

Econometric Analysis – Dr. Sobel

Econometrics Session 2:

4. Perform Ordinary Least Squares (OLS) Regression Analysis in gretl

What it is, what it does, and why we do it: Regression analysis is basically fitting and estimating the trend line in the X-Y graphs. There are many different types of regressions, but the most basic and easiest is Ordinary Least Squares (or OLS). This method fits the line through the data that minimizes the sum of the squared differences between all points and the trend line. Doing a regression allows us to include many different variables, so we can examine the relationship between say, X and Y, but we can also “control” for other things that might impact the relationship or variables.

The OLS regression model is only valid when certain assumptions are met. One is that the “errors” between the points and the trend line are “random” or “normally distributed”, and another is that the dependent variable is a “continuous” variable (meaning it can take on a wide range of numbers including the possibility of non-whole numbers). One example of when this is violated is when the dependent variable is a count variable (takes the values zero, one, two, three, etc.). For these cases you cannot use OLS regressions. We will examine some of these in part 3 of the econometric section.

A regression has a “dependent” variable and also “independent” variables (sometimes called “endogenous” and “exogenous” variables, respectively). The dependent variable is the one being explained or predicted and the independent variables are the ones we are using to explain it. The regression allows us to see how a change in any one of the independent variables impacts the dependent variable, holding constant the other variables. For every regression you must specify which one is the dependent variable and also which one or many to include as independent variables.

How to run a basic OLS regression in gretl: Click the button in the main window that looks like the greek letter beta (β). This will open a window where you can specify the “dependent” variable and “independent” variables. Highlight the variable you want to be the dependent variable and click the blue arrow, then highlight anything you want to use as independent (explanatory) variables and hit the green arrow to include them in the bottom list.

Let’s try to explain/predict a student’s college GPA (colgpa) using their high school GPA (hsgpa) using an OLS regression

RUNNING A BASIC OLS REGRESSION IN GRETL EXAMPLE (explaining college GPA with high school GPA):

The screenshot shows three windows from the gretl software. The main window on the left lists variables: const, auto-generated constant, colgpa (GPA at college), hsgpa (High school GPA), vsat (Verbal SAT score), msat (Math SAT score), dsci (1 for a science major), dsoc (1 for a social science major), dhum (1 for a humanities major), darts (1 for an arts major), dcam (1 if student lives on campus), and dpub (1 if student attended public high school). A red circle highlights the β button at the bottom. The middle window, 'Specify Model', shows 'colgpa' in the 'Dependent variable' field and 'const' and 'hsgpa' in the 'Independent variables' field. A red box highlights the 'const' variable in the independent list with the text: 'This is the "constant" and it should automatically be there (always include it)'. The right window, 'Model 1', shows the regression results:

	coefficient	std. error	t-ratio	p-value
const	0.920577	0.204631	4.499	8.83e-06 ***
hsgpa	0.524173	0.0571206	9.177	1.95e-18 ***

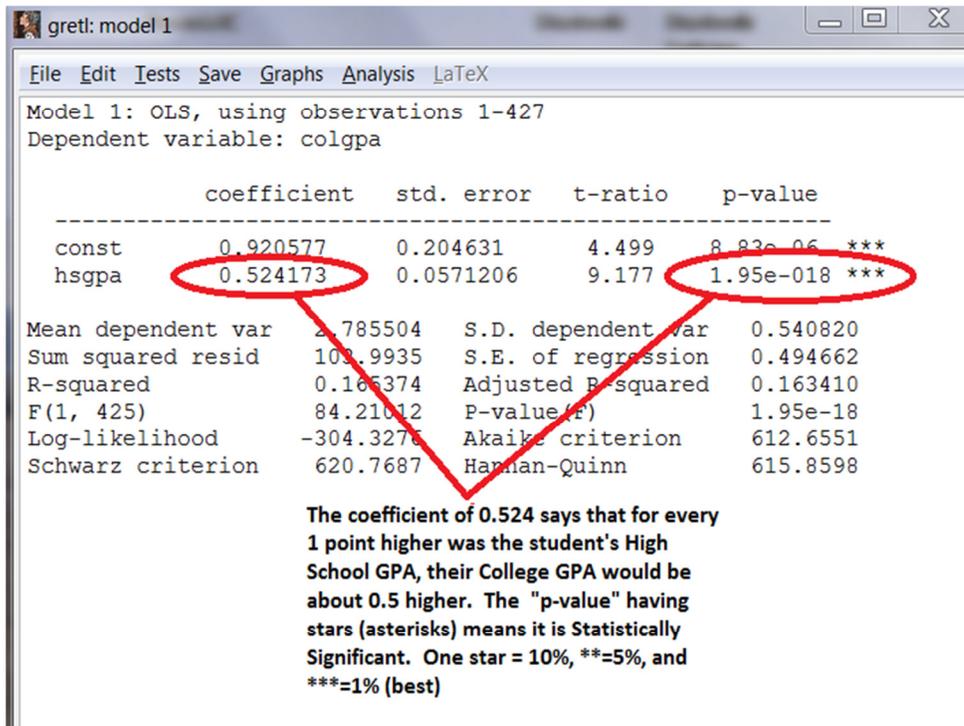
Other statistics shown include: Mean dependent var: 2.785504, S.D. of dependent var: 0.540820, Sum squared resid: 103.9935, S.E. of regression: 0.494662, R-squared: 0.165374, Adjusted R-squared: 0.163410, F(1, 425): 84.21012, P-value(F): 1.95e-18, Log-likelihood: -304.3276, Akaike criterion: 612.6551, Schwarz criterion: 620.7687, Hannan-Quinn: 615.8598.

****NOTE:** You should ALWAYS include a “constant” as an independent variable. This is basically allowing the trend line to have a y-intercept. If you do not include it, it will force the trend line to go through the origin (0,0 point), which can mess up your regression. gretl will normally automatically include a constant for you in the list (but make sure it is indeed there as “const” in the independent variable list). But we never care if the constant is significant and never interpret it.

