Regression analysis in practice with GRETL

Prerequisites

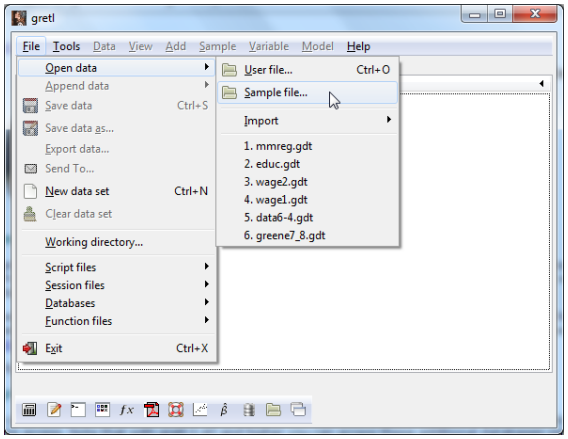
You will need the GNU econometrics software GRETL installed on your computer (http://gretl.sourceforge.net/), together with the sample files that can be installed from http://gretl.sourceforge.net/gretl_data.html.

This is no econometrics textbook, hence you should have already read some econometrics text, such as Gujarati's Basic Econometrics (my favorite choice for those with humanities or social science background) or Greene's Econometric Methods (for those with at least BSc in Math or related science).

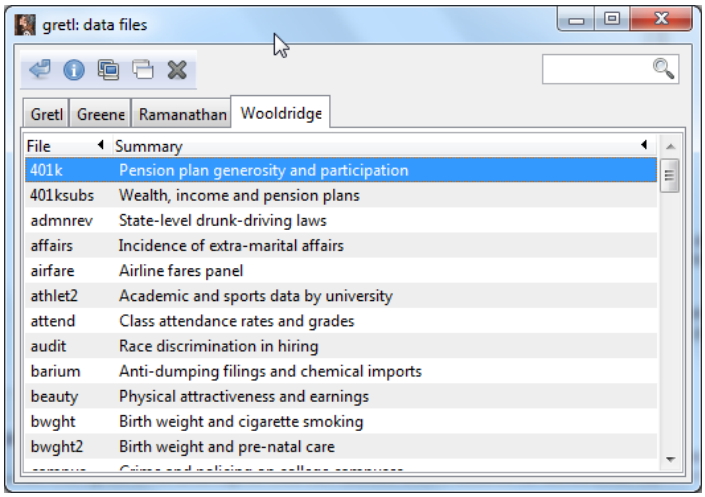
1. Introduction to GRETL

1.1 Opening a sample file in GRETL

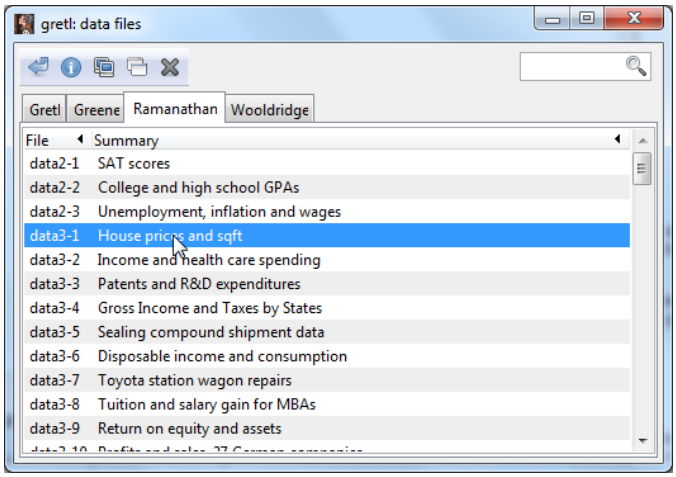
Now we open data3-1 of the Ramanathan book (Introductory econometrics wit applications, 5th ed.). You can access the sample files in the File menu under Open data/Sample file…



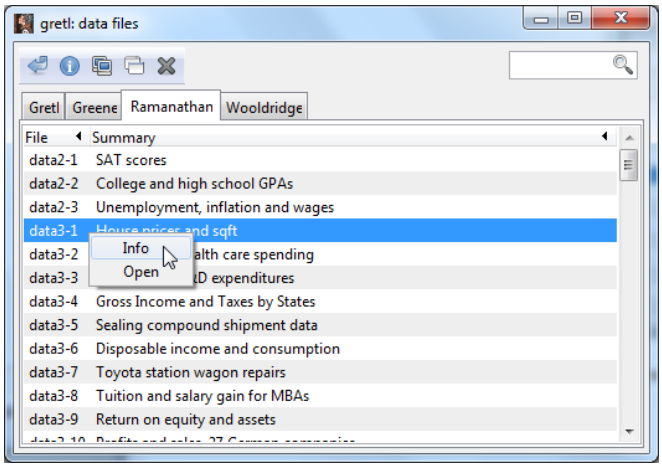
Once you click on the sample files… you are shown a window with the sample files installed on your computer. The sheets are named after the author of the textbook which the sample files are taken from.



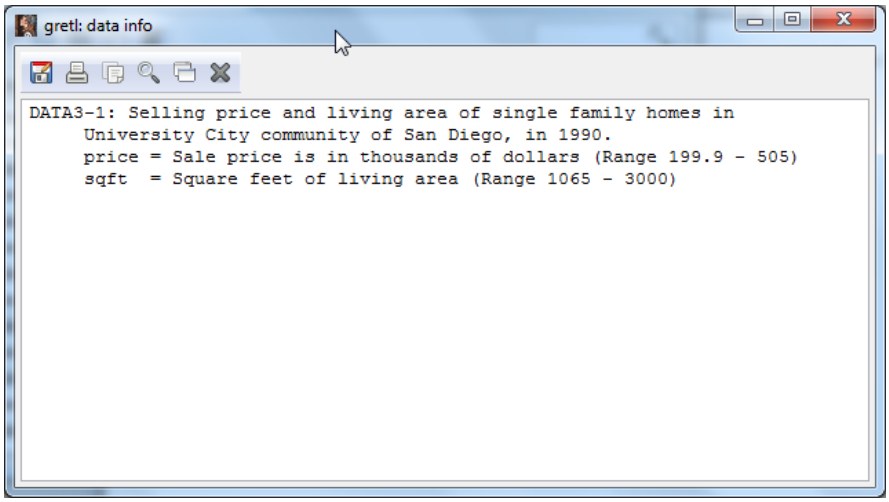
Choose the sheet named “Ramantahan” and choose data3-1.



You can open the dataset by left-clicking on its name twice, but if you right click on it, you will have the option to read the metadata (info).

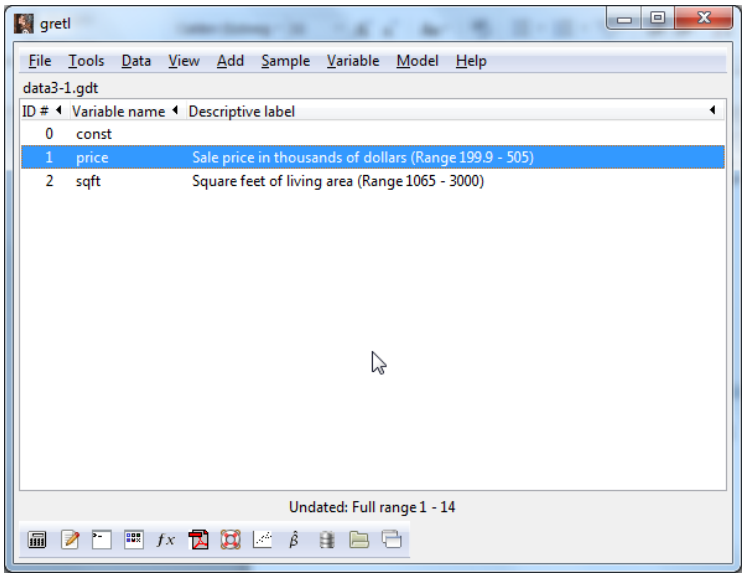


This may prove useful because sometime it gives you the owner of the data (if it was used for an article), and the measurement units and exact description of the variables.

