Prof. Dr. Peter von der Lippe **(Übungsblatt E)**

Ergänzung der Aufgabe "Mindestlöhne" zu einer multiplen Regression

Das Beispiel "Mindestlöhne" zur einfachen multiplen Regression ergab die folgenden Parameter $\hat{\alpha}$ = 7,1028, $\hat{\beta}$ = - 0,08175, r^2 = 0,008275 sowie die geschätzten Standardabweichungen der Schätzung von $\hat{\alpha}$ und $\hat{\beta}$ in Höhe von $\hat{\sigma}_{\hat{\beta}}$ = 0,400223 und $\hat{\sigma}_{\hat{\alpha}}$ = 2,9256.
Mit den Regressoren x_2, x_3 und x_4 ist der Datensatz wie folgt angereichert worden

y = Arbeitslosenquote
x1 = Höhe des Mindestlohns, x2 = Arbeitskosten 2005 in € je Stunde,
x3 = Personalzusatzkosten in Prozent des Direktentgelts
x4 = Jahressollarbeitsstunden

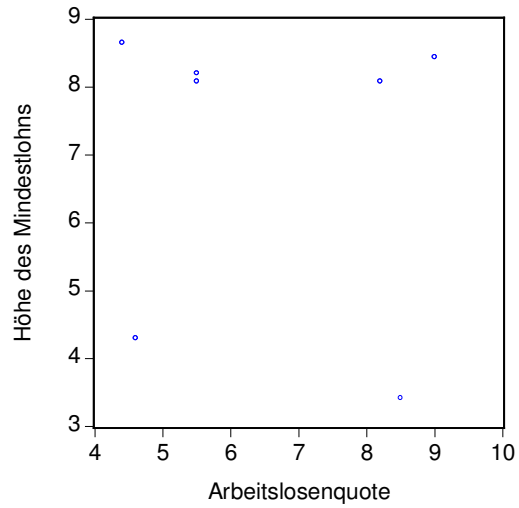
Die Daten sind als eine *Gruppe* zusammengefasst worden:

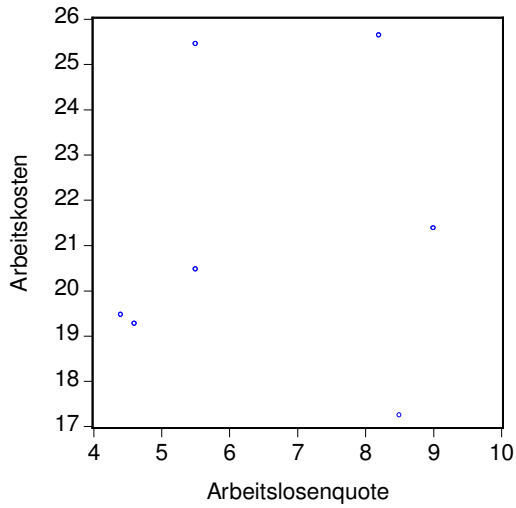
obs	Y	X1	X2	X3	X4
1 Irland	4.4	8.65	19.47	40	1817
2 Frankreich	9	8.44	21.38	93	1610
3 U.K.	5.5	8.2	20.47	46	1704
4 Belgien	8.2	8.08	25.64	90	1709
5 Niederlande	5.5	8.08	25.45	83	1736
6 USA	4.6	4.3	19.27	45	1920
7 Spanien	8.5	3.42	17.25	85	1754

