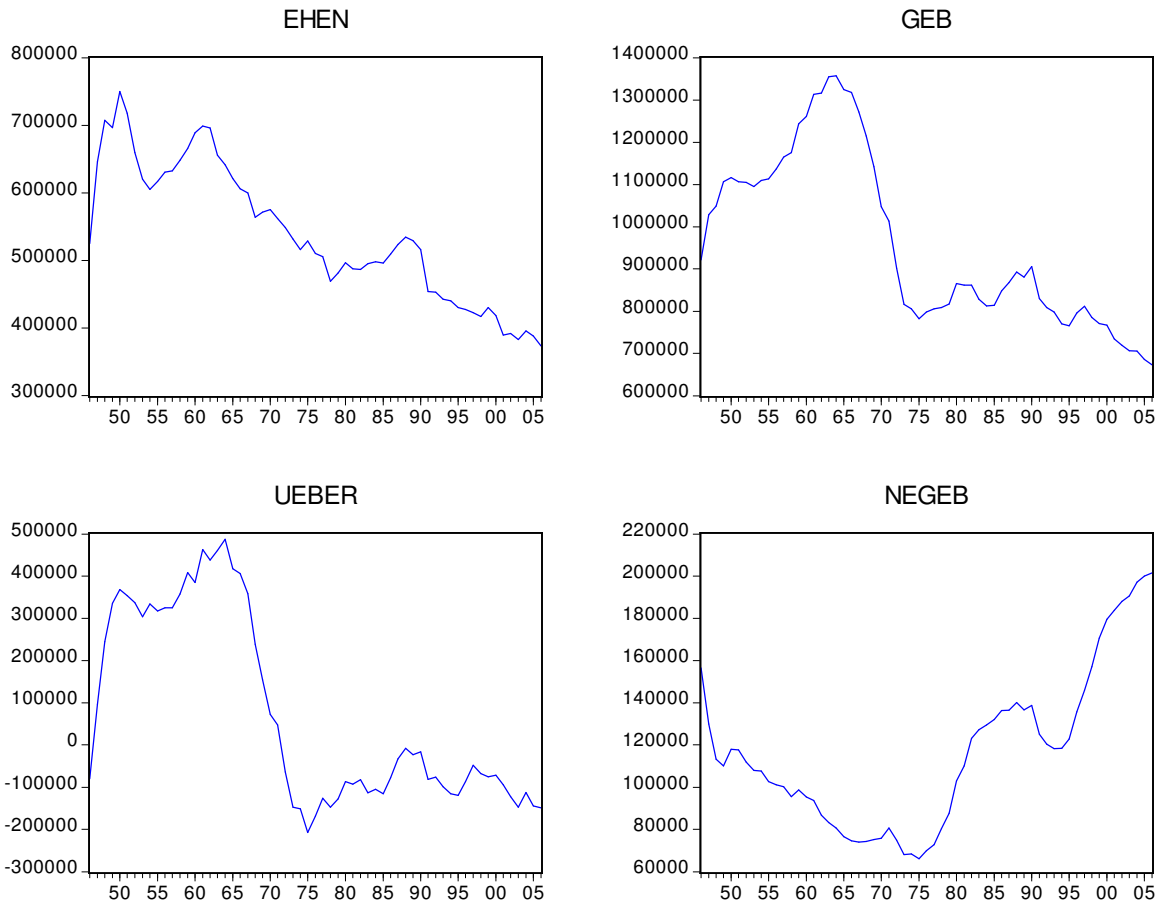
Es empfiehlt sich, zunächst die Korrelationen zwischen den Variablen und die graphische Darstellung der Zeitreihen zu betrachten. Es zeigt sich (nicht überraschend), dass die nicht-ehelichen Geburten negativ korreliert sind mit den Geburten insgesamt und den beiden anderen Variablen EHEN und UEBER. Sie haben außerdem einen positiven Trend ab Mitte der 70erJahre. Schon an der Graphik zeigt sich, dass die Anzahl der Geburten (GEB) und der Geburtenüberschuss (UEBER) sehr hoch miteinander korreliert sind (0,9578). Es ist leicht, mit EViews die folgende Korrelationsmatrix und Zusammenstellung der Zeitreihen zu erhalten:

Correlation Matrix

	EHEN	GEB	UEBER	NEGEB
EHEN	1.000000	0.872612	0.866170	-0.593161
GEB	0.872612	1.000000	0.957846	-0.582012
UEBER	0.866170	0.957846	1.000000	-0.413271
NEGEB	-0.593161	-0.582012	-0.413271	1.000000

Die Zahlen zeigen, dass wir wohl ein Problem mit möglicher Kollinearität haben werden.

