Prof. Dr. Peter von der Lippe
Dr. Katrin Nihalani

Mehr Ärzte = höhere Lebenserwartung?

Ein Beispiel für einfache Regression und Scheinkorrelation

1. Datensatz und einfache Regression

Der hier vorgestellten Berechnung liegt ein Datensatz mit Zeitreihen der Bundesrepublik Deutschland von 1992 bis 2006 zu Grunde. Die benutzten Variablen lauten „**leben** = Lebenserwartung[1] in Jahren“, „**arzt** = Anzahl der Ärzte je 1000 Einwohner“, „**aus** = Gesundheitsausgaben in Prozent des BIP“, „**pkaus** = pro Kopf Gesundheitsausgaben“ und „**year** = Zeitvariable“ (1992, …., 2006, also T = 15 Jahre). Ziel ist es mit dem Statistikprogramm EViews einen linearen Trend für die jeweiligen Variablen zu bestimmen.

Auf der linken Seite des Screenshots sieht man das gebildete EViews Workfile mit dem Namen "NIHALANI" bestehend aus der Liste der Variablen und auf der rechten Seite eine geschätzte Regressionsfunktion (nämlich EQ02). Die einfache Regression (der Gleichung ist der Name "EQ02" gegeben worden) wird in EViews mit dem Befehl "leben c arzt" eingegeben.

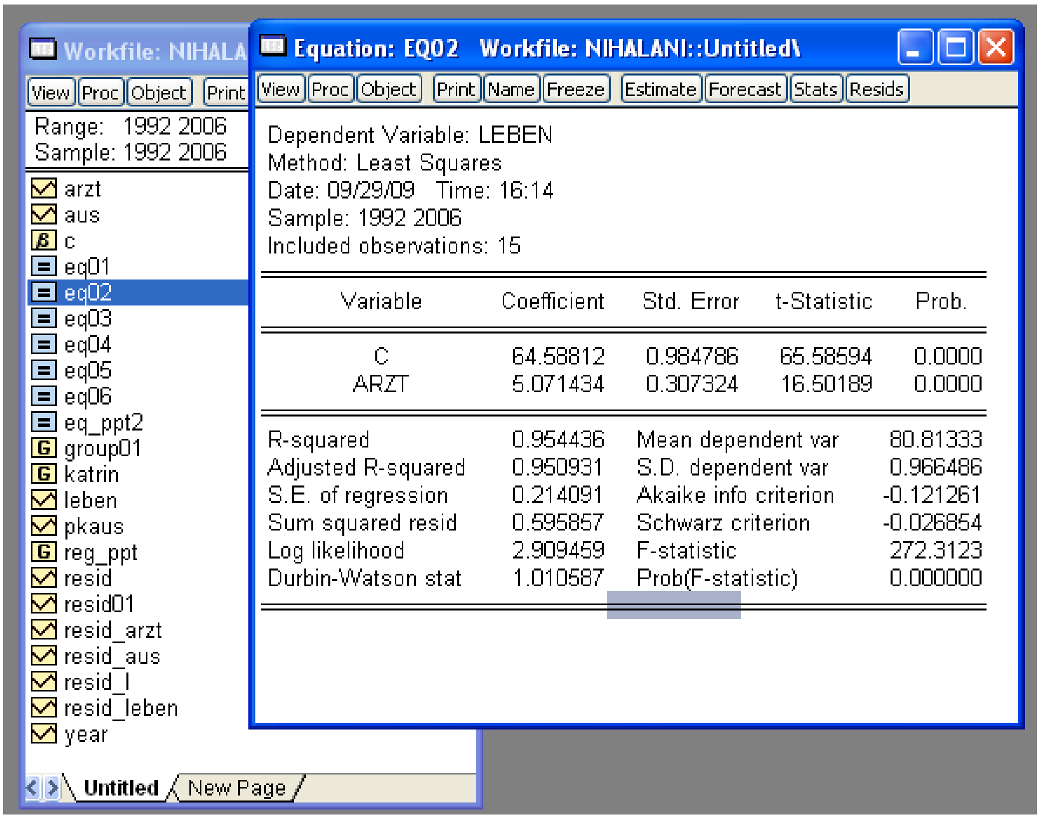


Abbildung 1: EViews Workfile

[1] Genauer: einer "nulljährigen" Frau.

