Statistical methods for Education Economics

Massimiliano Bratti
http://www.economia.unimi.it/bratti

Course of *Education Economics*
Faculty of Political Sciences, University of Milan
Academic Year 2007-08

November 26, 2008

The econometric software GRETL used in these notes is *open source* and freely downloadable at http://gretl.sourceforge.net/. The user guide is downloadable at the same address.

Contents

Introduction

Suppose that we want to answer the following question: Do university graduates earn more *on average* than high school diplomats?

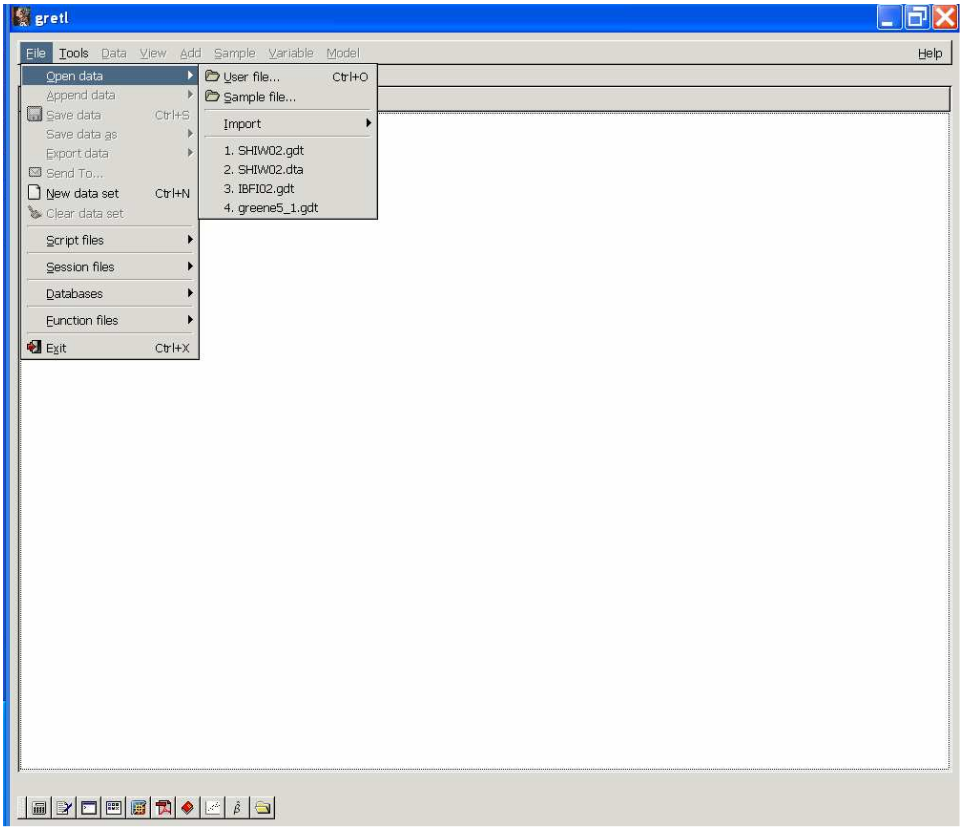
In order to give an answer to this question we need to have data simultaneously reporting both individuals' education and income. An example is the Bank of Italy's Survey of Household Income and Wealth[1] (SHIW, hereafter).

With these data, a simple way to answer is by computing the mean of yearly income of individuals with a university a degree and that of individuals with only secondary schooling.

We use the SHIW 2002 data to have an idea of the differences in income between graduate workers and workers with a high school diploma in the same year (2002). We focus on individuals aged 18-65 (working age). *yl* is yearly labour income in € and *degree* is a dummy variable that takes on value one for individuals with a degree and zero otherwise.

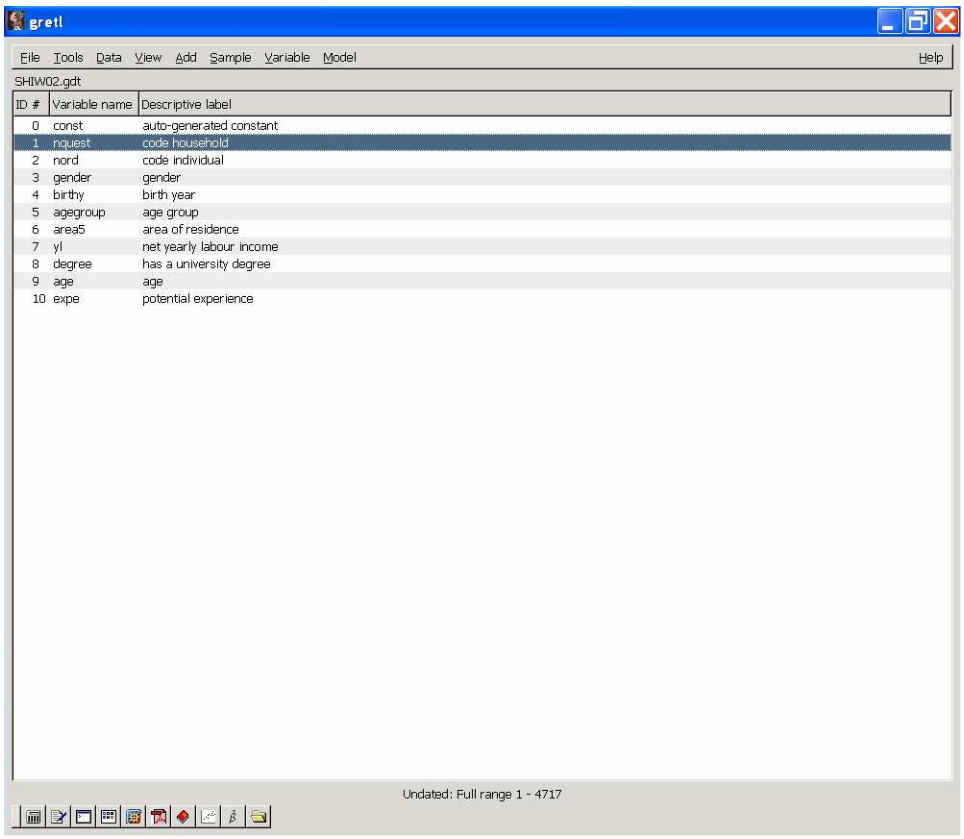
Before going on, we have to open the data file after having launched the econometric software GRETL, in the following way. Data files in GRETL have the extension .gdt, the file to be open is called SHIW02.

[1] *Indagine sui Bilanci delle Famiglie Italiane*, in Italian.