How to interpret the OLS regression results (“coefficient estimates”, “statistical significance”, & “goodness of fit”)

The “coefficient” estimate is the impact of a one unit change in that independent variable on the dependent variable. It is the slope of the line in the X-Y plot. Sometimes we only care if the coefficient estimate is positive or negative (does that variable positively or negatively impact the other variable). But we ALWAYS care if it is “statistically significant” (using p-value and the asterisks beside it). In research papers we report the t-ratio or standard (std.) error that gives us the p-value, and put the asterisks/stars beside the coefficient estimate. You can round it to fewer decimal places.

COEFFICIENT ESTIMATES FROM A REGRESSION, HOW TO INTERPRET THEM:



Model 1: OLS, using observations 1-427
Dependent variable: colgpa

	coefficient	std. error	t-ratio	p-value
const	0.920577	0.204631	4.499	8.83e-06 ***
hsgpa	0.524173	0.0571206	9.177	1.95e-018 ***

Mean dependent var	2.785504	S.D. dependent var	0.540820
Sum squared resid	103.9935	S.E. of regression	0.494662
R-squared	0.166374	Adjusted R-squared	0.163410
F(1, 425)	84.21012	P-value(F)	1.95e-18
Log-likelihood	-304.3276	Akaike criterion	612.6551
Schwarz criterion	620.7687	Hannan-Quinn	615.8598

The coefficient of 0.524 says that for every 1 point higher was the student's High School GPA, their College GPA would be about 0.5 higher. The "p-value" having stars (asterisks) means it is Statistically Significant. One star = 10%, **=5%, and *=1% (best)**

If one wanted to “predict” a student’s college GPA from their high school GPA one would use the equation:

$$\text{Predicted college GPA} = 0.920577 + 0.524173 * \text{High School GPA}$$

where the first number comes from the coefficient estimate for the “constant” (const)

Statistical significance is very important. This is shown by the stars or asterisks to the far right for each variable. Three stars is the “best”, two is “good”, and one is just “okay”. If there are no stars, the variable is said to be insignificant, and so the coefficient might as well be zero (meaning you could exclude it from the regression, it’s not an important predictor or doesn’t explain the other variable very well).

These stars/asterisks are based on the number in the column titled “p-value”. This is the probability value for the statistical test. A p-value of 1% (0.01) or lower gives three stars, a p-value greater than 1% (0.01) but less than 5% (0.05) gives two stars, and a p-value greater than 5% (0.05) but less than 10% (0.10) gives one star. If the p-value is higher than 10% (0.10) the variable is said to be “insignificant”. In writing up your results, for example, for a variable with two stars we usually would say that “the variable is significant at the 5% level”. Note the example has one using scientific notation, the “e-018” means there the true decimal is 18 places to the left (so the p-value is like 0.00000000000000000195)

Optional: The p-value is based on a “t-test”. It is calculated by dividing the coefficient estimate by the standard error. This gives the “t-ratio” reported in the results. If the t-ratio is larger (in absolute value) than roughly 1.645 the variable will be significant. T-ratios of 2 or more give statistical significance at the 5% level. This is called “Hypothesis testing” and it is important to specify your hypothesis and for some tests to know what the “null” is. Sometimes we use “Standard errors” (std. error) to construct a “confidence interval” for the estimate using the standard error times two (the impact is 0.524173 plus or minus 2*0.0571206, that is 0.52 ± 0.114241 , or that the estimated impact of high school GPA on college GPA probably is in the range 0.409932 to 0.638414, with our best estimate being 0.524173).

How good the model is (goodness of fit) is measured by the R-squared or Adjusted R-squared (adjusted is better)

- R-squared ranges from zero to one (one would be a perfect fit)
- interpreted as “the percent of the variation explained”
- gretl will also save and report the “Log-likelihood” or “lnL” but you don’t have to show this

R-SQUARED FROM A REGRESSION, HOW GOOD THE REGRESSION FITS THE DATA:

Model 1: OLS, using observations 1-427
Dependent variable: colgpa

	coefficient	std. error	t-ratio	p-value	
const	0.920577	0.204631	4.499	8.83e-06	***
hsgpa	0.524173	0.0571206	9.177	1.95e-018	***

Mean dependent var	2.785504	S.D. dependent var	0.540820
Sum squared resid	109.8935	S.E. of regression	0.506846
R-squared	0.165374	Adjusted R-squared	0.163410
F(1, 425)	81.21012	P-value (F)	1.95e-18
Log-likelihood	-314.3276	Akaike criterion	612.6551
Schwarz criterion	620.7687	Hannan-Quinn	615.8598

The "R-squared" tells us how good the model is at fitting the data. This shows that a student's high school GPA alone predicts about 16 percent of the variation in their college GPA. Normally (but not always) people report the "Adjusted R-squared" (which adjusts for the number of variables in the regression). You can report both if you like. Just make sure to tell which you report.

So, the data also includes the student’s math and verbal SAT scores. So which is the best predictor of a student’s college GPA? is it their high school GPA? their verbal SAT score? their math SAT score? To find out let’s do a regression for each and compare the adjusted R-squareds:

Model 1: OLS, using observations 1-427
Dependent variable: colgpa

	coefficient	std. error	t-ratio	p-value	
const	1.62845	0.151348	10.76	4.85e-024	***
msat	0.00204309	0.000263713	7.747	6.96e-014	***

Mean dependent var	2.785504	S.D. dependent var	0.540820
Sum squared resid	109.1797	S.E. of regression	0.506846
R-squared	0.123752	Adjusted R-squared	0.121690
F(1, 425)	60.02224	P-value (F)	6.96e-14
Log-likelihood	-314.7177	Akaike criterion	633.4355
Schwarz criterion	641.5491	Hannan-Quinn	636.6402