Aus Abbildung 1 kann die geschätzte Regressionsgerade abgelesen werden:

$$\hat{y}_t = 64{,}5881 + 5{,}0714 \cdot x_t \text{ (wobei y = leben, x = arzt und T = 15).}$$

Man sieht ferner, dass alle Regressionskoeffizienten *signifikant* sind, insbesondere ist H_0: β = 0 zu verwerfen. Die Prüfgröße ist t = 16,5018; sie ist t-verteilt mit T-2 = 15-2 = 13 Freiheitsgraden. Anders gesagt: $\hat{\beta}$ = 5,0714 ist signifikant verschieden von β = 0. Dies impliziert auch die Prüfgröße F = 272,31, die ebenfalls signifikant ist.[2] Auch das relativ hohe *Bestimmtheitsmaß* von R^2 = 0,9544 verleitet zu der Annahme, die Schätzfunktion könne als Politikempfehlung dienen weil sie den Daten gut angepasst ist, etwa nach dem Motto: *ein Arzt je 1000 Einwohner mehr steigert die Lebenserwartung der Frauen um 5,07 Jahre.*[3] Es wird gezeigt, dass R² nicht das einzige Beurteilungskriterium sein kann und dass es insbesondere notwendig ist, die Möglichkeit einer Scheinkorrelation zu bedenken und die Plausibilität des Ergebnisses einzuschätzen.

"Ein Arzt mehr" ist übrigens relativ viel, wenn man die Mittelwerte der Variablen betrachtetet: es gilt nämlich für x = arzt: $\bar{x}$ = 3,2, x_{min} = 2,84 und x_{max} = 3,45 (entsprechend gilt für die Lebenserwartung y: $\bar{y}$ = 80,8, y_{min} = 79,3 und y_{max} = 82,4 Jahre).

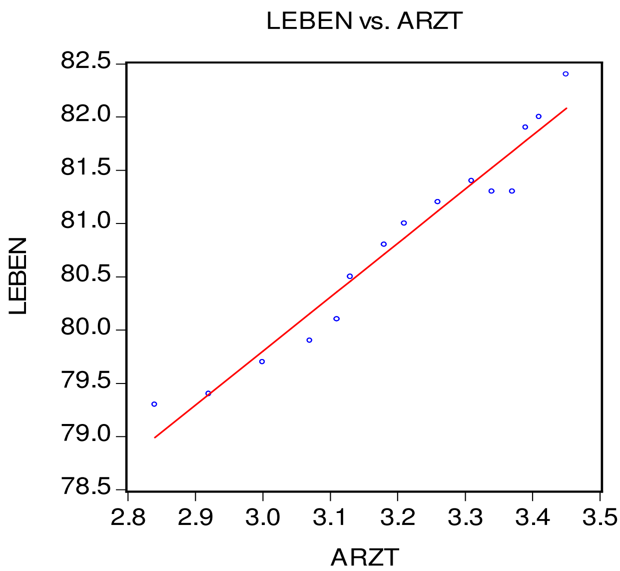


Abbildung 2

Auch eine einfache Überlegung (ohne jede Regressionsrechnung) scheint das Ergebnis $\hat{\beta}$ = 5,0714 zu bestätigen. Die „Arztdichte" ist von 1992 bis 2006 um Δx = 0,61 = 3,45 – 2,84 gestiegen und die Lebenserwartung um Δy = 3,1 und dies ist genau 3,1/0,61 = 5,082 mal so viel wie Δx.

Die Güte der Anpassung zeigt sich auch am nebenstehenden Streuungsdiagramm, welches mit EViews generiert wurde.