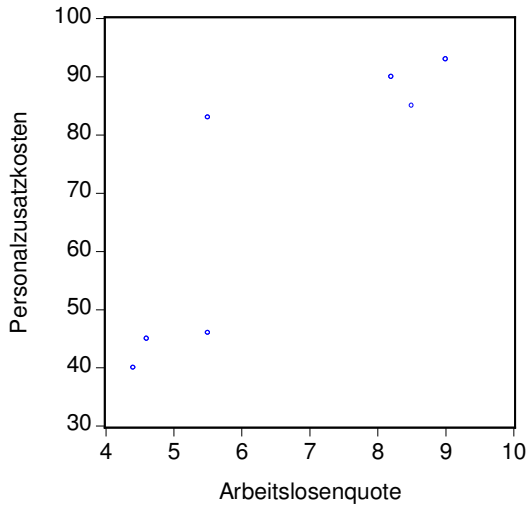
Und für die einfache Regression **nur mit x_1** (wie in der Übungsaufgabe zur einfachen Regression, download A, die mit den dortigen Ergebnissen zu vergleichenden Zahlen sind mit Fettdruck hervorgehoben) erhält man den folgenden Ausdruck

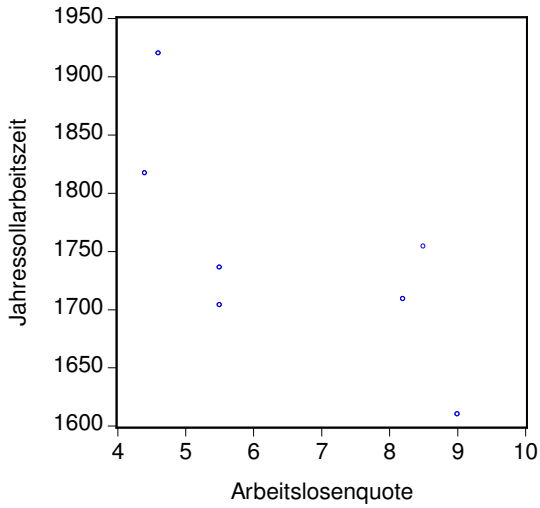
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	**7.102804**	**2.925650**	2.427769	0.0595
X1	**-0.081750**	**0.400224**	-0.204259	0.8462
R-squared	**0.008275**	Mean dependent var		6.528571
Adjusted R-squared	-0.190070	S.D. dependent var		1.964446
S.E. of regression	2.143020	Akaike info criterion		4.597266
Sum squared resid	22.96268	Schwarz criterion		4.581811
Log likelihood	-14.09043	F-statistic		0.041722
Durbin-Watson stat	2.788567	Prob(F-statistic)		0.846205

Wie man sieht, erhält man die gleichen Werte, wie in der anfänglichen Übungsaufgabe (Download A) mit Excel, bzw. dem Taschenrechner errechnet. Ferner ist leicht zu erkennen, dass die Steigung nicht signifikant von 0 verschieden ist, weil der t-Wert nur- 0,204259 beträgt. Auch der Ordinatenabschnitt ist nicht auf dem 5% Niveau gesichert (prob-value hiervon 5,95% also weniger als 5%).
Die mit EViews erzeugten Streuungsdiagramme (jeweils y als abhängige Variabel auf der Abszisse) Befehl "Multiple Graph (Scatter plot)" für die Gruppe "Mindestlohn" der Variablen y, x_1, …,x_4 zeigt, dass bei x_1 und x_2 mit der abhängigen Variable y kaum korrelieren









Die Korrelationstabelle, die ebenfalls leicht mit EViews zu berechnen ist, zeigt – wie schon die Graphik – eine betragsmäßig recht hohe Korrelation zwischen y und x3 und x4

	Y	X1	X2	X3	X4
Y	1.000000	-0.090969	0.103105	0.864993	-0.709782
X1	-0.090969	1.000000	0.605727	0.014235	-0.531127
X2	0.103105	0.605727	1.000000	0.455431	-0.372796
X3	0.864993	0.014235	0.455431	1.000000	-0.667893
X4	-0.709782	-0.531127	-0.372796	-0.667893	1.000000