Model 2: OLS, using observations 1-427
Dependent variable: colgpa

	coefficient	std. error	t-ratio	p-value	
const	1.99740	0.141279	14.14	1.81e-037	***
vsat	0.00157055	0.000277004	5.670	2.64e-08	***

Mean dependent var	2.785504	S.D. dependent var	0.540820
Sum squared resid	115.8373	S.E. of regression	0.522071
R-squared	0.070319	Adjusted R-squared	0.068132
F(1, 425)	32.14621	P-value (F)	2.64e-08
Log-likelihood	-327.3551	Akaike criterion	658.7102
Schwarz criterion	666.8238	Hannan-Quinn	661.9150

- A) Adjusted R-squared from a regression of College GPA on just High School GPA: 0.1634
- B) Adjusted R-squared from a regression of College GPA on just Math SAT Score: 0.1217
- C) Adjusted R-squared from a regression of College GPA on just Verbal SAT Score: 0.0681

Based on this, the verbal SAT score is the worst predictor, and the high school GPA is the best predictor. However, each is statistically significant at the best 1% level (three stars for each one), so all are important predictors of college GPA, but the high school GPA just explains it more closely.

Regressions with multiple variables: Now let's see if using all three at the same time helps. That is, can we do even better at predicting college GPA if we use information on high school GPA and information on verbal and math SAT scores. So run a regression including all three as independent variables at the same time:

A REGRESSION WITH MULTIPLE INDEPENDENT VARIABLES:

gretl: model 3

Model 3: OLS, using observations 1-427
Dependent variable: colgpa

	coefficient	std. error	t-ratio	p-value	
const	0.423249	0.219749	1.926	0.0548	*
hsgpa	0.398349	0.0605865	6.575	1.44e-010	***
msat	0.00101521	0.000293603	3.458	0.0006	***
vsat	0.000737453	0.000280682	2.627	0.0089	***

Mean dependent var	2.785504	S.D. dependent var	0.540820
Sum squared resid	97.16385	S.E. of regression	0.479272
R-squared	0.220187	Adjusted R-squared	0.214657
F(3, 423)	39.81265	P-value (F)	1.11e-22
Log-likelihood	-289.8245	Akaike criterion	587.6490
Schwarz criterion	603.8761	Hannan-Quinn	594.0584

All three are still significant. The adjusted R-squared now rises to 0.214657, meaning we are doing an even better job of predicting/explaining college GPA (we now explain 21% of the variance in it using these three variables). If we did want to do an equation to predict a student's college GPA it would now be: COLGPA = 0.42 + .0398 HSGPA + 0.0010*msat + 0.0007*vsat. Note that because the coefficient on math SAT is bigger, this means a one unit higher math SAT score has a greater impact on college performance than a one unit higher verbal SAT score.

Note that the data also includes information about the student's major, whether they live on campus, and whether they went to a public (versus private) high school. Perhaps some majors are easier and some harder, or that living on campus matters, and we should control for this to do an even better job.

Note that these new variables are "dummy" or "indicator" variables that only take the value of zero/one. This is fine for an independent variable (if it was the dependent variable we would need to use something different from OLS, we will get to this in part 3). The coefficient on a dummy variable is easy to interpret, as it's the effect of going from zero to one, or rather the effect of having that characteristic versus not having it.

A REGRESSION WITH MULTIPLE INDEPENDENT VARIABLES (including "dummy" or "indicator" variables):

gretl: model 4

Model 4: OLS, using observations 1-427
Dependent variable: colgpa

	coefficient	std. error	t-ratio	p-value	
const	0.367296	0.224302	1.638	0.1023	
hsgpa	0.405914	0.0634178	6.401	4.17e-010	***
msat	0.00108585	0.000302752	3.587	0.0004	***
vsat	0.000725891	0.000289903	2.504	0.0127	**
dsci	-0.0273225	0.0573192	-0.4767	0.6338	
dsoc	0.0561481	0.0727785	0.7715	0.4409	
dhum	-0.00405912	0.141771	-0.02863	0.9772	
darts	0.228650	0.188921	1.210	0.2269	
dcam	-0.0407050	0.0521617	-0.7804	0.4356	
dpub	0.0294027	0.0630396	0.4664	0.6412	

Mean dependent var	2.785504	S.D. dependent var	0.540820
Sum squared resid	96.20445	S.E. of regression	0.480319
R-squared	0.227887	Adjusted R-squared	0.211223

None of these indicator/dummy variables have stars by them. None of them are significant. So apparently living on campus, public vs. private high school, nor major really matters once you are already accounting for the student's High School GPA, and math and verbal SAT scores. Note that the adjusted R-squared is lower than the model with only those three variables.

Keeping/saving your results – What to report or show in a paper, how to copy it into a word file

FROM A WINDOW OF REGRESSION RESULTS:

- File menu, then “Save to session as icon” or you can “print” or save as a word document
- Save to session as icon is best, as you can then make a table of many results easily
it will save them as icons named “model 1”, “model 2”, etc.

SAVING YOUR REGRESSION RESULTS:

To save your results for later you can save them as an icon. Then these results will be an icon in the icon view

You can also copy and paste the results right into a document file using “Edit” menu, then “copy” and then paste

Generally, we run many different regressions and show them in a table all at the same time. In gretl this is very easy as long as you save each set of results as an icon. Once several (up to six maximum) models are saved as icons, open the icon view. Then you can either drag & drop the models onto the icon called “Model table” or you can right click the model and choose “Add to model table”. The order in which they appear in the table depends on the order in which you put them in the model table. Note that gretl’s model table will show everything you need to put in your research paper and more (we usually don’t need/show “lnL” but you can leave it in, and the full name for it is “Log-likelihood” if you want to put that in your table in the paper). Once created, click on the “Model table” icon to view the table of results.