Neben R^2 ist auch der „Standard Error of Regression (S.E. of regression)“ 0,214091 (siehe Abbildung 1) von Interesse. Er errechnet sich als

[2] F-Test und t-Test laufen im Falle der einfachen Regression auf das gleiche hinaus. Der Koeffizient "Durbin Watson" (abgekürzt DW oder einfach d) von d < 2 weist auf eine positive Autokorrelation der Störgrößen in Höhe von etwa $\rho \approx +0{,}4947$ hin, denn es gilt $d \approx 2(1-\rho)$.

[3] Dass man so einfach nicht folgern kann sieht man auch an der Interpretation der Konstanten C (das ist der geschätzte Koeffizient α). Man könnte meinen: hätte man in Deutschland überhaupt keine Ärzte, würden die Frauen im Durchschnitt immerhin noch 64,58 Jahre alt werden. Der Ordinatenabschnitt (intercept) C ist aber im Allgemeinen für die Interpretation wenig interessant.

$\hat{\sigma} = \sqrt{\frac{S_{\hat{u}\hat{u}}}{T-2}} = \sqrt{\frac{0{,}595857}{13}}$ = 0,21409 und ist auch mit diesem Wert im „Equation Output“ angegeben.

2. Haupteinwände gegen das bisher erhaltene Ergebnis der einfachen Regression

Insgesamt existieren zwei Haupteinwände gegen die bisher erhaltenen Ergebnisse der Einfachregression:

1. Es muss überprüft werden, ob die bei der Regressionsanalyse (implizit) gemachten (**Modell-)Annahmen** bei den vorhandenen Daten gelten.
2. Es muss geprüft werden, ob nicht evtl. eine **Scheinkorrelation** vorliegt (spurious correlation).

zu 1:

Die Modellannahmen betreffen die Gleichung (A-Annahmen), die Störvariablen (B-Annahmen) sowie die Variablen selber (C-Annahmen).

A-Annahmen: Gleichung

A1: Es darf keine Fehlspezifikation vorliegen bzw. es muss überprüft werden, ob „omitted variables“ existieren.[4] Man kann versuchen, sich gegen einen Spezifikationsfehler – also omitted variables – im Falle einer Einfachregression abzusichern, indem man das Modell mit weiteren Regressoren berechnet (→ Abschnitt 3).

A2 wäre verletzt, wenn der Zusammenhang (in der Grundgesamtheit) nichtlinear wäre. In dem hier vorgeführten Beispiel besteht jedoch kein Anlass dies anzunehmen (vgl. Abbildung 2). Das gilt auch für Annahme

A3, denn bei der relativ kurzen Zeitreihe von 1992 bis 2006 ist ein Strukturbruch wenig wahrscheinlich.

B-Annahmen: Störvariablen u_t (u_1, …, u_T)

B1: $E(u_t) = 0$ ($\forall t$)[5]

B2: $E(u_t)^2 = \sigma^2$ ($\forall t$); Homoskedastizität

B3: $E(u_t, u_{t-k}) = 0$ mit k = 1, 2, … (keine Autokorrelation)

B4: Die Größen u_t (also u_1, u_2, …, u_T) sind normalverteilt.[6]

---

[4] Bei einer multiplen Regression muss ferner überprüft werden, ob überflüssige Regressoren in die Gleichung aufgenommen sind. Die Aufnahme überflüssiger Variablen ist jedoch im Vergleich zu „omitted variables“ weniger gravierend.

[5] Symbol bedeutet: für alle t = 1, 2, …, T. Die Annahme ist wichtig für die Erwartungstreue, kann aber kaum mit statistischen Tests geprüft werden.

