Tutorial 6

keywords: hypothesis testing, t-test, test statistic, critical value, confidence intervals, R-squared, interpretation of coefficients, multiple linear regression, omitted variable bias

estimated reading time: 33 minutes

Quang Bui

August 28, 2018

2018 - Semester 1. Question 1

A sample of 25 employees was taken. Each employee was asked to assess his own job satisfaction (X) on a scale from 1 to 10. The number of days (Y) an employee was absent from work was also registered. Fitting a linear regression model based on least squares method gave the sample regression line,

$$\hat{Y} = 13.6 - 1.2X$$

Also found were:

$$\bar{X} = 6.0 \quad \sum_{i=1}^{n}(X_i - \bar{X})^2 = 130 \quad SSR = 80.6 \quad ESS = 186.9$$

(a) Test at the 1% level that job satisfaction has no effect on absenteeism using the t-test based on the sample slope estimate.

For the simple linear regression model of absenteeism (Y),

$$Y = \beta_0 + \beta_1 X + u$$

β_1 measures the true effect of job satisfaction on absenteeism (not holding any variable(s) constant).

If job satisfaction does not have a true effect on absenteeism then,

$$\beta_1 = 0$$

but if it does,

$$\beta_1 \neq 0$$

After we estimate our model,

$$\hat{Y} = 13.6 - 1.2X$$

we can perform this hypothesis test.

State the null and alternative hypothesis

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

The test statistic and its distribution under H_0

$$t = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \sim t_{n-k-1} \quad under\ H_0$$

$$
\begin{aligned}
n &= sample\ size = 25 \\
k &= number\ of\ regressors\ in\ the\ model = 1 \\
d.o.f &= n - k - 1 = 23
\end{aligned}
$$

Calculate the test statistic

$$t_{calc} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = ?$$

For a simple linear regression model the standard error of $\hat{\beta}_1$ is given by the formula,

$$[se(\hat{\beta}_1)]^2 = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-k-1} = \frac{SSR}{n-k-1} = \frac{80.6}{23} = 3.5043$$

$$[se(\hat{\beta}_1)]^2 = \frac{3.5043}{130} = 0.0270$$

$$\therefore se(\hat{\beta}_1) = (0.0270)^{1/2} = 0.1643$$

$$\therefore t_{calc} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{-1.2}{0.1643} = -7.3037$$

Critical value and rejection region

1% $significance\ level\ \therefore \alpha = 0.01$, two-sided t-test

Since we are performing a t-test, the critical value(s) (which bounds the rejection region) come from a t-distribution. The t-distribution of interest in this hypothesis test, is one with $degrees\ of\ freedom\ = n - k - 1 = 25 - 2 = 23$.

TABLE G.2 Critical Values of the t Distribution

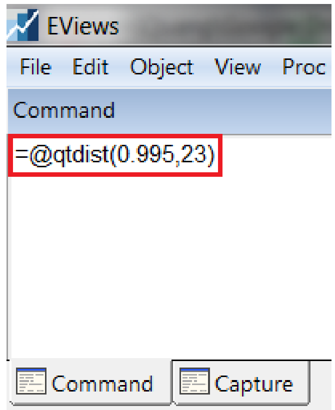
	Significance Level				
1-Tailed:	.10	.05	.025	.01	.005
2-Tailed:	.20	.10	.05	.02	$\boxed{.01}$
Degrees of Freedom					
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
$\boxed{23}$	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
90	1.291	1.662	1.987	2.368	2.632
120	1.289	1.658	1.980	2.358	2.617

From the statistics table:

$$+t_{crit} = t_{23,0.005} = 2.807$$
$$-t_{crit} = t_{23,0.005} = -2.807$$

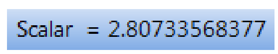
To obtain the critical value using EViews,

$$Command\ window := @qtdist(0.995, 23)$$



$(press\ Enter\ to\ execute)$

and the value appears in the bottom left corner,



From EViews:

$$+t_{crit} = t_{23,0.005} = 2.807$$
$$-t_{crit} = t_{23,0.005} = -2.807$$

For a two-sided t-test, we reject H_0 if,

$$t_{calc} > +t_{crit}$$
$$or$$
$$t_{calc} < -t_{crit}$$

Conclusion

Since $t_{calc} = -7.3037 < -t_{crit} = -2.807$, we reject the null at the 1% significance level and conclude that there is sufficient evidence from our sample to suggest that job satisfaction has a statistically significant impact on absenteeism.

(b) Compute the coefficient of determination and comment on it.

$$R^2 = \frac{ESS}{SSR}$$
$$SST = ESS + SSR = 186.9 + 80.6 = 267.5$$
$$\therefore R^2 = \frac{186.9}{267.5} = 0.699$$

Approximately 70% of the variability in absenteeism can be explained by job satisfaction.

Question 1

Hypothesis test on a single parameter, the meanings of the size of a test and a confidence interval:

Consider the classical linear regression model

$$y_i = \beta_0 + \beta_1 x_i + u_i \quad i = 1, 2, \ldots, n$$

A random sample of size $n = 22$ is drawn and the estimated model based on this sample is:

$$\hat{y}_i = \underset{(3.1)}{5.4} + \underset{(1.5)}{3.2} x_i \quad i = 1, 2, \ldots, 22$$
$$R^2 = 0.26$$

(a) Test $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ at the 5%.

For the simple linear regression model of y,

$$y_i = \beta_0 + \beta_1 x_i + u_i \quad i = 1, 2, \ldots, 22$$

β_1 measures the true impact that x has on y (not holding any variable(s) constant because there are no other independent variables other than x in this model).