Multiple Time Series Graph



Die Regression "Alle" in der GEB erklärt wird mit allem drei Regressoren ergab die folgende Regressionsfunktion:

Gleichung "Alle"

Dependent Variable: GEB Date: 02/18/08
Sample: 1946 2006
Included observations: 61

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1110171.	78245.79	14.18826	0.0000
EHEN	-0.108368	0.125030	-0.866736	0.3897
UEBER	0.819269	0.049666	16.49554	0.0000
NEGEB	-1.308104	0.185874	-7.037584	0.0000

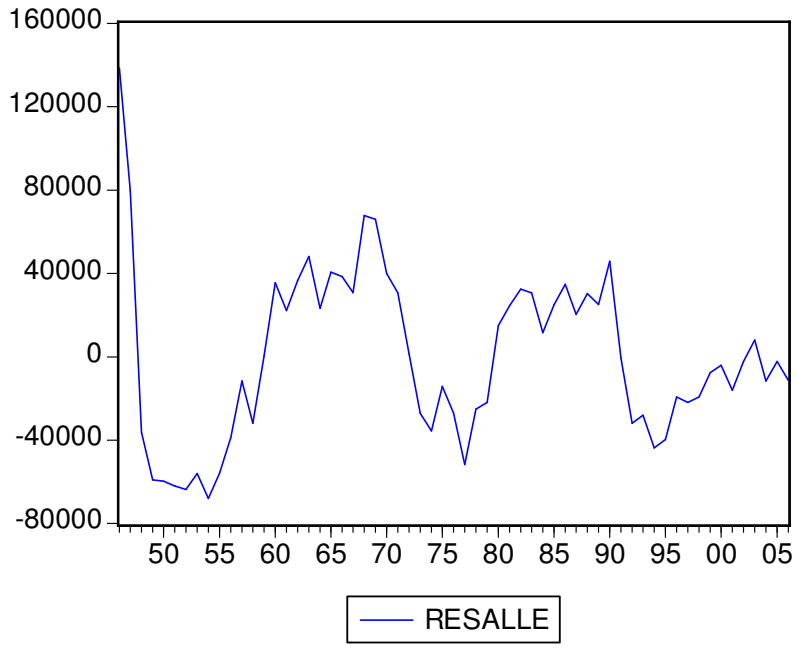
R-squared	0.959792	Mean dependent var	959278.8
Adjusted R-squared	0.957676	S.D. dependent var	203209.3
S.E. of regression	41805.72	Akaike info criterion	24.18278
Sum squared resid	9.96E+10	Schwarz criterion	24.32120
Log likelihood	-733.5748	F-statistic	453.5478
Durbin-Watson stat	0.368766	Prob(F-statistic)	0.000000

Es ist auffallend, dass sich die korrigierte multiple Bestimmtheit zwar von 0,757408 (mit nur EHEN als Regressor) auf 0,957676 erhöht hat, nun aber EHEN keinen signifikanten Erklärungsbeitrag mehr liefern. In dieser Konstellation "Alle" wirkt sich natürlich die hohe Korrelation zwischen UEBER und GEB stark aus.

Danach ist es auch nicht überraschend, dass R^2 nicht stark sinkt (nur von 0,9598 auf 0,9593) und das korrigierte R^2 sogar steigt von 0,9577 auf 0,9579, wenn man den Regressor EHEN he-