C-Annahmen: Variablen (y und x bzw x_1, x_2, …, x_K)

C1: x muss exogen[7] sein (darf nicht mit u korrelieren)

C2: Die Varianz von x darf nicht null oder ≈ 0 sein (bei multipler Regression: keine offene Kollinearität).

zu 2:

Die Variablen x und y sollten keinen gemeinsamen stochastischen oder deterministischen Trend haben. Auf die Möglichkeit eines stochastischen Trends soll hier nicht eingegangen werden (unit root test und cointegration – vgl. Download Nr. 5). Ein einfacher deterministischer Trend wäre ein linearer Trend. Es zeigt sich, dass beide Variablen einen linearen Trend besitzen.

leben = -345,4 + 0,213214 year (R^2 = 0,97335) [= EQ04] und

arzt = -78.83 + 0,041036 year (R^2 = 0,97158) [= EQ07].[8]

Man kann nun eine einfache Regression mit den Residuen (geschätzten Störgrößen $\hat{u}_t$) der beiden Regressionen berechnen, die hier resid_leben bzw. resid_arzt genannt werden. Das Ergebnis zeigt, dass die Korrelation praktisch verschwindet (R^2 sinkt von 0,9544 (vgl. Abbildung 1) auf 0,026574 (vgl. Abbildung 3)).

Dependent Variable: RESID_LEBEN
Method: Least Squares
Date: 10/07/09 Time: 12:40
Sample: 1992 2006
Included observations: 15

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-5.83E-14	0.041706	-1.40E-12	1.0000
RESID_ARZT	0.819344	1.375356	0.595732	0.5616

R-squared	0.026574	Mean dependent var	-6.06E-14
Adjusted R-squared	-0.048305	S.D. dependent var	0.157761
S.E. of regression	0.161527	Akaike info criterion	-0.684728
Sum squared resid	0.339181	Schwarz criterion	-0.590321
Log likelihood	7.135461	F-statistic	0.354897
Durbin-Watson stat	1.227221	Prob(F-statistic)	0.561583

Abbildung 3: EQ05

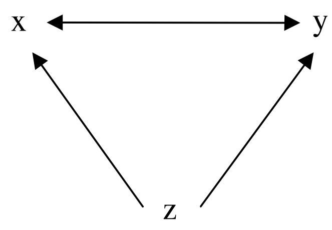
[6] Sind die Annahmen B2 und B3 nicht erfüllt (wohl aber B1) ist die Schätzung zwar erwartungstreu, nicht aber effizient. B4 ist wichtig für Konfidenzintervalle und Tests.

[7] Manche sprechen auch von „modellexogen“, d.h. nicht im Modell erklärt. Ist C1 nicht erfüllt, dann ist die Schätzung nicht konsistent.

[8] Eine ähnlich naive Interpretation dieser Gleichung wie die gem. Fußnote 3 würde bedeuteten: zu Christi Geburt (year = 0) hat es -79 Ärzte je 1000 Einwohner in Deutschland gegeben und alle 10 Jahre kommen etwa 0,4 Ärzte hinzu.

Aus dem Screenshot der Abbildung 3 wird zudem ersichtlich, dass der Regressor resid_arzt der geschätzten Gleichung resid_leben = C + $\hat{\beta} \cdot$(resid_arzt) nicht signifikant ist. (Ein t-Wert von 0,5957 oder größer hat eine Wahrscheinlichkeit von 0,5616.) und entsprechend ist auch die F-Statistik nicht signifikant.

Eine Scheinkorrelation liegt vor, wenn zwei Variablen miteinander korrelieren (symbolisiert durch den Doppelpfeil ↔) weil hinter ihnen eine gemeinsame Ursache z steht, etwa y = Geburt und x = Störche. Die dritte Variable z kann die Zivilisation (Übergang zu einer moderneren Lebensweise) oder einfach die Zeit (Trend) sein.



