Econometric Analysis – Dr. Sobel

Econometrics Session 1:

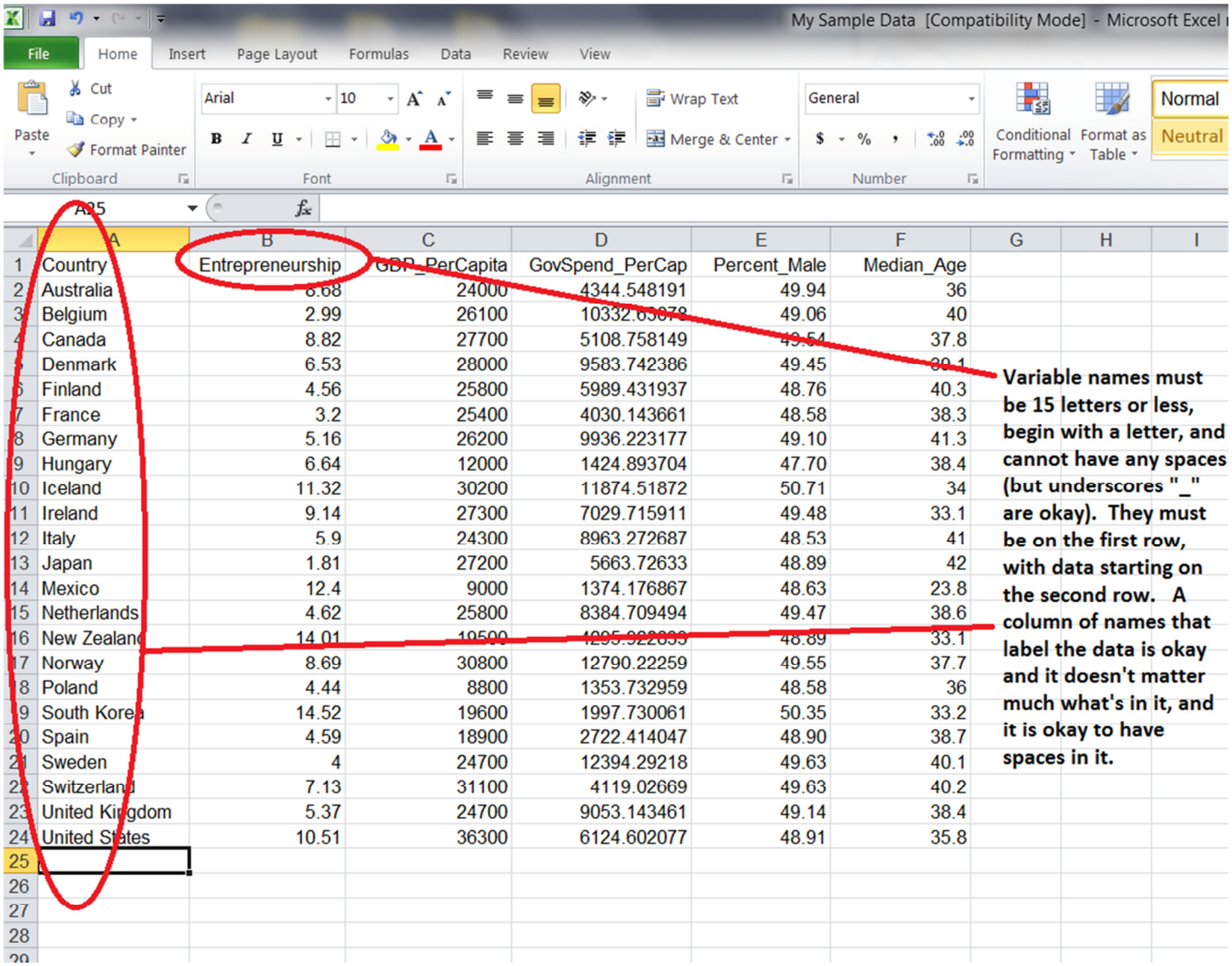
1. Building a data set

Which software - usually best to use Microsoft Excel (XLS format) but CSV is also okay

Variable names (first row only, 15 character max in gretl software we will be using, 1st must be letter, no spaces but _ symbol is ok)

First column is sometimes dates or labels (state names) and it is okay if this has spaces, but give it a title too

EXAMPLE DATA SET IN EXCEL:



	A	B	C	D	E	F	G	H	I
1	Country	Entrepreneurship	GDP_PerCapita	GovSpend_PerCap	Percent_Male	Median_Age			
2	Australia	8.68	24000	4344.548191	49.94	36			
3	Belgium	2.99	26100	10332.63678	49.06	40			
4	Canada	8.82	27700	5108.758149	49.54	37.8			
5	Denmark	6.53	28000	9583.742386	49.45	39.1			
6	Finland	4.56	25800	5989.431937	48.76	40.3			
7	France	3.2	25400	4030.143661	48.58	38.3			
8	Germany	5.16	26200	9936.223177	49.10	41.3			
9	Hungary	6.64	12000	1424.893704	47.70	38.4			
10	Iceland	11.32	30200	11874.51872	50.71	34			
11	Ireland	9.14	27300	7029.715911	49.48	33.1			
12	Italy	5.9	24300	8963.272687	48.53	41			
13	Japan	1.81	27200	5663.72633	48.89	42			
14	Mexico	12.4	9000	1374.176867	48.63	23.8			
15	Netherlands	4.62	25800	8384.709494	49.47	38.6			
16	New Zealand	14.01	19500	4295.922833	48.89	33.1			
17	Norway	8.69	30800	12790.22259	49.55	37.7			
18	Poland	4.44	8800	1353.732959	48.58	36			
19	South Korea	14.52	19600	1997.730061	50.35	33.2			
20	Spain	4.59	18900	2722.414047	48.90	38.7			
21	Sweden	4	24700	12394.29218	49.63	40.1			
22	Switzerland	7.13	31100	4119.02669	49.63	40.2			
23	United Kingdom	5.37	24700	9053.143461	49.14	38.4			
24	United States	10.51	36300	6124.602077	48.91	35.8			
25									
26									
27									
28									

Variable names must be 15 letters or less, begin with a letter, and cannot have any spaces (but underscores "_" are okay). They must be on the first row, with data starting on the second row. A column of names that label the data is okay and it doesn't matter much what's in it, and it is okay to have spaces in it.

The number of "observations" (data points) matters. More is better. Always report the number of observations you use in a research paper.

Missing observations – usually best to have none, but most programs can deal with them – leave cell blank

Types of datasets: cross section, time series, panel (usually must specify, and label)

– in gretl software we will be using, go to Data menu, Dataset structure to change or specify

Source information – how to document it
Descriptions: what is included, what year, how measured, etc.
Keep track of units of measurement (data in millions of dollars? Percentages: would 5% be as 0.05 or 5.0?)

EXAMPLE OF HOW TO DOCUMENT YOUR DATA IN YOUR RESEARCH PAPER:

Appendix 1: List of Variables and Sources

Table 5

Variable name	Minimum	Maximum	Average	Obs.	Source	Definition
Dependent variables						
Average business start-up rate	0.114	0.229	0.145	50	a	Average business births as a percentage of total businesses with 1–9 employees (2002–2003)
Net business creation rate	0.001	0.052	0.018	50	a	Average net business creation as a percentage of total businesses with 1–9 employees (2002–2003)
Venture capital per capita	0	379.39	46.635	48	b	Venture capital invested (by destination) per capita (2005)
Patents per capita	0.044	1.086	0.249	50	c	Number of patents granted (all types) per 1,000 people (2005)
Productive entrepreneurship	3.2	43	23.583	48	d	Index of entrepreneurial activity from Sobel (2008) based on several measures of entrepreneurship
Independent variables						
Cultural diversity	0.021	0.44	0.132	50	e	Probability that 2 randomly chosen individuals from a state were born in different countries, calculated like a standard Herfindahl–Hirshman Index (2000)
Median age	27.1	37.9	35.536	50	f	Median age of state population (2000)
Percent male	88.8	107.6	94.114	50	f	Percent of 18 years and over state population that is male (2000)
Percent college degree	14.8	33.2	23.776	50	f	Percent of state population with a bachelor's degree or higher (2000)
Population density	0.0011	1.1344	0.1819	50	f	Population density per square mile of land area in state (2000), in thousands
Median household income	29.696	55.146	41.371	50	e	Median household income (1999), in thousands
Economic freedom	5.6	7.8	6.684	50	f	Economic freedom of North America (EFNA) index (2005)

Data Sources: *a* Company Statistics Division of the United States Census Bureau (http://www.census.gov/csd/susb/susb.htm), *b* Venture Capital Association's (NVCA) Yearbook, *c* the US Patent and Trademark Office (USPTO), *d* Index constructed in Sobel (2008), *e* US Census 2000 (http://www.census.gov), *f* Fraser Institute (Karabegovic and McMahon 2005)

Transformations (getting units right, like per capita, logs, changes, percent changes)
- generally try to not have the units be to different (billions for one variable, hundreds for another)
- use per capita to adjust for differences due to like the size of states or changes thru time
    - or sometimes per something else (school spending per student)
- sometimes the effects are "percentage" based (e.g., a 10% weight loss) – use logs or percent changes
- sometimes when we use "time series" data we do the year to year changes as the variables

2. Reading the data into “econometric” software that can run “regressions”

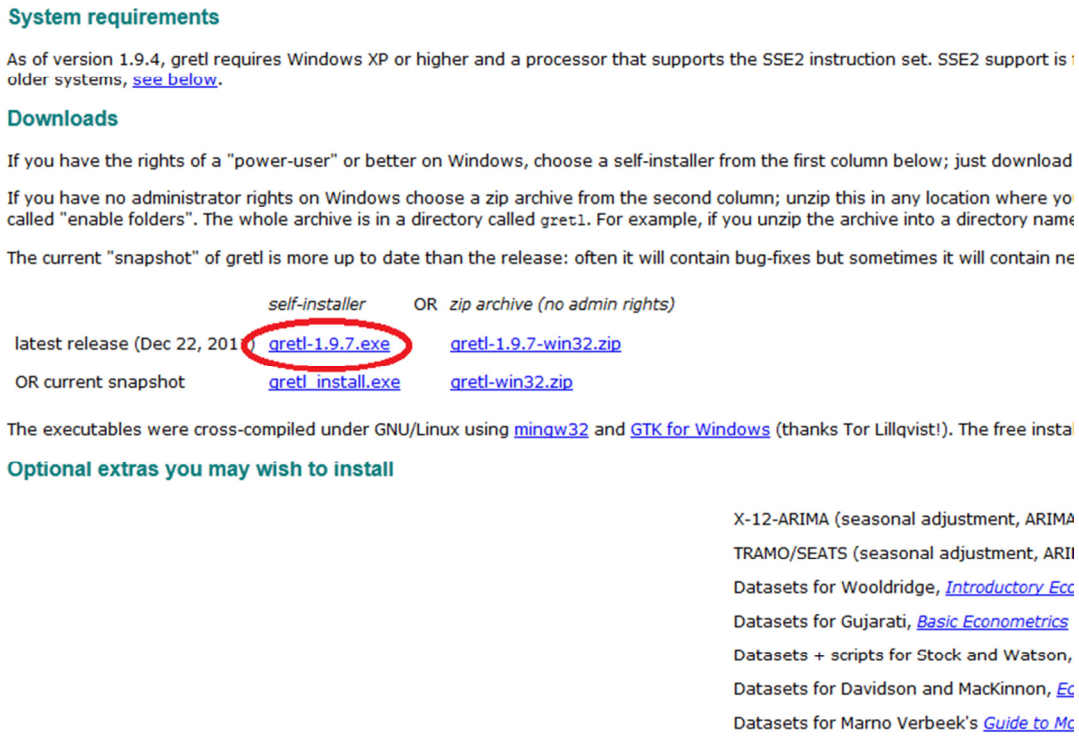
Everyone must obtain or use software called “gretl”, downloadable at http://gretl.sourceforge.net/

- will also be available on computers in Beaty 120 computer lab and Beatty Center Atrium
- if you know another program (like EViews) you are welcome to use it, but I can’t help much

DOWNLOAD SOFTWARE FROM GRETL WEBSITE EXAMPLE (homepage):

gretl is available for both Windows and Mac, and can even be installed on computers on which you do not have administrative rights. There are even older versions available for older operating systems.

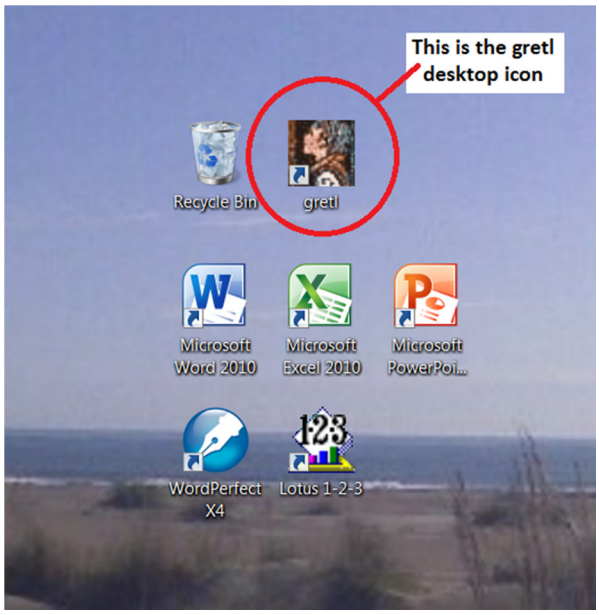
DOWNLOAD SOFTWARE FROM GRETL WEBSITE EXAMPLE (Windows download page):



choose to download “latest release” and use the one for “self-installer” if you do have administrative rights
do not worry about downloading any of the “optional extras”