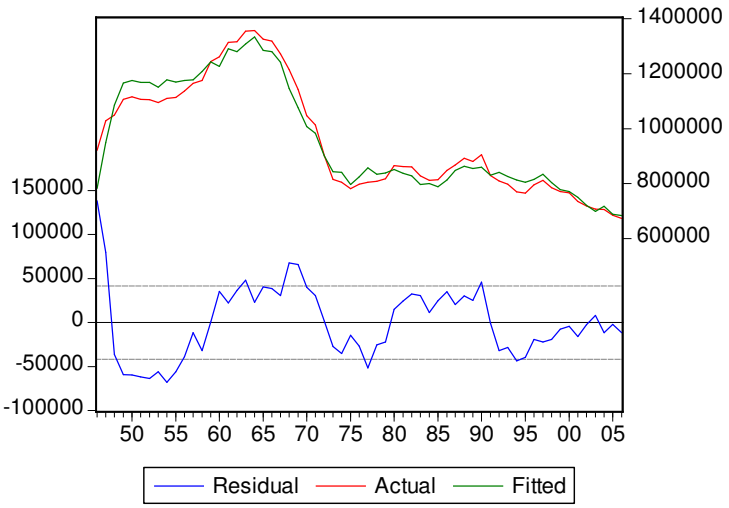
Wie nebenstehend gesagt ist UEBER faktisch (sachlich gesehen) kein echter Regressor. Die Auswahl der Regressoren war hier (für die Bildung eines Übungsbeispiels) mehr bestimmt von der Datenverfügbarkeit als von inhaltlichen Überlegungen. Der Regressor UEBER ist mehr oder weniger das gleiche wie der Regressand GEB.

Residuen bei der Regression mit allem Regressoren



Schon aus dem obigen Ausdruck ist erkennbar, dass eine positive Autokorrelation vorliegen dürfte (DW ist mit 0,368766 sehr klein). Man könnte auch hier wieder die verschiedensten Tests durchführen. Es zeigt sich auch an dem Bild: die Regressionsfunktion ist offenbar für die frühen Jahre (1946 bis 1950) den Daten nicht gut angepasst. Das wird noch deutlicher auf einer Abbildung auf der nächsten Seite. Man könnte somit diese Daten aus der Stichprobe herausnehmen und sehen ob die Anpassung dann besser ist.

Für weitere Analysen der Residuen muss man ihnen einen Namen geben. Befehle ***proc/make residuals***. (Für die obige Gleichung mit allen Regressoren habe ich den Namen "resalle" gewählt). Die von EViews angelegte Reihe resid wird bei jeder neuen Regressionsgleichung mit einer anderen Reihe von Residuen überschrieben. Wenn man die Residuen nicht benennt ist schnell nicht mehr klar, welche Reihe die automatisch erzeugte Reihe "resid" darstellt.



Dies ist eine weitere Darstellung von RESALLE, jetzt einschließlich y (actual) und $\hat{y}$ (fitted). Man erhält sie beim Equation output mit der Funktion ***resids***.

Wie bereits an der DW Statistik erkennbar ist, sind diese Residuen autokorreliert. Das wird auch bestätigt durch das Ergebnis des Breusch Godfrey Tests auf Autokorrelation (= serial correlation LM Test). Siehe unten S. 8 und 9.

Überlegungen zur Spezifikation (Mauer/Steinmetz, Kap. 13)

Die folgende Tabelle vergleicht verschiedene Regressionsfunktionen hinsichtlich der Güte der Anpassung (anhand von $\overline{R}^2$ = Adjusted R-squared, DW und dem F Wert, nicht signifikante Regressoren sind farblich markiert, • bedeutet, dass der Regressor in der Gleichung enthalten ist und – dass er nicht enthalten ist)

Gl.	EHEN	UEBER	NEGEB	$\overline{R}^2$	DW	F
1	•	•	•	0,95768	0,368766	453,5
2	-	•	•	0,95786	0,359726	682,9
3	•	-	•	0,75985	0,199046	95,9
4	•	•	-	0,92265	0,190396	356,9

Schließt man aus sachlichen Gründen die Variable UEBER als Regressor aus (man kann schlecht argumentieren, dass der abnehmende Geburtenüberschuss die Ursache für abnehmende Geburten ist) erhält man keine hinsichtlich der Güte der Anpassung gegenüber der einfachen Regression (nur mit Ehen, dort war das korrigierte R^2 0,757408) bessere Regressionsgleichung mit zwei Regressoren.