Das Ergebnis von oben (ein Zusammenhang verschwindet wenn man die trendbereinigten Werte miteinander korreliert) erhält man übrigens auch wenn man die Zeit als weiteren Regressor in die Regressionsgleichung einbezieht.[9] Man erhält mit unseren Daten:

Dependent Variable: LEBEN

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-280.8123	114.6213	-2.449915	0.0306
ARZT	0.819344	1.431516	0.572361	0.5776
YEAR	0.179592	0.059596	3.013475	0.0108
R-squared	0.974063	Mean dependent var		80.81333
Adjusted R-squared	0.969741	S.D. dependent var		0.966486
S.E. of regression	0.168122	Akaike info criterion		-0.551395
Sum squared resid	0.339181	Schwarz criterion		-0.409785
Log likelihood	7.135461	F-statistic		225.3338
Durbin-Watson stat	1.227221	Prob(F-statistic)		0.000000

Tabelle 1: EQ06

Dass hier eine Scheinkorrelation vorliegt wird auch daran deutlich, dass die abhängige (zu erklärende) Variable y = „leben" mit der Anzahl der privaten Hochschulen = hspriv (an denen meist gar nicht Medizin studiert werden kann) korreliert ist (mit $0{,}934104^{1/2} = 0{,}966491$), obgleich völlig klar ist, dass man diese Korrelation (anders als bei „arzt" als Regressor) nicht kausal interpretieren darf, denn warum sollte durch Gründung privater Hochschulen die Lebenserwartung steigen?

[9] Der Zusammenhang ist in der Ökonometrie bekannt unter dem Namen „Theorem von Frisch und Waugh" (vgl. Tintner, G. (1960), Handbuch der Ökonometrie, S. 276 f.). In einigen Fällen (farblich markiert) haben wir sogar die gleichen numerischen Werte wie in Abbildung 3.

Dependent Variable: LEBEN

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	78.84317	0.159632	493.9054	0.0000
HSPRIV	0.043332	0.003192	13.57500	0.0000
R-squared	0.934104	Mean dependent var		80.81333
Adjusted R-squared	0.929035	S.D. dependent va		0.966486

Tabelle 2

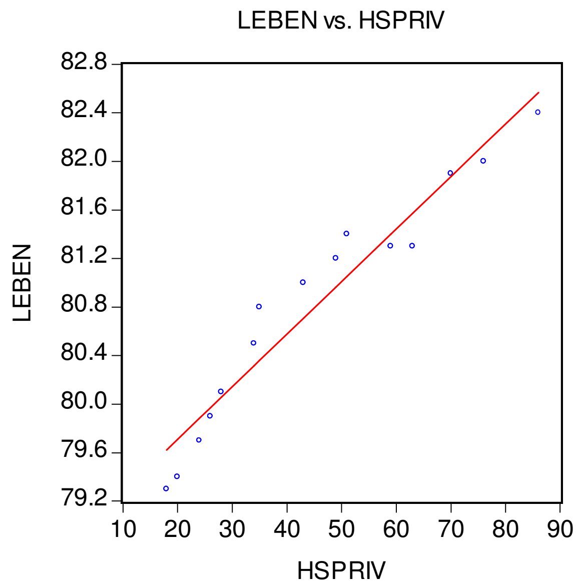


Abbildung 4

Der Zusammenhang zwischen der Lebenserwartung und der Anzahl privater Hochschulen ist in Abbildung 4 auch noch einmal grafisch dargestellt. Es gilt also nicht nur

(1) mehr Ärzte = höhere Lebenserwartung (mag plausibel sein) sondern auch

(2) mehr private Hochschulen = höhere Lebenserwartung.

Eine Hochschule mehr entspricht etwa 16 Tage (365 * 0,0433) mehr Lebenserwartung (10 Hochschulen mehr entspricht also einem halben Jahr).

3. Multiple Regression

Bei einer multiplen Regression werden mehr als ein Regressor herangezogen. Für unsere Analyse – die bisher nur aus *einer* abhängigen Variablen bestand – müssen wir also nun weitere Regressoren hinzunehmen. Es bietet sich an, zunächst die Variable **aus** = Gesundheitsausgaben in Prozent des BIP und dann auch **pkaus** = pro Kopf Gesundheitsausgaben zu betrachten und auch nichtlineare Zusammenhänge nicht a priori auszuschließen. In Abbildung 5 ist eine multivariate Regression mit den Regressoren „arzt“ und „aus“ durchgeführt.