If x does not truly impact y then,

$$\beta_1 = 0$$

but if it does,

$$\beta_1 \neq 0$$

After we estimate our model,

$$\hat{y}_i = \underset{(3.1)}{5.4} + \underset{(1.5)}{3.2} x_i \quad i = 1, 2, \ldots, 22$$

we can perform this hypothesis test.

State the null and alternative hypothesis

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

The test statistic and its distribution under H_0

$$t = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \sim t_{n-k-1} \quad under\ H_0$$

$$
\begin{aligned}
n &= sample\ size = 250 \\
k &= number\ of\ regressors\ in\ the\ model = 1 \\
d.o.f &= n - k - 1 = 248
\end{aligned}
$$

Calculate the test statistic

$$t_{calc} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{3.2}{1.5} = 2.1333$$

Critical value and rejection region

$5\%\ significance\ level\ \therefore \alpha = 0.05$, two-sided t-test

Since we are performing a t-test, the critical value(s) (which bounds the rejection region) come from a t-distribution. The t-distribution of interest in this hypothesis test, is one with $degrees\ of\ freedom\ = n - k - 1 = 250 - 2 = 248$. Since $d.o.f = 248$ is not in the statistics table, we take a conservative approach and choose the closest available degrees of freedom less than 248 i.e. $d.o.f = 120$.

TABLE G.2 Critical Values of the t Distribution

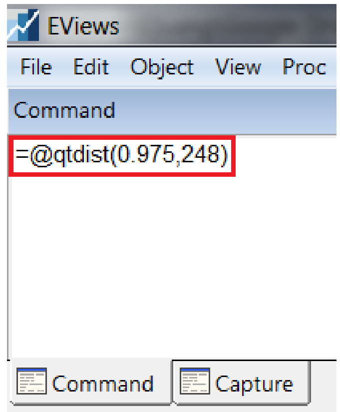
	Significance Level				
1-Tailed:	.10	.05	.025	.01	.005
2-Tailed:	.20	.10	$\boxed{.05}$	.02	.01
Degrees of Freedom 1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
90	1.291	1.662	1.987	2.368	2.632
$\boxed{120}$	1.289	1.658	1.980	2.358	2.617
∞	1.282	1.645	1.960	2.326	2.576

From the statistics table:

$$+t_{crit} = 1.980$$
$$-t_{crit} = -1.980$$

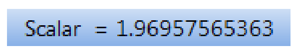
To obtain the critical value using EViews,

$$Command\ window := @qtdist(0.975, 248)$$



$(press\ Enter\ to\ execute)$

and the value appears in the bottom left corner,



From EViews:

$$+t_{crit} = t_{248,0.975} = 1.970$$
$$-t_{crit} = t_{248,0.025} = -1.970$$

For a two-sided t-test, we reject H_0 if,

$$t_{calc} > +t_{crit}$$

$$or$$

$$t_{calc} < -t_{crit}$$

Conclusion

Since $t_{calc} = 2.1333 > +t_{crit} = 1.970$, we reject the null at the 5% significance level and conclude that there is sufficient evidence from our sample to suggest that x has a statistically significant impact on y.

(b) Construct a 95% confidence interval for β_1

$$\hat{\beta}_1 \pm t_{crit} \times se(\hat{\beta}_1)$$

$$\hat{\beta}_1 \pm t_{n-k-1,1-\frac{\alpha}{2}} \times se(\hat{\beta}_1)$$

$$\hat{\beta}_1 \pm t_{248,0.975} \times se(\hat{\beta}_1)$$

$$3.2 \pm 1.970 \times 1.5$$

$$(0.245, 6.155)$$

We are 95% confident that the true impact of x on y is between 0.245 and 6.155.

(c) Suppose that you learn that y_i and x_i are independent. Would you be surprised? Explain.

If y_i and x_i are independent then x_i should not effect y_i, that is,

$$\beta_1 = 0$$

but we rejected

$$H_0 : \beta_1 = 0$$

and concluded that x_i has a statistically significant impact on y_i at the 5% significance level $\therefore$ we would be surprised.

From this sample, we found evidence at the 5% significance level to reject the null hypothesis that x_i has no impact on y_i. That is, there is only a 5% probability of rejecting H_0 when it is in fact true. We could have wrongly rejected H_0 when it is true (Type I error), but since we performed this test at the 5% significance level, this error occurs with only a 5% probability,

$$P(Type\ I\ error) = \alpha = 0.05$$

Given that y_i and x_i are independent $\therefore$ x_i has no true impact on y_i i.e. $\beta_1 = 0$, if we applied repeated sampling and performed the same hypothesis test across many more samples, we would find that H_0 is wrongly rejected in 5% of these samples.

(d) Suppose that you learn that y_i and x_i are independent and many samples of $n = 250$ are drawn, regressions estimated, and (a) and (b) answered. In what fraction of the samples would H_0 from (a) be rejected? In what fraction of samples would the value $\beta_1 = 0$ be included in the confidence interval from (b)?

The significance level, α, is the probability of rejecting the null hypothesis when it is true. Since we set $\alpha = 0.05$, the probability of rejecting $H_0 : \beta_1 = 0$ when it is true i.e. x_i has no true impact on y_i is 0.05.

Therefore, given that x_i does not help to explain y_i (because they are independent), we would reject $H_0 : \beta_1 = 0$ in 5% of the samples.