CREATING A TABLE OF RESULTS FROM SEVERAL REGRESSIONS:

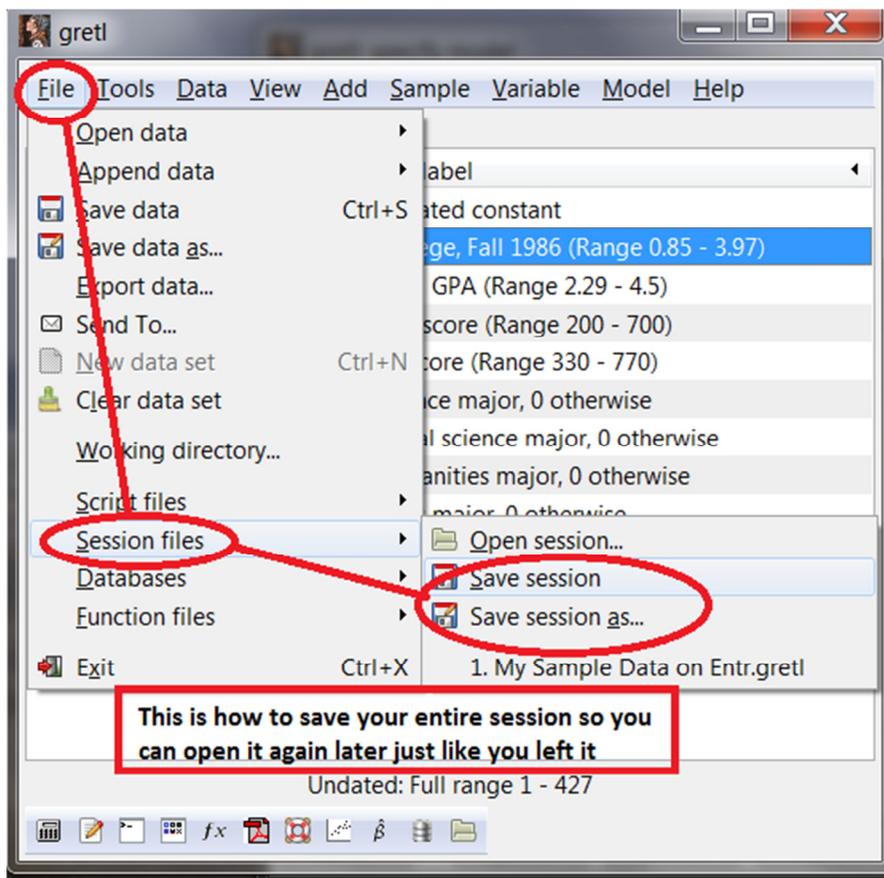
To create a table of results from several regressions using the same dependent variable, after saving each model result as an icon, then drag the icon for each model (in the order you want) onto the Model Table icon. Or you can right click the model's icon and select "add to model table". This table can be saved, or even copied and pasted into Microsoft Word or whatever software you are writing your paper using.

	(1)	(2)	(3)	(4)	(5)
const	0.9206** (0.2046)	1.628** (0.1513)	1.997** (0.1413)	1.402** (0.1696)	0.4232* (0.2197)
hsgpa	0.5242** (0.05712)				0.3983** (0.06059)
msat		0.002043** (0.0002637)		0.001695** (0.0002881)	0.001015** (0.0002936)
vsat			0.001571** (0.0002770)	0.0008444** (0.0002938)	0.0007375** (0.0002807)
Adj. R**2	0.1634	0.1217	0.0681	0.1364	0.2147
lnL	-304.3	-314.7	-327.4	-310.6	-289.8

Standard errors in parentheses
* indicates significance at the 10 percent level
** indicates significance at the 5 percent level

How to save the session to reopen later easily including any results or graphs you have done
“File” menu, “Session files”, “Save Session as”

SAVING YOUR ENTIRE SESSION SO YOU CAN REOPEN IT LATER WITH EVERYTHING THERE:



Frequently encountered basic problems in regressions:

One variable turns insignificant when you include additional variables – this can happen because the new variable simply is a better predictor (so given we have the new variable, the old one just doesn't matter in explaining the data), or it can happen because the two variables are measuring the same thing, a problem known as “multicollinearity”.

Multicollinearity – don't include two independent variables that are too closely correlated (meaning they measure roughly the same thing). If two variables have a high correlation coefficient (near 1, or let's say at least above 0.7 or 0.8), then don't include both variables. If you do, one or both will be insignificant, even if they matter. Check your correlation coefficients!

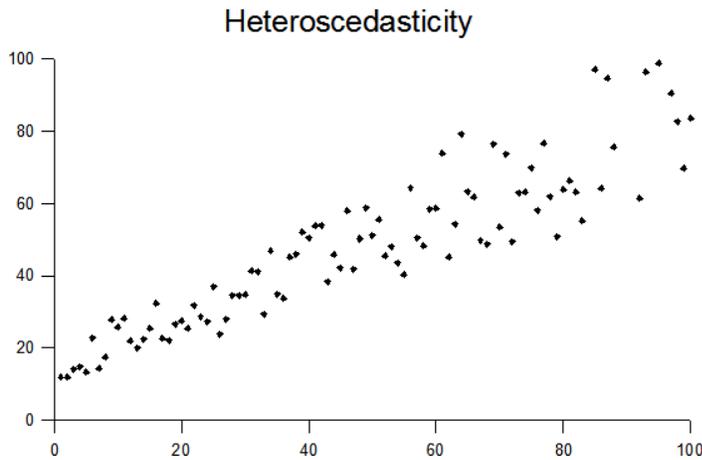
Outliers – one data point way off the line from the rest can really change your results. So always look at the graph of the data to see if there is a bad data point, then exclude that observation from the regression.

Selecting the sample / Working with subsamples / splitting the data – sometimes we may want to see if things are different among subsets of the data (men vs. women) or to exclude some observations that are outliers. To select to run the regression on only part (a subset) of your data: use the “Sample” menu from the top, “restrict based on criterion”, then use something like: $income > 50000$ or $year = 1995$ using one of your variables. To return to using the full sample, go to “Sample” menu, “restore full range”.