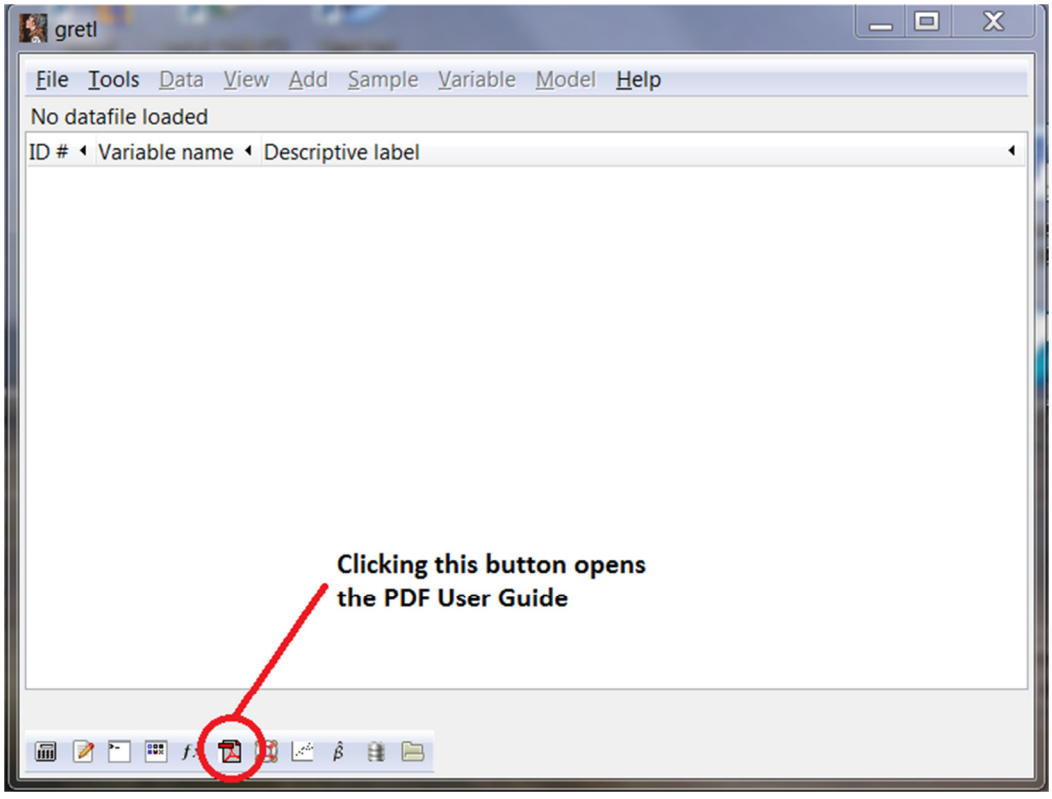
Once installed, open gretl by clicking on the gretl icon on your desktop:

GRETL DESKTOP ICON EXAMPLE:



When you open gretl you will get a main page that looks like this below:

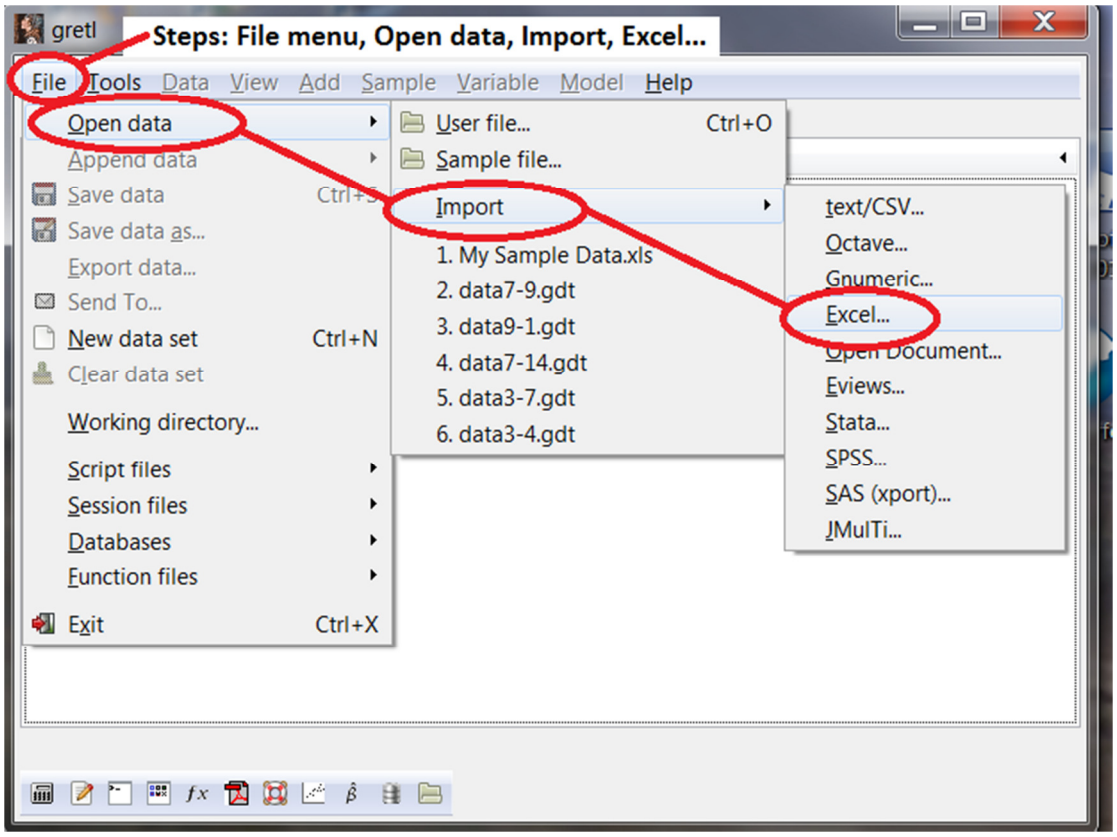
GRETL MAIN OPENINIG SCREEN EXAMPLE:



gretl has a built in user manual if you ever have questions or problems (PDF “users guide” button along bottom)

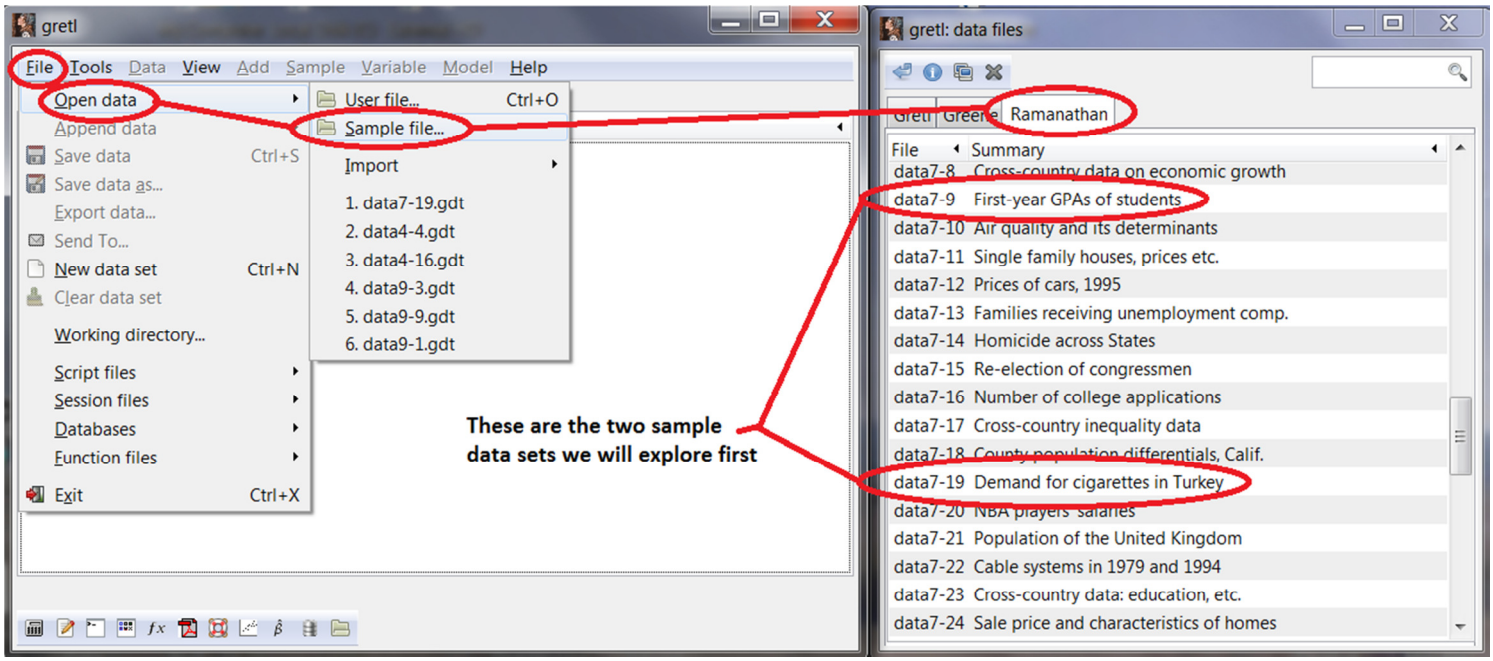
Reading data into gretl (File menu, Open Data, Import, then Excel if it's your data)

GRETL IMPORTING EXCEL DATA EXAMPLE:



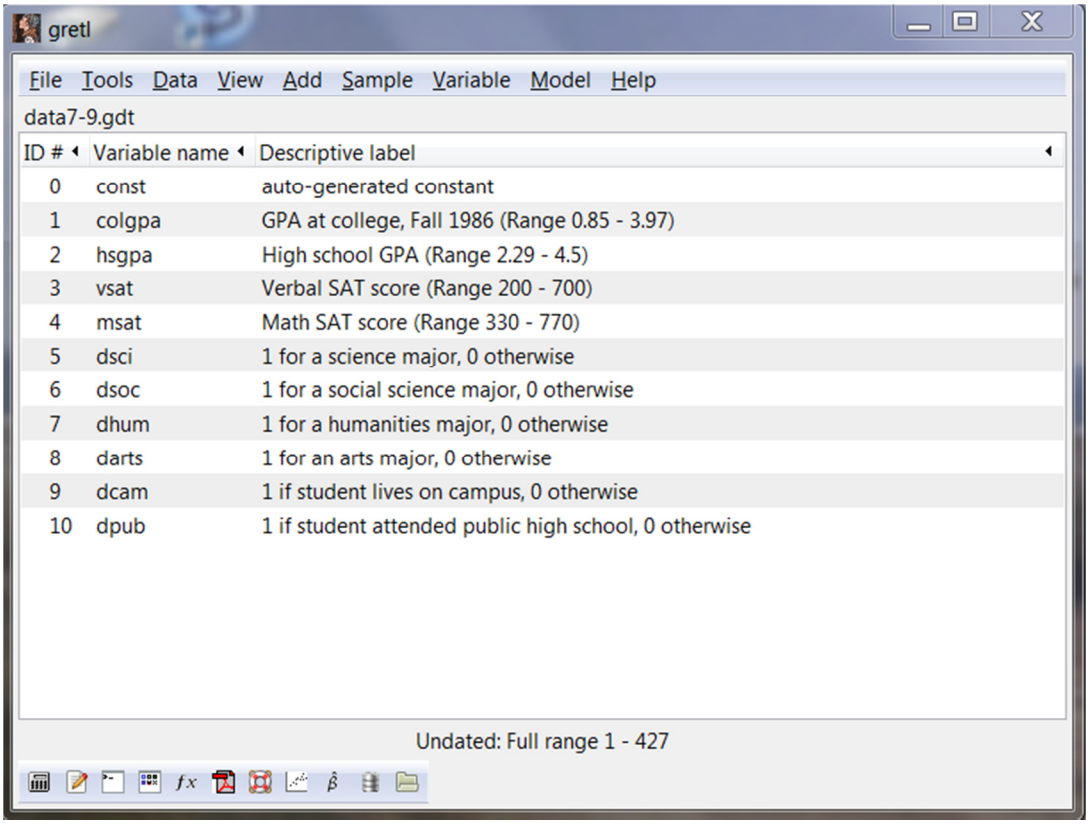
- For now, use sample dataset (File menu, Open Data, Sample file): Ramanathan data7-9 "First year GPAs of students", we may also use data7-19 "Demand for cigarettes in Turkey" but we will start with the GPA data

GRETL OPENING RAMANATHAN GPA SAMPLE DATA EXAMPLE:



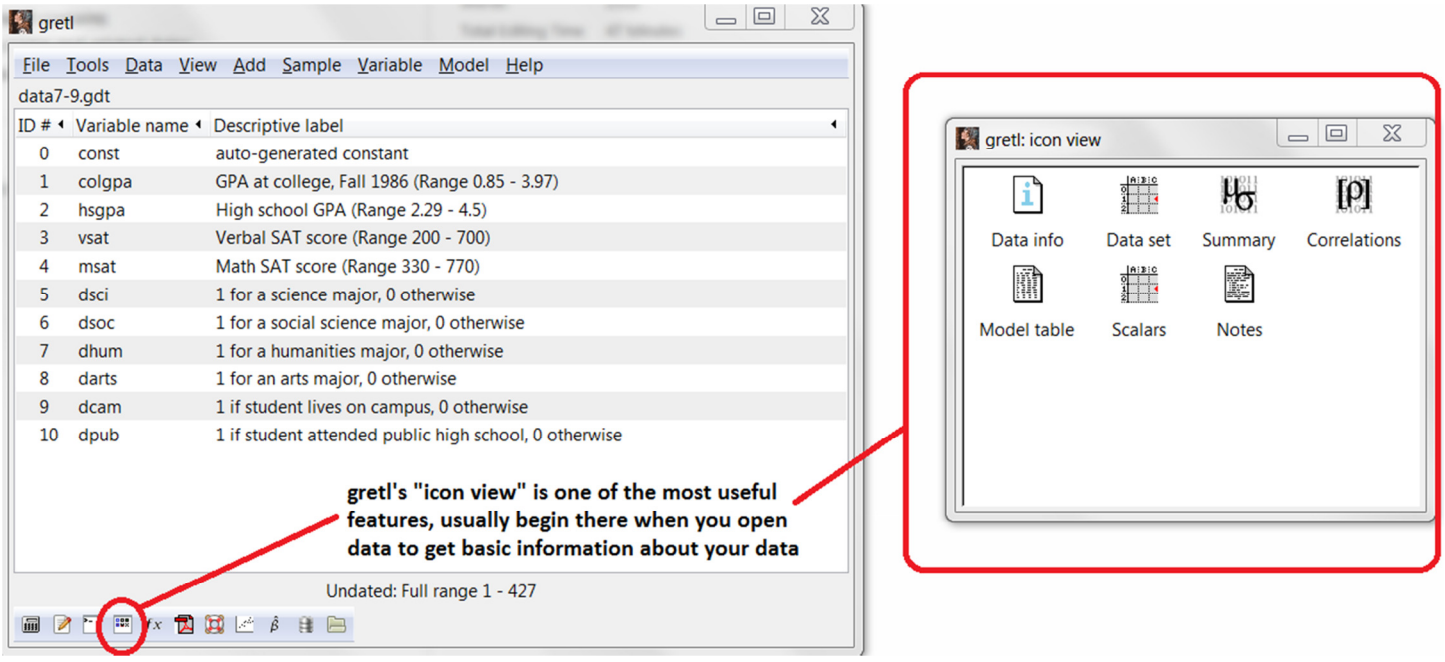
once you open the data, the main screen will look like this below:

GRETL MAIN SCREEN AFTER OPENING SAMPLE GPA DATA:



one of the most used features is the "icon view", so let's open it. From this window you can look at your data and get basic statistical information about your data:

OPENING GRETL'S ICON VIEW:



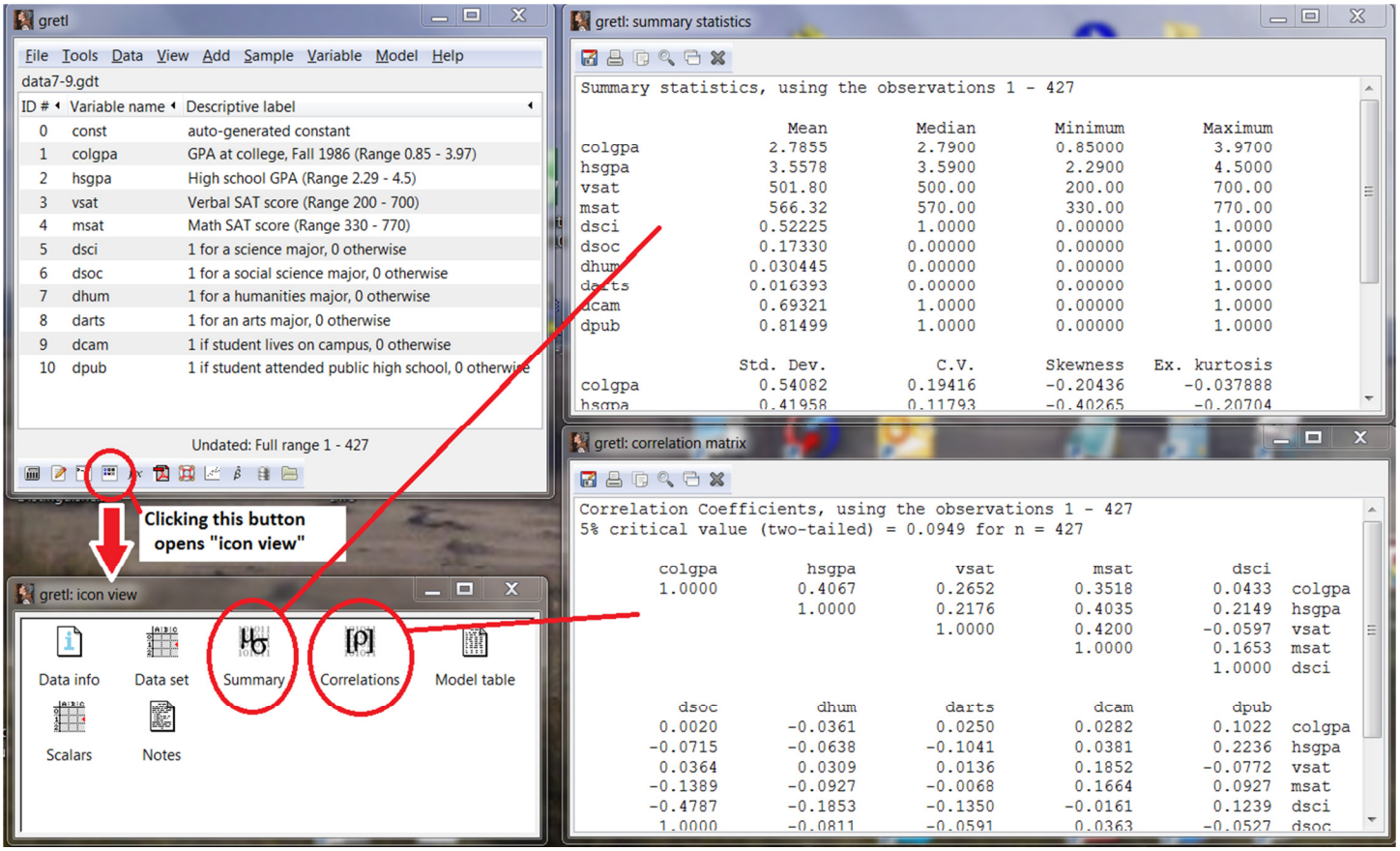
3. First steps: Examine basic descriptive statistics, and examine correlations, creating new variables

Note: these can be done in Excel too but are easier in gretl [Excel commands shown in brackets]

Examine "descriptive statistics" such as the mean (average), maximum, minimum of each variable

- in gretl open "session icon view" and click "Summary" (or can use "View" menu then "Summary Statistics")
- [in Excel for data in cells A1 to A20 the commands are: =AVERAGE(A1:A20); =MAX(A1:A20); =MIN (A1:A20)]

USING GRETL'S ICON VIEW TO GET SUMMARY STATISTICS AND CORRELATIONS:



Examine "correlation coefficients"

- in gretl either use "View" menu then "Correlation Matrix" or open "session icon view" and click "Correlations"
- [in Excel for data in cells A1 to A20 and B1 to B20 the command is: =CORREL(A1:A20,B1:B20)]

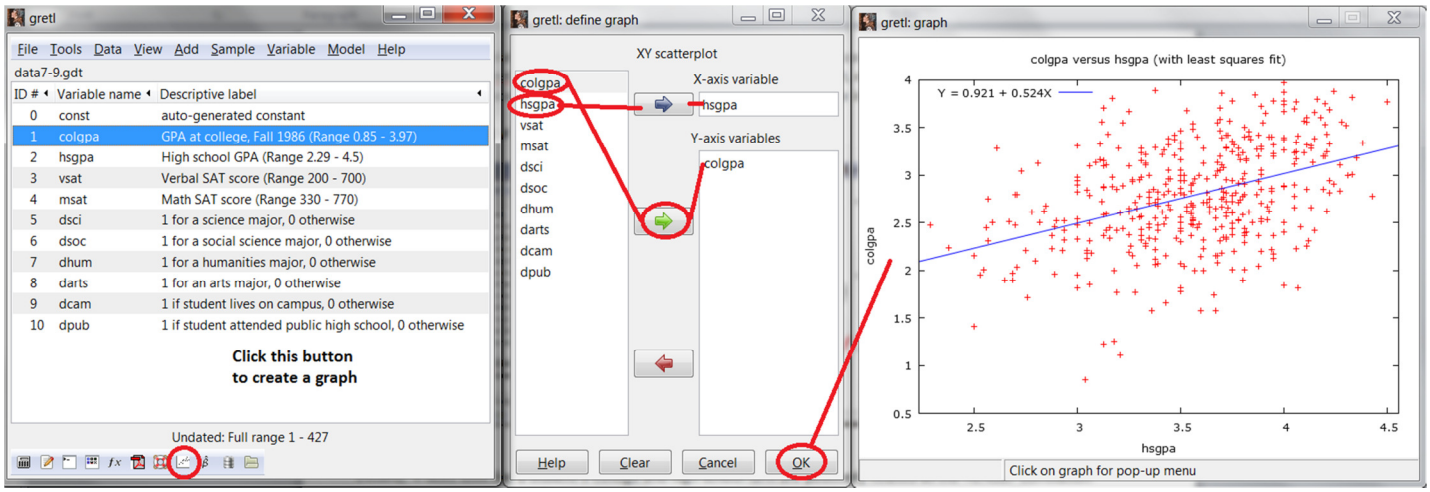
We will use these later, but for now just know the correlation coefficients measure whether two variables are positively or negatively related, and how strongly on a scale of zero to one. It shows the correlations between all of your variables, but most of the time you will only be interested in one or a few.

Example: what is the correlation between a student's college GPA (colgpa) and highschool GPA (hsgpa)? +0.4067

Graph key variables against each other as XY plot – examine visual correlations

- in gretl click X-Y graph icon along bottom to create a chart, and choose the variables
- in Excel you have to create a "Scatter" type chart from the insert menu, after highlighting the data

CREATING A GRAPH OF TWO VARIABLES IN GRETL:

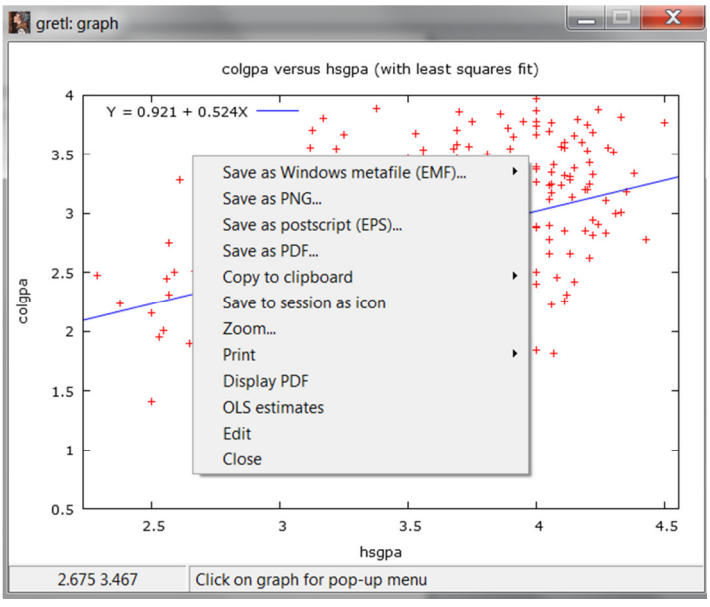


Visually, it does look like a student's college and high school GPA are positively related as the +0.4067 correlation coefficient suggested.

Showing trend lines or "least squares fit" (which is an Ordinary Least Squares regression we are about to perform) in the graph is helpful to show you the relationship. The blue line in the graph above is a trend line. gretl usually adds a trend line automatically (in Excel you have right click any one data point in the scatter chart you create and choose "Add trendline") In gretl it also gives you the equation of the trend line at the top of the chart.

Saving your graph: in gretl, click anywhere on the graph, can save as PDF, copy to clipboard (to say paste in your paper), and you can save it so you can reopen it later by choosing "Save to session as icon"

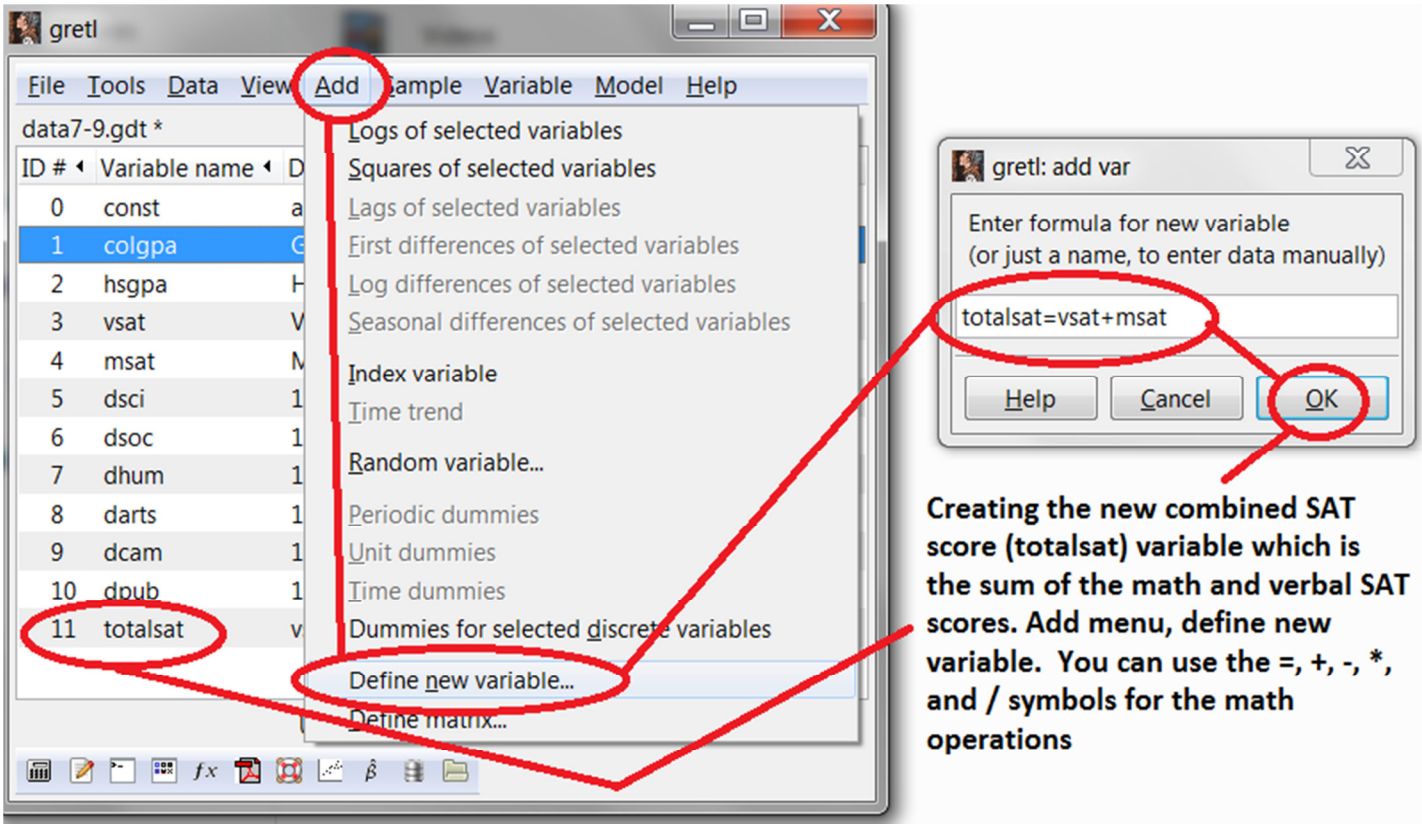
RIGHT CLICK IN THE GRAPH ANYWHERE TO BRING UP A MENU OF OPTIONS TO SAVE OR PRINT YOUR GRAPH:



Creating a new variable in gretl from your existing variables
- Add menu, Define new variable, do as equation (e.g., Z=X+Y if X and Y are in your data and Z is new)

In our sample data, there is both the student's verbal SAT (vsat) and math SAT (msat) score. But there is no variable for the student's combined math plus verbal score. So let's create a new variable named "totalsat" for that.

CREATING A NEW VARIABLE IN GRETL (COMBINED SAT SCORE EXAMPLE):



gretl can also create "logs" or "squared" versions of your variables automatically, these are the first two entries in the Add menu shown above where we picked "Define new variable".

gretl can also create random variables for you.

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.