Bemerkenswert ist, dass die Arbeitslosenquote steigt mit zunehmendem Anteil der Personalnebenkosten an den gesamten Arbeitskosten (r = 0,864993) und auch steigt mit abnehmender Arbeitszeit (Verkürzung der Arbeitszeit *erhöht* also tendenziell die Arbeitslosenquote, wie die negative Korrelation – 0,709782 zeigt).
Die beachtlich hohe Korrelationen zwischen x_1 und x_2 in Höhe von r_{12} = 0,6057, aber auch r_{14} = - 0,5311 und r_{34} = - 0,6679 sprechen dafür, dass man bei den Daten ein Problem mit Multikollinearität hat (vgl. auch Fußnote 2).
Für die "group" der fünf Variablen y, x_1,..,x_4 kann man neben der multiple graph und der Korrelationstabelle auch mit view → descriptive statistics die folgende Tabelle mit Kennzahlen der Verteilung der jeweiligen Variable bestimmen. Einiges davon ist bei der geringen Zahl

der Daten (observations) für jede Variable (nur n = T = 7) im Grunde natürlich wenig sinnvoll und nur hier als Übungsbeispiel zu vertreten.

Das gilt für den **Jarque-Bera Test** auf Normalverteiltheit (vgl. auch Seite 5 unten) der entsprechenden Variablen (er orientiert sich an den beiden Momenten, Schiefe [skewness] und Wölbung [Kurtosis])[1] und natürlich auch für die nachfolgend beschriebenen Regressionen [in der Praxis würde man derartige Rechnungen mit nur t = 7 Wertetupeln nicht durchführen]).

	Y	X1	X2	X3	X4
Mean	6.528571	7.024286	21.27571	68.85714	1750.000
Median	5.500000	8.080000	20.47000	83.00000	1736.000
Maxim um	9.000000	8.650000	25.64000	93.00000	1920.000
Minimum	4.400000	3.420000	17.25000	40.00000	1610.000
Std. Dev.	1.964446	2.185985	3.178615	23.85672	97.46623
Skewness	0.200260	-0.963872	0.426804	-0.251796	0.447346
Kurtosis	1.264124	2.037254	1.821517	1.163242	2.686566
Jarque-Bera	0.925657	1.354231	0.617595	1.057958	0.262125
Probability	0.629500	0.508081	0.734329	0.589206	0.877163
Sum	45.70000	49.17000	148.9300	482.0000	12250.00
Sum Sq. Dev.	23.15429	28.67117	60.62157	3414.857	56998.00

Farblich markiert sind Angaben, die bei den Regressionsrechnungen jeweils für die dependent variable y wieder erscheinen. Es gelten dabei die folgenden Zusammenhänge 6,528 = 45,7/7 und für die Standardabweichung (Std.Dev) , 1,96444 = $\sqrt{23.15429/6}$.

Man kann nun leicht alle einfachen Regressionen berechnen, was hier nur kurz summarisch wiedergegeben werden soll (Regressoren x_i mit i = 1, 2, 3, 4)

	i = 1	i = 2	i = 3	i = 4
C (const.) = $\hat{\alpha}$	7.102804	5.172865	1.624113	31.56366
X_i (Regr. Koeff = $\hat{\beta}_i$)	-0.081750	0.063721	0.071227*	-0.014306
Korrelation r_{yi}	- 0.090969	0.103105	0.864993	- 0.709782
$R^2 = r^2_{y.i}$	0.008272	0.8258	0.748213	0.073990

* dies ist der einzige Regressionskoeffizient der auf dem 5% Niveau signifikant ist (prob 0.011945)

Die größten einfachen Korrelationskoeffizienten sind r_{13} und r_{14}, d.h. neben den Lohnebenkosten hat vor allem die Arbeitszeit einen Einfluss auf die Arbeitslosenquote. Auf der nächsten Seite findet man die Regression von y auf x_3 und x_4. Die Ergebnisse ermöglichen es, partielle Korrelationen auszurechnen und die Rekursionsformeln zu verifizieren:

$$r_{y4.3} = \frac{r_{y4} - r_{y3}r_{34}}{\sqrt{(1-r^2_{y3})(1-r^2_{34})}} = \frac{-0.7097 - [0.865 \cdot (-0.6679)]}{\sqrt{0.25187 \cdot 0.55392}} = \frac{-0.13206}{0.37346} = -0.3536$$

$$R^2_{y.34} = R^2_{y.3} + r^2_{y4.3}(1 - R^2_{y.3}) = 0.748213 + 0.125041 \cdot 0.251787 = 0.779697 .$$