Dass NEGEB in der Regression von GEB auf EHEN und NEGEB nicht signifikant ist, zeigt sich auch daran, dass der Erklärungsgewinn durch NEGEB im Vergleich zur Regression nur auf EHEN nicht signifikant ist (der Regressionskoeffizient für NEGEB in Gl. 3 ist nicht signifikant). Das gleiche Ergebnis erhielte man auch mit dem in den downloads wiederholt beschriebenen F-Test auf Signifikanz hinzu kommender Regressoren mit der Prüfgröße:

$$(*) \qquad F = \frac{(S^0_{\hat{u}\hat{u}} - S_{\hat{u}\hat{u}})/L}{S_{\hat{u}\hat{u}}/(T-K-1)} \sim F_{L,\,T-K-1}$$

Man kann dies auch mit EViews durch den ***redundant variables*** Test (ein spezieller coefficient test) feststellen wenn man ausgehend von der Gleichung "Alle" eingibt ***View/Coefficientl Tests/….*** Es können so eine oder mehrere Variablen daraufhin überprüft werden, ob sie redundant sind. Man erhält

Redundant Variables: NEGEB

F-statistic	49.52758	Prob. F(1,57)	0.000000
Log likelihood ratio	38.14651	Prob. Chi-Square(1)	0.000000

wenn man als potenziell redundante Variable NEGEB eingibt und es folgt dann der übliche Ausdruck der Koeffizienten, t- Werte, R^2 usw. jetzt für die Gleichung ohne NEGEB also von Gl. 4 (mit den Regressoren EHEN und UEBER. Zu den Freiheitsgraden: L = 1 (Anzahl der

Restriktionen) T-K-1 = 61 – 3 -1 = 57. Der F-Wert (entspricht dem von Gl. *) ist hier signifikant prob(F) < 0,01) so dass man NEGEB nicht als redundant bezeichnen kann, obgleich sich mit NEGEB (wie unten gezeigt) adj. R^2 nur wenig erhöht (von 0,924856 auf 0,959263). Prüft man ob UEBER redundant ist, so erhält man einen größeren F-Wert

Redundant Variables: UEBER

F-statistic	272.1027	Prob. F(1,57)	0.000000

Bei UEBER *und* NEGEB (L = 2 Restriktionen) erhält man

Redundant Variables: NEGEB UEBER

F-statistic	140.5887	Prob. F(2,57)	0.000000

Oder bei EHEN und NEGEB

Redundant Variables: EHEN NEGEB

F-statistic	30.00046	Prob. F(2,57)	0.000000

Die Veränderung der multiplen Bestimmtheit durch das Hinzukommen weiterer Regressoren kann auch mit den Rekursionsformeln verdeutlicht werden. Wir beginnen den Aufbau (bottom up, "Maurer") mit x_1 = EHEN, fügen dann hinzu x_2 = UEBER und x_3 = NEGEB

Nur x_1 = EHEN $\quad R^2_{y.1} = r^2_{y1}$	$= 0{,}761451$
Plus UEBER $\quad R^2_{y.12} = r^2_{y1} + \dfrac{(r_{y2} - r_{y1}r_{12})^2}{1 - r^2_{12}}$ $= r^2_{y1} + r^2_{y2.1}(1 - r^2_{y1})$	$= 0{,}924856$, somit ist die part. Bestimmtheit $r^2_{y2.1} = \dfrac{0{,}924856 - 0{,}761451}{1 - 0{,}761451} = 0{,}6845$
Plus NEGEB $\quad R^2_{y.123} = R^2_{y.12} + r^2_{y3.12}(1 - R^2_{y.12})$	$= 0{,}959792$. Berechn. der partielle Bestimmth. $r^2_{y3.12} = \dfrac{0{,}959792 - 0{,}924856}{1 - 0{,}924856} = 0{,}4649$

Würde man den Aufbau mit UEBER beginnen, weil dieser Regressor mit GEB am höchsten korreliert $R^2_{y.1} = r^2_{y1} = (0{,}957846)^2 = 0{,}917469$, so erhielte man bei Hinzukommen eines zweiten Regressors x_2 zu x_1 = UEBER

X_2 =	$R^2_{y.12}$	$r^2_{y2.1}$
EHEN	0,924856	0,0895
NEGEB	0,959263	0,5064

Als beste Regressionsfunktion mit zwei Regressoren erhält man auch nach dem Kriterium eines möglichst großen $\overline{R}^2_{y.12}$ die Gleichung mit UEBER und NEGEB, oben Gl. 2 (die hinsichtlich der inhaltlichen Interpretation natürlich nichts hergibt).