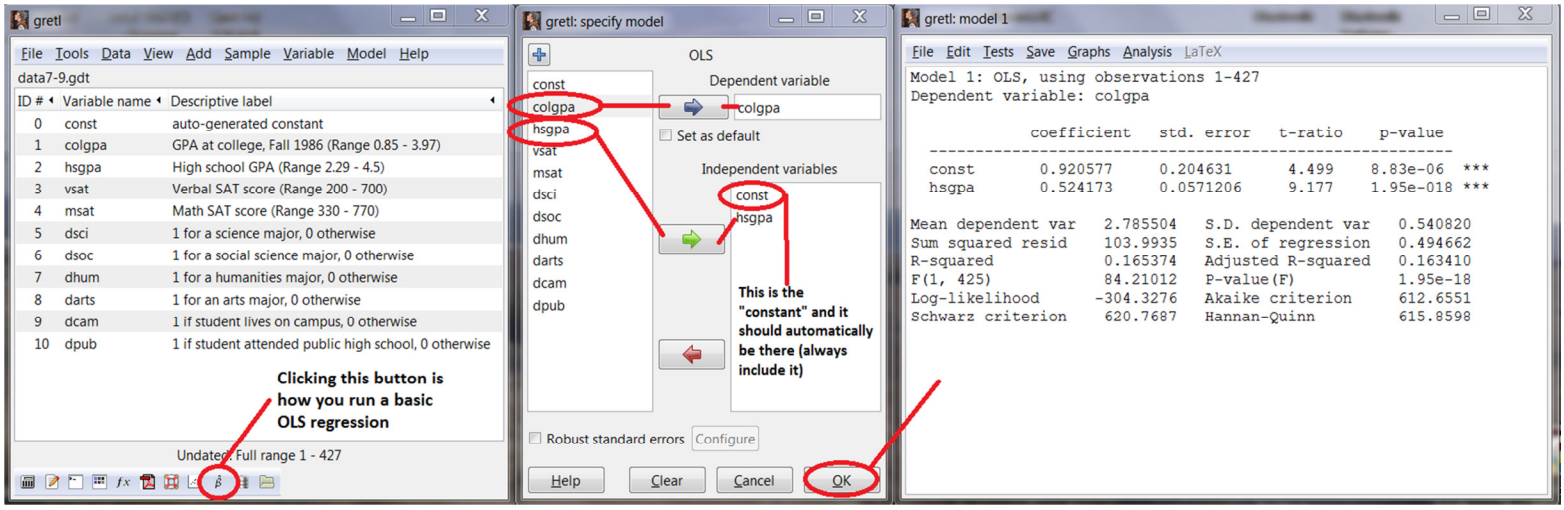
The OLS regression model is only valid when certain assumptions are met. One is that the "errors" between the points and the trend line are "random" or "normally distributed", and another is that the dependent variable is a "continuous" variable (meaning it can take on a wide range of numbers including 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):

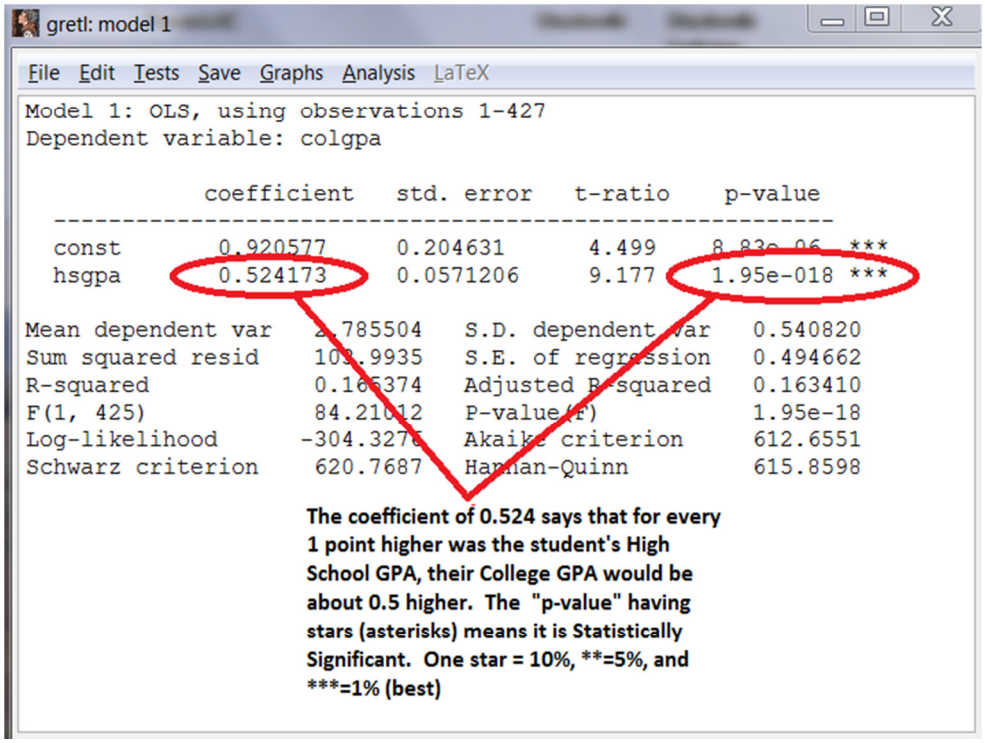


**NOTE: You should ALWAYS include a "constant" as an independent variable. This is basically allowing the trend line to have a y-intercept. If you do not include it, it will force the trend line to go through the origin (0,0 point), which can mess up your regression. gretl will normally automatically 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:



If one wanted to "predict" a student's college GPA from their high school GPA one would use the equation:

Predicted college GPA = 0.920577 + 0.524173 * 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).

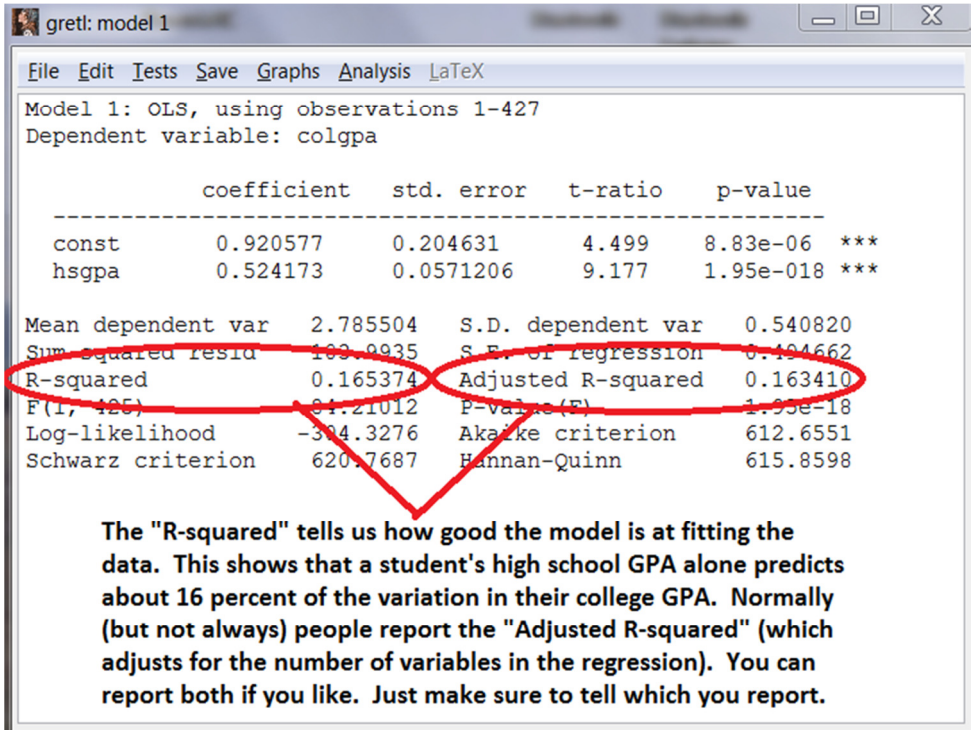
These stars/asterisks are based on the number in the column titled "p-value". This is the probability value for the statistical test. A p-value of 1% (0.01) or lower gives three stars, a p-value greater than 1% (0.01) but less than 5% (0.05) gives two stars, and a p-value greater than 5% (0.05) but less 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).

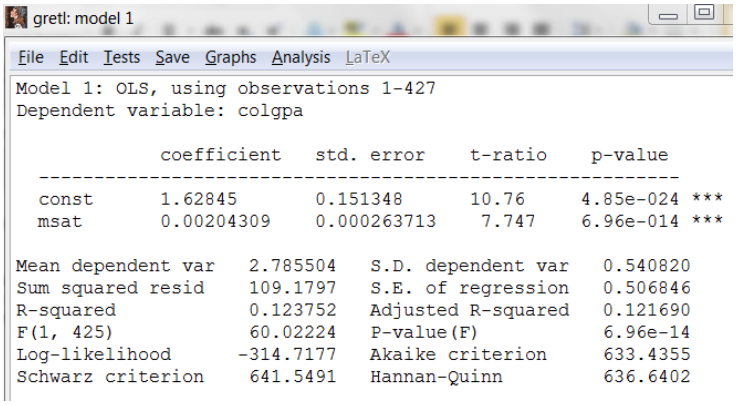
How good the model is (goodness of fit) is measured by the R-squared or Adjusted R-squared (adjusted is better)

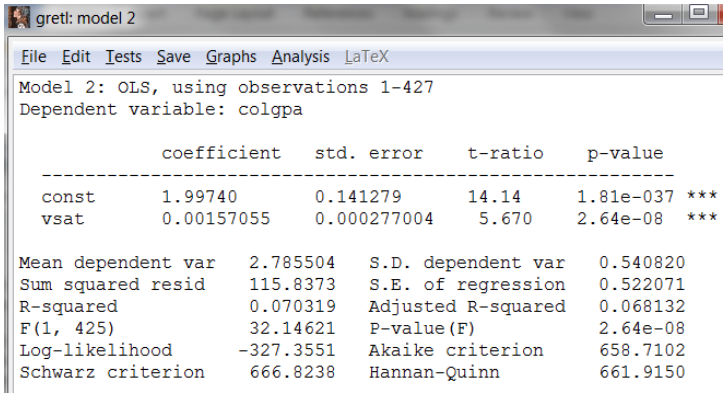
- R-squared ranges from zero to one (one would be a perfect fit)
- interpreted as "the percent of the variation explained"
- gretl will also save and report the "Log-likelihood" or "lnL" but you don't have to show this

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



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:





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

File Edit Tests Save Graphs Analysis LaTeX

```
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

File Edit Tests Save Graphs Analysis LaTeX

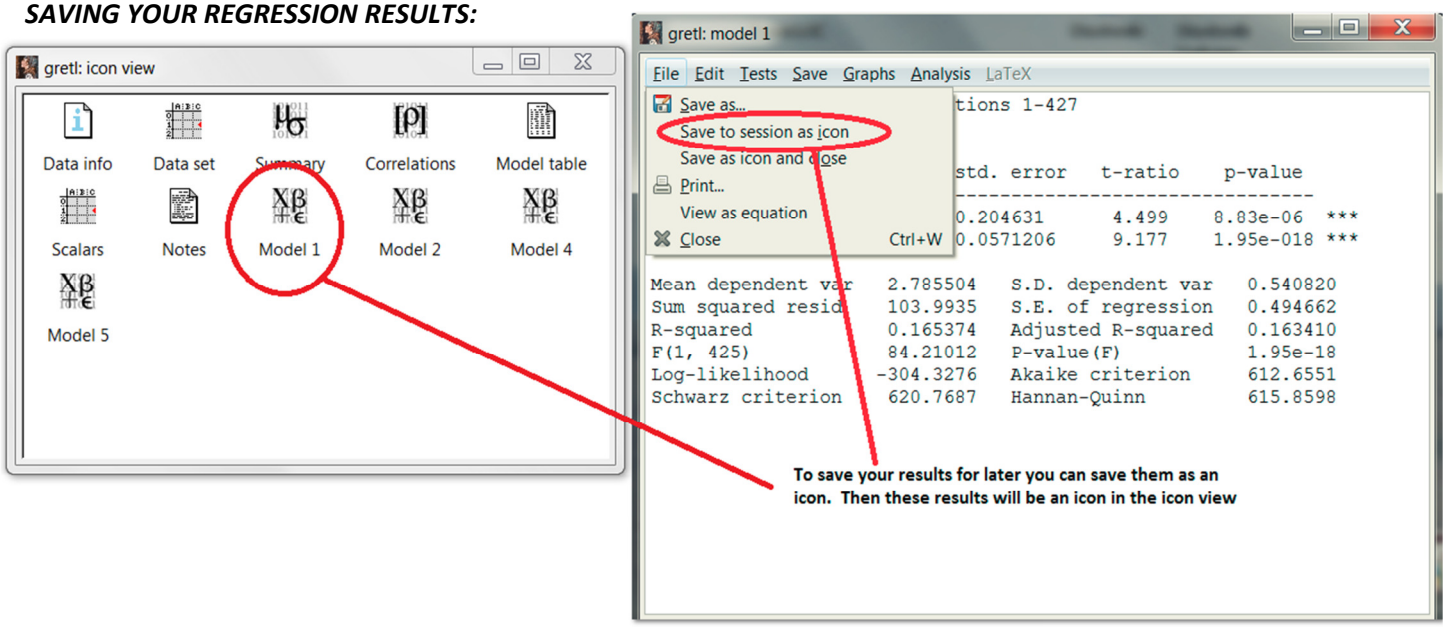
```
Model 4: OLS, using observations 1-427
Dependent variable: colgpa

             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.