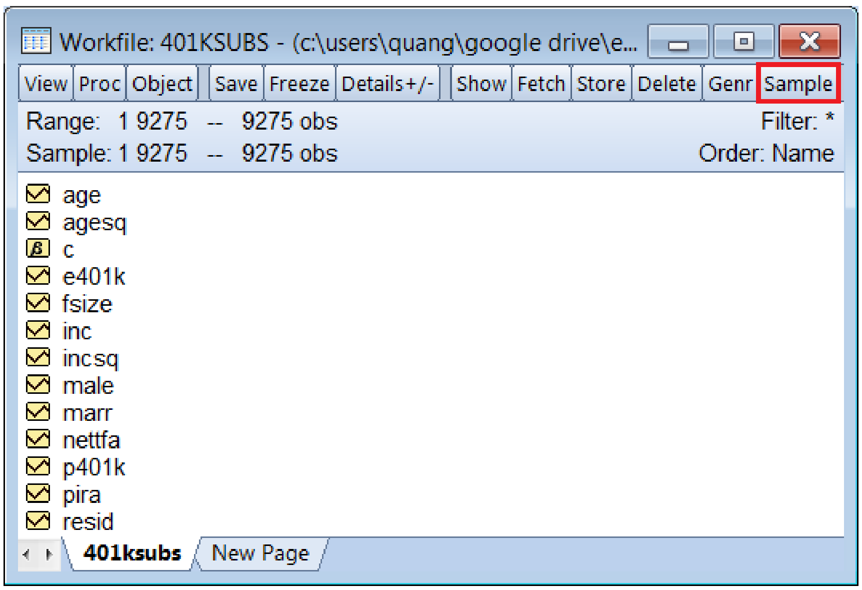
The true value of β_1 will lie in 95% of the confidence intervals. If we learn that y_i and x_i are independent, then the true value of β_1 is 0, so $\beta_1 = 0$ would lie in 95% of the confidence intervals.

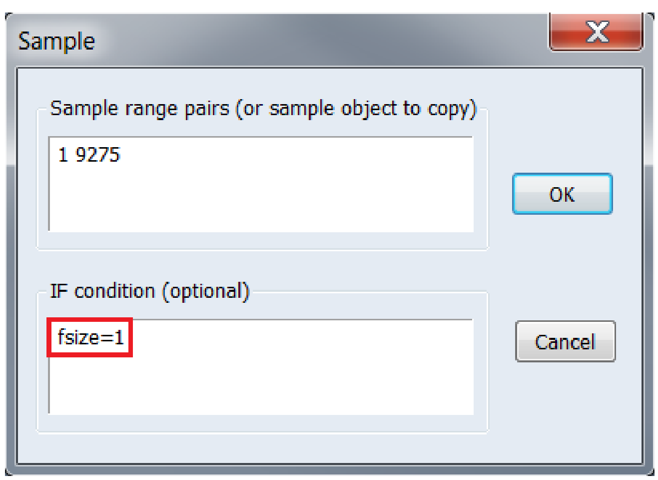
Question 3

File $401KSUBS.wf1$ contains information on net financial wealth ($nettfa$), age of the survey respondent (age), annual family income (inc), family size ($fsize$), and participation in certain pension plans for people in the United States. The wealth and income variables are both recorded in thousands of dollars.

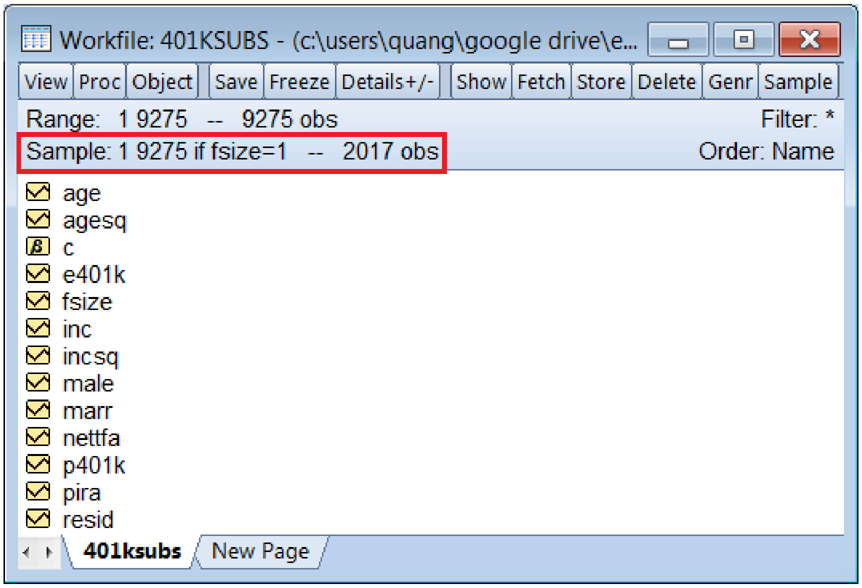
(a) How many single-person households are there in the data set?

We need to change our sample to include only single-person households. In EViews, click on *Sample* and type $fsize = 1$ in the *IF condition* dialog box,





This tells EViews to change the current working sample, to a sample of only single-person households,



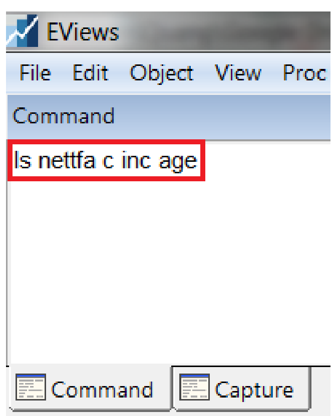
Using the data only for single-person households, estimate the model

$$nettfa_i = \beta_0 + \beta_1 inc_i + \beta_2 age_i + u_i \quad i = 1, 2, \ldots, 2017$$

Report the estimated equation (including standard errors of coefficients). Interpret the slope coefficients. Are there any surprises in the slope estimates.