Autokorrelation der Residuen bei der Regression "Alle"

Bei dem "**Breusch-Godfrey** Serial Correlation LM **Test**"[5] wird die Regressionsfunktion $\hat{u}_t = \gamma_0 + \gamma_1 x_{1t} + ... + \gamma_K x_{Kt} + \sum_{s=1}^{s=p} \alpha_s \hat{u}_{t-s} + v_t$ geschätzt, wobei $\hat{u}_t$ die Residuen (geschätzten Störgrößen) der ursprünglichen Regression ("Alle") sind und für die Störgröße v_t die Standardannahmen B1 bis B4 gelten sollten. Die Anzahl s der Lags muss gewählt und angegeben werden (hier s = 4). Die maßgebliche Prüfgröße ist wieder TR^2 (= Obs*R-squared). Sie ist asymptotisch χ^2 verteilt mit s Freiheitsgraden und ist im Beispiel signifikant (prob value < 0,01), d.h. die Nullhypothese (H_0 lautet Residuen sind nicht autokorreliert "up to lag order s") wird verworfen. Man erhält das folgende Ergebnis

[5] Output im Layout verändert und insbesondere verkürzt.

F-statistic	17.03664	Prob. F(4,53)	0.000000
Obs*R-squared	34.31332	Prob. Chi-Square(4)	0.000001

Dependent Variable: RESID

Date: 04/10/08 Time: 13:18 Sample: 1946 2006 Included observations: 61

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	29586.09	54808.06	0.539813	0.5916
EHEN	-0.055032	0.089067	-0.617864	0.5393
UEBER	0.017581	0.035895	0.489782	0.6263
NEGEB	-0.011353	0.128011	-0.088687	0.9297
RESID(-1)	0.900890	0.134545	6.695807	0.0000
RESID(-2)	-0.346373	0.178288	-1.942768	0.0574
RESID(-3)	0.298666	0.178799	1.670406	0.1007
RESID(-4)	-0.174399	0.140438	-1.241820	0.2198

R-squared	0.562513	Durbin-Watson stat	1.559409
Adjusted R-squared	0.504732	F-statistic	9.735225
Sum squared resid	4.36E+10	Prob(F-statistic)	0.000000

White Test auf Heteroskedastitztät (Fortsetzung von oben Seite 3 oben)

Geschätzt wird die Gleichung $\hat{u}_t^2 = \gamma_0 + \gamma_1 x_t + \gamma_2 z_t + \gamma_3 x_t^2 + \gamma_4 z_t^2 + \delta_1 x_t z_t ... + v_t$ (bei zwei Regressoren x und z, der Koeffizient δ zum interaction term $x_t z_t$ heißt **cross term**.) Man kann bei der Ausführung des Tests wählen "with.." oder "without cross terms". Dieser Test wird von White als recht allgemeiner Test auf Fehlspezifikation (misspecification) beschrieben. Im Handbuch zu EViews heißt es zum White Heteroskedasticity Test:

> " … since the null hypothesis underlying the test assumes that errors are both homoskedastic and independent of the regressors, and that the linear specification of the model is correct. Failure of any of these conditions could lead to a significant test statistic. Conversely, a non-significant test statistic implies that none of the three conditions is violated." (p. 363)

Die maßgebliche Prüfgröße ist auch hier wieder Obs*R-squared, also 25.42402, die χ^2 verteilt ist mit 9 Freiheitsgraden, weil es im folgenden 9 Regressoren (ohne das Absolutglied C) gibt. Nach dem Ausdruck ergibt sich, dass H_0 anzulehnen ist (Modellannahmen *nicht* erfüllt):

F-statistic	4.049627	Prob. F(9,51)	0.000583
Obs*R-squared	25.42402	Prob. Chi-Square(9)	0.002536

Test Equation: Dependent Variable: RESID^2

Method: Least Squares, Date: 04/10/08 Time: 13:34, Sample: 1946 2006 Included observations: 61

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.10E+11	4.94E+10	2.217284	0.0311
EHEN	-154748.7	155938.7	-0.992369	0.3257
EHEN^2	-0.028089	0.140633	-0.199731	0.8425
EHEN*UEBER	0.023517	0.120246	0.195572	0.8457
EHEN*NEGEB	1.657799	0.402621	4.117520	0.0001
UEBER	24239.86	63449.76	0.382032	0.7040
UEBER^2	-0.008725	0.027551	-0.316692	0.7528
UEBER*NEGEB	-0.282069	0.128603	-2.193328	0.0329
NEGEB	-1249701.	319402.6	-3.912619	0.0003
NEGEB^2	1.862580	0.535243	3.479878	0.0010