Dependent Variable: LEBEN
Method: Least Squares
Date: 09/29/09 Time: 15:56
Sample: 1992 2006
Included observations: 15

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	69.13018	1.597900	43.26315	0.0000
ARZT	7.026329	0.650739	10.79745	0.0000
AUS	-1.050916	0.326367	-3.220049	0.0074

R-squared	0.975557	Mean dependent var	80.81333
Adjusted R-squared	0.971483	S.D. dependent var	0.966486
S.E. of regression	0.163211	Akaike info criterion	-0.610684
Sum squared resid	0.319656	Schwarz criterion	-0.469074
Log likelihood	7.580134	F-statistic	239.4642
Durbin-Watson stat	1.236946	Prob(F-statistic)	0.000000

Abbildung 5: EQ01

Der kritische Punkt an dieser Regressionsgleichung ist, dass der Regressionskoeffizient für die Variable „aus“ negativ ist (-1,0509). Es dürfte schwer zu vermitteln sein, dass mehr Gesundheitsausgaben in Prozent des BIP zu einer Verkürzung der Lebenserwartung führen. Konkret gesprochen bedeutet das (da „aus“ zwischen 9,6 und 10,8 schwankt), dass eine Steigerung der Gesundheitsausgaben auf 11,5 % (Zunahme um 1 %-Punkt - von 10,5 auf 11,5 % - ergibt wegen -1,0509 eine Verkürzung der Lebensdauer um ungefähr ein Jahr (-1)) des BIP bei Konstanz der Arztdichte die Lebenserwartung um gut ein Jahr verringert.

Andererseits würde dies „neutralisiert“ werden, wenn die Arztdichte um 0,14 = 1/7,026 steigen würde, was nicht aus der Welt gegriffen ist. Die Arztdichte ist von 2001 bis 2006 genau um diesen Wert gestiegen (von 3,31 auf 3,45). Es ist auch klar, dass die Regressoren „arzt“ und „aus“ untereinander stark korrelieren, so dass eine Zunahme von „aus“ um 1 %-Punkt durchaus mit einer Zunahme von „arzt“ um 0,14 einhergehen kann.[10]

Wenn man den negativen Koeffizienten für „aus“ als nicht sehr schön und wenig interpretierbar empfindet, kann man es

- mit einem anderen Regressor versuchen (etwa mit „pkaus“ statt „aus“, so in "Equation: EQ08" in Abbildung 6) oder auch
- weitere Regressoren hinzufügen (vgl. „Equation EQ09“ in Abbildung 6). In EQ09 wurde die Gleichung[11]

  leben c arzt aus pkaus (in der Schreibweise von EViews)
  (oder $y_t = \alpha + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{13t} + u_t$ (mit x_1 = arzt, x_2 = aus und x_3 = pkaus)

  geschätzt.

[10] Genau genommen gilt sogar arzt = -1,608 + 0,467·aus (EQ12), so dass eine Zunahme von „aus“ um 1 „arzt“ nicht um 0,14 sondern sogar um 0,467 erhöht. Die Korrelation r zwischen „arzt“ und „aus“ ist beträchtlich r = 0,9329 (weil $r^2 = R^2 = 0,8704$). Der negative Koeffizient für „aus“ ist also nicht so ein Problem, wie dies zunächst erscheinen mag, weil eine isolierte Erhöhung von „aus“ ohne eine mindestens diese überkompensierende Erhöhung von „arzt“ nicht vorkommen dürfte.

[11] Auf die weitere Möglichkeit, nichtlineare Zusammenhänge zu betrachtet gehen wir in Abschnitt 4 ein.