To estimate an model from the Command window,

$$ls\ nettfa\ c\ inc\ age$$



$(press\ Enter\ to\ execute\ code)$

Dependent Variable: NETTFA
Method: Least Squares
Date: 04/07/18 Time: 18:24
Sample: 1 9275 IF FSIZE=1
Included observations: 2017

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	−43.03981	4.080393	−10.54796	0.0000
INC	0.799317	0.059731	13.38200	0.0000
AGE	0.842656	0.092017	9.157631	0.0000

R-squared	0.119343	Mean dependent var	13.59498
Adjusted R-squared	0.118469	S.D. dependent var	47.59058
S.E. of regression	44.68275	Akaike info criterion	10.43854
Sum squared resid	4021048.	Schwarz criterion	10.44688
Log likelihood	−10524.27	Hannan-Quinn criter.	10.44160
F-statistic	136.4648	Durbin-Watson stat	1.959509
Prob(F-statistic)	0.000000		

Table 1: Regression output of $nettfa$ on a constant, inc, and age.

$$\widehat{nettfa} = \underset{(4.0804)}{-43.0398} + \underset{(0.0597)}{0.7993}inc + \underset{(0.0920)}{0.8427}age$$

Interpretations of the estimated coefficients:

$\hat{\beta}_1 = 0.7993$ - The model estimates that for an additional \$1,000 in income, net financial wealth is predicted to increase by approximately \$800, on average, holding the age of the individual constant. ($nettfa$ and inc are measured in \$'000.)

$\hat{\beta}_2 = 0.8427$ - The model estimates that if a person ages by 1 year, his/her net financial wealth is predicted to increase by approximately \$843, on average, holding income constant. ($nettfa$ is measured in \$'000.)

(c) Does the intercept in (b) have an interesting meaning? Explain.

$\hat{\beta}_0 = -43.0398$. The estimated intercept coefficient represents the predicted net financial wealth for an individual aged 0 with no income. The population of interest is single-person households and there are clearly no one with those characteristics in this population.

(d) Test the hypothesis that $H_0 : \beta_2 = 1$ against $H_1 : \beta_2 < 1$. Do you reject the null hypothesis at the 1% significance level?

We are testing the null hypothesis that aging by one year increases net financial wealth by \$1,000, against the alternative hypothesis that it increases net financial wealth by less than \$1,000. ($nettfa$ is measured in \$'000.)

State the null and alternative hypothesis

$$H_0 : \beta_2 = 1$$
$$H_1 : \beta_2 < 1$$

The test statistic and its distribution under H_0

$$t = \frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)} = \frac{\hat{\beta}_2 - 1}{se(\hat{\beta}_2)} \sim t_{n-k-1} \quad under\ H_0$$
$$n = sample\ size = 2017$$
$$k = no.\ of\ regressors\ in\ the\ model = 2$$
$$d.o.f = n - k - 1 = 2014$$

Calculate the test statistic

$$t_{calc} = \frac{\hat{\beta}_2 - 1}{se(\hat{\beta}_2)} = \frac{0.8427 - 1}{0.0920} = -1.7098$$

Critical value and rejection region

1% $significance\ level\ \therefore \alpha = 0.01$, one-sided t-test on the left tail

TABLE G.2 Critical Values of the *t* Distribution

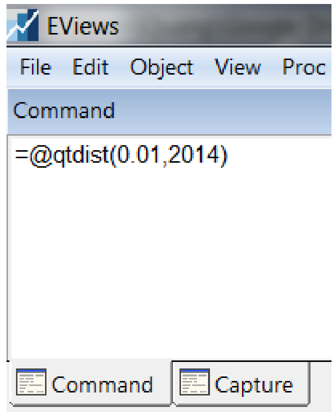
	Significance Level				
1-Tailed:	.10	.05	.025	.01	.005
2-Tailed:	.20	.10	.05	.02	.01
Degrees of Freedom					
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
90	1.291	1.662	1.987	2.368	2.632
$\boxed{120}$	1.289	1.658	1.980	2.358	2.617
∞	1.282	1.645	1.960	2.326	2.576

Since we are performing a one-sided t-test on the left tail, where the rejection region lies, we will compare t_{calc} with $-t_{crit}$. From the statistics table:

$$-t_{crit} = -2.358$$

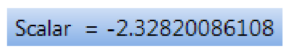
To obtain the critical value using EViews,

$$Command\ window := @qtdist(0.01, 2014)$$



$(press\ Enter\ to\ execute)$

and the value appears in the bottom left corner,



From EViews:

$$-t_{crit} = -2.3282$$

For a one-sided t-test on the left tail, we reject H_0 if,

$$t_{calc} < -t_{crit}$$

Conclusion

Since $t_{calc} = -1.7098 > -t_{crit} = -2.3282$, we do not reject the null at the 1% significance level and conclude that there is insufficient evidence from our sample to suggest that aging by 1 year increases net financial wealth by less than $1,000.

(e) If you omit *age* from the model and rerun the regression, is the estimated coefficient on *inc* much different from the estimate in part (b)? Why or why not?

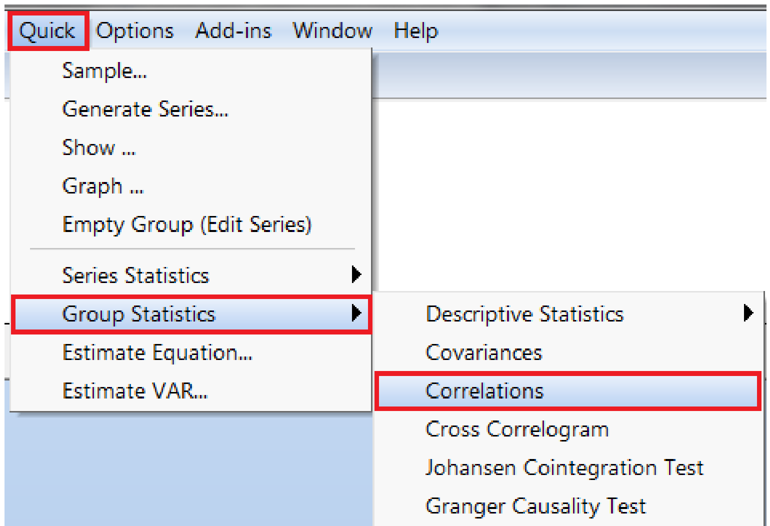
The estimated coefficient on *inc* represents the estimated change in net financial wealth for a unit increase in income, holding age constant. If *age* were uncorrelated with *inc*,