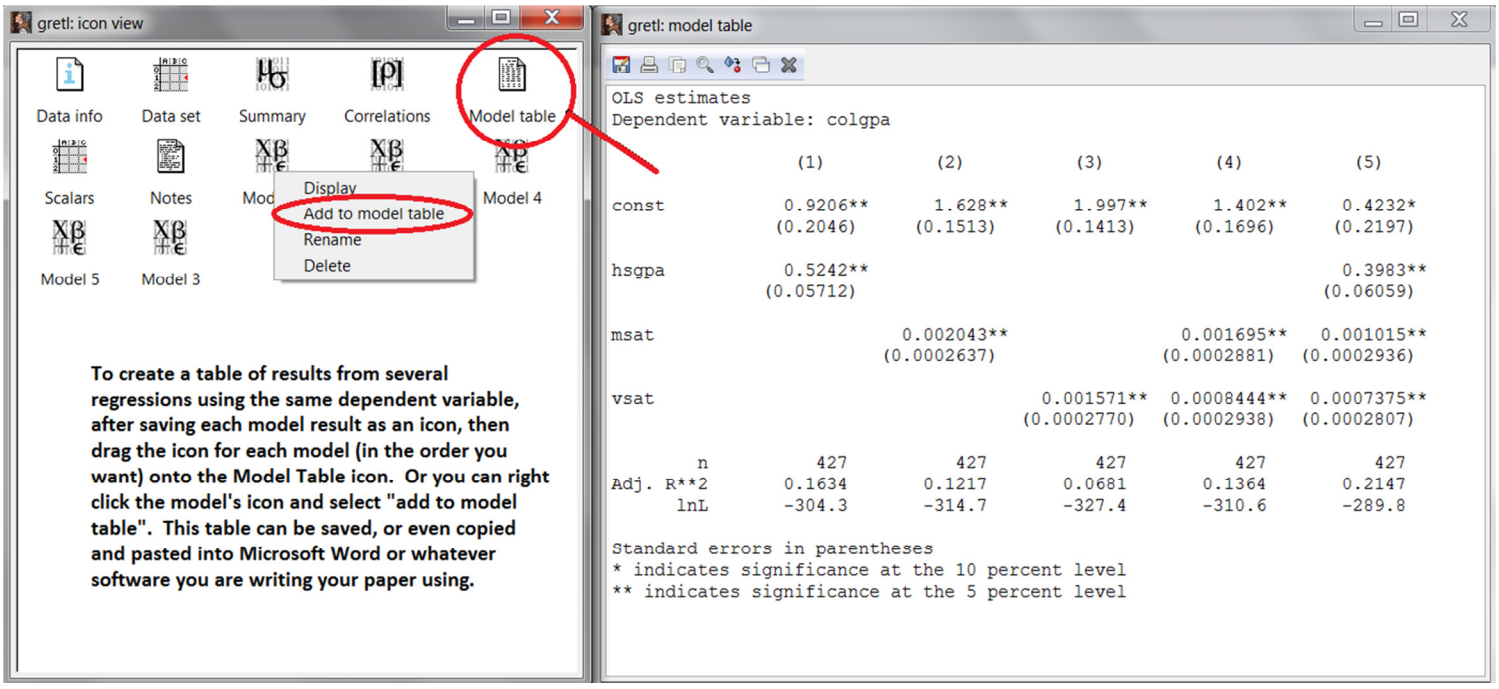
SAVING YOUR REGRESSION RESULTS:



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.

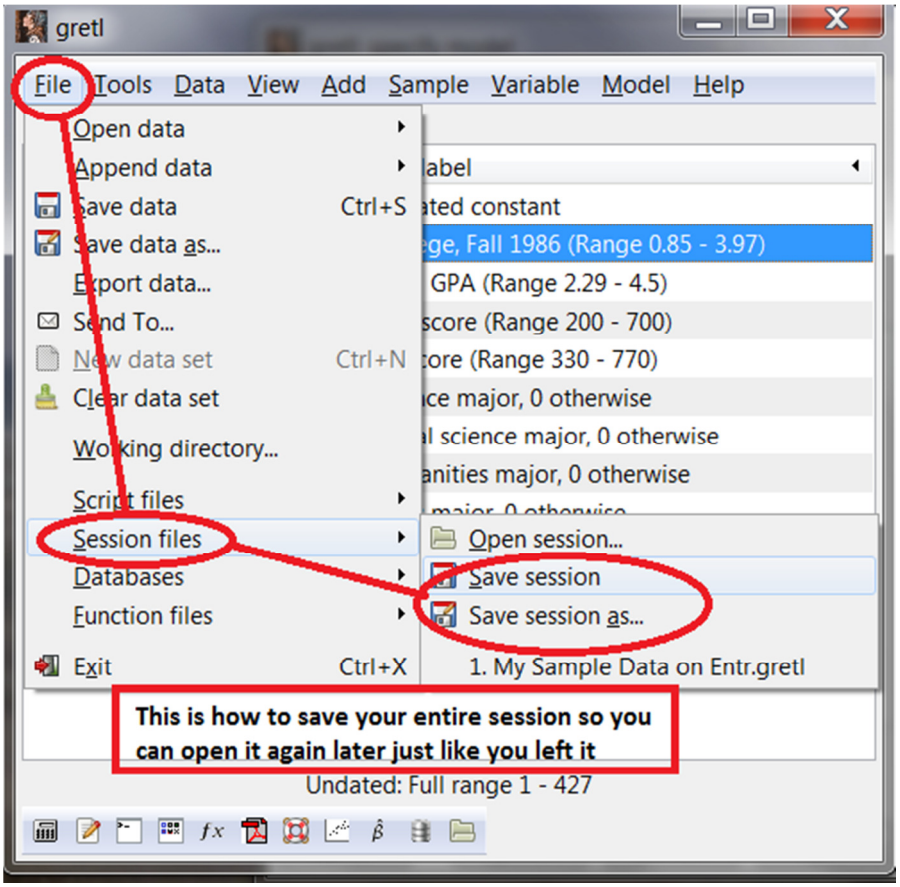
CREATING A TABLE OF RESULTS FROM SEVERAL REGRESSIONS:



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".

Multicollinarity – 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!

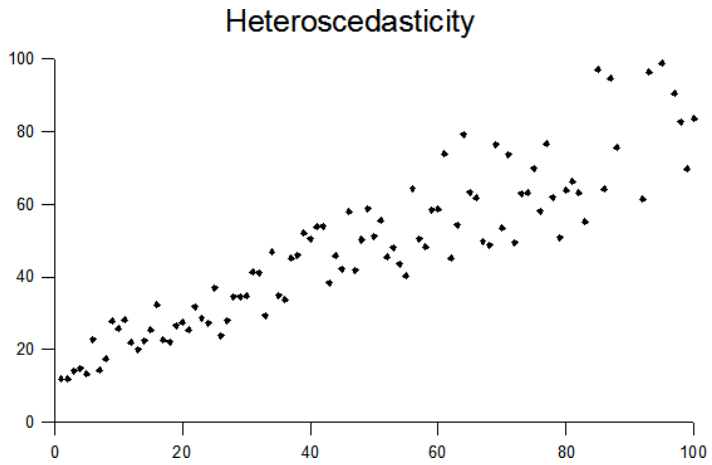
Outliers – one data point way off the line from the rest can really change your results. So always look at the graph of the data to see if there is a bad data point, then exclude that observation from the regression.

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".

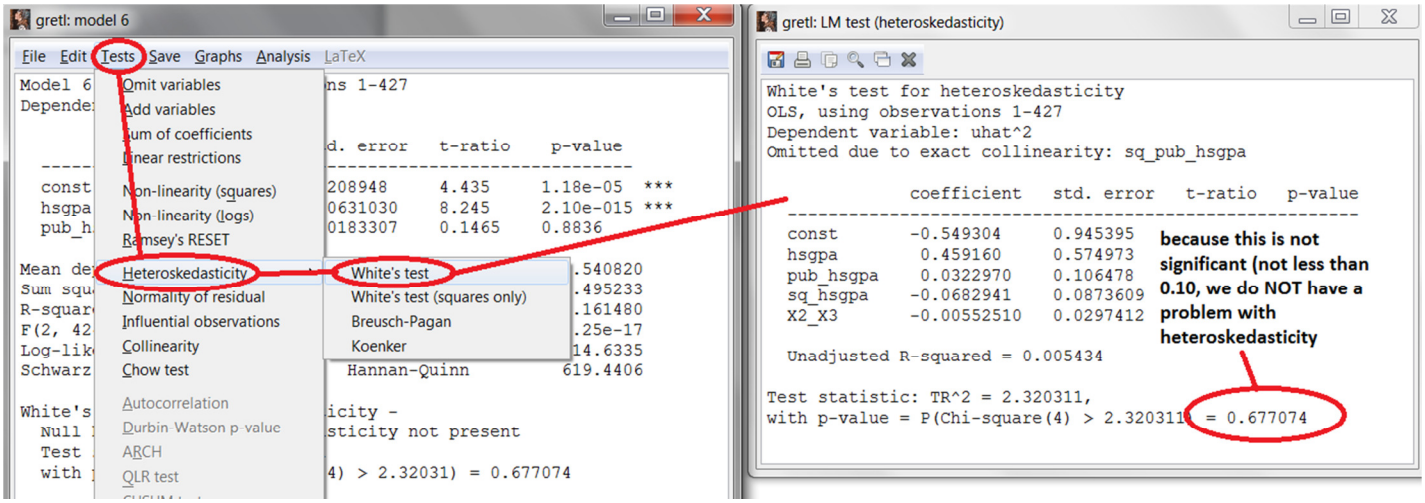
Note: After running a regression, in the results table there is a "tests" menu that has some tests for frequent problems including "influential observations" (outliers) and heteroskedasticity:

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:

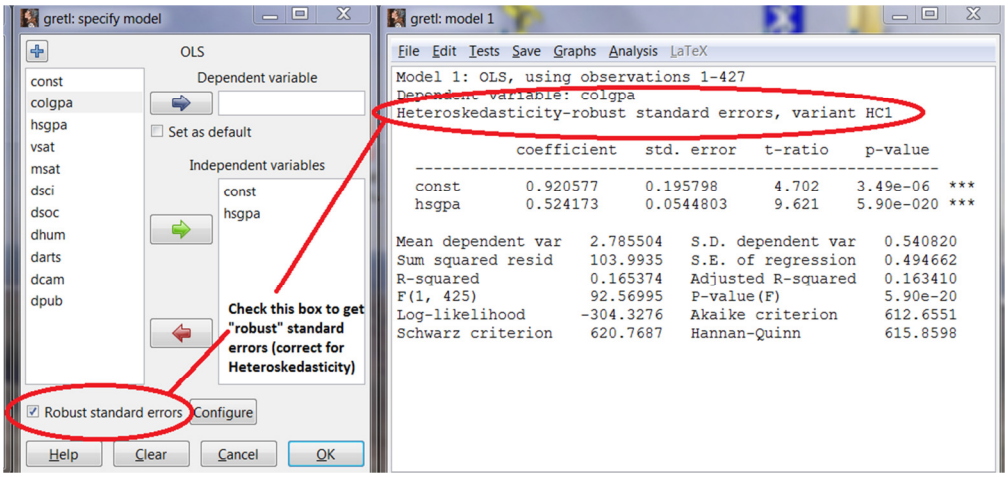
HETEROSCEDASTICIY IN THE DATA:



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



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

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

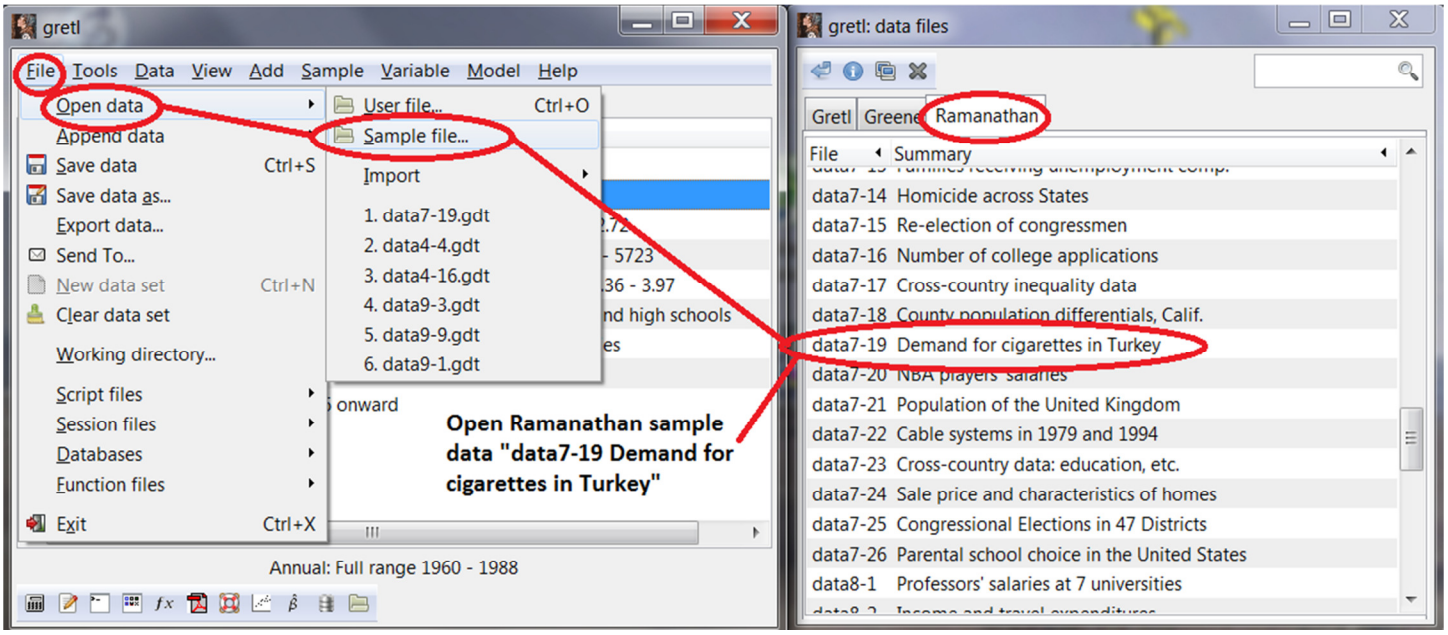
****IF YOU DO THIS, PLEASE NOTE IN YOUR RESULTS TABLE: "ROBUST STANDARD ERRORS"****

Econometric Analysis – Dr. Sobel

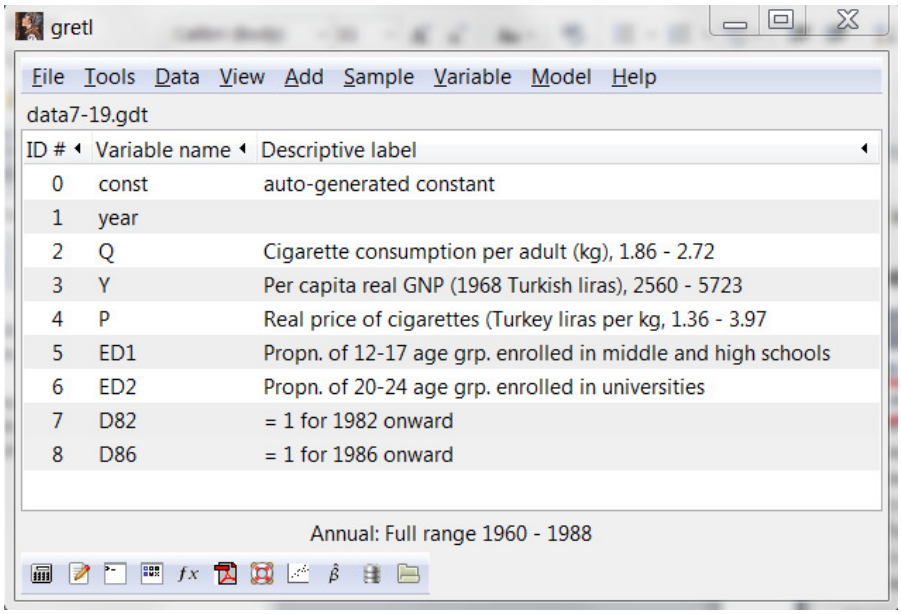
Econometrics Application to the Law of Demand:

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

OPENING RAMANATHAN SAMPLE DATA 7-19:



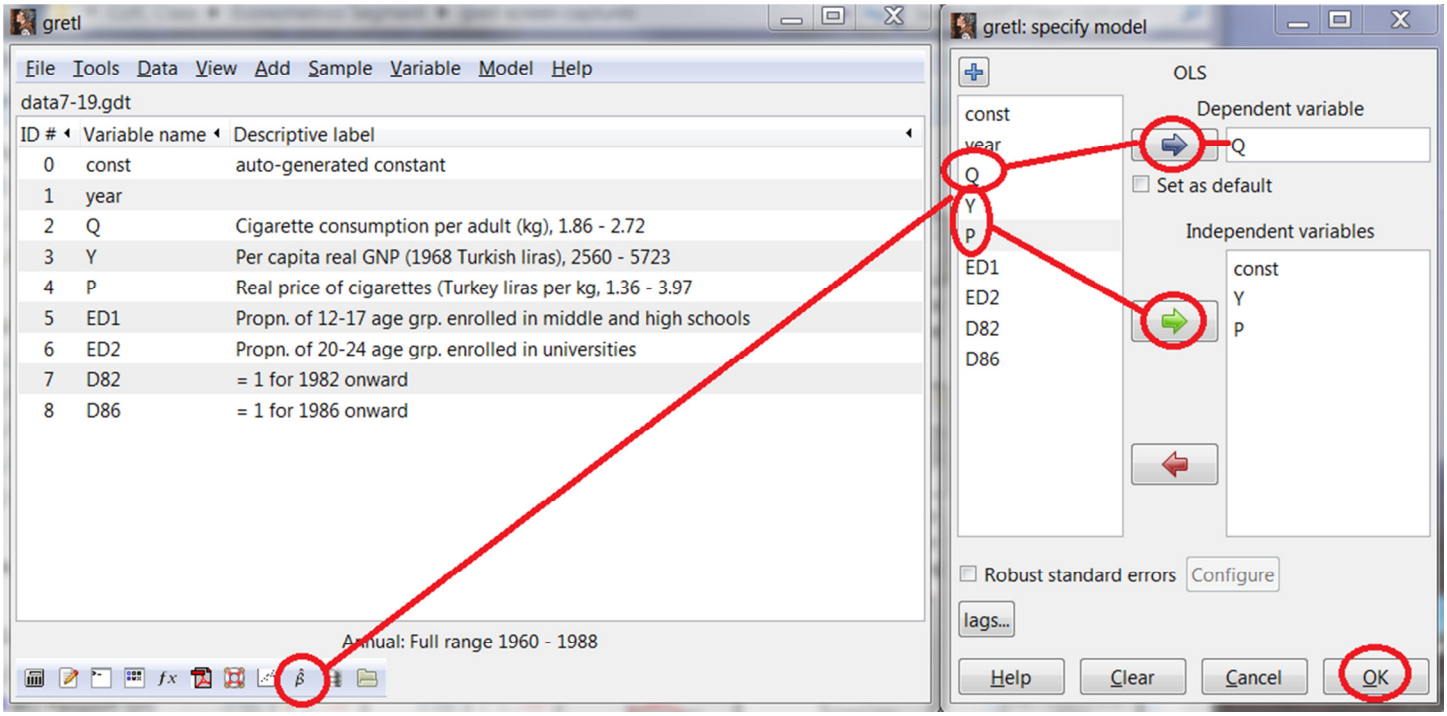
2. Here is the data that will be in your data set:

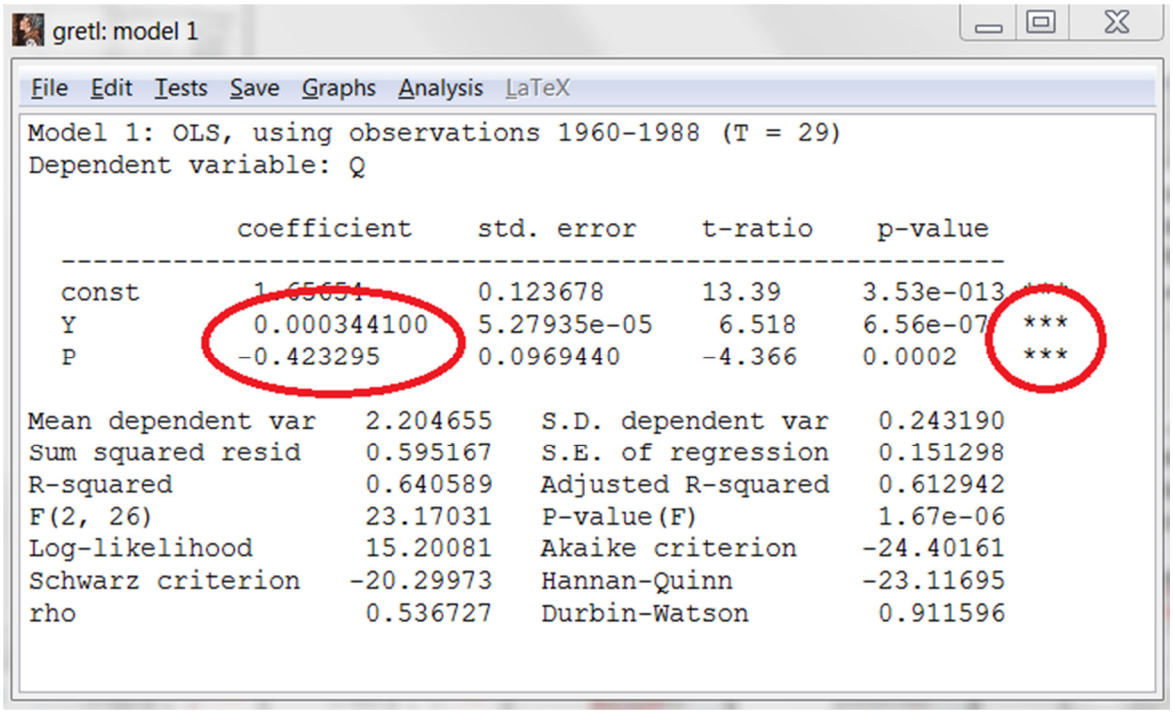


We have everything we need to estimate a demand curve: Q is the quantity demanded, Y is income, and P is price.

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):

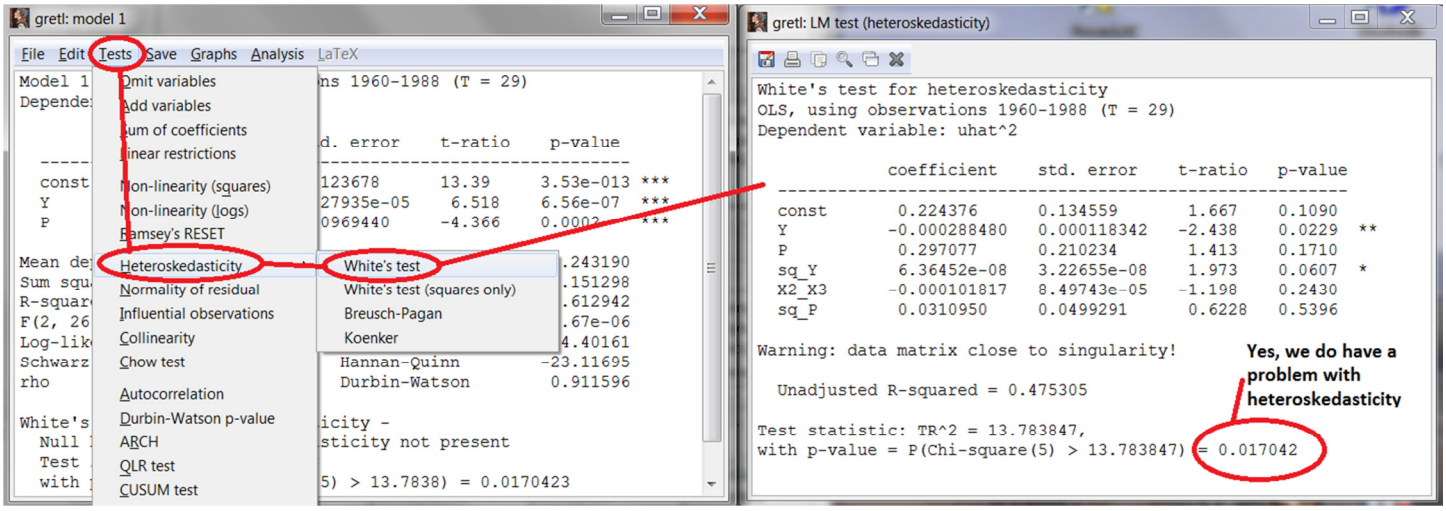




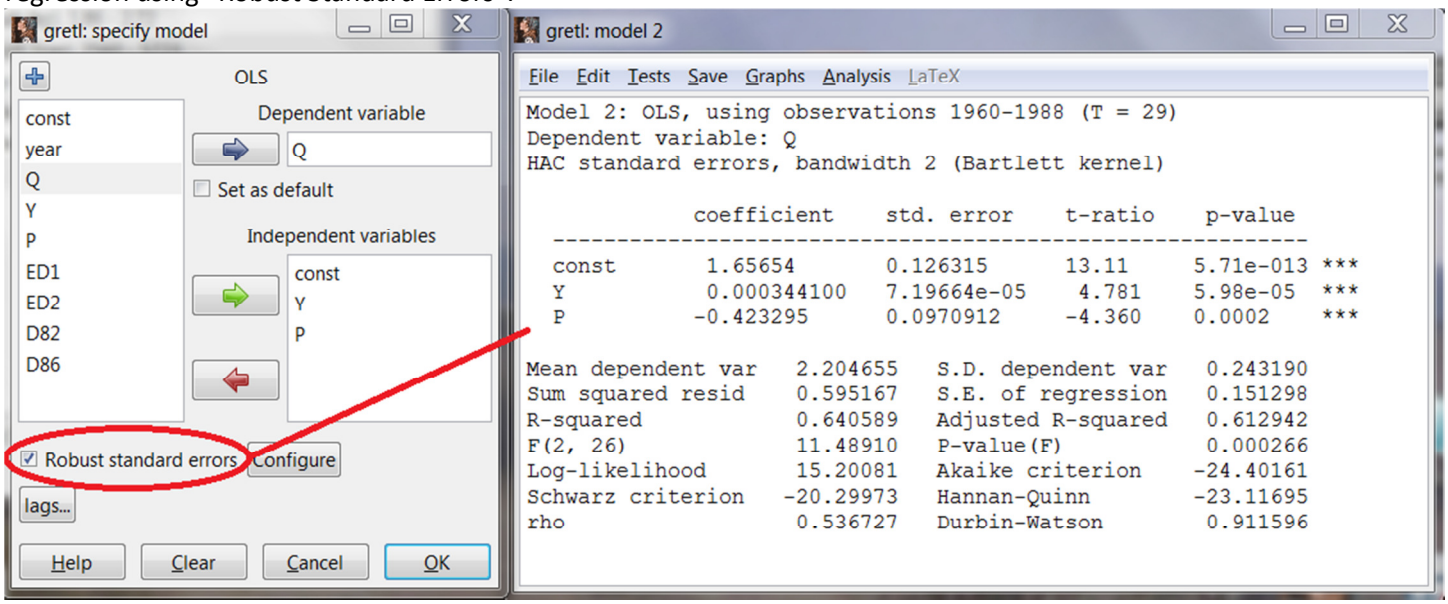
Our hypotheses are confirmed, income (Y) has a positive and significant coefficient, and price (P) has a negative and significant coefficient.

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.

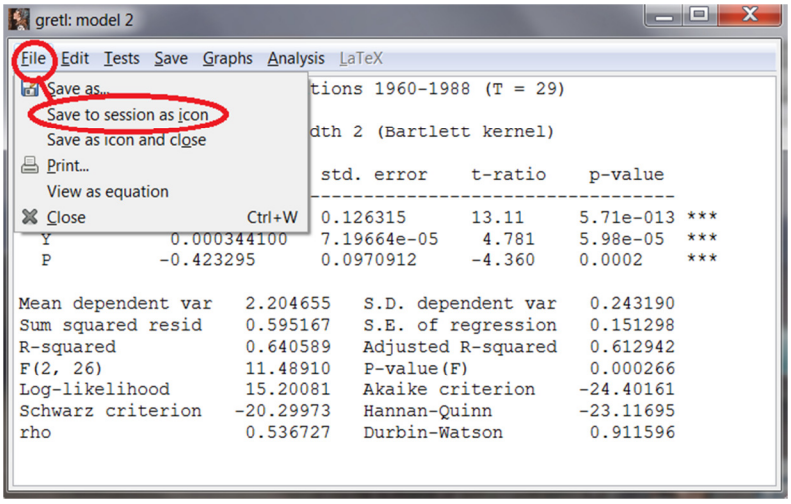
However, we need to check to be sure our model doesn't suffer from any common problems, most importantly heteroskedasticity. So run the White's test:



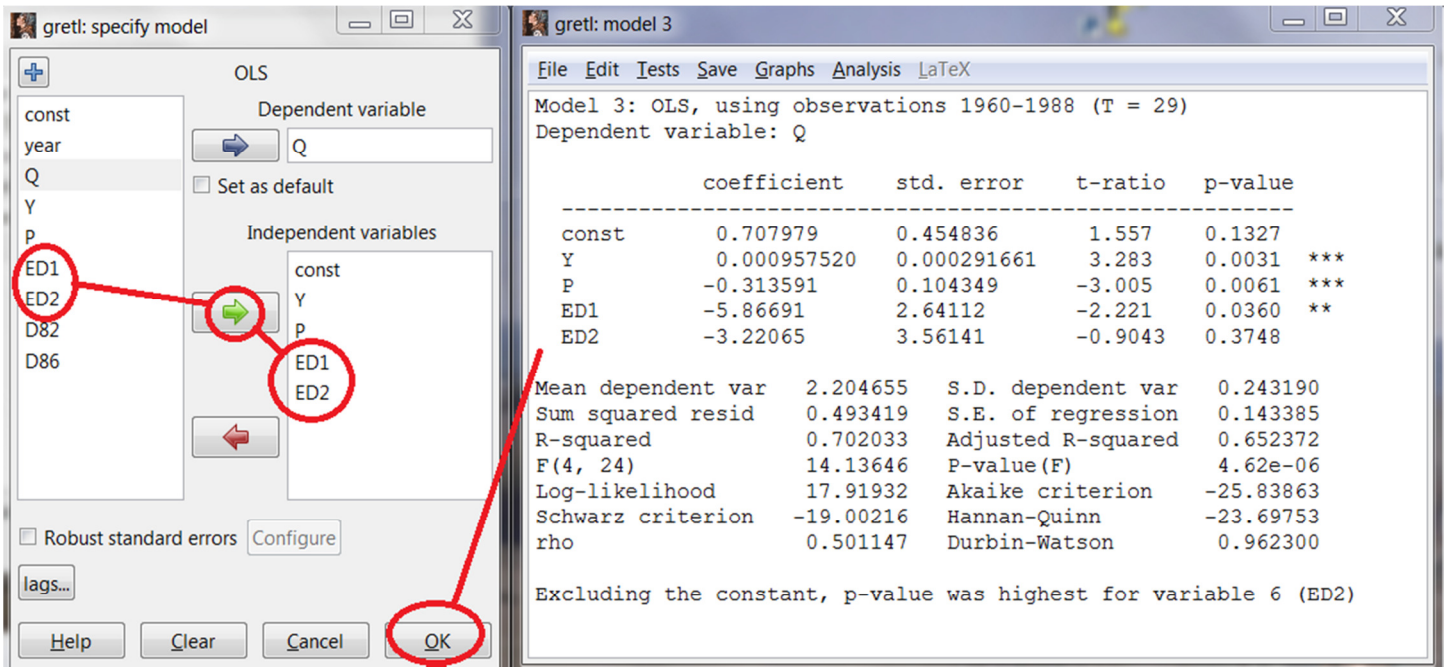
The test p-value is smaller than 0.10 (10%), so yes we have a problem with heteroskedasticity. We need to re-run the regression using "Robust Standard Errors":



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:

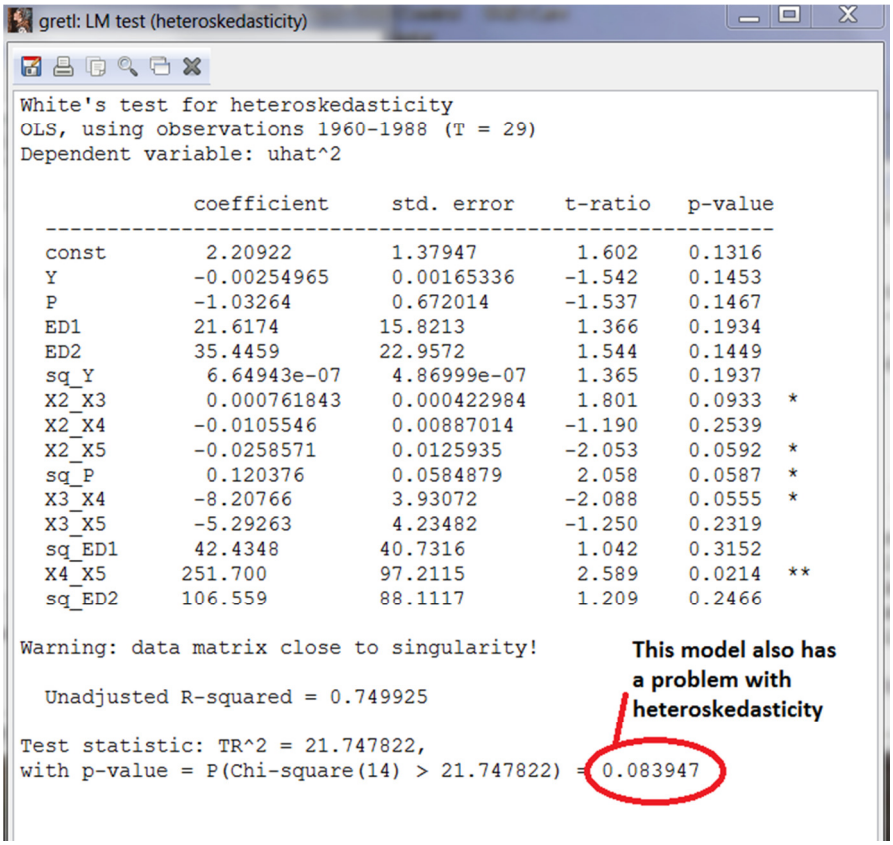


Now, let's try including two of the other variables in the data set, measuring educational enrollment, ED1 and ED2 which are the proportions of the 12-17 age group enrolled in middle and high schools, and the proportion of the 20-24 age group enrolled in universities.

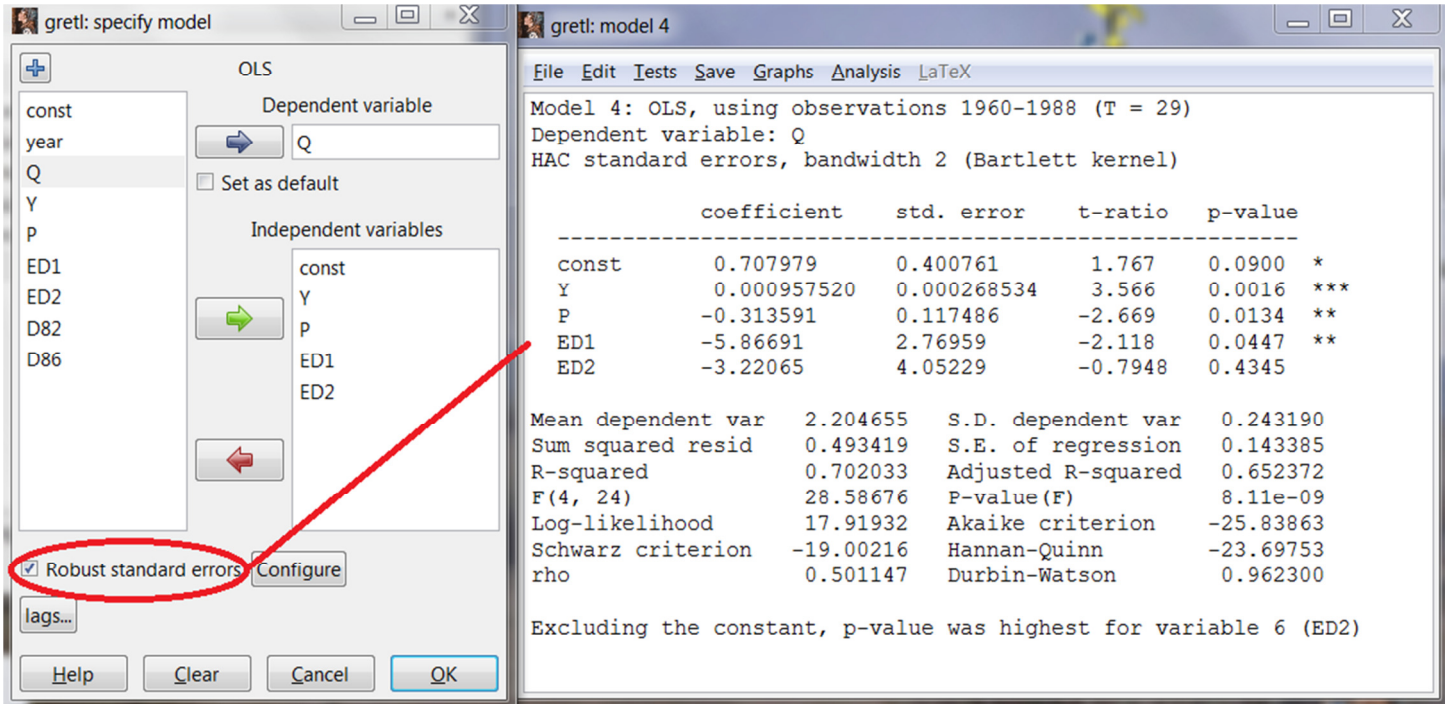


Again our hypotheses are confirmed, income (Y) has a positive and significant coefficient, and price (P) has a negative and significant coefficient. The new variables on educational enrollment suggest that a higher proportions of young individuals enrolled in school does REDUCE the demand for cigarettes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

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

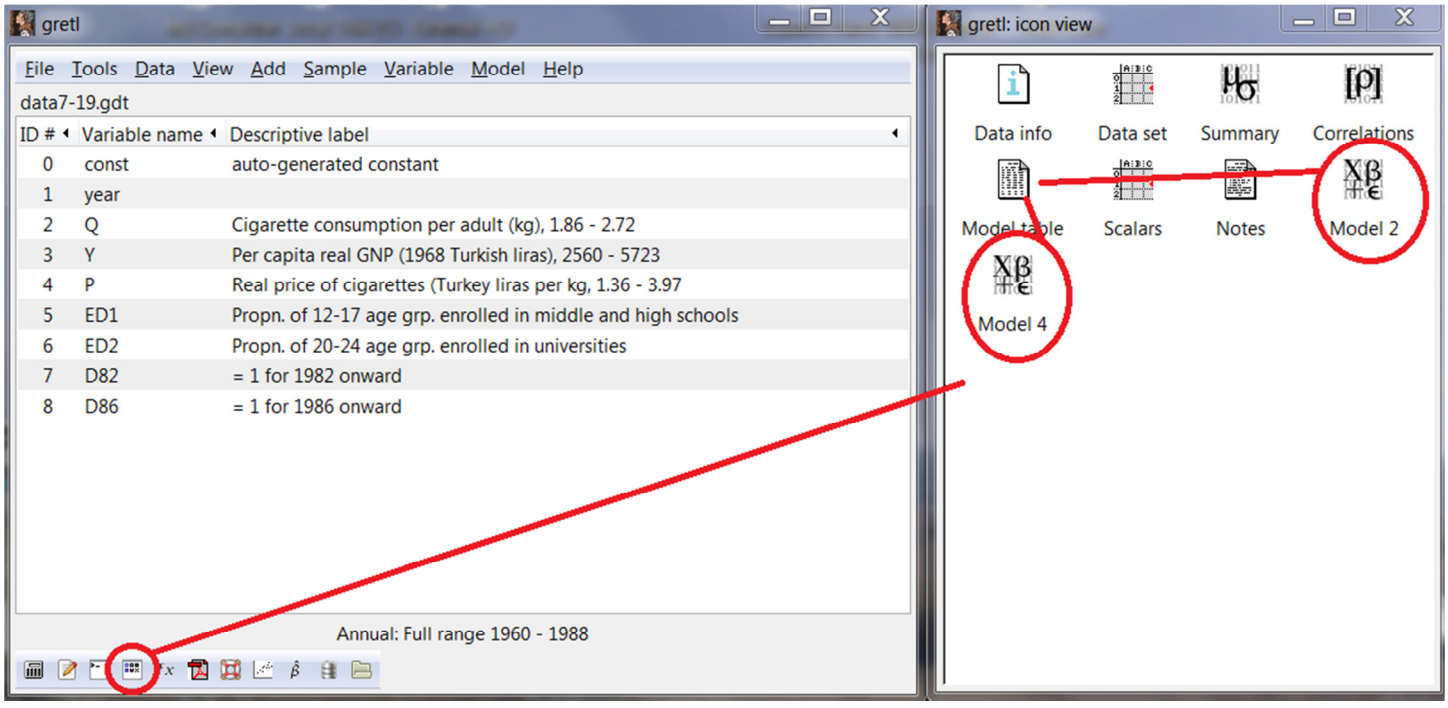


So, now re-run this extended model with “Robust Standard Errors”:



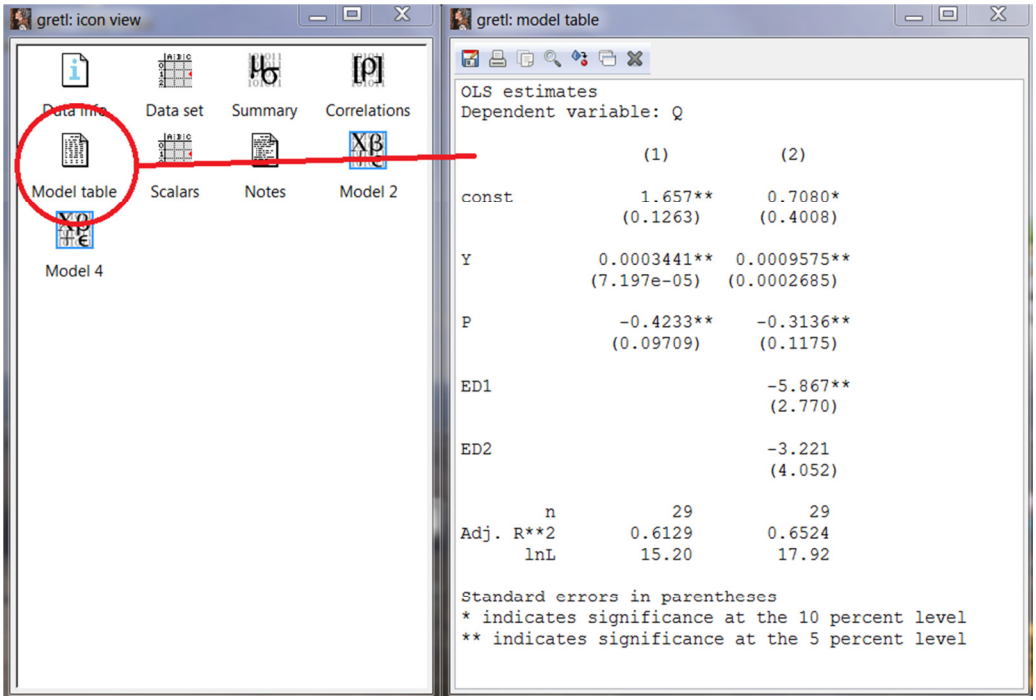
And again, our hypotheses are confirmed, these are just better estimates. So now, let’s save this model as an icon as well.