Equation: EQ08 Workfile: NIHALANI::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: LEBEN
Method: Least Squares
Date: 10/07/09 Time: 16:59
Sample: 1992 2006
Included observations: 15

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	68.91462	2.620964	26.29361	0.0000
ARZT	3.040401	1.187931	2.559408	0.0250
PKAUS	0.000820	0.000465	1.761210	0.1036

R-squared	0.963795	Mean dependent var	80.81333
Adjusted R-squared	0.957760	S.D. dependent var	0.966486
S.E. of regression	0.198635	Akaike info criterion	-0.217839
Sum squared resid	0.473470	Schwarz criterion	-0.076229
Log likelihood	4.633794	F-statistic	159.7211
Durbin-Watson stat	1.220260	Prob(F-statistic)	0.000000

Equation: EQ09 Workfile: NIHALANI::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: LEBEN
Method: Least Squares
Date: 10/07/09 Time: 17:00
Sample: 1992 2006
Included observations: 15

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	72.02130	2.227328	32.33530	0.0000
ARZT	5.307830	1.158836	4.580311	0.0008
AUS	-0.958018	0.306725	-3.123373	0.0097
PKAUS	0.000624	0.000359	1.735713	0.1105

R-squared	0.980812	Mean dependent var	80.81333
Adjusted R-squared	0.975579	S.D. dependent var	0.966486
S.E. of regression	0.151036	Akaike info criterion	-0.719420
Sum squared resid	0.250930	Schwarz criterion	-0.530606
Log likelihood	9.395649	F-statistic	187.4231
Durbin-Watson stat	1.115907	Prob(F-statistic)	0.000000

Abbildung 6: EQ08 – rechts und EQ09 links

Bei EQ08 ist zwar der Koeffizient bei „pkaus“ (anders als bei „aus“) positiv (+0,000820), er ist aber nicht signifikant auf einem Signifikanzniveau von 10 % oder weniger. Bei EQ09 (rechter Teil des Screenshots) ist der Koeffizient von „aus“ wieder negativ. Das kommt übrigens nicht daher, dass „aus“ und „leben“ negativ miteinander korreliert wären,[12] wie die folgende Korrelationsmatrix zeigt:[13]

	ARZT	AUS	LEBEN	PKAUS
ARZT	1	0.932940	0.97695	0.97076
AUS	0.93294	1	0.85911	0.890598
LEBEN	0.97695	0.85911	1	0.97161
PKAUS	0.97077	0.89059	0.971612	1

Tabelle 3

4. Transformationen und Nichtlinearität

Man kann zahlreiche Transformationen z. B. Bildung von Differenzen (genauer: Rückwärtsdifferenzen)[14] $\Delta x_t = x_t - x_{t-1}$ oder Logarithmen $\log(x_t)$ bilden und als Regressoren einführen. In Tabelle 4 ist ein Beispiel zur Schätzung des Modells in Differenzen aufgeführt.

[12] Vielmehr ist die Korrelation + 0,85911.

[13] Betrachtet man die Variablen als Gruppe (group), dann kann man für diese die Korrelationsmatrix aufstellen oder auch unit root Tests etc. durchführen.

[14] In EViews definiert man die Differenzen einer Variable, etwa leben als d(leben). Eine Vowärtsdifferenz wäre $\nabla x_t = x_{t+1} - x_t$.

Dependent Variable: LEBEN

Included observations: 14 after adjustments

Variable	**Coefficient**	**Std. Error**	**t-Statistic**	**Prob.**
C	0.152604	0.172649	0.883900	0.3975
D(ARZT)	1.566191	2.232076	0.701675	0.4989
D(AUS)	-0.844313	0.346178	-2.438959	0.0349
D(PKAUS)	0.000516	0.001132	0.456162	0.6580
R-squared	0.485134	Mean dependent var		0.221429
Adjusted R-squared	0.330674	S.D. dependent var		0.176193
S.E. of regression	0.144148	Akaike info criterion		-0.801001
Sum squared resid	0.207785	Schwarz criterion		-0.618413
Log likelihood	9.607008	F-statistic		3.140837
Durbin-Watson stat	1.766855	Prob(F-statistic)		0.073865