Note: After running a regression, in the results table there is a “tests” menu that has some tests for frequent problems including “influential observations” (outliers) and heteroskedasticity:

Correcting for “heteroskedasticity” - one of the assumptions of the basic OLS model is that the errors of the points around the line are uniform or random. Problems occur when the scatter of points looks like a cone:

HETEROSCEDASTICITY IN THE DATA:



Testing for this from the model results window: “Tests”, “Heteroskedasticity”, “White's test for heteroskedasticity”

White's test for heteroskedasticity
 OLS, using observations 1-427
 Dependent variable: uhat^2
 Omitted due to exact collinearity: sq_pub_hsgpa

	coefficient	std. error	t-ratio	p-value
const	-0.549304	0.945395		
hsgpa	0.459160	0.574973		
pub_hsgpa	0.0322970	0.106478		
sq_hsgpa	-0.0682941	0.0873609		
x2_x3	-0.00552510	0.0297412		

Unadjusted R-squared = 0.005434
 Test statistic: $TR^2 = 2.320311$,
 with p-value = $P(\text{Chi-square}(4) > 2.320311) = 0.677074$

because this is not significant (not less than 0.10, we do NOT have a problem with heteroskedasticity)

If it is significant then you need to correct for it (e.g., a p-value smaller than 0.10) using the check box for “Robust standard errors” when you pick the variables for the regression:

Model 1: OLS, using observations 1-427
 Dependent variable: colgpa
 Heteroskedasticity-robust standard errors, variant HC1

	coefficient	std. error	t-ratio	p-value
const	0.920577	0.195798	4.702	3.49e-06 ***
hsgpa	0.524173	0.0544803	9.621	5.90e-020 ***

Mean dependent var 2.785504 S.D. dependent var 0.540820
 Sum squared resid 103.9935 S.E. of regression 0.494662
 R-squared 0.165374 Adjusted R-squared 0.163410
 F(1, 425) 92.56995 P-value (F) 5.90e-20
 Log-likelihood -304.3276 Akaike criterion 612.6551
 Schwarz criterion 620.7687 Hannan-Quinn 615.8598

****YOU SHOULD ALWAYS CHECK YOUR REGRESSION FOR THIS****

****SOME PEOPLE JUST ALWAYS CHECK THIS BOX****

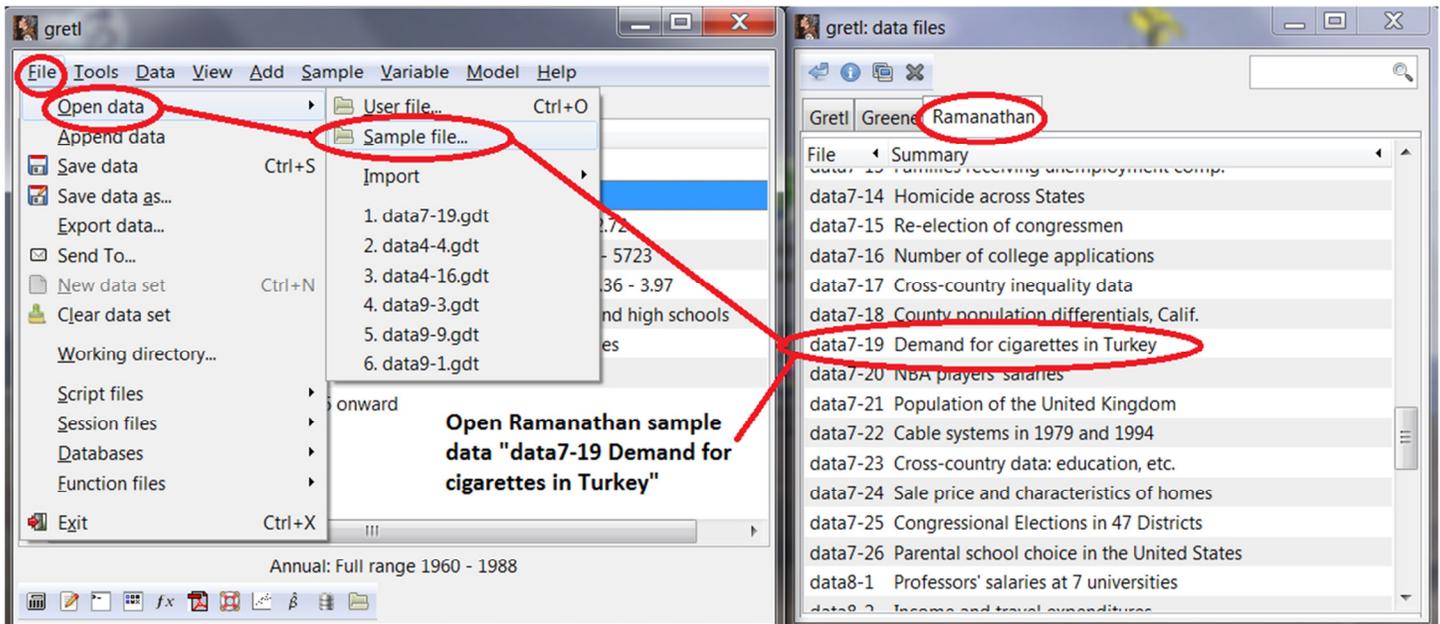
****IF YOU DO THIS, PLEASE NOTE IN YOUR RESULTS TABLE: “ROBUST STANDARD ERRORS”****

Econometric Analysis – Dr. Sobel

Econometrics Application to the Law of Demand:

1. Read in the Sample Data Set Ramanathan data7-19 “Demand for cigarettes in Turkey”

OPENING RAMANATHAN SAMPLE DATA 7-19:



2. Here is the data that will be in your data set:

The screenshot shows the gretl software interface displaying the data set 'data7-19.gdt'. The data set is shown as a table with the following variables and descriptive labels:

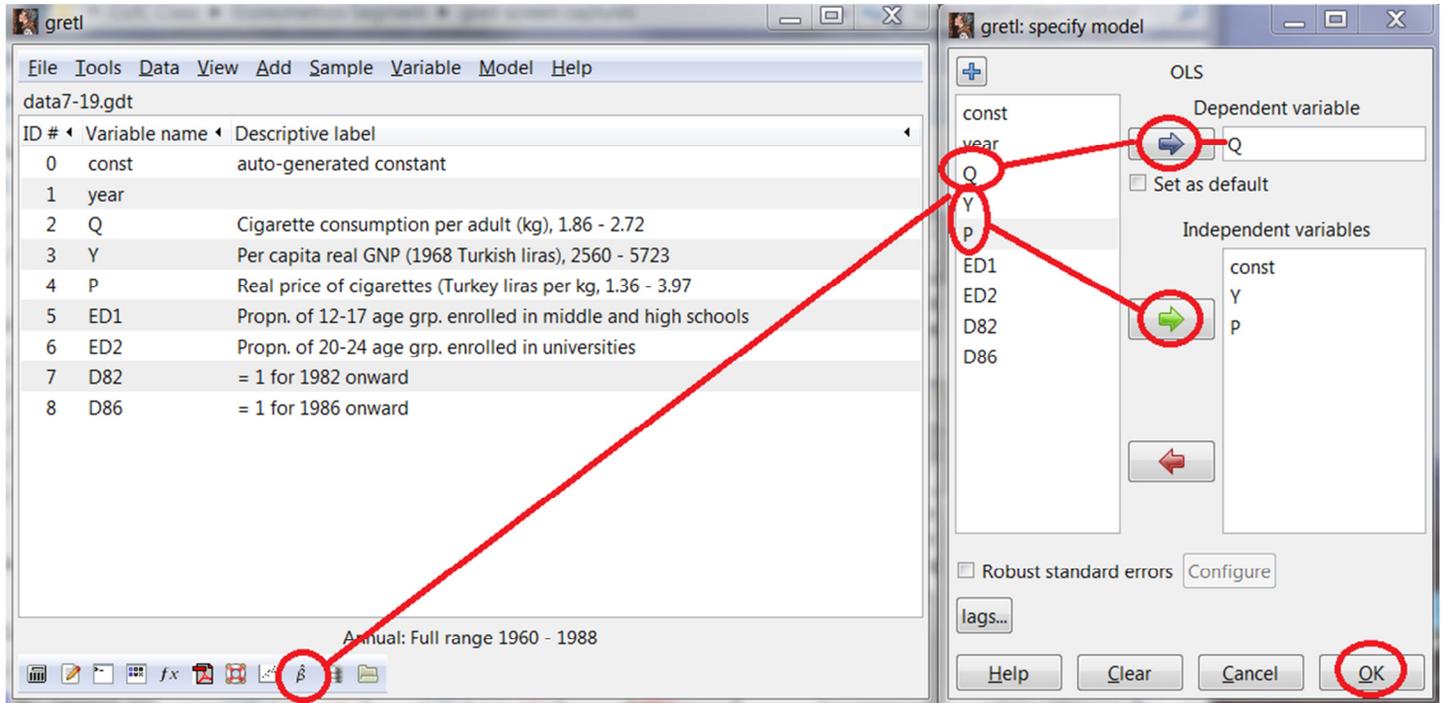
ID #	Variable name	Descriptive label
0	const	auto-generated constant
1	year	
2	Q	Cigarette consumption per adult (kg), 1.86 - 2.72
3	Y	Per capita real GNP (1968 Turkish liras), 2560 - 5723
4	P	Real price of cigarettes (Turkey liras per kg, 1.36 - 3.97
5	ED1	Propn. of 12-17 age grp. enrolled in middle and high schools
6	ED2	Propn. of 20-24 age grp. enrolled in universities
7	D82	= 1 for 1982 onward
8	D86	= 1 for 1986 onward

Annual: Full range 1960 - 1988

We have everything we need to estimate a demand curve: Q is the quantity demanded, Y is income, and P is price.

Based on economic theory, our hypothesis should be that the coefficient on price is NEGATIVE and significant (the law of demand), and the coefficient on income should be POSITIVE and significant (if cigarettes are a normal good).

3. Let's estimate the demand curve, select Q as the dependent variable, and as independent variables have the constant (const), income (Y), and price (P):



gretl: model 1

Model 1: OLS, using observations 1960-1988 (T = 29)
Dependent variable: Q

	coefficient	std. error	t-ratio	p-value
const	1.65654	0.123678	13.39	3.53e-013
Y	0.000344100	5.27935e-05	6.518	6.56e-07 ***
P	-0.423295	0.0969440	-4.366	0.0002 ***

Mean dependent var	2.204655	S.D. dependent var	0.243190
Sum squared resid	0.595167	S.E. of regression	0.151298
R-squared	0.640589	Adjusted R-squared	0.612942
F(2, 26)	23.17031	P-value (F)	1.67e-06
Log-likelihood	15.20081	Akaike criterion	-24.40161
Schwarz criterion	-20.29973	Hannan-Quinn	-23.11695
rho	0.536727	Durbin-Watson	0.911596

Our hypotheses are confirmed, income (Y) has a positive and significant coefficient, and price (P) has a negative and significant coefficient.

If we wished to interpret the coefficients, they would say that for every 10,000 Turkish Liras of higher income, cigarette consumption is 3.4 kg higher, and that for every 1 Turkish Lira per kg the price increases, cigarette consumption falls by 0.42 kg.