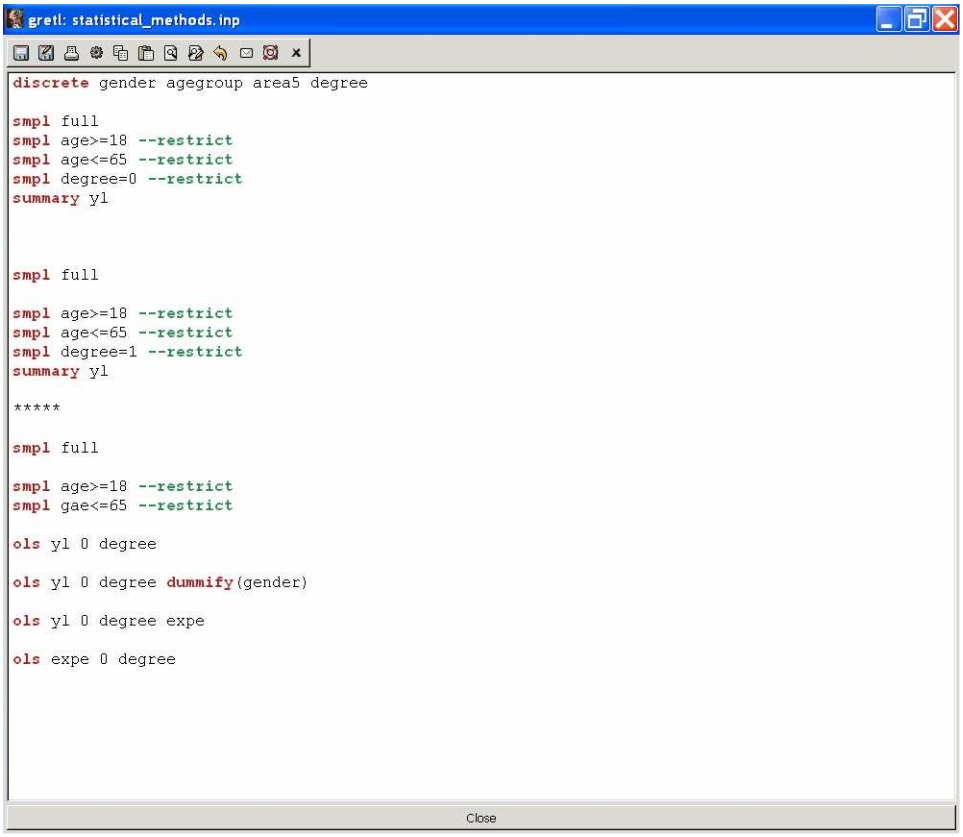


The command in the menu loads the data. GRETL will ask you if you want to attribute to the data a *time-series* or *panel* format, please answer no, since here we are using *cross-section* data, i.e. data that refer to several individuals observed in the same year. *time-series* data refer only to one individual and several years, a *panel data* refers to many individuals and many years.

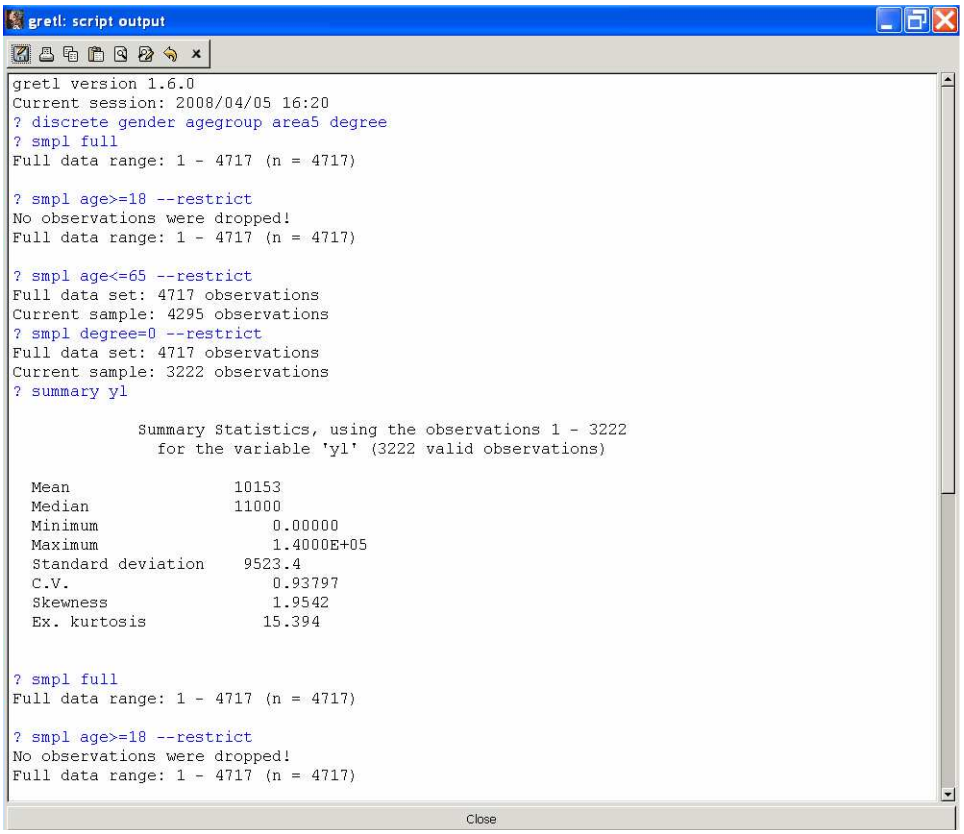
The list of variables and labels in the data file is:

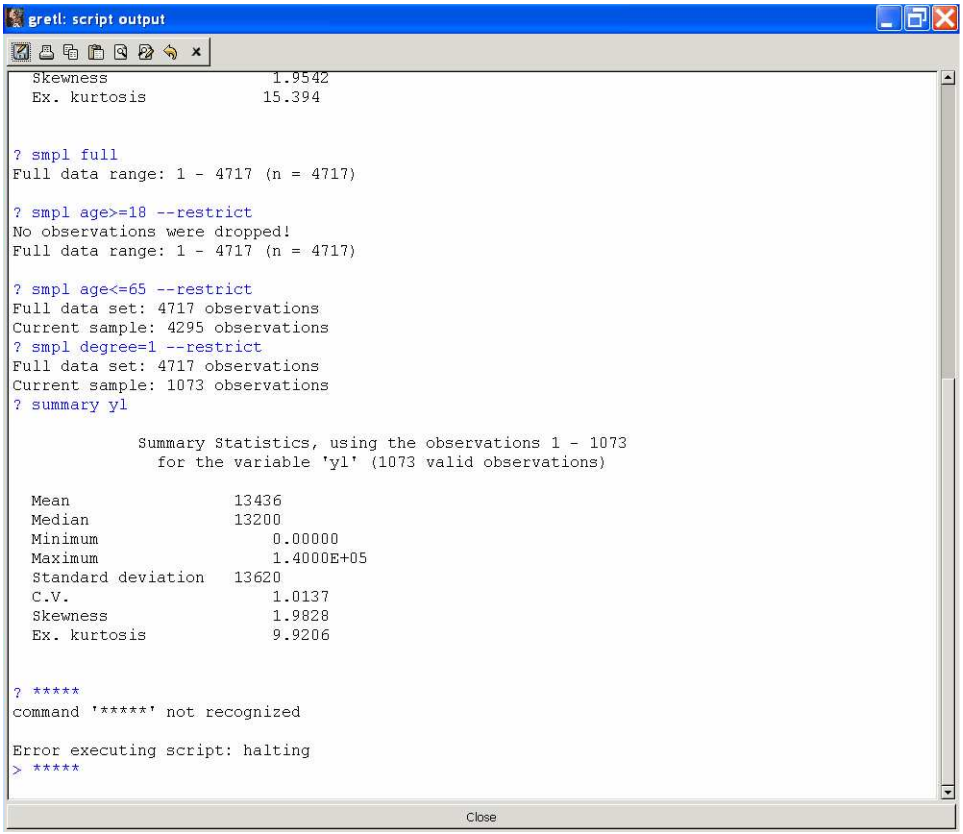


Then, we open a *script* file (or syntax file) in which we can write GRETL commands. In the example we use the command `discrete` that declares a variable to be discrete, `smpl` that selects the sample and `summary` that shows the sample descriptive statistics. The commands syntax is the following



And the descriptive statistics are:


gretl: script output
gretl version 1.6.0
Current session: 2008/04/05 16:20
? discrete gender agegroup area5 degree
? smpl full
Full data range: 1 - 4717 (n = 4717)
? smpl age>=18 --restrict
No observations were dropped!
Full data range: 1 - 4717 (n = 4717)
? smpl age<=65 --restrict
Full data set: 4717 observations
Current sample: 4295 observations
? smpl degree=0 --restrict
Full data set: 4717 observations
Current sample: 3222 observations
? summary yl
Summary Statistics, using the observations 1 - 3222
for the variable 'yl' (3222 valid observations)
Mean 10153
Median 11000
Minimum 0.00000
Maximum 1.4000E+05
Standard deviation 9523.4
C.V. 0.93797
Skewness 1.9542
Ex. kurtosis 15.394
? smpl full
Full data range: 1 - 4717 (n = 4717)
? smpl age>=18 --restrict
No observations were dropped!
Full data range: 1 - 4717 (n = 4717)
Close



What can we observe from the data tabulation?

- Individuals without a degree earn on average 10,153 € while those with a degree 13,436 €. Hence, individuals with a degree earn on average 3,283 € more than those with secondary schooling only.
- Given that the result is obtained from a *sample survey* and not from the whole population we have to evaluate if: 1) the two measures are representative of the whole Italian pupulation; 2) if the two means are statistically different. We will see later on how to answer these questions.

Hence, if we do want to estimate the 'causal' effect of higher education (i.e. having a degree) on income (in the sense that it is educations that causes the increase in income) we run into two problems:

- the problem of **statistical inference**. Since in general we do not have data for all individuals in the population, in what conditions can we infer something on the whole population starting from our sample data? It is clear that the sample must be *representative* (random sample, big enough, etc.).

- the problem of *identification*. It is correct on the basis of the simple comparison of mean income to infer anything on the *causal* effect of education on income? What could happen?

  Individuals with a degree could systematically differ from those without a degree with respect to both observable and unobservable characteristics (in the data). An example of the former is family wealth. Children of richer parents could both demand more education and have access to better jobs (i.e. jobs paying higher wages) through, for instance, social networks. As to the latter (unobservable characteristics), abler individuals (i.e. higher IQ which is not usually measured in the data) could demand more education (for them it is easier to achieve higher education) and also have higher incomes (if intelligence is rewarded by employers in terms of higher wages). In both cases we could observe a *spurious* positive correlation between higher education and income. The solution to the problem, we'll see, it is different in the two cases.

Covariance

Before introducing the concept of **linear regression** it is important to review some statistical concepts:

- **Mean** of income W (wages). $\overline{W} = \frac{1}{n}\sum_i^n W_i$
- **Variance** of income W. $var(W) = \frac{1}{n}\sum_i^n (W_i - \overline{W})^2$, is a measure of dispersion of the distribution
- **Covariance** between two random variables S (schooling) and W (wages). $cov(SW) \equiv \sigma_{SW} = \frac{1}{n}\sum_i^n (S_i - \overline{S})(W_i - \overline{W})$. It is a measure of linear association between the two random variables. If $\sigma_{SW} > 0$ the association is positive, i.e. if S increases also W increases (i.e. they move in the same direction). If $\sigma_{SW} < 0$ the association is negative, if S increases W decreases (i.e. they move in opposite directions).

Let us consider S a a dummy variable for having a degree (*schooling*) and W the level of yearly income (or *wage*).

Linear regression

In what follows, S is a variable which takes value one in case an individual has a degree and zero otherwise, while W is the level of yearly labour incomes.

We aim to study at the empirical level the relation between W and S. From an empirical viewpoint we often consider $w = ln(W)$ instead of W

where $ln(\cdot)$ is the 'natural logarithm'.
Let us define:

- **the expected value of W conditional on $S = s_0$**: $E(w|S = s_0)$, is the average value of incomes (W) among the individuals with education s_0. $E(w|S)$ is a function, or depends on, S, in the sense that if S varies, also w varies. Examples: $E(w|S = 0)$ is the average income among individuals without a degree, $E(w|S = 1)$ is the average income among those who have a degree.

For simplicity, we can assume that W is a linear function of S, that is:

$$W_i = \alpha + \beta S_i + \epsilon_i \tag{1}$$

where i is the subscript for the individual, α is the intercept, β the slope coefficient and ϵ_i a stochastic error term (also called residual or disturbance), which enters the relation since the linear relation that we will be able to find will be only imperfect (i.e. we are likely to omit some relevant variables). W_i the variable appearing on the left-hand-side (LHS) is called the **dependent variable** (or the variable that we want to explain), while the one (or those) appearing on the right-hand-side (RHS) is the **independent variable**, or **explanatory variable**. The RHS variables are sometimes also called **regressors** or **covariates**.