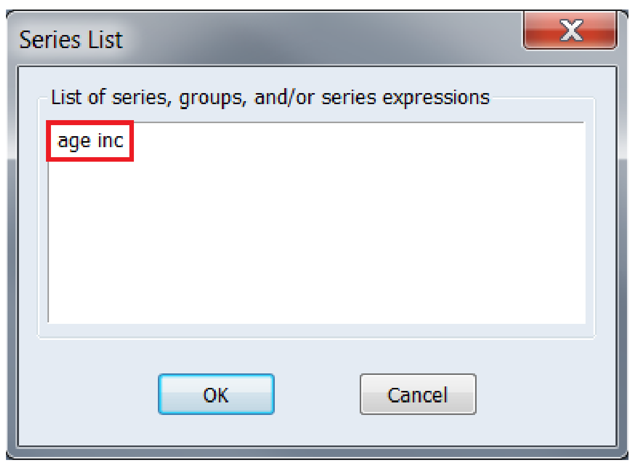
$$\widehat{corr}(age, inc) = 0$$

then whether or not we hold *age* constant, will not impact the effect of income on net financial wealth. Put differently, if *age* were strongly correlated with *inc* or could be a proxy for *inc*, then the effect of income on net financial wealth should change after controlling for the effect of age on net financial wealth constant.

To obtain the sample correlation coefficient of age and inc,

$$Quick \rightarrow \ Group\ Statistics \rightarrow Correlations$$





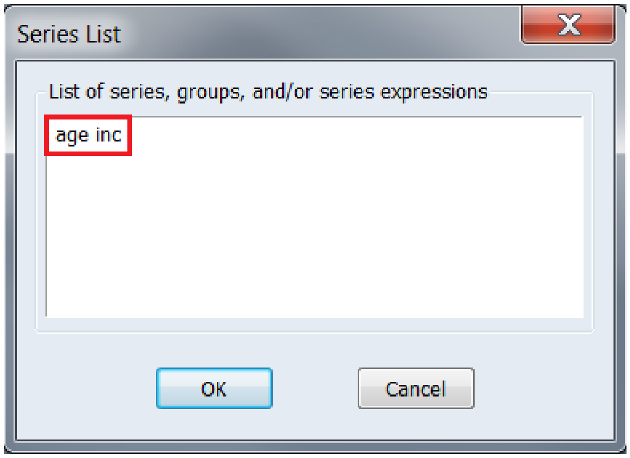
	AGE	INC
AGE	1.000000	0.039059
INC	0.039059	1.000000

$$\widehat{corr}(age, inc) = 0.0391$$

age and inc have a very weak linear relationship.

Rerunning the model of $nettfa$ without age using the Command window,

$$Command\ window : ls\ nettfa\ c\ inc$$



(*press Enter to execute code*)

Dependent Variable: NETTFA
Method: Least Squares
Date: 04/09/18 Time: 18:26
Sample: 1 9275 IF FSIZE=1
Included observations: 2017

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	−10.57095	2.060678	−5.129843	0.0000
INC	0.820681	0.060900	13.47589	0.0000

R-squared	0.082673	Mean dependent var	13.59498
Adjusted R-squared	0.082218	S.D. dependent var	47.59058
S.E. of regression	45.59223	Akaike info criterion	10.47834
Sum squared resid	4188483.	Schwarz criterion	10.48390
Log likelihood	−10565.41	Hannan-Quinn criter.	10.48038
F-statistic	181.5995	Durbin-Watson stat	1.914495
Prob(F-statistic)	0.000000		

$$\widehat{nettfa} = \underset{(2.0607)}{-10.5710} + \underset{(0.0609)}{0.8207}inc$$

As we can see, the estimated coefficient of *inc* is now 0.82 which is not that much different from that value of 0.79 obtained in part (b). This implies that there is no significant omitted variable bias for the coefficient on *inc* after *age* has been removed.

If age, which is assumed to belong in the model of net financial wealth, is omitted from the model,

$$nettfa = \beta_0 + \beta_1 inc + v$$

it is then captured by the error term v,

$$v = \beta_2 age + u$$

If age is also correlated with income, then we will have an omitted variable bias problem i.e. the OLS estimator will be a biased estimator and we will have biased estimates.

That is, estimating

$$nettfa = \beta_0 + \beta_1 inc + v$$

with the OLS estimator,

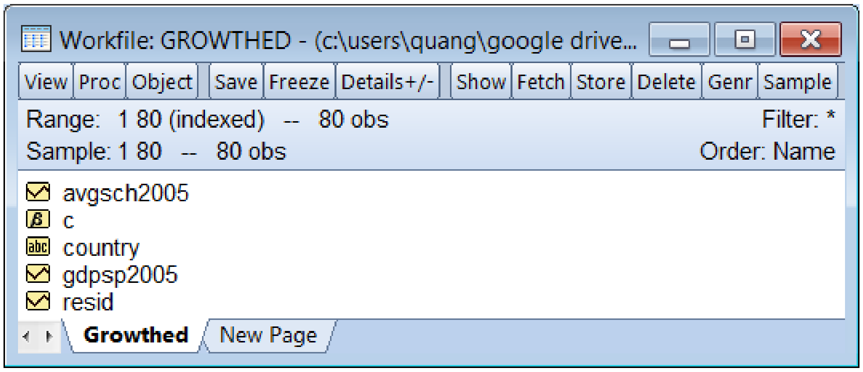
$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$$

will produce biased estimates,

$$E(\hat{\boldsymbol{\beta}}) \neq \boldsymbol{\beta}$$

Question 4

File $GROWTHED.wf1$ contains observations on GDP per capita (in US dollars) in 2005 and 'Average years spent in education in 2005' for 80 countries.