However, we need to check to be sure our model doesn't suffer from any common problems, most importantly heteroskedasticity. So run the White's test:

gretl: model 1

Model 1: OLS, using observations 1960-1988 (T = 29)
Dependent variable: uhat^2

	d. error	t-ratio	p-value
const	123678	13.39	3.53e-013 ***
Y	27935e-05	6.518	6.56e-07 ***
P	0969440	-4.366	0.0002 ***

White's test for heteroskedasticity
OLS, using observations 1960-1988 (T = 29)
Dependent variable: uhat^2

	coefficient	std. error	t-ratio	p-value
const	0.224376	0.134559	1.667	0.1090
Y	-0.000288480	0.000118342	-2.438	0.0229 **
P	0.297077	0.210234	1.413	0.1710
sq_Y	6.36452e-08	3.22655e-08	1.973	0.0607 *
X2_X3	-0.000101817	8.49743e-05	-1.198	0.2430
sq_P	0.0310950	0.0499291	0.6228	0.5396

Warning: data matrix close to singularity!
Unadjusted R-squared = 0.475305
Test statistic: $TR^2 = 13.783847$, with p-value = $P(\text{Chi-square}(5) > 13.783847) = 0.017042$

Yes, we do have a problem with heteroskedasticity

The test p-value is smaller than 0.10 (10%), so yes we have a problem with heteroskedasticity. We need to re-run the regression using "Robust Standard Errors":

gretl: specify model

OLS

Dependent variable: Q

Independent variables: const, Y, P

Robust standard errors

gretl: model 2

Model 2: OLS, using observations 1960-1988 (T = 29)
Dependent variable: Q
HAC standard errors, bandwidth 2 (Bartlett kernel)

	coefficient	std. error	t-ratio	p-value
const	1.65654	0.126315	13.11	5.71e-013 ***
Y	0.000344100	7.19664e-05	4.781	5.98e-05 ***
P	-0.423295	0.0970912	-4.360	0.0002 ***

Mean dependent var 2.204655 S.D. dependent var 0.243190
Sum squared resid 0.595167 S.E. of regression 0.151298
R-squared 0.640589 Adjusted R-squared 0.612942
F(2, 26) 11.48910 P-value(F) 0.000266
Log-likelihood 15.20081 Akaike criterion -24.40161
Schwarz criterion -20.29973 Hannan-Quinn -23.11695
rho 0.536727 Durbin-Watson 0.911596

Even after adjusting for heteroskedasticity, our hypotheses are both confirmed again, but these are better estimates. So now, let's save this model as an icon:

gretl: model 2

File Edit Tests Save Graphs Analysis LaTeX

Save as...
Save to session as icon
Save as icon and close

Print...
View as equation

Close Ctrl+W

	std. error	t-ratio	p-value
const	0.126315	13.11	5.71e-013 ***
Y	7.19664e-05	4.781	5.98e-05 ***
P	0.0970912	-4.360	0.0002 ***

Mean dependent var 2.204655 S.D. dependent var 0.243190
Sum squared resid 0.595167 S.E. of regression 0.151298
R-squared 0.640589 Adjusted R-squared 0.612942
F(2, 26) 11.48910 P-value(F) 0.000266
Log-likelihood 15.20081 Akaike criterion -24.40161
Schwarz criterion -20.29973 Hannan-Quinn -23.11695
rho 0.536727 Durbin-Watson 0.911596

Now, let's try including two of the other variables in the data set, measuring educational enrollment, ED1 and ED2 which are the proportions of the 12-17 age group enrolled in middle and high schools, and the proportion of the 20-24 age group enrolled in universities.

gretl: specify model

OLS

Dependent variable: Q

Independent variables: Y, P, ED1, ED2

gretl: model 3

Model 3: OLS, using observations 1960-1988 (T = 29)
Dependent variable: Q

	coefficient	std. error	t-ratio	p-value
const	0.707979	0.454836	1.557	0.1327
Y	0.000957520	0.000291661	3.283	0.0031 ***
P	-0.313591	0.104349	-3.005	0.0061 ***
ED1	-5.86691	2.64112	-2.221	0.0360 **
ED2	-3.22065	3.56141	-0.9043	0.3748

Mean dependent var 2.204655 S.D. dependent var 0.243190
Sum squared resid 0.493419 S.E. of regression 0.143385
R-squared 0.702033 Adjusted R-squared 0.652372
F(4, 24) 14.13646 P-value(F) 4.62e-06
Log-likelihood 17.91932 Akaike criterion -25.83863
Schwarz criterion -19.00216 Hannan-Quinn -23.69753
rho 0.501147 Durbin-Watson 0.962300

Excluding the constant, p-value was highest for variable 6 (ED2)

Again our hypotheses are confirmed, income (Y) has a positive and significant coefficient, and price (P) has a negative and significant coefficient. The new variables on educational enrollment suggest that a higher proportions of young individuals enrolled in school does REDUCE the demand for cigarettes (they are both negative), although the second one for university enrollment is not significant.

But again, we need to check for problems with heteroskedasticity:

gretl: LM test (heteroskedasticity)

White's test for heteroskedasticity
OLS, using observations 1960-1988 (T = 29)
Dependent variable: uhat^2

	coefficient	std. error	t-ratio	p-value
const	2.20922	1.37947	1.602	0.1316
Y	-0.00254965	0.00165336	-1.542	0.1453
P	-1.03264	0.672014	-1.537	0.1467
ED1	21.6174	15.8213	1.366	0.1934
ED2	35.4459	22.9572	1.544	0.1449
sq_Y	6.64943e-07	4.86999e-07	1.365	0.1937
X2_X3	0.000761843	0.000422984	1.801	0.0933 *
X2_X4	-0.0105546	0.00887014	-1.190	0.2539
X2_X5	-0.0258571	0.0125935	-2.053	0.0592 *
sq_P	0.120376	0.0584879	2.058	0.0587 *
X3_X4	-8.20766	3.93072	-2.088	0.0555 *
X3_X5	-5.29263	4.23482	-1.250	0.2319
sq_ED1	42.4348	40.7316	1.042	0.3152
X4_X5	251.700	97.2115	2.589	0.0214 **
sq_ED2	106.559	88.1117	1.209	0.2466

Warning: data matrix close to singularity!