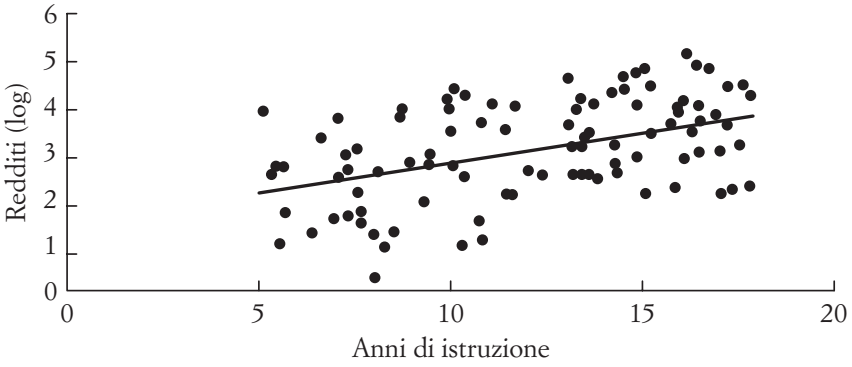
What does the relation (1) say?

- an individual's income linearly depend on education;
- the income W_i of individuals with the same level of education only differ by the stochastic term ϵ_i;
- for each year of education income increases by β units (euros in our case);
- $E(w_i|S_i) = \alpha + \beta E(S_i|S_i) + E(\epsilon_i|S_i)$. In the classical linear model we assume that the last addend is zero, i.e. that S_i does not give any information on ϵ_i, i.e. on all variables which do affect income and have been omitted from the regression. Then if $E(w_i|S_i) = \alpha + \beta S_i$ [note that we have used the fact that $E(S_i|S_i) = S_i$] $E(w_i|S_i)$ is called the regression function of W_i on S_i. We will see that the assumption that $E(\epsilon_i|S_i) = 0$ is not always a credible one and that it might often fail.

With these assumptions we can now compute (estimate) the parameters (or coefficients) of the linear regression, i.e. α and β. From a graphical point of view we put education on the horizontal axis and income on the vertical one. Estimating a regression is equivalent to drawing the straight line that better approximates the linear relation existing between education

and income as represented by the dots in the graph (*Redditi*=incomes, *Anni di istruzione*=years of education).

FIG. 19.1. DIAGRAMMA DI DISPERSIONE DEI REDDITI E DELL'ISTRUZIONE (ESPRESSA IN ANNI DI SCUOLA COMPIUTI).



Nota: È rappresentata la retta interpolante.

il Mulino

How to estimate the parameters? OLS

The method that is used to 'draw this line' is the method of **ordinary least squares** (OLS, hereafter).

In general there exist various methods to estimate α and β. One of these criteria is to minimize the sum of the squares of the distances between W_i (actual incomes) and those predicted $\hat{W}_i = \hat{\alpha} + \hat{\beta}S_i$. But the distances are the **estimated residuals**.
Hence, OLS is based on the following criterion:

$$\underset{\alpha,\beta}{\text{Min}} \sum_{i}^{n} \hat{\epsilon_i}^2 \tag{2}$$

where $\hat{\epsilon_i}$ are the estimated residuals, that is $\hat{\epsilon_i} = w_i - \hat{\alpha} - \hat{\beta}S_i$.

In the simple case we are considering here, in which the depend variable W_i only depends on one explanatory variable S_i (*bivariate regression*), it is possible to show that:

- $\hat{\beta} = \frac{cov(w_i,S_i)}{var(S_i)}$, that is using OLS we estimate a positive coeffic
- $\hat{\alpha} = \overline{w} - \hat{\beta}\overline{S}$, where a bar over the variables stands for mean (i.e. expected value).

When we are including many variables in the RHS we have the case of *multivariate regression*. In this case too, it is possible to use OLS estimation.

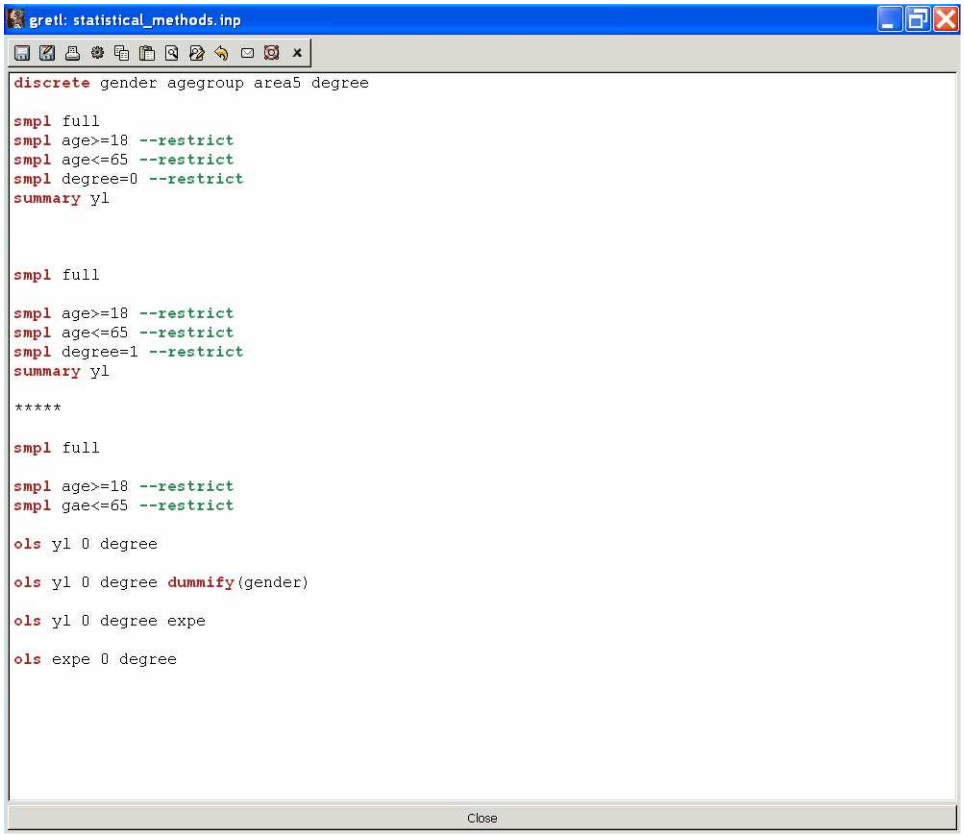
OLS: An example

Let's go back to our SHIW 2002 data and estimate, for instance, the premium in terms of yearly labour income, of graduates with respect to high school diplomats using a linear regression estimated with OLS.