(a) From an economic point of view, what direction would you expect the relationship between GDP and education to have?

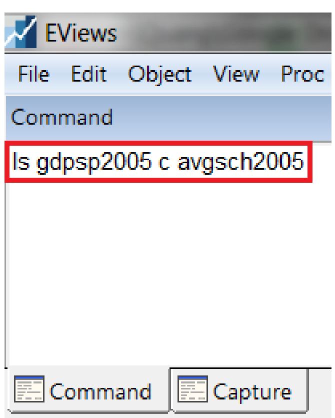
We would expect countries with higher levels of education on average to be more productive which in turn leads to higher output (income) per worker $\therefore$ a positive correlation between GDP and measures of a country's education.

(b) Given your conclusion in (a), run the relevant regression, report the estimated model and interpret the estimates for the intercept and slope coefficients.

$$gdpsp2005 = \beta_0 + \beta_1 avgsch2005 + u$$

To estimate the model of $gdpsp2005$ from the Command window,

$$Command\ window: ls\ gdpsp2005\ c\ avgsch2005$$



(*press Enter to execute code*)

Dependent Variable: GDPSP2005
Method: Least Squares
Date: 04/08/18 Time: 17:30
Sample: 1 80
Included observations: 80

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-9255.898	1663.450	-5.564278	0.0000
AVGSCH2005	2734.208	206.1562	13.26280	0.0000

R-squared	0.692795	Mean dependent var	10976.04
Adjusted R-squared	0.688856	S.D. dependent var	10636.55
S.E. of regression	5933.098	Akaike info criterion	20.23916
Sum squared resid	$2.75E+09$	Schwarz criterion	20.29871
Log likelihood	-807.5665	Hannan-Quinn criter.	20.26304
F-statistic	175.9018	Durbin-Watson stat	1.540989
Prob(F-statistic)	0.000000		

$$\widehat{gdpsp2005} = \underset{(1663.450)}{-9255.898} + \underset{(206.1562)}{2734.208} avgsch2005$$

Interpretations of the estimated coefficients:

$\hat{\beta}_0 = -9255.898$

The model estimates that in a country where people on average have no education ($avgsch2005 = 0$), the level of GDP is per capita is expected to be -$9,255.90. This result does not make much economic sense.

$\hat{\beta}_1 = 2734.208$

The model estimates that when average year of education increases by 1 year, the countries level of GDP per capita is expected to increase by $2,734.208.

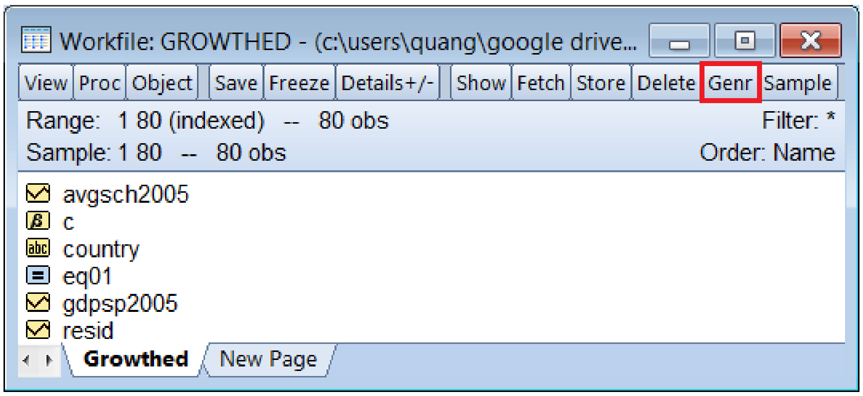
(c) How could you run your regression again to address a strange/meaningless result from your regression in part (b)?

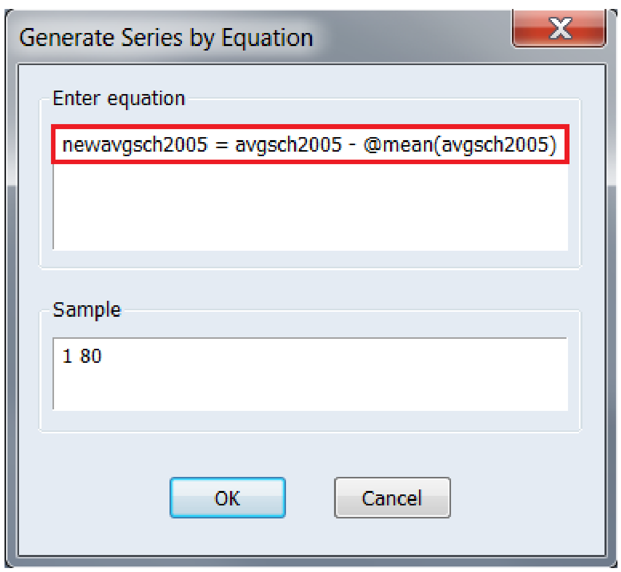
We could transformation the independent variable such that the estimated intercept represents GDP per capita for a country whose average level of education is the sample of country's mean average level of education ($\overline{avgsch2005}$),