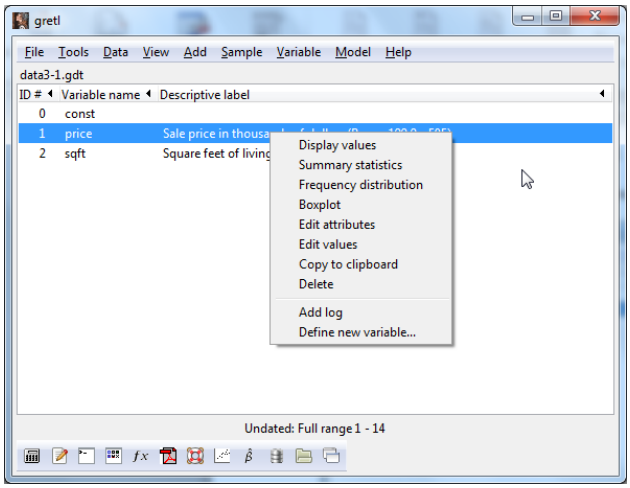


1.2 Basic statistics and graphs in GRETL

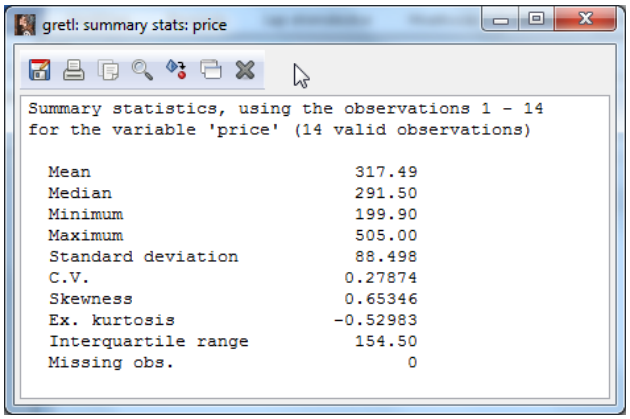
We have now our variables with descriptions in the main window.



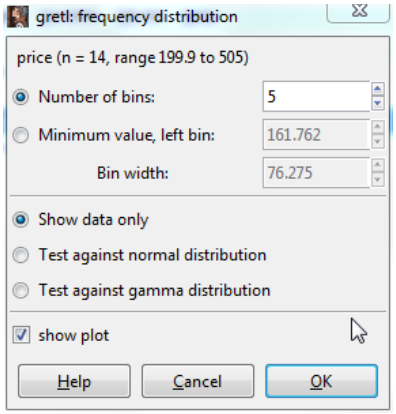
You can access to basic statistics and graphs my selecting one (or more by holding down ctrl) of the variables by right-click.



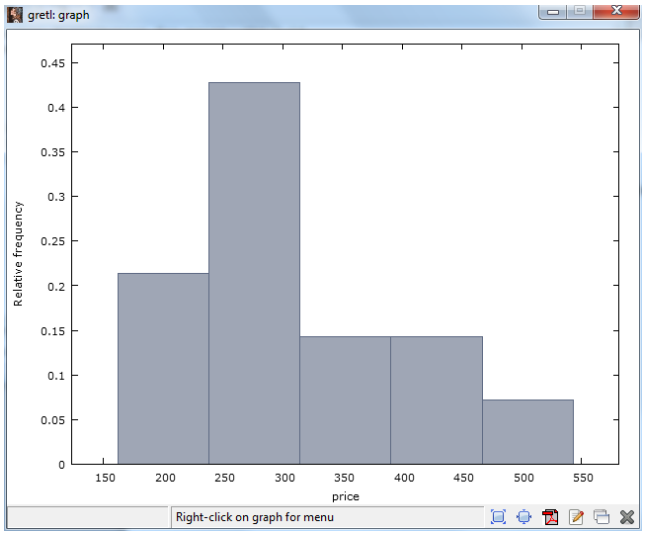
Summary statistics will yield you the expected statistics.



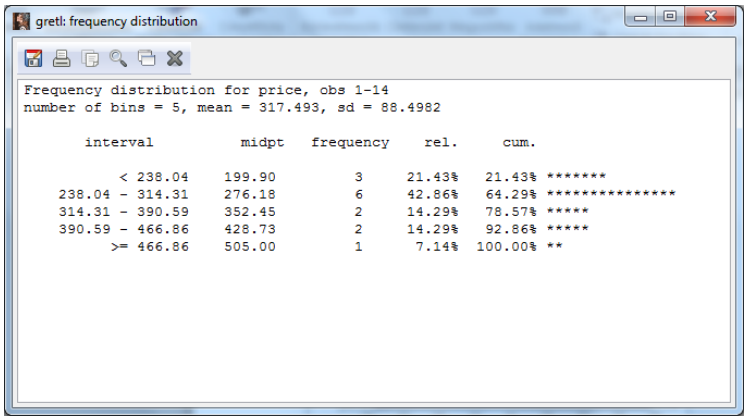
Frequency distribution should give you a histogram, but first you need to choose the number of bins.



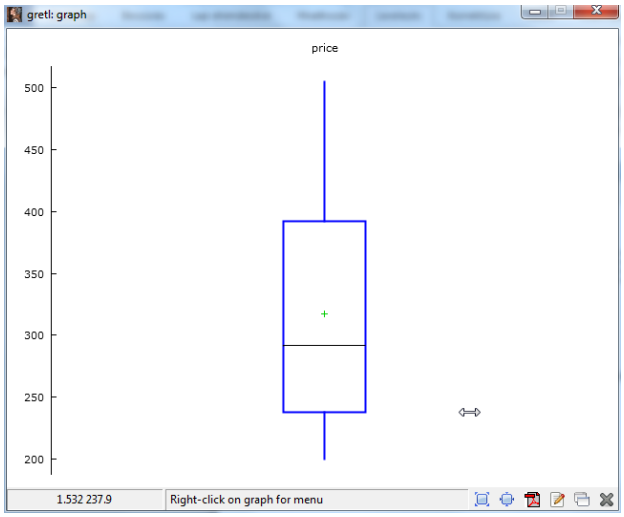
With 14 observations 5 bins look enough.



You also have a more alphanumerical version giving you the same information:

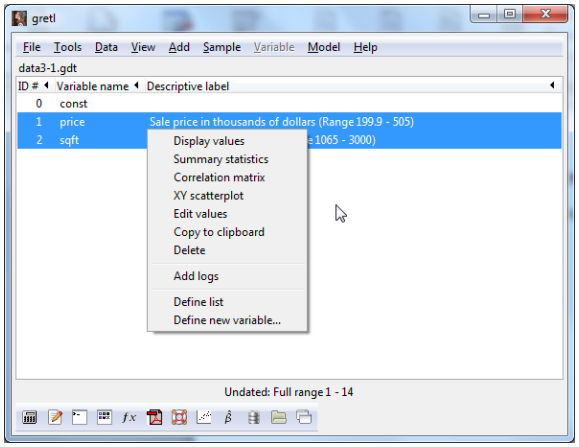


You also ask for a boxplot.

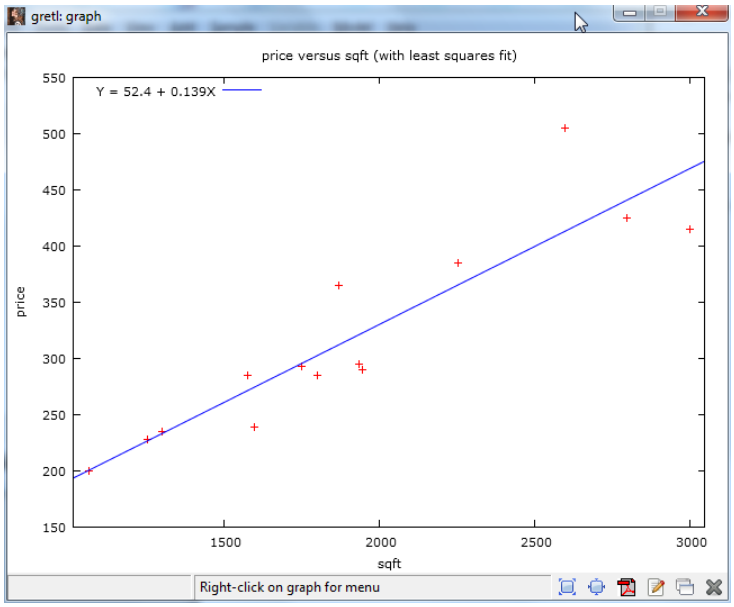


Showing that the prices are skewed to the right (the typical price being less than the average) which is often referred to as positive skew.

If you choose both variables, you have different options as they are now treated as a group of possibly related variables.