Unadjusted R-squared = 0.749925

Test statistic: $TR^2 = 21.747822$,
with p-value = $P(\text{Chi-square}(14) > 21.747822) = 0.083947$

This model also has a problem with heteroskedasticity

So, now re-run this extended model with “Robust Standard Errors”:

The OLS specification window shows the following settings:

- Dependent variable: Q
- Independent variables: const, Y, P, ED1, ED2
- Robust standard errors

The model output window displays the following results:

	coefficient	std. error	t-ratio	p-value	
const	0.707979	0.400761	1.767	0.0900	*
Y	0.000957520	0.000268534	3.566	0.0016	***
P	-0.313591	0.117486	-2.669	0.0134	**
ED1	-5.86691	2.76959	-2.118	0.0447	**
ED2	-3.22065	4.05229	-0.7948	0.4345	

Additional statistics from the output window:

- Mean dependent var: 2.204655
- Sum squared resid: 0.493419
- R-squared: 0.702033
- F(4, 24): 28.58676
- Log-likelihood: 17.91932
- Schwarz criterion: -19.00216
- rho: 0.501147
- S.D. dependent var: 0.243190
- S.E. of regression: 0.143385
- Adjusted R-squared: 0.652372
- P-value(F): 8.11e-09
- Akaike criterion: -25.83863
- Hannan-Quinn: -23.69753
- Durbin-Watson: 0.962300

Excluding the constant, p-value was highest for variable 6 (ED2)

And again, our hypotheses are confirmed, these are just better estimates. So now, let’s save this model as an icon as well.

Now, open the icon view by clicking on the button on the lower part of the main screen:

The main window displays the following variable list:

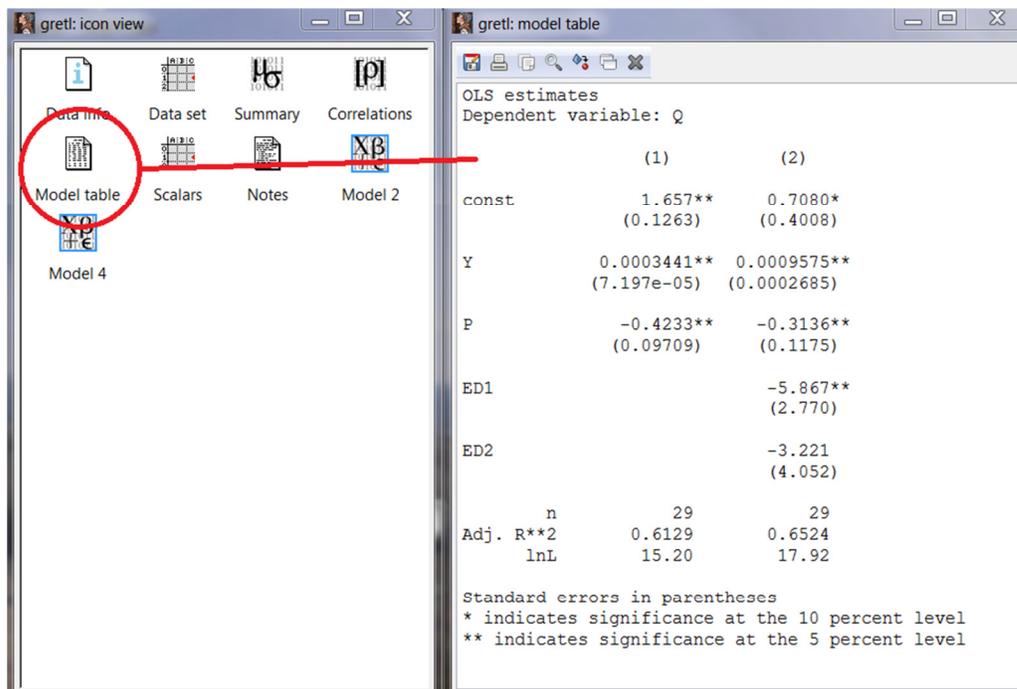
ID #	Variable name	Descriptive label
0	const	auto-generated constant
1	year	
2	Q	Cigarette consumption per adult (kg), 1.86 - 2.72
3	Y	Per capita real GNP (1968 Turkish liras), 2560 - 5723
4	P	Real price of cigarettes (Turkey liras per kg), 1.36 - 3.97
5	ED1	Propn. of 12-17 age grp. enrolled in middle and high schools
6	ED2	Propn. of 20-24 age grp. enrolled in universities
7	D82	= 1 for 1982 onward
8	D86	= 1 for 1986 onward

The icon view window shows the following icons:

- Data info
- Data set
- Summary
- Correlations
- Model table
- Scalars
- Notes
- Model 2
- Model 4

Now, let’s create a model table that we could put in a paper with these two sets of results. First drag and drop the Model 2 icon onto the Model Table icon. Then next, drag and drop the Model 4 icon onto the Model Table icon.

Now, double click the Model Table icon to open up the table of results we created:



These results are ready to be put into a paper, or you can print them. In the table in your paper, however, you do need to make a note that the standard errors that are reported in parenthesis are Robust Standard Errors. You should also be sure to better label the variables with longer descriptions. I saved it as RTF (Word) format and then opened it and pasted it below and added descriptions and updated the standard error note. I did add borders to the table.

OLS estimates
Dependent variable: Q (Cigarette Consumption Per Adult)

	(1)	(2)
Constant	1.657** (0.1263)	0.7080* (0.4008)
Y (Income)	0.0003441** (7.197e-05)	0.0009575** (0.0002685)
P (Price)	-0.4233** (0.09709)	-0.3136** (0.1175)
ED1 (School Enrollment Middle/High)	--	-5.867** (2.770)
ED2 (School Enrollment University)	--	-3.221 (4.052)
n (Observations)	29	29
Adj. R ²	0.6129	0.6524
lnL	15.2	17.92

Robust standard errors in parentheses
* indicates significance at the 10 percent level
** indicates significance at the 5 percent level