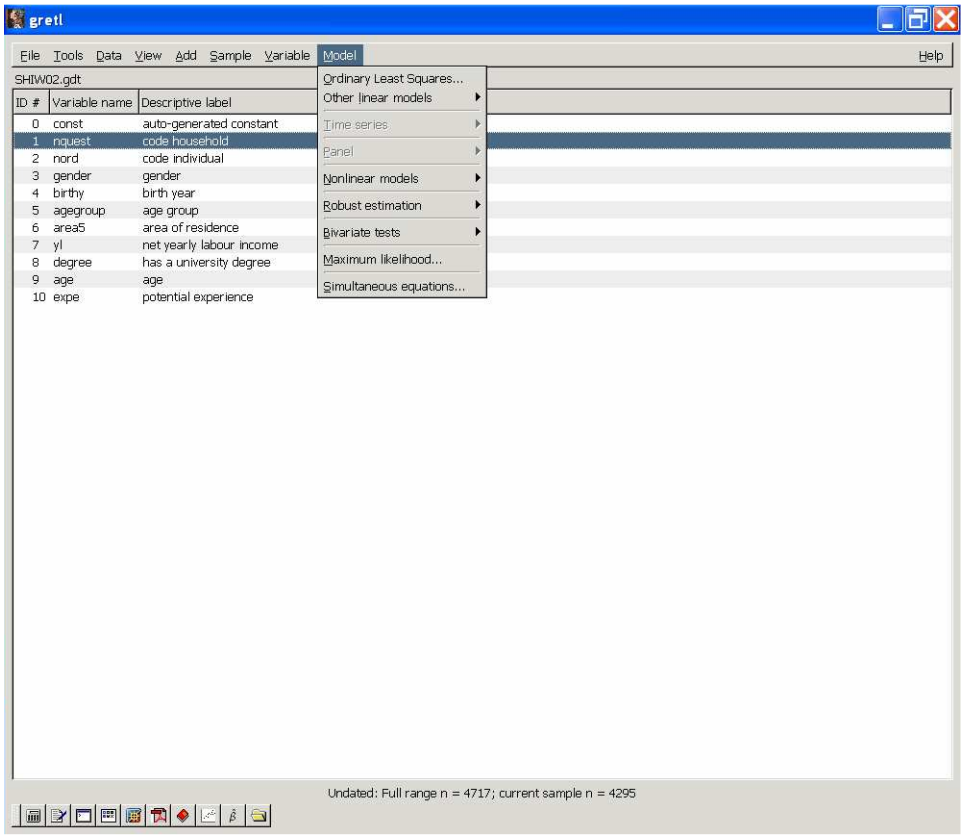
We have already said that S_i is a variable that takes value 1 if an individual has a university degree and 0 otherwise (i.e. it is a *dummy* variable). The conditional mean remains $E(w_i|S_i) = \alpha + \beta S_i$. In particular:

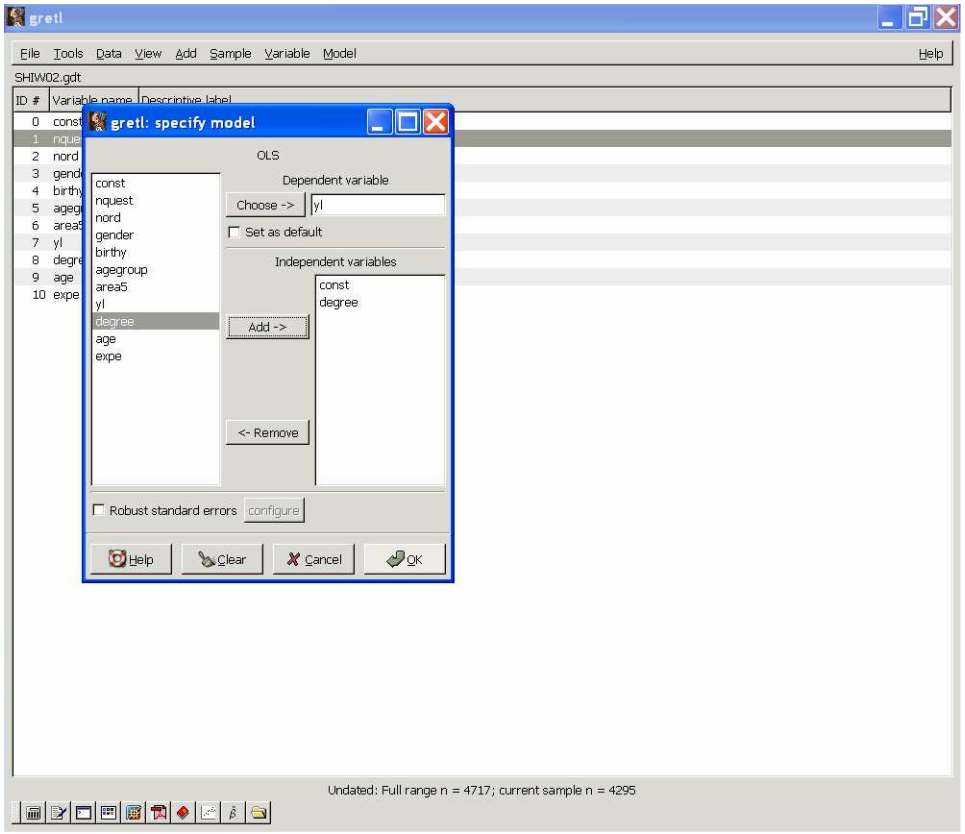
- for individuals without a degree $E(w_i|S_i = 0) = \alpha + \beta \cdot 0 = \alpha$;
- for individuals with a degree $E(w_i|S_i) = \alpha + \beta \cdot 1 = \alpha + \beta$

The command to estimate a linear regression with OLS in GRETL is:



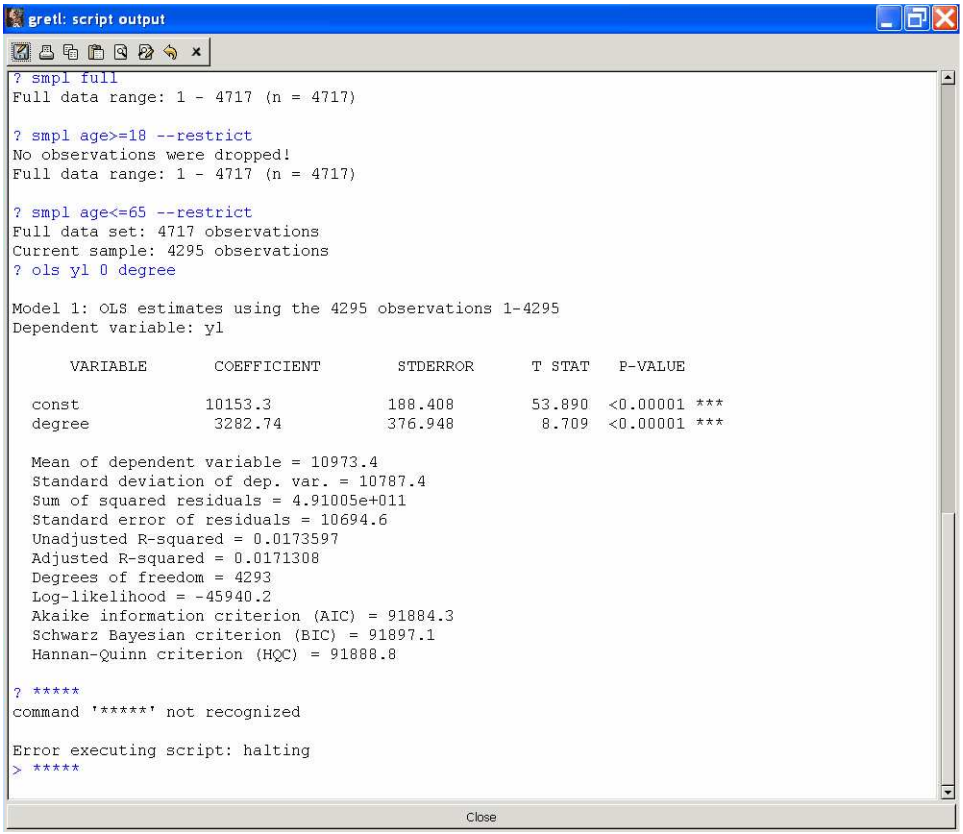
For the lazy ones there is the possibility of using GRETL's menu to estimate the regression:


gretl
File Tools Data View Add Sample Variable Model Help
SHIW02.gdt
ID # Variable name Descriptive label
0 const auto-generated constant
1 nquest code household
2 nord code individual
3 gender gender
4 birthy birth year
5 agegroup age group
6 area5 area of residence
7 yl net yearly labour income
8 degree has a university degree
9 age age
10 expe potential experience
Ordinary Least Squares...
Other linear models
Time series
Panel
Nonlinear models
Robust estimation
Bivariate tests
Maximum likelihood...
Simultaneous equations...
Undated: Full range n = 4717; current sample n = 4295



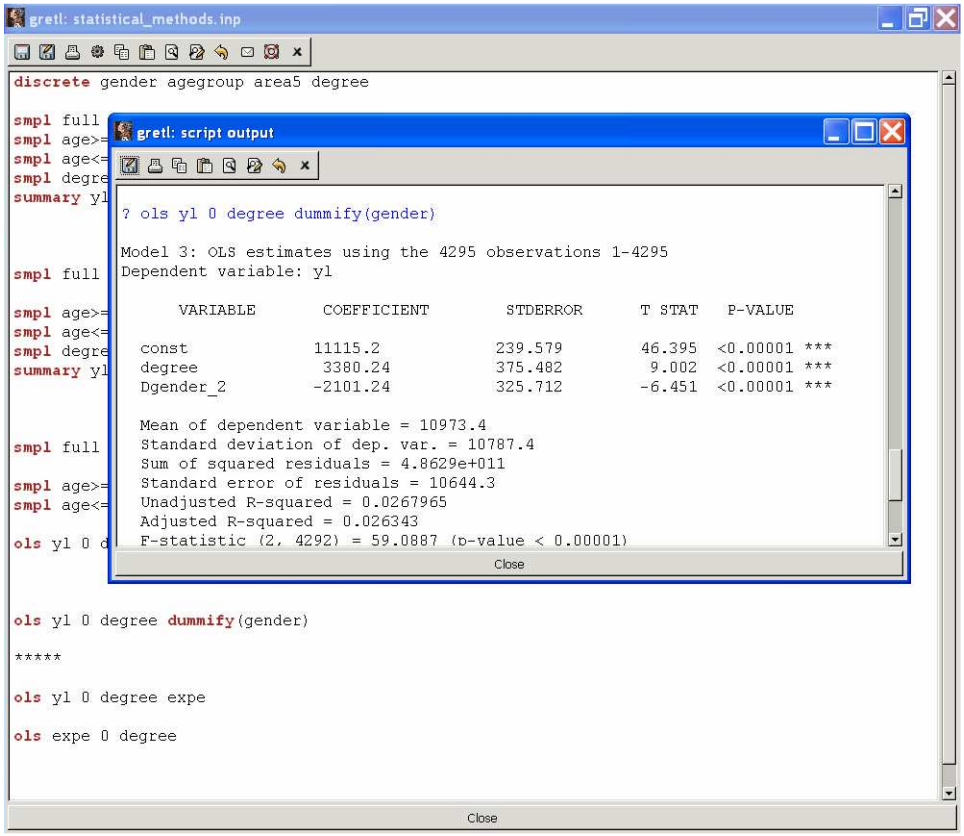
but in general it is a good rule to use script files since they allows to check what has been done (also a lot of time ago!) and correct errors without having to run all commands from the beginning. Moreover, a script file ensures 'replicability' of our work by others, or with other data.

The estimation's result is:



we get $\hat{\beta} = 3,283$. So the mean difference in yearly income between graduates and non-graduates is 3,283€, in favour of the former. To be noted that the constant is the average income of individuals without a degree [indeed if $S_i = 0$ we obtain the average income of a high school worker - $E(W_i|S_i = 0)$, this happens since in the RHS we only have an explanatory variable: education]. Then we might wonder: so much trouble for nothing? We already estimated the same difference in in income by simply comparing the means???? The answer is not. The advantage of using regression methods and OLS become clear when we use a *multivariate* regression, i.e. when we include several regressors in the RHS.

An example is the following in which we have included also a *categorical* variable for gender (=1 for males and =2 for females). We use the command `dummify` that creates a dummy variable for gender called D_sex2, which takes on value 0 for males and 1 for females (we sometimes say that the omitted category are males - the category for which the value of the dummy is zero). We now re-estimate the regression:



and obtain that on average women have a penalization of 2,101 € in their yearly labour income.

Inference and hypothesis testing

Once we have obtained an estimate of β using OLS, namely $\hat{\beta}$, we might wonder whether our estimate is a good estimate of the parameter in the population. In particular, we might be interested to know if $\hat{\beta}$ is close to the 'true' value β in the population (or how much the difference $\hat{\beta} - \beta$) is large).

Whether an estimate is good or not depends on the property of an estimator. Then, it is important to clarify two concepts:

- *estimator*. It is the *procedure* that allows to obtain an estimate of β. For instance the estimator of minimum least squares;
- *estimate*. It is a *particular value* of $\hat{\beta}$ that has been obtained by applying a given *estimator* to a given sample.

It is clear that from m different random samples of size n drawn from the same population we can obtain m different estimates of the 'true' parameter β.