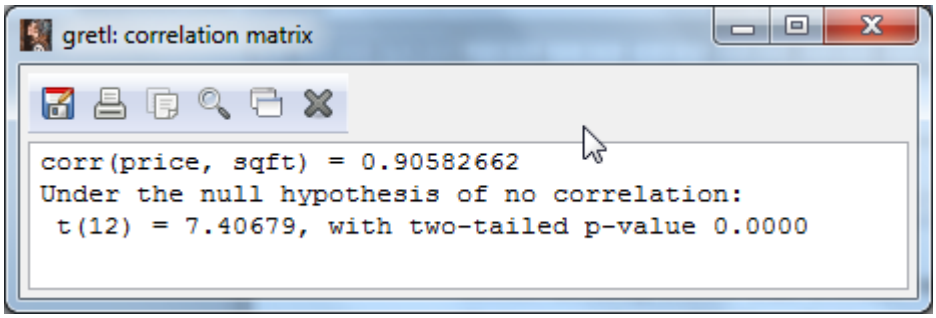


The option scatterplot will give you the following:



Which already shows that a linear relationship can be assumed between the price of houses and their area. A graphical analysis may be useful as long as you have a two-variate problem. If you have more than two variables, such plots are not easy to understand anymore, so you should rely on 2D representation of problems like different residual plots.

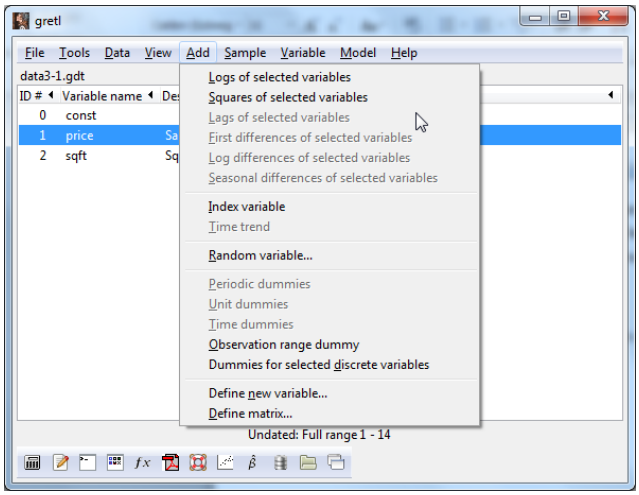
You can also have the correlation coefficient estimated between the two variables:



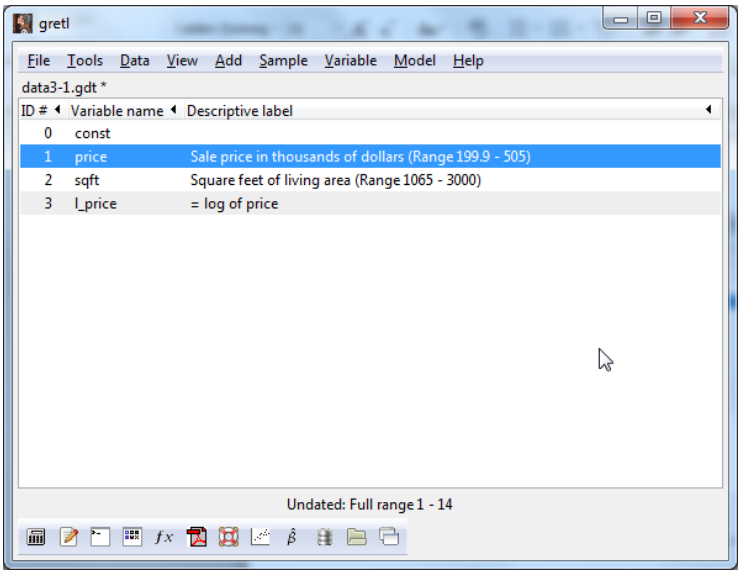
With a hypothesis test with the null hypothesis that the two variables are linearly independent or uncorrelated. This is rejected at a very low level of significance (check out the p-value: it is much lower than any traditional level of significance, like 0.05 (0.01) or 5% (1%)).

1.3 transforming variables

Transforming variables can be very useful in regression analysis. Fortunately, this is very easily done in GRETL. You simple choose the variables that you wish to transform and choose the Add menu. The most often required transformations are listed (the time-series transformations are now inactive since our data is cross-sectional), but you can always do you own transformation by choosing “Define new variable” .



Let us transform our price into a natural logarithm of prices. This is done by first selecting price, then by left-clicking the Logs of selected variables in the Add menu. You will now have the transformed price variable in you main window as well.

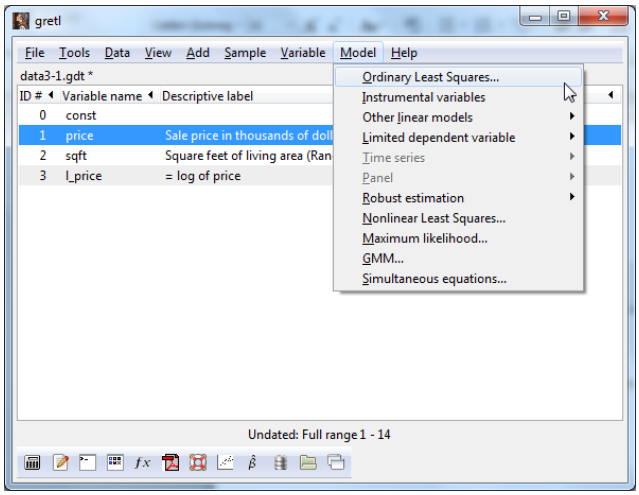


Now that you know the basics of GRETL, we can head to the first regression.

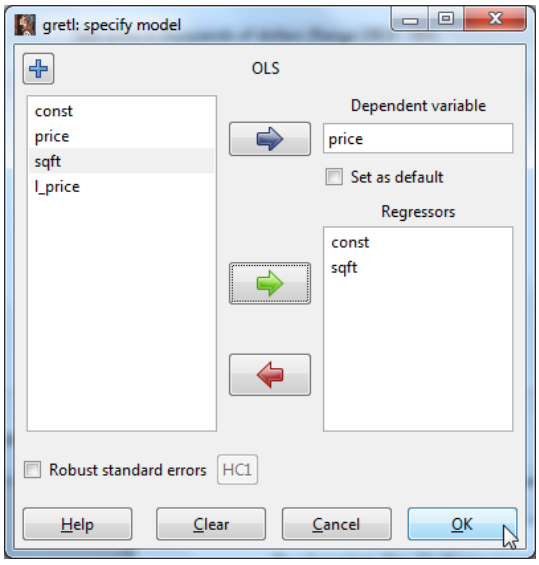
2. First linear regression in GRETL

2.1 Two-variate regression

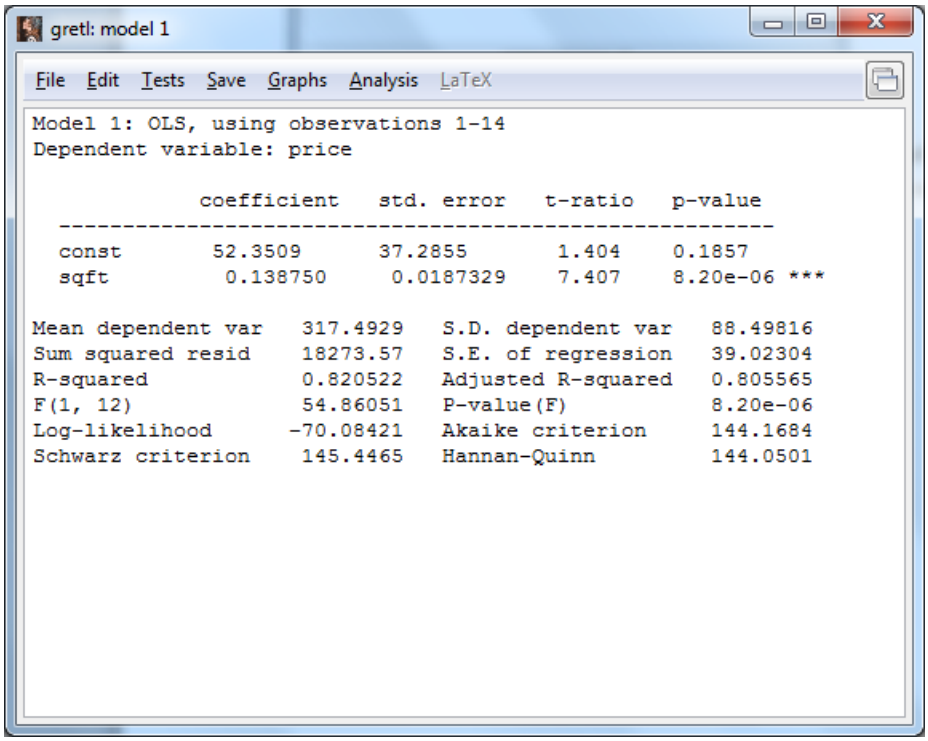
You can estimate a linear regression equation by OLS in the Model menu:



By choosing the Ordinary Least Squares you get a window where you can assign the dependent and explanatory variables. Let our first specification be a linear relationship between price and area:



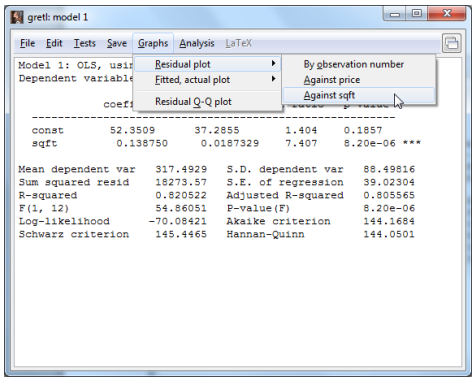
After left-clicking OK, we obtain the regression output:



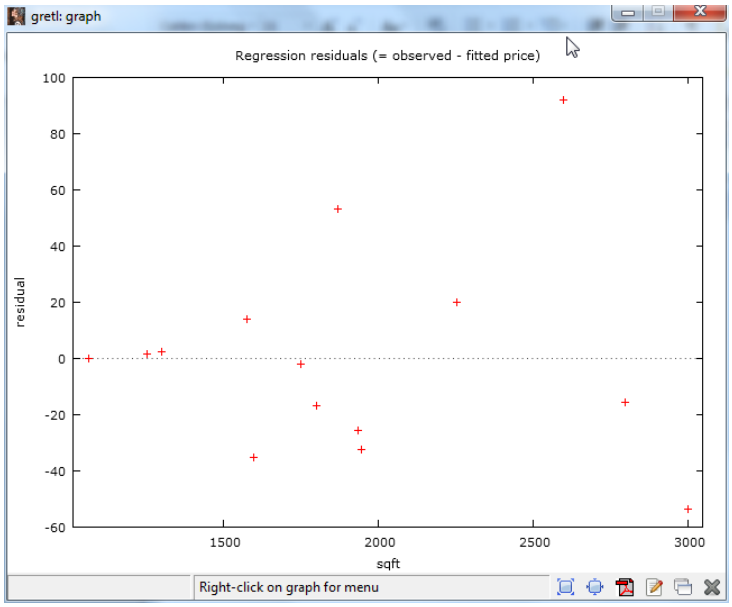
The intercept does not seem to be statistically significant (i.e. the population parameter is not different from zero at 10% level of significance), while the slope parameter (the coefficient of the area) is significant at even 1%. The R^2 is also quite high (0.82) signifying a strong positive relationship between the area of houses and their prices. If a house had one square feet larger living area, its sale price was on average higher by 138.75 dollar.

2.2 Diagnostic checks -heteroscedasticity

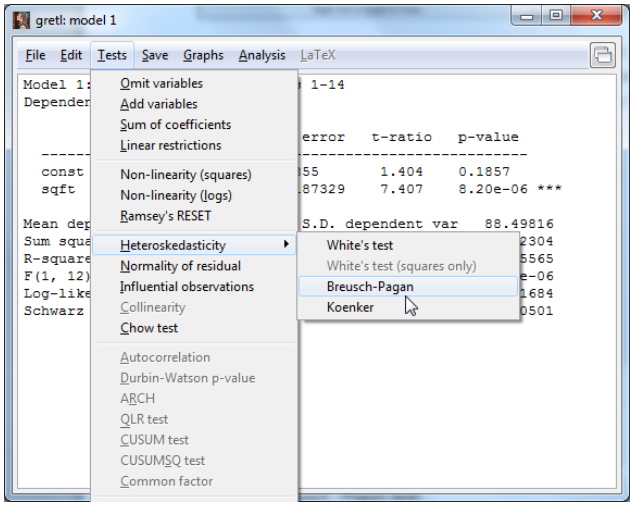
But this does not mean that we should necessarily believe our results. The OLS is BLUE (Best Unbiased Linear Estimator) only if the assumptions of the classical linear model are fulfilled. We cannot test for exogeneity (that is difficult to test statistically anyway), and since we have cross-sectional data, we should also not care much about serial correlation. We can test heteroscedasticity of the residual though. Heteroscedasticity means that the variance of the error is not constant. From estimation point of view what really matters is that the residual variance should not be dependent on the explanatory variables. Let us look at this graphically. Go to the Graph menu in the regression output.



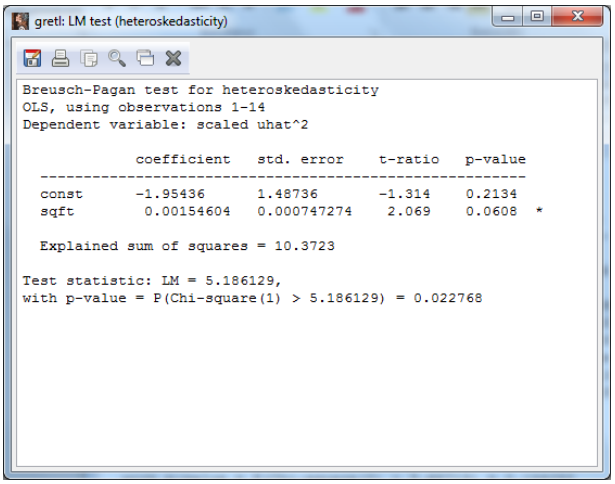
Plotting the residuals against the explanatory variable will yield:



You can observe that while the average of the residual is always zero, its spread around its mean does seem to depend on the area. This is an indication of heteroscedasticity. A more formal test is a regression of the square of the residuals on the explanatory variable(s). This is the Breusch-Pagan test:

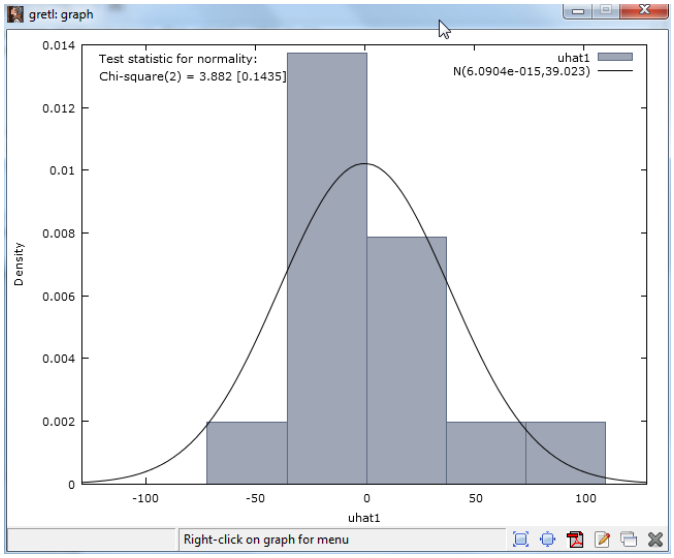


What you obtain after clicking on the Breush-Pagan test under Tests menu is the output of the test regression. You can observe that the squared residuals seem to depend positively on the value of area, so the prices of larger houses seem to have a larger variance than those of smaller houses. This is not surprising, and as you can see, heteroscedasticity is not an error but rather a characteristic of the data itself. The model seems to perform better for small houses than for bigger ones.



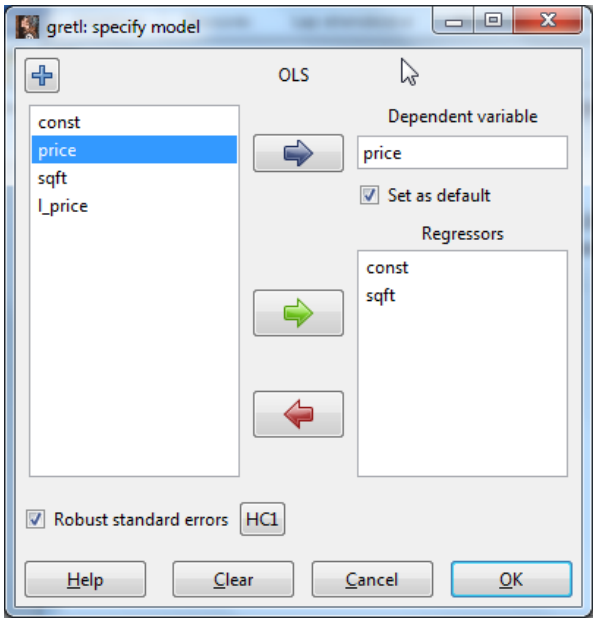
```
gretl: LM test (heteroskedasticity)

Breusch-Pagan test for heteroskedasticity
OLS, using observations 1-14
Dependent variable: scaled uhat^2

             coefficient   std. error    t-ratio   p-value
  -------------------------------------------------------
  const      -1.95436      1.48736       -1.314    0.2134
  sqft        0.00154604   0.000747274    2.069    0.0608   *

  Explained sum of squares = 10.3723

Test statistic: LM = 5.186129,
with p-value = P(Chi-square(1) > 5.186129) = 0.022768
```