[1] Die Werte für die prob values (0,6295 0,58081 …) sind durchwegs größer als 0,05, so dass die Hypothese der Normalverteiltheit nicht auf dem 5%Niveau verworfen werden kann.

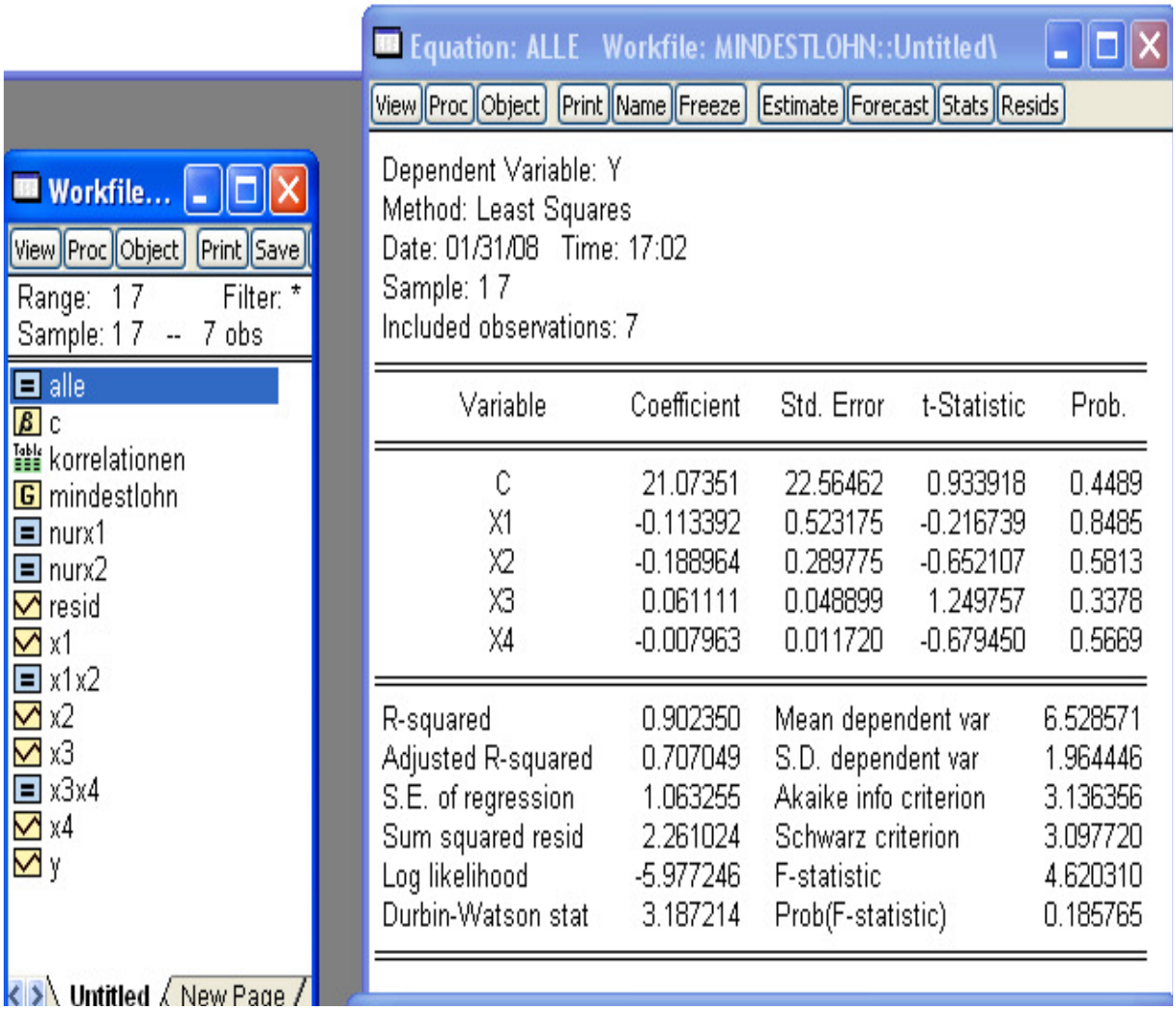
Wie man prüft, ob hinzugekommene Regressoren signifikant sieht steht auf Seite 5 oben.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	10.93603	12.39506	0.882289	0.4275
X3	0.058115	0.025965	2.238211	0.0888
X4	-0.004805	0.006355	-0.756080	0.4917