R-squared	0.416787	Durbin-Watson stat	0.984233
Adjusted R-squared	0.313867	F-statistic	4.049627

3. Multiple Regression (Teil 2):

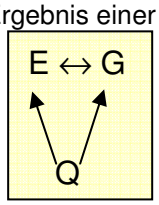
Geburten in Abhängigkeit von weiteren Einflussfaktor (KIGAUS, KIGED und QUOTE) Kürzerer Datensatz (1965 – 2003)

Die hinzugekommenen Variablen sind der Kindergeldbezug bei Haushalten allgemein (= KIGED) und speziell von Ausländern (= KIGAUS) sowie die Erwerbsquote von Frauen (=QUOTE).

Korrelationstabelle

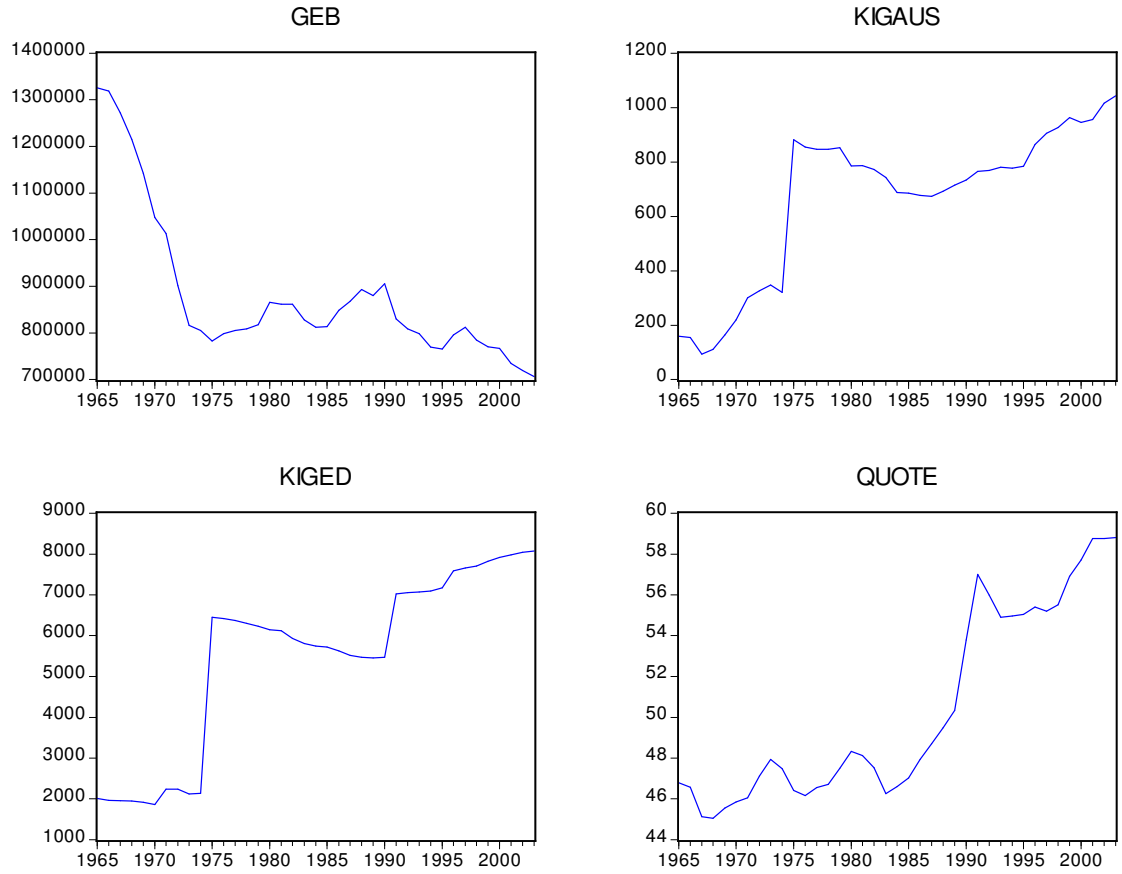
	GEB	EHEN	KIGAUS	KIGED	NEGEB	QUOTE	UEBER
GEB (G)	1	0.827906	-0.860826	-0.794122	-0.541028	-0.551073	0.960937
EHEN (E)		1	-0.863298	-0.895699	-0.765194	-0.841886	0.677220
KIGAUS	-	-	1	0.972416	0.688142	0.646865	-0.755858
KIGED	-	-		1	0.738383	0.732547	-0.653824
NEGEB	-	-			1	0.807896	-0.3261697
QUOTE (Q)	-	-				1	-0.333598