You can also check the normality of the residuals under the Tests menu:

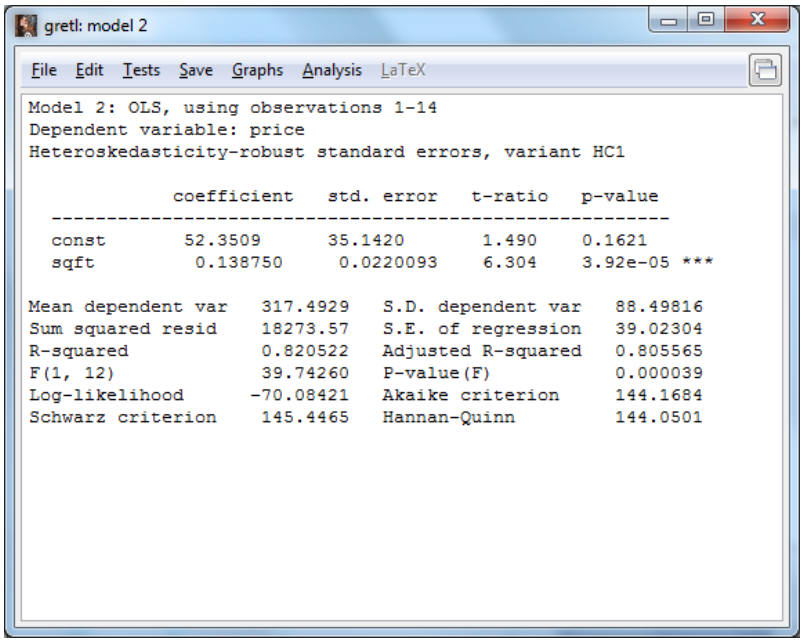


Even though normality itself is not a crucial assumption, with only 14 observations we cannot expect that the distribution of the coefficients is close to normal unless the dependent variable (and the residual) follows a normal distribution. Hence this is a good news.

Heteroscedasticity is a problem though inasmuch as it may affect the standard errors of the coefficients, and may reduce efficiency. There are two solutions. One is to use OLS (since it is still unbiased), but have the standard errors corrected for heteroscedasticity. This you can achieve by reporting heteroscedasticity robust standard errors, which is the popular solution. Go back to the Model menu, and OLS, and have now robust standard errors selected:



While the coefficients did not change, the standard errors and the t-statistics did.



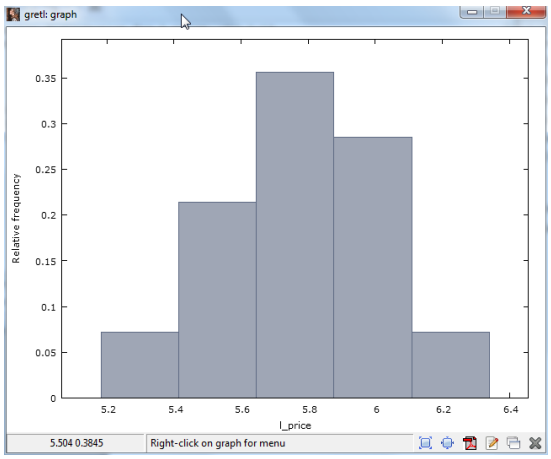
An alternative way is to transform you data so that the homoscedasticity assumption becomes valid again. This requires that observation with higher residual variance are given lower weight, while observations where the residual variance was lower are given a relatively higher weight. The resulting methodology is called the Weighted Least Squares (WLS) or sometime it is also referred to the Feasible Generalized Least Squares (FGLS). You can achieve this option in the Model menu under “Other linear models” as “heteroscedasticity corrected”.

```
gretl: model 3
File  Edit  Tests  Save  Graphs  Analysis  LaTeX

Model 3: Heteroskedasticity-corrected, using observations 1-14
Dependent variable: price

             coefficient   std. error   t-ratio    p-value
  -------------------------------------------------------
  const       51.5813      8.89621       5.798    8.50e-05  ***
  sqft         0.139439    0.00809705   17.22     7.94e-010 ***

Statistics based on the weighted data:

Sum squared resid    39.51385   S.E. of regression   1.814613
R-squared            0.961110   Adjusted R-squared   0.957869
F(1, 12)             296.5631   P-value(F)           7.94e-10
Log-likelihood      -27.12830   Akaike criterion     58.25659
Schwarz criterion    59.53471   Hannan-Quinn         58.13828

Statistics based on the original data:

Mean dependent var   317.4929   S.D. dependent var   88.49816
Sum squared resid    18279.82   S.E. of regression   39.02971
```

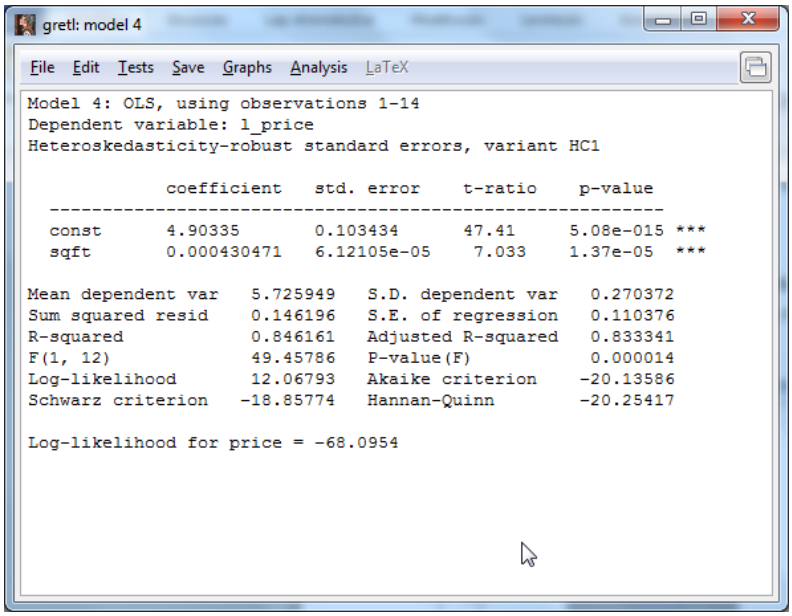
You may wonder which one is better? To transform your data or rather to have only you standard errors corrected and stick with the OLS. You can see that there may be significant changes in the t-statistics, while the coefficients are basically the same. Hence it is often the case that you do not gain much by reweighting your data. Nevertheless, theoretically both are correct ways to treat heteroscedasticity. The majority of articles report robust statistics and does not do WLS, partly for convenience, and partly because there is some degree of distrust toward data that has been altered: do not forget that once you weight your data it is not the same data anymore, but, in this particular case, all your variables (including the dependent variable) will be divided by the estimated standard error of the residual for that particular observation.

2.3 Alternative specifications

We may also try a different specification: since the prices are skewwed to the right, a logartihmic transformation may just bring it more to the center. For a visuals look of the idea, let us look at the histogram of log prices.



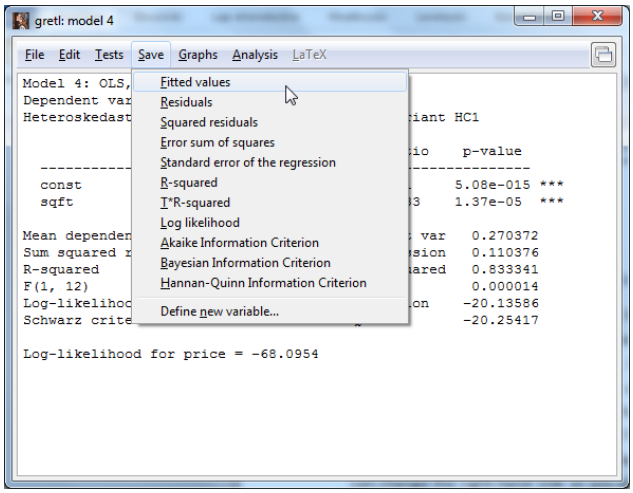
Indeed it look much more centered than prices. This may have improving effects on the model, even though there is no guarantee. We need to estimate the alternative model and compare it with the original one.