An important property of the estimator OLS is that:

- it is *consistent*. If we draw from a population some samples which are increasing in size we will obtain estimates of the true parameter of increasing precision (i.e. $n \to +\infty$ then $\beta \to \hat{\beta}$, i.e. the difference between the two goes to zero).

However, this is an asymptotic property, which holds for very big samples. What about the case of small samples? The probability theory helps us to evaluate the goodness of an estimate.

In this case we use the concept of:

- *confidence interval of level x%.* Let us assume to be able to apply the OLS to all possible samples of size n, the c.i. of level x% is the interval in which in x% of cases the true value falls.

From the results of our regression we observe that a standard error is associated to each estimate, e.g. $\sigma_{\hat{\beta}} = 376.95$, which is a measure of the variability of the estimate to changing the particular sample extracted from the population. The higher the standard error, the lower the precision of the estimate. In particular, the c.i. of level 95% can be obtained as $(\hat{\beta} - 1.96\sigma_{\hat{\beta}}, \hat{\beta} + 1.96\sigma_{\hat{\beta}})$.

In our case to the estimate $\hat{\beta} = 3282.74$ is associated a 95% c.i. $(2543.73, 4021.75)$. Hence, by applying OLS to all possible samples of size n extracted from the population, in 95% the true value will fall in this interval.

Moreover, the larger $\sigma_{\hat{\beta}}$ the larger the c.i. and the estimate obtained will be less precise and of little use.

Another problem is that of evaluating whether there exists or not a **statistically significant** relationship between two variables, in our case between yearly income and education. This is equivalent to testing the hypothesis $H_0 : \beta = 0$. This hypothesis can be tested using the c.i., if the x% c.i. includes zero then the null hypothesis H_0 cannot be rejected at the 1-x% significance level. In case the 95% c.i. does not include zero then H_0 can be rejected at the 5% level of significance (5%=100%-95%). The significance level is the probability of **type I error**, tha is the probability of falsely rejecting H_0 when it is true (false positive). When we reject H_0 we can conclude that there exists a statistically significant relationship (at the 5% significance level in our case) between income and education.

In our case the 95% c.i. is $(2543.73, 4021.75)$, which excludes zero. Another way of assessing statistical significance is by looking at the $p - value$

in the estimation output, which is the probability mass to the right of the estimate of β obtained, in our case is 0.00. A p-value less or equal to 5% (0.05) implies that the particular estimate obtained falls in the rejection area of the null hypothesis (i.e. the relation is statistically significant at 5%). In particular, our results suggest the existence of a positive relation (correlation or causation?) between education and incomes.

The significance level can also be evaluated using the statistics t-Student that is obtained as $t = \frac{\hat{\beta}}{\sigma_{\hat{\beta}}}$. If $|t| > 1.96$ then the coefficient is statistically significant at 5%.

Identification

Up to now we have considered the problem of statistical inference. Now we introduce the issue of Identification. Under which conditions is it possible to identify the **causal effect** of education on income?

The estimated coefficient $\hat{\beta}$ is positive and significant. However, is it possible to conclude that education causes an increase in income? Put it in another way, can we expect that if one random individual in the population with a high school diploma gets a university degree s/he will have an increase in income of 3283 €?

Omitted variables and spurious correlations

If the estimated relationship is **causal** the answer to the question above is yes, while if the relationship is a spurious correlation the answer is no.

What is a **spurious correlation** and which variables could be responsible for it?

A spurious correlation is a correlation that does not reflect a causal relationship.

A spurious correlation may be determined by variables that have been omitted from the regression, which are correlated both with the fact of possessing a degree and the level of income. Some examples are:

- *unobserved ability* (IQ). Abler individuals (innate ability) could both get more education (since it is easier for them) and be paid a higher salary irrespective of education. This is the case for instance of the **signaling theory** of education elaborated by Spence. In this case a low ability individual should not expect an increase in income due to

education if the employer can observe ability, since what causes the increase in income (what is rewarded by the employer) is ability. In the very same way, 3,283€ may not be the average increase in income caused by education;

- *family wealth*. Richer individuals could both get more education (e.g., education is a consumption good or family wealth reduces the direct costs of education) and find better jobs, which pay higher wages, through social networks. Hence, also in this case, the 3,283€ difference in income could not be the effect of education but may be partly attributable to social networks.

Remedies for the omitted variables problem

There exists different remedies for the omitted variables problem depending on the fact whether the omitted variables are observable or not:

- *omitted variables are observable* (e.g. family wealth). In this case the remedy is obvious, if the omitted variable is available we can include it in the regression. For instance, we can include family background, measured in terms of parents' education or jobs among the covariates in the regression explaining incomes;

- *omitted variables are unobservable* (e.g. IQ). In this case it is possible to use the **instrumental variables** (IVs) estimator or the **within-group estimator**.

Omitted variables are observable

Let us assume that we have omitted from the regression an important determinant of income that is job experience, which will be indicated as X_i. In this case the 'true' model is:

$$w_i = \delta + \gamma S_i + \lambda X_i + \epsilon_i \tag{3}$$

If we take the conditional expectations of W_i on S_i we have:

$$E(w_i|S_i) = \delta + \gamma S_i + \lambda E(X_i|S_i) \tag{4}$$

while we have estimated the following regression (wrongly omitting experience):

$$E(w_i|S_i) = \alpha + \beta S_i \tag{5}$$

Let us assume that $E(X_i|S_i)$ can be expressed in turn as a linear function of S_i:

$$E(X_i|S_i) = \psi + \rho S_i \tag{6}$$

with $\rho \neq 0$.

By substituting (6) in (4) we obtain:

$$E(w_i|S_i) = (\delta + \lambda\psi) + (\gamma + \lambda\rho)S_i \tag{7}$$

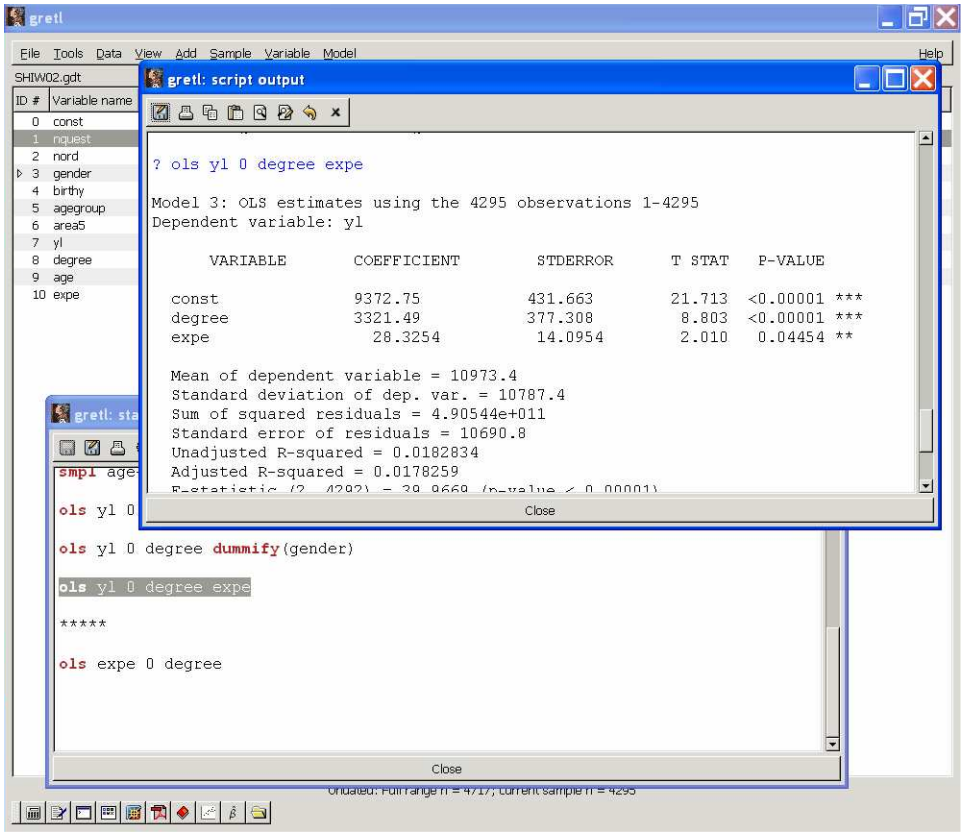
Hence if we omit X_i from the regression we obtain $\beta = \gamma + \lambda\rho$. The difference $\beta - \gamma = \lambda\rho$ is the so-called **omitted variables bias**, whose sign depends on the correlation between S_i and X_i and on the correlation between X_i and W_i. What kind of bias should we expect? In general we could expect $\lambda > 0$ since labour incomes are increasing in job experience. By contrast, ρ could be positive or negative. Indeed, if individuals with greater education have a lower probability of being unemployed, they will have a higher job experience($\rho > 0$). However, individuals with more education spend more time in the educational system and enter the labour market later, so they could have less experience ($\rho < 0$). Therefore, the sign of ρ must be determined empirically (i.e. using the data).

Sign of bias:

- if $\lambda > 0$ and $\rho > 0$ then *bias* is positive ($\lambda\rho > 0$);
- if $\lambda > 0$ and $\rho < 0$ then *bias* is negative ($\lambda\rho < 0$).

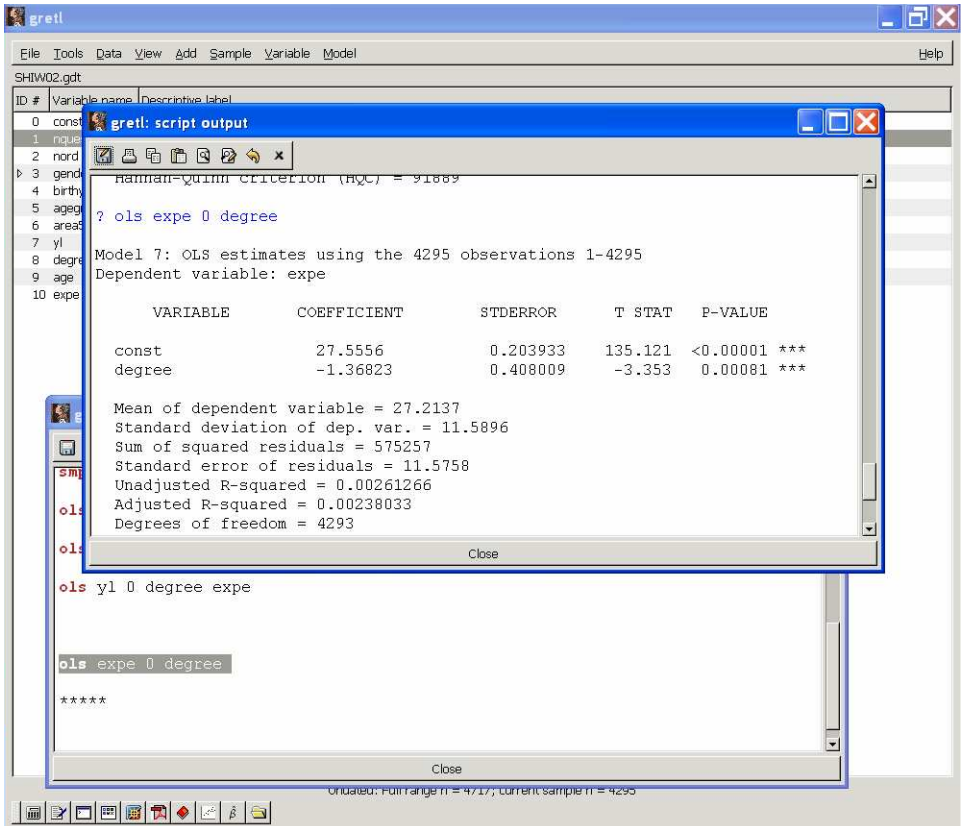
What happens in our SHIW sample if we do include job experience among the regressors. Since we do not have actual job experience, we have used 'potential experience' (like it is often done in the literature), which is given by $EXPE = age - 13 - 6$ if the individual has a high school diploma and $EXPE = age - 17 - 6$ if the individual has a university degree. Individuals in Italy start schooling at age 6. Obtaining a high school diploma requires 13 years of schooling (in total) and a university degree 17 years (in total), i.e. four more years of schooling.

Then we obtain :



The *R-squared* shows the percentage of the variance of incomes explained by the covariates included in the regression. *Adj.R-squared* is the R-squared adjusted for the number of covariates included (since the R-squared always increases when we add more regressors). The adjusted R-squared goes from 1.71% in the model without job experience to 1.78% in the model with job experience. The model explaining more of the variance in incomes is better, i.e. the one with a higher adjusted R-squared.

We see that when including potential experience the coefficient of education falls to $\hat{\gamma} = 3321.49$, and it remains statistically significant at 5%. Hence, the bias is negative, that is $\hat{\beta} - \hat{\gamma} < 0$. Indeed, $\hat{\lambda} = 28.33 > 0$. Moreover:



that is $\hat{\rho} = -1.37 < 0$. Hence, individuals with higher education generally have about one year and 5 months less potential experience. It is important to notice that the difference is less than the legal duration of tertiary education (4 years). A possible explanation is that probably workers with higher education partly recover their initial disadvantage in experience due to a late entry in the labour market by remaining less unemployed.