Es ist auffallend, dass mit zunehmender Erwerbsbeteiligung von Frauen (steigender QUOTE Q) die Eheschließungen (E) aber auch die Geburten (G) abnehmen, so dass Q mit E und G negativ korreliert ist r_{EQ} = - 0,841889 und r_{GQ} = - 0,551073. Man könnte meinen, dass deshalb der hohe Wert für r_{EG} nur Ergebnis einer Scheinkorrelation ist (aufgrund der gemeinsamen Abhängigkeit von Q).



Dazu müsste jedoch doch für die partielle Korrelation $r_{EG.Q} = \frac{r_{EG} - r_{EQ} r_{GQ}}{\sqrt{(1 - r_{EQ}^2)(1 - r_{GQ}^2)}}$ gelten $r_{EG.Q} = 0$, also $r_{EG} = r_{EQ}\, r_{GQ}$. Das ist jedoch nicht der Fall. Vielmehr ist $r_{EG.Q}$ = 0,8164.

Bei den beiden Kindergeldvariablen KIGED und KIGAUS fällt der offenbar rechtlich bedingte Niveauanstieg 1975 auf. Sehr deutlich wird auch, dass QUOTE und GEB negativ korreliert sind.



Für die Regressionsgleichung mit **allen** Regressoren erhält man

Dependent Variable: GEB
Sample: 1965 2003
Included observations: 39

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1104208.	142143.4	7.768270	0.0000
EHEN	0.146602	0.161780	0.906181	0.3716
KIGAUS	-20.64474	63.43238	-0.325461	0.7470
KIGED	-1.445993	8.132879	-0.177796	0.8600
NEGEB	-0.442639	0.140172	-3.157819	0.0035
QUOTE	-3801.013	1435.584	-2.647711	0.0125
UEBER	0.866552	0.042562	20.35997	0.0000

R-squared	0.990785	Mean dependent var	881329.9
Adjusted R-squared	0.989057	S.D. dependent var	161299.8
S.E. of regression	16873.50	Akaike info criterion	22.46602
Sum squared resid	9.11E+09	Schwarz criterion	22.76461
Log likelihood	-431.0875	F-statistic	573.4149
Durbin-Watson stat	1.263608	Prob(F-statistic)	0.000000

Der große Wert für R^2 (0,990785) und das signifikante F bei gleichzeitig drei (gelb markiert) nichtsignifikanten Regressoren ist ein Hinweis auf Multikollinearität. Man könnte versuchen, statt Niveauwerte von Geburten (oder auch der Regressoren) jeweils mit Wachstumsraten zu arbeiten. Der Erfolg derartiger Versuche dürfte aber gering sein. Wie die beiden Streuungsdiagramme (mit Excel erzeugt) auf der nächsten Seite (Zeitraum 1960 bis 2004) zeigen ist die Korrelation zwischen den *Wachstumsraten* der Erwerbsquote von Frauen und den Wachstumsraten der Geburten mit r = -0,033166 (denn $r^2 = (-0{,}033166)^2 = -0{,}0011$) deutlich geringer als zwischen den entsprechenden *Niveauwerten*.