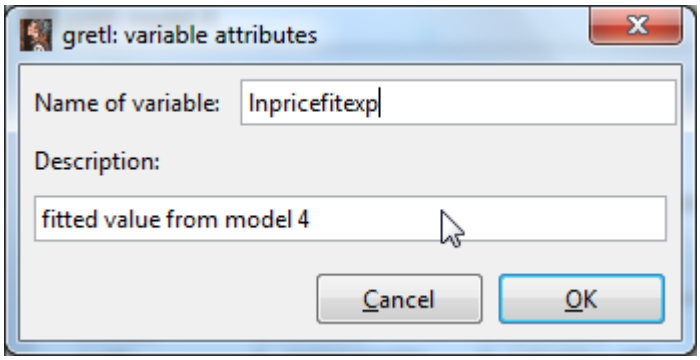
```
gretl: model 4
File  Edit  Tests  Save  Graphs  Analysis  LaTeX

Model 4: OLS, using observations 1-14
Dependent variable: l_price
Heteroskedasticity-robust standard errors, variant HC1

              coefficient   std. error    t-ratio    p-value
  ---------------------------------------------------------
  const       4.90335       0.103434      47.41      5.08e-015 ***
  sqft        0.000430471   6.12105e-05    7.033     1.37e-05  ***

Mean dependent var   5.725949   S.D. dependent var   0.270372
Sum squared resid    0.146196   S.E. of regression   0.110376
R-squared            0.846161   Adjusted R-squared   0.833341
F(1, 12)             49.45786   P-value(F)           0.000014
Log-likelihood       12.06793   Akaike criterion    -20.13586
Schwarz criterion   -18.85774   Hannan-Quinn        -20.25417

Log-likelihood for price = -68.0954
```

We obtain now a statistically significant constant term, saying that a building with null area would sell at exp(4.9)=134.3 thousands dollar (this is a case when the constant seemingly has no deep economic meaning, even though you may say that this is the price of location, or effect of fixed cost factors on price), and every square feet additional area would increase the price by 0.04% on average (do not forget to multiply by 100% if log is in the left-hand side). You may be tempted to say that this log-lin (or exponential) specification is better than the linear specification, simply, because its R^2 exceeds that of the original specification. This would be a mistake though. All goodness of fit statistics, including R^2, the log-likelihood, or the information criteria (Akaike, Schwarz and Hannan-Quinn) are dependent on the measurement unit of the dependent variable. Hence, if the dependent variable does not remain the same, you cannot use these for a comparison. They can only be utilized for model selection (i.e. telling which specification describes the dependent variable better) if the left-hand side of the regression remains the same, albeit you can change the right-hand side as you please.

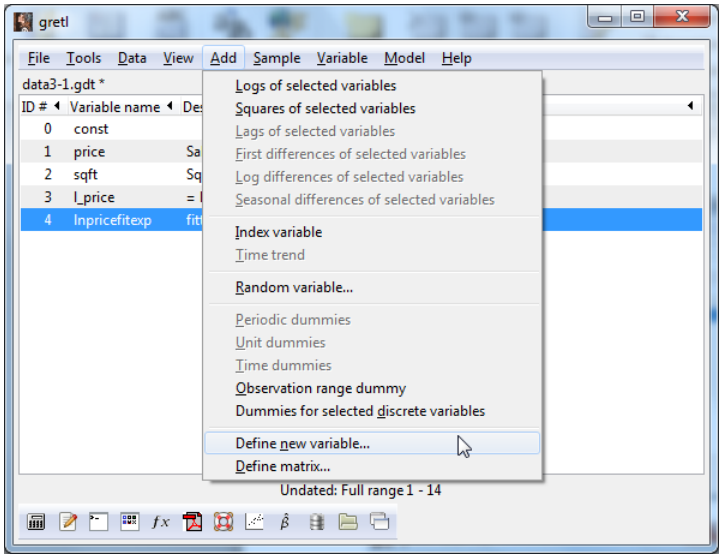
In such situations when the dependent variable has been transformed, the right way to compare different models is to transform the fitted values (as estimated from the model) to the same units as the original dependent variable, and look at the correlation between the original variable and the fitted values from the different specifications. Hence, now, we should save the fitted values from this regression, than take its exponential, so that it is in thousand dollars again, and look at the correlation with the dependent variable. Saving the fitted values is easy in GRETL:

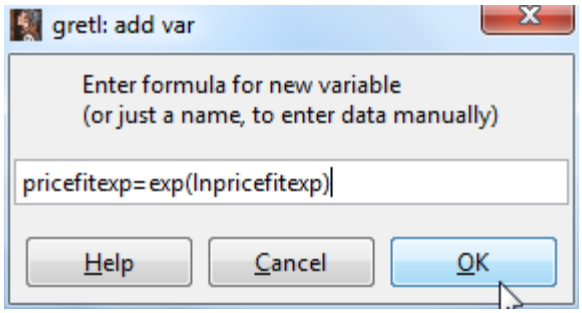


Let us call the fitted values lnpricefitexp:

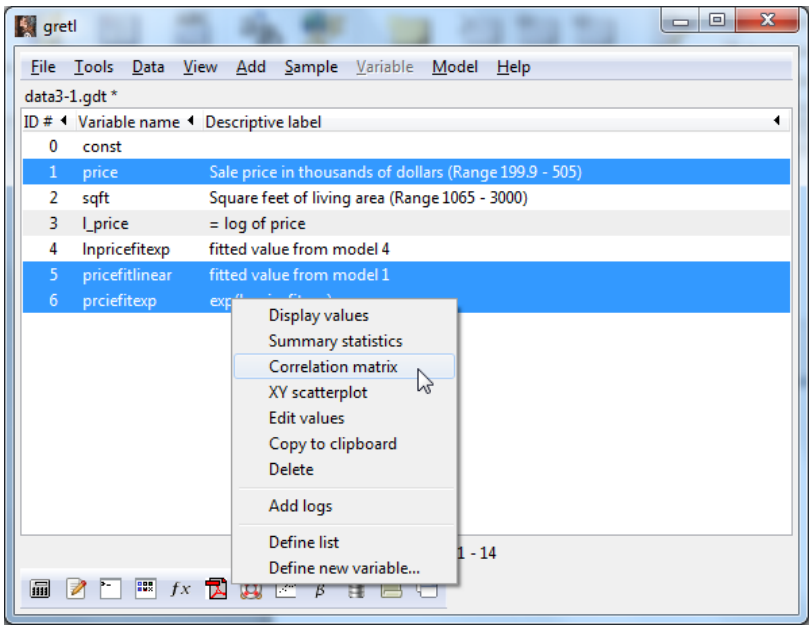


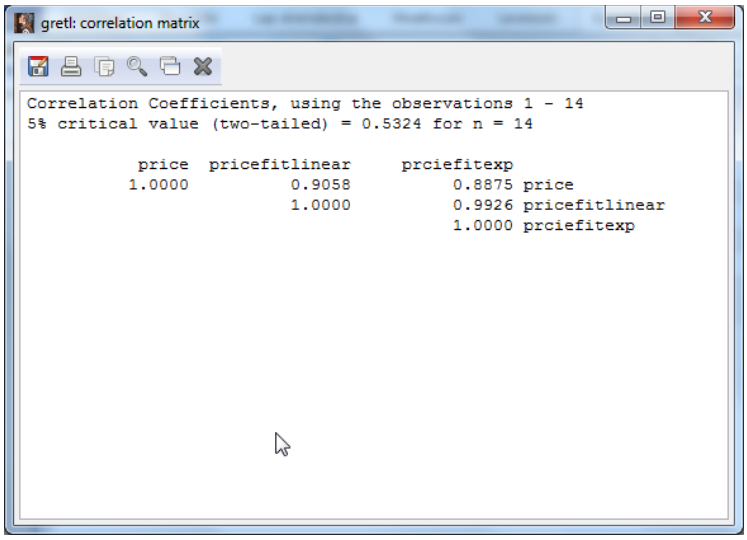
Now we have the fitted values from the exponential model as a new variable. Let us take the exponential of it:





Let us also save the fitted values from the linear model as pricefitlinear. We can now estimate the correlations:





We can observe that the linear correlation coefficient between price and the fitted price from the linear model is 0.9058, while the correlation between the price and the fitted price from the exponential specification is 0.8875. Hence the linear model seems better.

3. Multivariate regression

3.1 Some important motivations behind multivariate regressions

Life is not two-dimensional so two-variate regression are rarely useful. We need to continue into the realm of multivariate regressions.