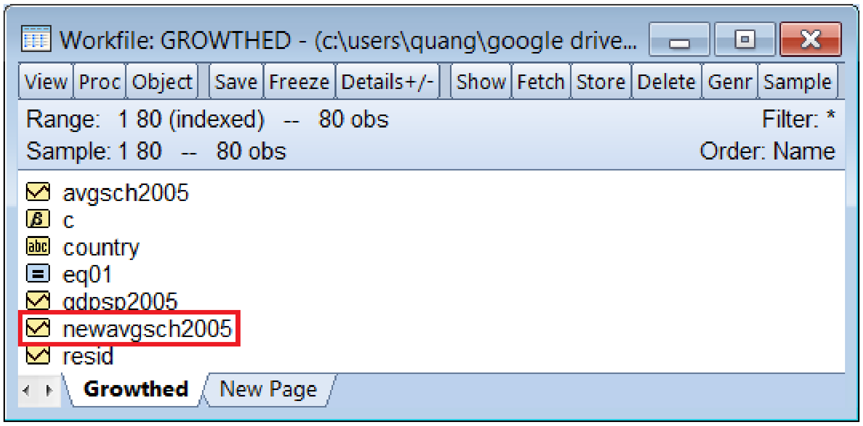
$$newavgsch2005 = avgsch2005 - \overline{avgsch2005}$$

To generate the variable $newavgsch2005$,

$$Genr \rightarrow newavgsch2005 = avgsch2005 - @mean(avgsch2005) \rightarrow OK$$





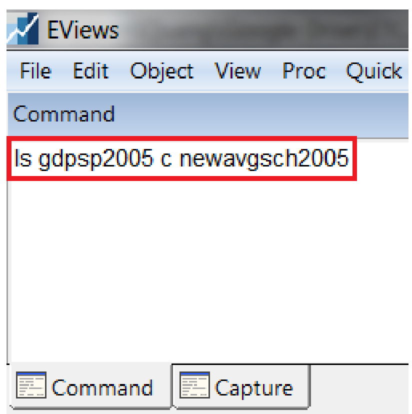


We want to run the following regression:

$$gdpsp2005 = \beta_0 + \beta_1 newavgsch2005 + u$$

to do this from the Command window,

$$Command\ window : ls\ gdpsp2005\ c\ newavgsch2005$$



Dependent Variable: GDPSP2005
Method: Least Squares
Date: 04/08/18 Time: 18:18
Sample: 1 80
Included observations: 80

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	10976.04	663.3405	16.54662	0.0000
NEWAVGSCH2005	2734.208	206.1562	13.26280	0.0000

R-squared	0.692795	Mean dependent var	10976.04
Adjusted R-squared	0.688856	S.D. dependent var	10636.55
S.E. of regression	5933.098	Akaike info criterion	20.23916
Sum squared resid	$2.75E+09$	Schwarz criterion	20.29871
Log likelihood	-807.5665	Hannan-Quinn criter.	20.26304
F-statistic	175.9018	Durbin-Watson stat	1.540989
Prob(F-statistic)	0.000000		

$$\widehat{gdpsp2005} = \underset{(663.3405)}{10976.04} + \underset{(206.1562)}{2734.208} newavgsch2005$$

The estimated slope coefficient, $\hat{\beta}_1$, remains the same, but the estimated intercept coefficient changes. This estimated intercept coefficient is now interpreted as the estimated level of GDP per capita for a country where the people's average year of education equals to the sample mean average year of education.

When,

$$avgsch2005 = \overline{avgsch2005}$$

then,

$$\begin{aligned} newavgsch2005 &= avgsch2005 - \overline{avgsch2005} \\ &= \overline{avgsch2005} - \overline{avgsch2005} \\ &= 0 \end{aligned}$$

$$\therefore \widehat{gdpsp2005} = 10976.04 + 2734.208 \times 0 = 10976.04$$

(d) What is the coefficient of determination in the regression of part (b) and how would you interpret it?

$$R^2 = 69.3\%$$

Approximately 70% of the variability in GDP per capita can be explained by the country's average level of education.

(e) Is there a statistically significant relationship between education and GDP per capita?

Dependent Variable: GDPSP2005
Method: Least Squares
Date: 04/08/18 Time: 17:30
Sample: 1 80
Included observations: 80

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	−9255.898	1663.450	−5.564278	0.0000
AVGSCH2005	2734.208	206.1562	13.26280	0.0000

R-squared	0.692795	Mean dependent var	10976.04
Adjusted R-squared	0.688856	S.D. dependent var	10636.55
S.E. of regression	5933.098	Akaike info criterion	20.23916
Sum squared resid	$2.75E+09$	Schwarz criterion	20.29871
Log likelihood	−807.5665	Hannan-Quinn criter.	20.26304
F-statistic	175.9018	Durbin-Watson stat	1.540989
Prob(F-statistic)	0.000000		

The p-value for a test of statistical significant is reported in the regression output. Here, the p-value is 0.0000 which is less than α at any reasonable level of significance, therefore we would reject $H_0 : \beta_1 = 0$ and conclude that there is sufficient evidence from our sample to suggest that education has a statistically significant effect on GDP per capita.

(f) What is the 95% confidence interval for the slope coefficient? Comment on it.