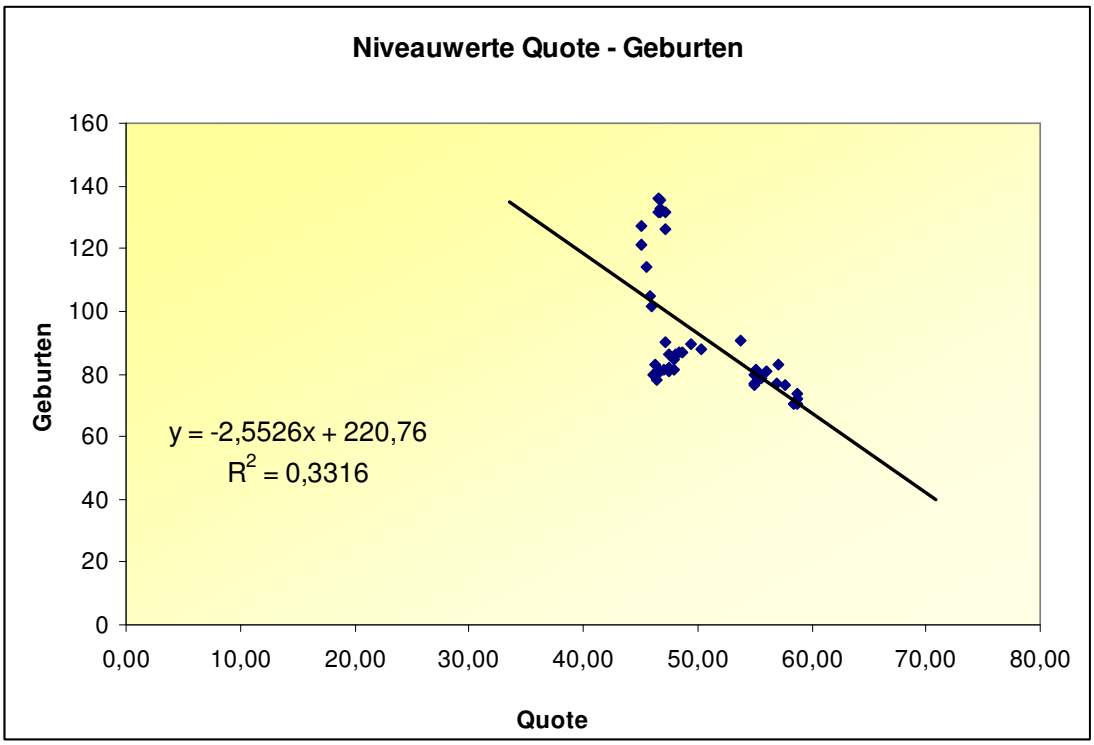
Nimmt man die sachlich nicht sinnvolle, aber formal viel "erklärende" Variablen Geburtenüberschuss (UEBER) heraus, so erhält man einen überraschend geringen Erklärungswert der Variable QUOTE und es fällt auf, dass jetzt gerade die drei vorher nicht signifikanten Regressoren EHEN, KIGAUS und KIGED einen signifikanten Erklärungsbeitrag leisten. Hinzu kommt, dass die zweite Schätzung (ohne UEBER) hinsichtlich des Durbin Watson Koeffizienten schlechter ist als die erste.

Dependent Variable: GEB

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-317920.6	455385.8	-0.698135	0.4900
EHEN	2.105528	0.478423	4.400972	0.0001
KIGAUS	-886.0776	173.1954	-5.116056	0.0000
KIGED	94.34500	24.40243	3.866213	0.0005
NEGEB	0.526801	0.484960	1.086278	0.2852
QUOTE	3387.389	5118.563	0.661785	0.5127

R-squared	0.871410	Mean dependent var	881329.9
Adjusted R-squared	0.851926	S.D. dependent var	161299.8
S.E. of regression	62068.72	Akaike info criterion	25.05051
Sum squared resid	1.27E+11	Schwarz criterion	25.30644
Log likelihood	-482.4849	F-statistic	44.72579
Durbin-Watson stat	0.458367	Prob(F-statistic)	0.000000

Die Suche nach einer geeigneten Regressionsfunktion für die Geburten in Deutschland soll hier abgebrochen werden, weil hier das Ziel nur darin bestand, ein Demonstrationsbeispiel für ökonometrische Konzepte und für Berechnungen mit der entsprechenden Statistik-Software zu entwickeln.



Die Korrelation von $-\sqrt{0{,}3316} = -0{,}5785$ entspricht trotz geringfügiger Abweichung hinsichtlich des Beobachtungszeitraums (1960 – 2004 statt 1965 – 2003) ziemlich gut dem oben angegebenen Wert von - 0,551073 (das Abschneiden der Werte vor 2005 war nötig wegen der Kindergeldreihen, die erst 1965 beginnen).