As you have seen in the lecture notes on OLS, multivariate regressions has the great advantage that the coefficients of the explanatory variables can be interpreted as net or ceteris paribus effects. In other words, the coefficient of variable x can be seen as the effect of x on the dependent variable y with all other explanatory variables fixed. This is similar to the way of thinking behind comparative statics. But beware! This is only true for variables which are included in the regression. If you omit variables that are important and correlated with the variables that you actually included in your regression, the coefficients will reflect the effect of the omitted variable too. Let us take the two-variate regression from section 2 as an example. It is quite obvious that house prices depend not only on the area of houses, but also on the quality of buildings, their distance from the city centre, or the number of rooms, etc. By including the area only as explanatory variable, you do not really measure the effect of area on the sale prices, but rather the total effect of area, including part of the effect of other factors that are not in your model but are related to area. If, for example, bigger houses are usually farther from the city centre, then the coefficient of the area will not only reflect that bigger houses are more valuable, but also that bigger houses are further away from the centre and hence their prices should be somewhat lowered because of this. The total effect of area on house prices should then be lower than the net effect, which would be free of the distance effect). You can observe this simply, by introducing new variables into a specification. You will have a big chance that important additional explanatory variables will change the coefficient of other explanatory variables.

Be therefore very well aware of the problem of omitted variables and the resulting bias. You can only interpret a coefficient as the net effect of that particular factor, if you have included all important variables in your regression, or you somehow removed the effect of those omitted variables (panel analysis may offer that).

3.2 Estimating a Mincer equation

We need now a different sample file: open wage2 of the Wooldridge datafiles.

You will find observations on socio-economic characteristics of 526 employees. The task is to find out how these characteristics affect their wages. These type of models are called Mincer equation, after Jacob Mincer's empirical work. The basic specification is as follows:

$$\ln wage_i = \beta_0 + \beta_1 educ_i + \sum_{j=2}^{k} \beta_j X_{ji} + u_i$$

where the coefficient of the education (usually but not exclusively expressed as years of education) is the rate of returns to education, and X denote a number of other important variables affecting wage such as gender, experience, race, job category, or geographical position.

Let us estimate the coefficient with all available variables.

gretl: model 2

File Edit Tests Save Graphs Analysis LaTeX

```
Model 2: OLS, using observations 1-526
Dependent variable: lwage
Heteroskedasticity-robust standard errors, variant HC1

              coefficient    std. error    t-ratio     p-value
  ---------------------------------------------------------------
  const        0.893125      0.123116       7.254     1.54e-012 ***
  educ         0.0467910     0.00898946     5.205     2.83e-07  ***
  exper        0.0254056     0.00506079     5.020     7.18e-07  ***
  tenure       0.0223215     0.00663245     3.366     0.0008    ***
  nonwhite    -0.00426773    0.0566830     -0.07529   0.9400
  female      -0.267974      0.0366919     -7.303     1.11e-012 ***
  married      0.0562608     0.0389164      1.446     0.1489
  numdep      -0.0215152     0.0130174     -1.653     0.0990    *
  smsa         0.138730      0.0394695      3.515     0.0005    ***
  northcen    -0.0584407     0.0441015     -1.325     0.1857
  south       -0.0444269     0.0433845     -1.024     0.3063
  west         0.0545441     0.0556100      0.9808    0.3271
  construc    -0.0528537     0.0905886     -0.5834    0.5599
  ndurman     -0.107439      0.0557239     -1.928     0.0544    *
  trcommpu    -0.0961487     0.0690703     -1.392     0.1645
  trade       -0.303270      0.0508095     -5.969     4.52e-09  ***
  services    -0.309147      0.0746331     -4.142     4.03e-05  ***
  profserv    -0.0951315     0.0557141     -1.707     0.0883    *
  profocc      0.224838      0.0466222      4.823     1.88e-06  ***
  clerocc      0.0383129     0.0529188      0.7240    0.4694
  servocc     -0.0944223     0.0556291     -1.697     0.0902    *
  expersq     -0.000529441   0.000107528   -4.924     1.15e-06  ***
  tenursq     -0.000373420   0.000232230   -1.608     0.1085

Mean dependent var   1.623268   S.D. dependent var   0.531538
Sum squared resid    66.66293   S.E. of regression   0.364048
R-squared            0.550576   Adjusted R-squared   0.530919
F(22, 503)           33.52886   P-value(F)           4.85e-84
Log-likelihood      -203.0952   Akaike criterion     452.1903
Schwarz criterion    550.2922   Hannan-Quinn         490.6015

Log-likelihood for wage = -1056.93

Excluding the constant, p-value was highest for variable 5 (nonwhite)
```

What we find is a reasonable R^2, and a lot of statistically insignificant variables. We will discuss how to reduce our model later, but let us first review the interpretation of the coefficients.

Since we have log wages on the left-hand side, the effect of explanatory variables should be interpreted as relative effects. For example the educ coefficient is 0.047, that is, if we have two employees who has the same gender, work in the same field, has the same experience and work at their present employer for the same time, the one with one year more education will have 4.7% higher salary on average. The female dummy is statistically significant and negative -0.268. The interpretation is, that if all other factors are the same, being woman will cause the wage to be exp(-0.268)-1=-0.235, that is 23.5% lower than for a man. We do not find a comparable result for race.

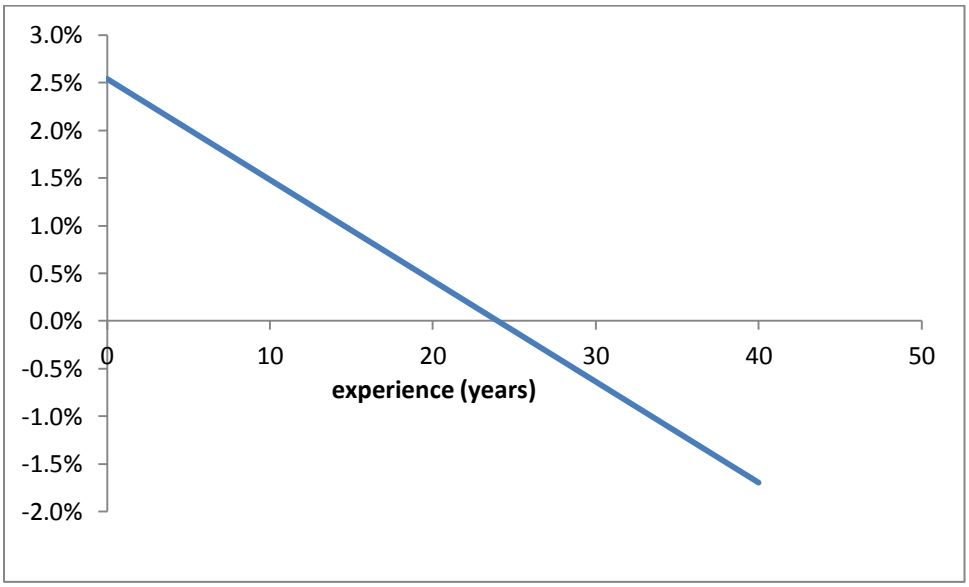
Another interesting feature of the model is the presence of squared explanatory variables, or quadratic function forms. You can see that it is not only experience that is included but also its square. This functional form is often used to capture non-linearities, i.e., when the effect of an explanatory variable depends on its own values as well. For example, there is reason to believe that experience has a large effect on wages initially, but at later phases this effect fades away. This is simply because the first few years are crucial to learn all those skills that are necessary for you to be an effective employee, but once you have acquired those skills, your efficiency will not improve much simply by doing the same thing for a longer time. This is what we find here as well. The positive coefficient of the experience suggests that at low levels of experience, any further years have a

positive impact on wages (2.5% in the first year), but this diminishes as shown by the squared experience. If we have a specification as follows:

$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + u_i$ then the marginal effect of x can be calculated as follows.

$$\frac{dy}{dx} = \beta_1 + 2\beta_2 x_i$$

We can use above expression to plot a relationship between the effect of experience on wage and experience:



After a while, it may even be that the effect of experience is negative in the wage, even though this may simple be because we force a quadratic relationship onto out data.