R-squared	0.779697	Mean dependent var	6.528571
Adjusted R-squared	0.669546	S.D. dependent var	1.964446
S.E. of regression	1.129264	Akaike info criterion	3.378536
Sum squared resid	5.100950	Schwarz criterion	3.355355
Log likelihood	-8.824878	F-statistic	7.078422
Durbin-Watson stat	2.057805	Prob(F-statistic)	0.048533

Wie man sieht hat sich R^2 nicht wesentlich erhöht, nämlich von $R^2_{y.3} = 0.748213$ zu $R^2_{y.34} = 0.779697$. Auch bei Hinzukommen von zwei weiteren Regressoren vergrößert sich R^2 nicht wesentlich zu $R^2_{y.1234} = 0.90235$

Die gelb markierte Zahl (und die entsprechende Zahl in der Gleichung "alle" wird für den F-Test auf der nächsten Seite benötigt) Betrachtet man alle vier (K = 4) Regressoren so erhält man mit E-Views die Gleichung ("Alle" genannt), die im Computerausdruck wie folgt aussieht



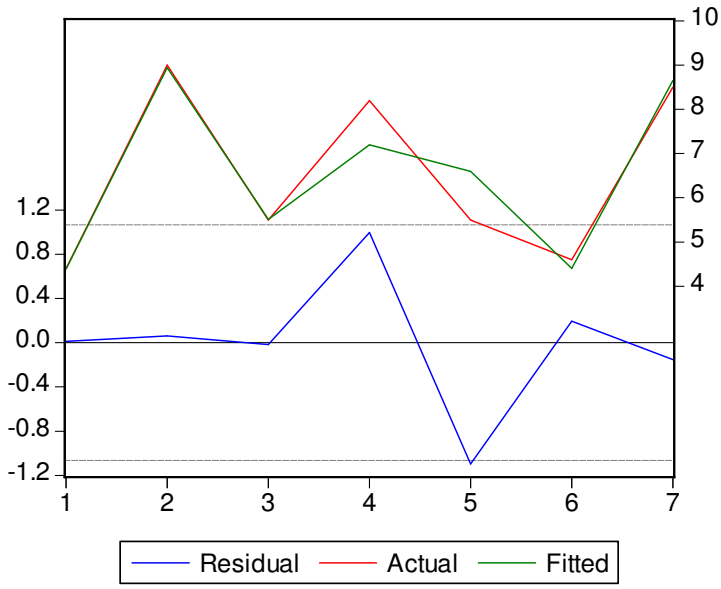
Links sieht man das sehr wichtige workfile window (nicht löschen!!), in dem alle Variablen verzeichnet sind, auch berechnete Gleichungen (Symbol =) oder benannte Tabellen (wie hier etwa "Korrelatio-

nen"). Es ist wichtig, ein Objekt jeweils zu benennen wenn es für spätere Berechnungen wieder benutzt werden soll und nicht verloren gehen soll.