Tabelle 4: EQ10

Wie man sieht ist dieses Modell nicht besser, sondern mit $R^2 = 0,4851$ deutlich schlechter als das entsprechende Modell in den Niveaugrößen (vgl. EQ09 in Abbildung 6 mit $R^2 = 0,9808$). Ein nichtlineares Modell kann des Weiteren häufig durch Logarithmieren (Logarithmus-Transformation) linearisiert werden (vgl. Abbildung 7).

Dependent Variable: LOG(LEBEN)
Method: Least Squares
Date: 09/29/09 Time: 16:49
Sample: 1992 2006
Included observations: 15

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.390969	0.076648	57.28757	0.0000
LOG(ARZT)	0.277373	0.028326	9.792349	0.0000
LOG(AUS)	-0.137838	0.045769	-3.011616	0.0108

R-squared	0.971295	Mean dependent var	4.392075
Adjusted R-squared	0.966510	S.D. dependent var	0.011968
S.E. of regression	0.002190	Akaike info criterion	-9.232740
Sum squared resid	5.76E-05	Schwarz criterion	-9.091130
Log likelihood	72.24555	F-statistic	203.0193
Durbin-Watson stat	1.208070	Prob(F-statistic)	0.000000

Abbildung 7: EQ03

Mit y = leben, x_1 = arzt und x_2 = aus, wäre dies das Modell $\hat{y} = 80.72 \cdot x_1^{0,277} \cdot x_2^{-0,138}$ wobei offen bleibt, wie sinnvoll das dieser Spezifikation zugrundeliegende ökonomische Modell ist. Es ist quasi eine Cobb-Douglas-Produktionsfunktion mit den output „Lebenserwartung" und den inputs „Artzdichte" und „Ausgaben für das Gesundheitswesen". Der negative Koeffizient für log(aus) ist ein Verhältnis von Wachstumsraten und auch hier schwer zu vermitteln: eine Zunahme des Ausgabenanteils um 1 % führt zur Abnahme der Lebenserwartung um 3 %.

Anhang:

Liste der Gleichungen

EQ	y (dep. var.)	Regressoren		
01	leben	arzt	aus	
02	leben	arzt		
03	log(leben)	log(arzt)	log(aus)	
04	leben	year		
05	resid(leben)	resid(arzt)		
06	leben	arzt	year	
07	arzt	year		
08	leben	arzt	pkaus	
09	leben	arzt	aus	pkaus
10	d(leben)	d(arzt)	d(aus)	d(pkaus)
11	leben	hspriv		
12	arzt	aus		

Verzeichnis der Daten

obs	LEBEN	ARZT	AUS	PKAUS	HSPRIV	YEAR	HSALL*
1992	79.3	2.84	9.6	1976	18	1992	318
1993	79.4	2.92	9.6	1992	20	1993	317
1994	79.7	3	9.8	2128	24	1994	329
1995	79.9	3.07	10.1	2274	26	1995	330
1996	80.1	3.11	10.4	2399	28	1996	337
1997	80.5	3.13	10.2	2413	34	1997	340
1998	80.8	3.18	10.2	2483	35	1998	348
1999	81	3.21	10.3	2592	43	1999	346
2000	81.2	3.26	10.3	2671	49	2000	354
2001	81.4	3.31	10.4	2808	51	2001	358
2002	81.3	3.34	10.6	2937	59	2002	362
2003	81.3	3.37	10.8	3088	63	2003	368
2004	81.9	3.39	10.6	3160	70	2004	374
2005	82	3.41	10.7	3348	76	2005	379
2006	82.4	3.45	10.5	3464	86	2006	389

* = "alle (wiss.) Hochschulen" in Deutschland (in der Analyse wurde davon nur die Teilmenge HSPPRIV der privaten Hochschulen benutzt)