3.3 Model selection

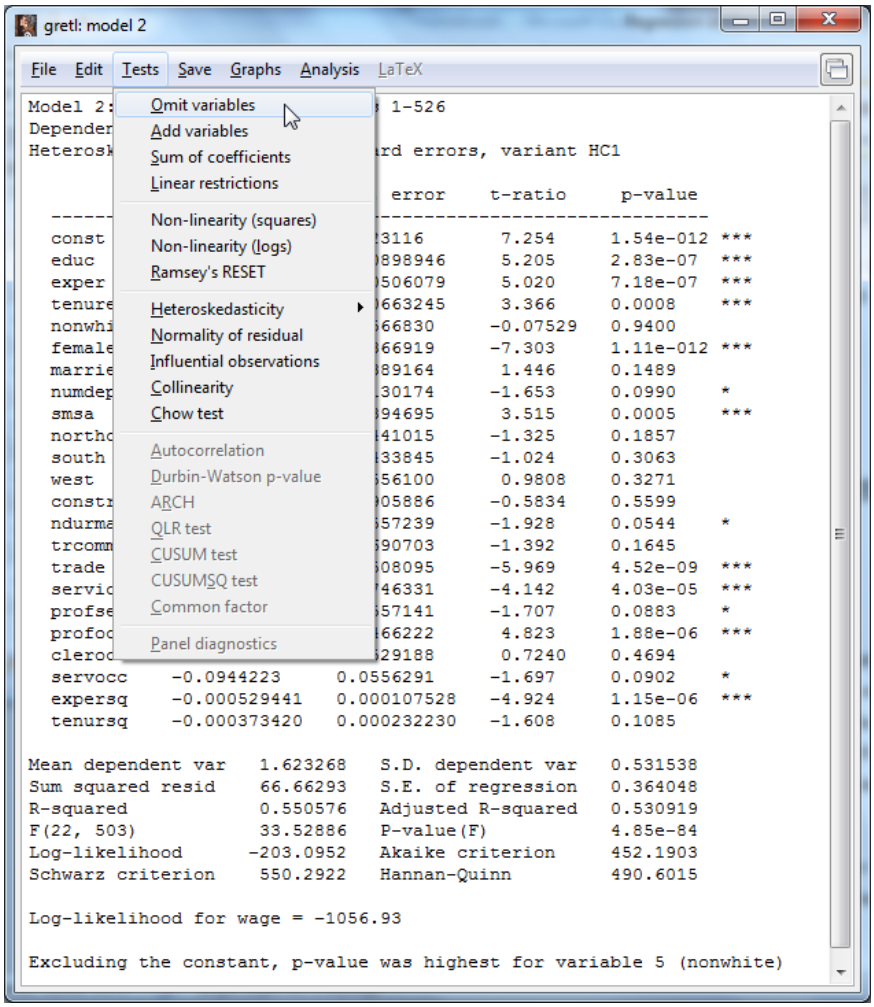
Should we or should we not omit the variables that are not significant at at least 10%? This is a question that is not easy to answer.

Statistically speaking, if we include variables that are not important (their coefficients are statistically not significant) we will still have unbiased results, but the efficiency of the OLS estimator will reduce. Hence, we can have that by including non-essential variables in our regression, an important variable will look statistically insignificant. Hence purely statistically speaking removing insignificant variables is a good idea. Yet, very often you will find that insignificant coefficients are still reported. The reason is that sometimes having a particular coefficient statistically insignificant is a result by its own right, or the author wishes to show that the results are not simply due to omitted variable bias, and hence leave even insignificant variables in the specification to convince the referees (who decide if an article is published or not). It may be a good idea though to report the original (or unrestricted) specification and a reduced (or restricted) specification as well.

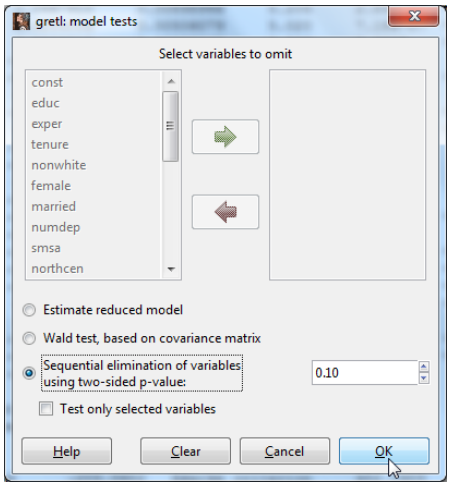
But which way is the best? Should you start out with a single explanatory variable and keep adding new variables, or rather should you start with the most complex model, and reduce it by removing

insignificant variables until you have all variables statistically significant. This question can be answered very simply. It should be the second way. Do not forget, that having unnecessary variables does not cause a bias in your parameter estimates, while omitting important variables does. Hence if you start out with single variable, you statistics on which you base your decision if you should add or remove a variable will be biased too. Having less efficient but unbiased estimates is now the lesser bad.

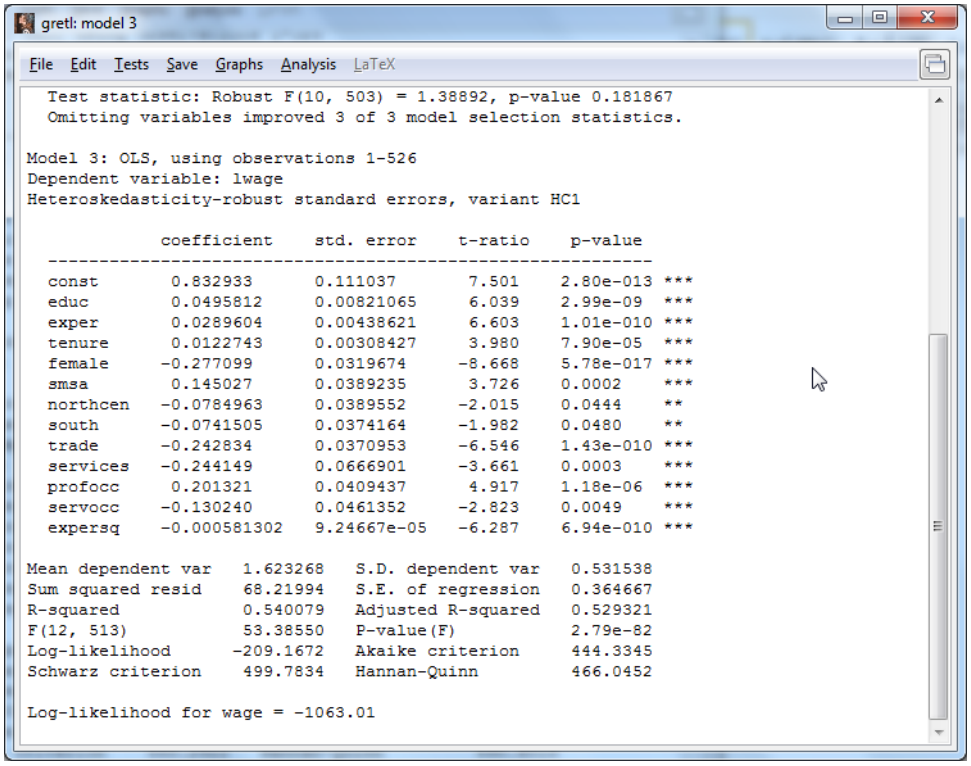
The standard method is to reduce you model by a single variable in each step. It is logical to remove the variable with the highest p-value. The process can also be automatized such as in GRETL. The omit variables option in the Tests menu allows you to choose automatic removal of variables.



This option allows you to assign the p-value at which a variables should be kept in the specification. Choosing this value 0.10 means that only variables that are significant at at least 10% will be retained in the specification.



The resulting specification is:




```
  Test statistic: Robust F(10, 503) = 1.38892, p-value 0.181867
  Omitting variables improved 3 of 3 model selection statistics.

Model 3: OLS, using observations 1-526
Dependent variable: lwage
Heteroskedasticity-robust standard errors, variant HC1

             coefficient    std. error    t-ratio    p-value
  ---------------------------------------------------------
  const       0.832933      0.111037       7.501    2.80e-013 ***
  educ        0.0495812     0.00821065     6.039    2.99e-09  ***
  exper       0.0289604     0.00438621     6.603    1.01e-010 ***
  tenure      0.0122743     0.00308427     3.980    7.90e-05  ***
  female     -0.277099      0.0319674     -8.668    5.78e-017 ***
  smsa        0.145027      0.0389235      3.726    0.0002    ***
  northcen   -0.0784963     0.0389552     -2.015    0.0444    **
  south      -0.0741505     0.0374164     -1.982    0.0480    **
  trade      -0.242834      0.0370953     -6.546    1.43e-010 ***
  services   -0.244149      0.0666901     -3.661    0.0003    ***
  profocc     0.201321      0.0409437      4.917    1.18e-06  ***
  servocc    -0.130240      0.0461352     -2.823    0.0049    ***
  expersq    -0.000581302   9.24667e-05   -6.287    6.94e-010 ***

Mean dependent var   1.623268   S.D. dependent var   0.531538
Sum squared resid    68.21994   S.E. of regression   0.364667
R-squared            0.540079   Adjusted R-squared   0.529321
F(12, 513)           53.38550   P-value(F)           2.79e-82
Log-likelihood      -209.1672   Akaike criterion     444.3345
Schwarz criterion    499.7834   Hannan-Quinn         466.0452

Log-likelihood for wage = -1063.01
```