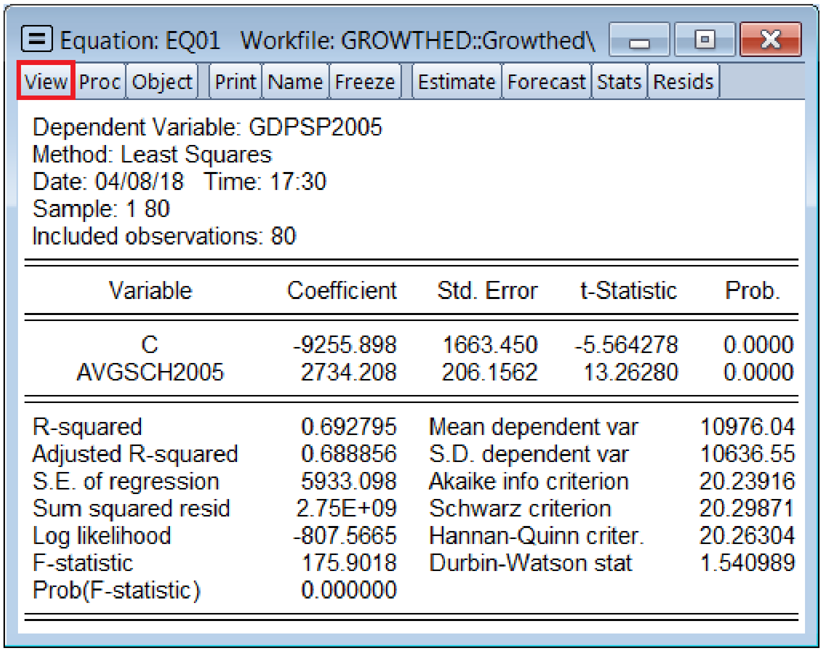
$$\hat{\beta}_1 \pm t_{crit} \times se(\hat{\beta}_1)$$

$$\hat{\beta}_1 \pm t_{n-k-1,1-\frac{\alpha}{2}} \times se(\hat{\beta}_1)$$

$$\hat{\beta}_1 \pm t_{78,0.975} \times se(\hat{\beta}_1)$$

To obtain the 95% CI of β_1 in EViews,

$$View \rightarrow Coefficient\ diagnostics \rightarrow Confidence\ intervals \rightarrow 0.95$$

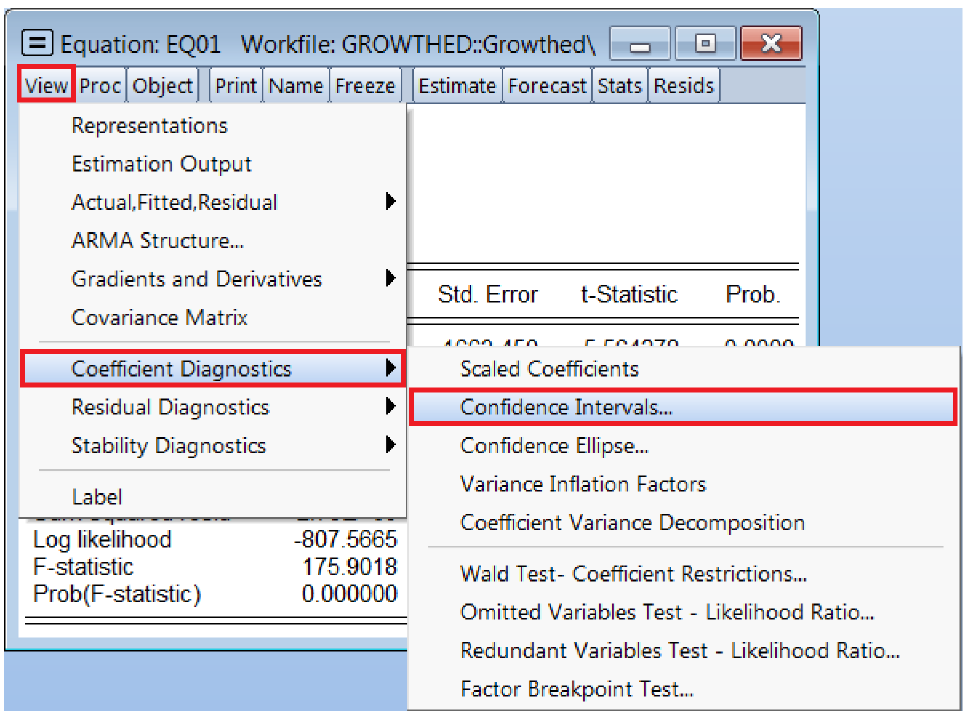

Equation: EQ01 Workfile: GROWTHED::Growthed\

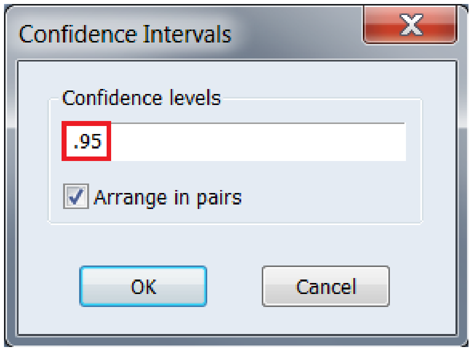
View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: GDPSP2005
Method: Least Squares
Date: 04/08/18 Time: 17:30
Sample: 1 80
Included observations: 80

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-9255.898	1663.450	-5.564278	0.0000
AVGSCH2005	2734.208	206.1562	13.26280	0.0000

R-squared	0.692795	Mean dependent var	10976.04
Adjusted R-squared	0.688856	S.D. dependent var	10636.55
S.E. of regression	5933.098	Akaike info criterion	20.23916
Sum squared resid	2.75E+09	Schwarz criterion	20.29871
Log likelihood	-807.5665	Hannan-Quinn criter.	20.26304
F-statistic	175.9018	Durbin-Watson stat	1.540989
Prob(F-statistic)	0.000000		





Coefficient Confidence Intervals
Date: 04/08/18 Time: 19:07
Sample: 1 80
Included observations: 80

Variable	Coefficient	95% CI Low	95% CI High
C	−9255.898	−12567.57	−5944.224
AVGSCH2005	2734.208	2323.782	3144.633

$$(2323.782, 3144.633)$$

We are 95% confident that the true population parameter β_1 lies between 2323.782 and 3144.633.