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



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

Econometric Analysis – Dr. Sobel

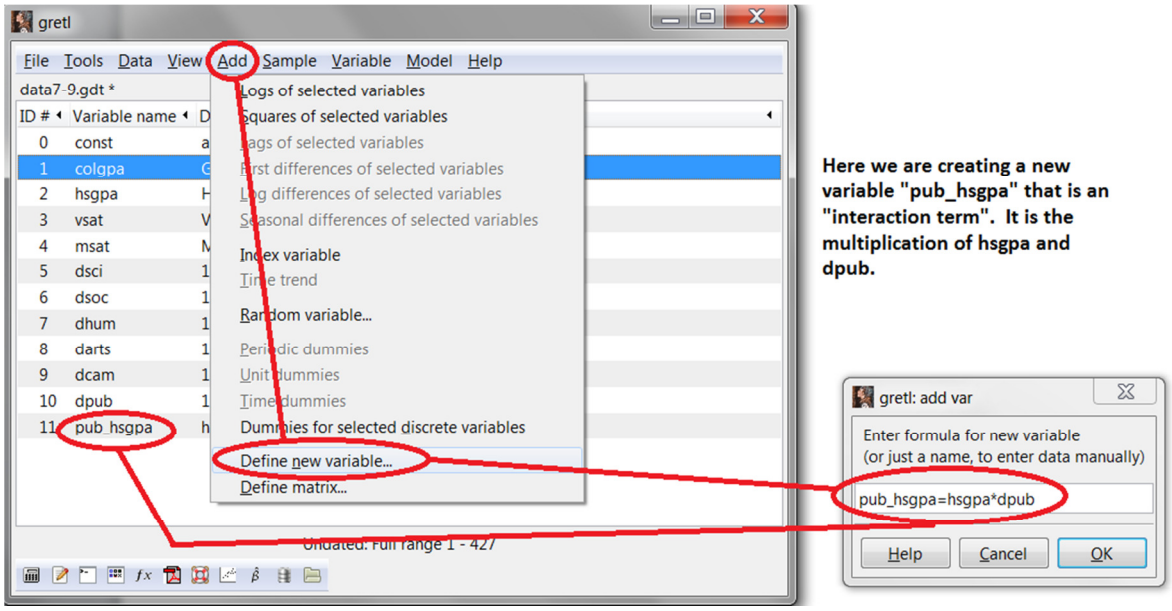
Econometrics Session 3 (time permitting and mostly for reference if your paper needs these adjustments):

5. Advanced Adjustments & Special Cases (time permitting and mostly for reference if your paper needs these adjustments)

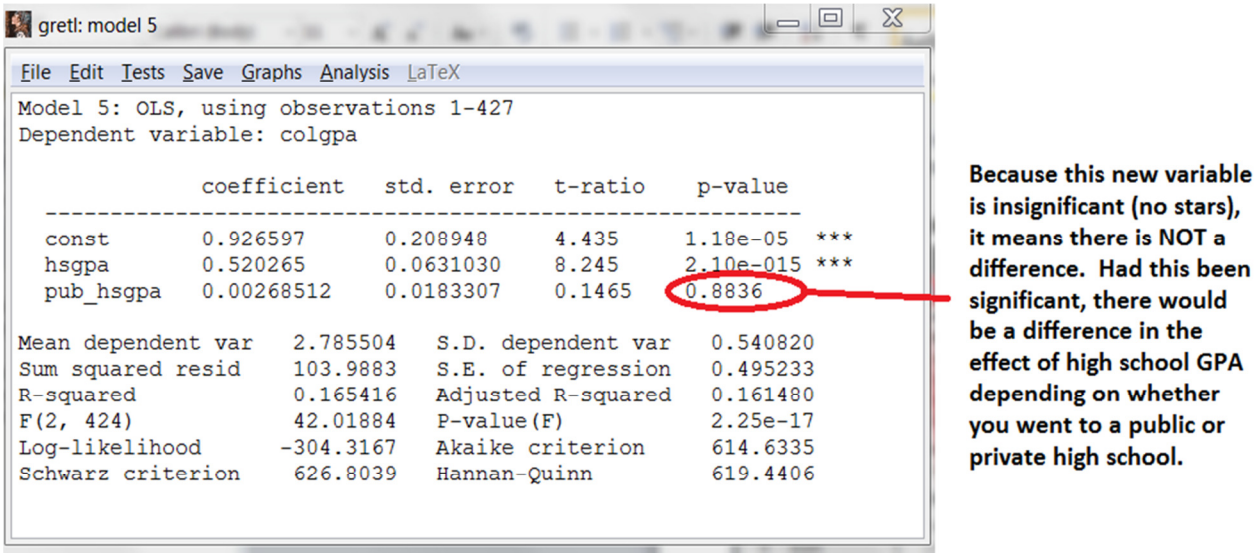
Interaction terms – when the effect of one variable depends on another variable (e.g. different for men/women)
- for two variables X and Y, we do this by including both X, Y, and also a new variable X*Y (X times Y)

For example, let's ask whether the impact of a person's high school GPA on their college GPA depends on whether they went to a private or public high school. In other words, if you went to a private school your high school GPA might matter more or less than it does if you went to a public high school. We would do this by including both the high school GPA as a variable AND including a new variable that we have to create that is the high school GPA times the zero/one variable indicating they went to a public high school. So let's first create the new variable pub_hsgpa = hsgpa*dpub

CREATING A NEW VARIABLE THAT IS AN INTERACTION TERM:



Now run the regression including this new variable as well:

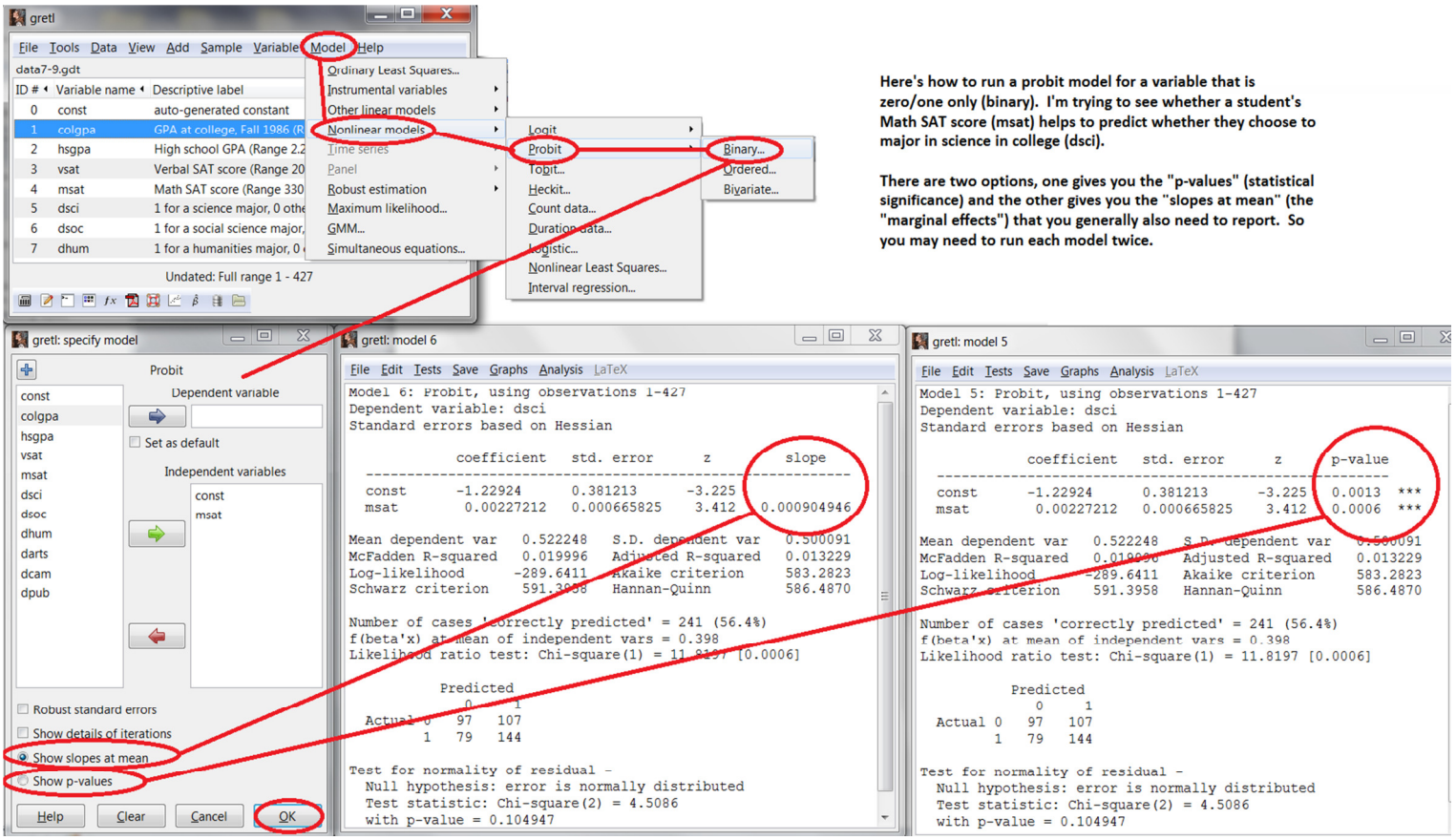


"Discrete", "Count", or Dummy Variables as the dependent variable (probit / logit / tobit)

- When the dependent variable is a zero/one variable we use "Probit" or "Logit" (usually Probit) but you can see which gives the best results. Run using "Model" menu, "Nonlinear models", "Probit" (or "Logit"), then "Binary". Note that the coefficients cannot be interpreted the same. Choosing "Show slopes at mean" will give you the "marginal effects" coefficients you can interpret the same (how a one unit change effects the PROBABILITY it is a one, and note these are in decimal form so 0.35 means it increases the probability of a one by 35%). But you will also need to run it the other way selecting "show p-values" to get the significance to show in your table you report as well.

Let's try to see whether a student's math SAT score impacts the PROBABILITY they choose to major in science (dsci). We have a dummy/indicator variable dsci that is equal to one if they live on campus. Let's run a Probit model using this as the dependent variable.

RUNNING A PROBIT MODEL FOR A DEPENDENT VARIABLE THAT IS ZERO/ONE:



The answer is "yes" students with higher math SAT scores are more likely to major in science disciplines (we know this from the stars to the right in the version that came from "Show p-values" to the right above).

The "show slopes at mean" coefficients are the ones to interpret. The coefficient on msat is 0.00227212. So for every 1 point higher on the math SAT they are 0.00227212 (or 0.227212%) more likely to major in science. Let's multiply those so they make sense. For every 100 points higher on the math SAT, they are roughly 22.7% (or 23%) more likely to major in science. This is really "percentage points" so for example if a student's odds of majoring in science were 60% normally, a 100 point increase in their math SAT would increase their probability of majoring in science to 83%.

- When the dependent variable is a count (0, 1, 2, 3...) variable we use “Poisson” or “Negative binomial” model. Run using “Model” menu, “Nonlinear models”, “Count data” and then you can pick Poisson or one of two negative binomial models. These are harder to interpret and if you need to do this you should see me.

- When the dependent variable is censored (or “truncated”) by being cut off at some high or low value we use a “Tobit” model (“Model” menu, “Nonlinear models”, “Tobit”) but this is unusual. For this you should also see me.

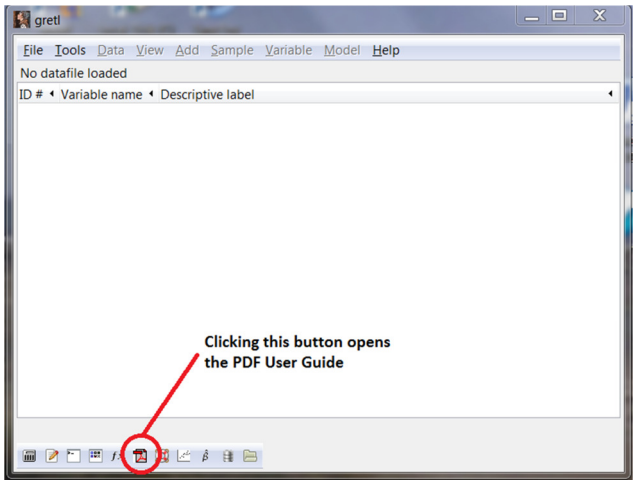
Panel Data Analysis & Fixed Effects – when you have both cross sectional and time series data (e.g., all states for a number of years) you generally include “fixed effects” which are simply dummy variables for each time period and/or for each state. This controls for anything specific to all states in a given year (like a recession) or for anything specific to one state in all years (like Hawaii is just always different). You can create and include these dummy variables yourself (although you must EXCLUDE one of the years or states as that will be in the constant). Alternatively, if you make sure you tell gretl your data is “Panel Data” (go to Data menu, Dataset structure to change or specify) you can then have it make these up for you by going to “Model” menu, selecting “Panel” then “Fixed and random effects”.

Special problems/issues with time series data (you would again have to see me for help with these)

- The problem of “nonstationary” data (will give significant results even when not)
  The EASIEST way to overcome this is to run your model in “Change” form were you use the year to year changes in ALL of the variables for the model rather than the raw data. You can, however, test for this as a problem using “Variable menu” then “Unit Root Tests”.
- In time series models we generally want to predict the future based on the past, which is called “autoregressive” or “AR” model, use “Model” then “Time series” then “Autoregressive estimation”.
- For some sets of two time series variables we just want to see if they tend to hang together through time, which is called a “Cointegration test” (found under “Model” menu, then “Time series”)
- Granger Causality (Chicken Egg paper we will be covering) – can only be done on time series data, then go to “Model”, “Time series”, “Vector autoregression”, here you put BOTH variables as “Endogenous” (dependent variables), you can run your model without any “Exogenous” (independent) variables. Again, this is one you would need to see me for help with.

List of gretl User Manual References & Pages for Important Commands

GRETL USER GUIDE (PDF "users guide" button along bottom):



1) How to run a basic OLS regression and details on all the menu commands for OLS regressions is in Chapter 2, Pages 5-12

2) Saving sessions, creating a model table, and the icon view is in Chapter 3, Section 3.4 on pages 16-18

3) Data files, how to read them in, rules on variable names, types of data is in Chapter 4, on pages 19-23

4) Dealing with missing values in the data is in Chapter 4, Section 4.6, on pages 27-28

5) Using subsamples of data, setting the sample is in Chapter 6, pages 36-38

6) Creating graphs and graph options is in Chapter 7, pages 40-41

7) Heteroskedasticity, robust standard errors, White's correction is in Chapter 15, pages 113-114

8) Using "panel" data and fixed effects is in Chapter 16, pages 121-125

9) Time series models are in Chapters 23-25, on pages 173-210. See Section 24.2 on Page 188 for Granger causality

10) Probit, Logit, and Tobit models are in Chapter 19, pages 234-253