Zum Testen der Hypothese H_0: $\beta_1 = \beta_2 = 0$ (die zu x_3 und x_4 hinzugekommenen Regressoren x_1 und x_2 liefern [zusammengenommen][2] keinen signifikanten Erklärungsbeitrag) ist wie folgt vorzugehen

$$F = \frac{(S^0_{\hat{u}\hat{u}} - S_{\hat{u}\hat{u}})/L}{S_{\hat{u}\hat{u}}/(T-K-1)} = \frac{(5{,}100950 - 2{,}261024)/2}{2{,}261024/2} = 1{,}256$$ was zu vergleichen ist mit dem Tabellenwert der F Verteilung bei 95% Signifikanzniveau und 2 Zähler- Freiheitsgrade und 2 Nenner- Freiheitsgrade. Der Tabellenwert ist 19,0 (da 1,25 < 19 ist H_0 , die Hpothese der Irrelevanz der hinzugekommenen Regressoren) nicht zu verwerfen

Mit "Resids" erhält man die folgende (farbige) Graphik der Residuen zur Beurteilung der Güte der Anpassung. Die Abszisse ist bei Querschnittsdaten, die hier vorliegen nicht von großem Interesse. Man sieht aber, dass sich die Arbeitslosenquote (y) bei den Einheiten 1,2 und 3 (Irland, Frankreich und England) recht gut erklären lässt, Die Einheiten 4 (Belgien) und 5 (NL) haben höhere (bzw. niedrigere) Arbeitslosenquote als aufgrund der Regressoren zu erwarten ist,



Ausgehend von **Resids** kann man auch mit **View** eine ganze Reihe von statistischen Tests (= residual tests) für Annahmen über die Störgrößen durchführen, etwa

Heteroskedastizität (B2)

White Heterosk. Test

Autokorrelation (B3)

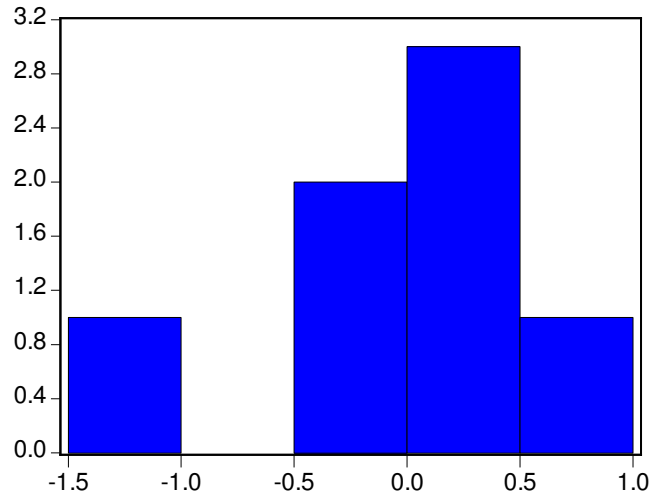
Correlogram Q Statistics

Squared residuals

Serial Correlation LM Test (= Breusch Godfrey Test)

ARCH LM Test

Normalverteiltheit (B4) → Histogram Normality Test, das ist die folgende Abbildung:



Series: Residuals	
Sample 1 7	
Observations 7	
Mean	-5.23e-16
Median	0.011607
Maximum	0.997262
Minimum	-1.095848
Std. Dev.	0.613871
Skewness	-0.249073
Kurtosis	3.331731
Jarque-Bera	0.104474
Probability	0.949104

[2] Dass jeder Regressor für sich genommen nicht signifikant ist sieht man an den t-Werten -0.2167 und -0.6521 und den dazugehörigen prob values (0.8485 und 0.5813), die alle größer als 0.05 also 5% sind. Wären die einzelne Regressoren (t Test) nicht signifikant, die Gesamtregression (F Test) dagegen schon, so spräche das für das Vorliegen von Multikollinearität.