Economics 167, Econometrics

Computer Lab 2

1. Open *STATA*.
2. Click *File*, then *Open*.
3. Locate the *STATA* dataset file *wages1.dta* and open it.
4. The dataset contains four variables taken from the US National Longitudinal Survey (NLS) and related to 1987:
   - *exper*;
   - *male*;
   - *school*;
   - *wage*.
5. Give a quick look at the data by clicking on *Data Browser* in the *STATA* command bar. The variable *exper* describes the years of work experience of each individual in the sample. The variable *male* is a dummy variable. A dummy variable is a discrete random variable that takes the value of one if a given characteristic is found in the individual we observe and a value of zero otherwise. In this case, the dummy variable *male* equals one if the individual is a man, zero if the individual is a woman. The variable *school* indicates the years of schooling for each individual in the sample. Finally, the variable *wage* describes the hourly wage rates for the individuals we are observing.
6. It is always a good idea to look at the data and examine them before running any kind of econometric analysis! We can obtain descriptive statistics for each of these variable by typing *Summarize* in the *STATA* command window:

Variable	Obs	Mean	Std. Dev.	Min	Max
exper	3294	8.043412	2.290661	1	18
male	3294	.5236794	.4995148	0	1
school	3294	11.63054	1.657545	3	16
wage	3294	5.757585	3.269186	.0765556	39.80892

   We have 3294 observations in the sample, i.e., $N = 3294$. For each variable, *STATA* reports the mean value, the minimum and the maximum values, and the standard deviation.
7. Calculate the cross-correlations of all the variables in the dataset. In this econometric exercise we are interested in estimating a wage equation for the individuals in the sample. In other words, we want to describe the wage rate as a function of individual characteristics. At this point, we are interested in the correlation coefficient of the

pairs of variables *wage-school*, *wage-exper*, and *wage-male*. Type *corr wage exper male school*:

(obs=3294)

	wage	exper	male	school
wage	1.0000			
exper	0.0514	1.0000		
male	0.1782	0.1295	1.0000	
school	0.2825	-0.1936	-0.1191	1.0000

8. Calculate the average hourly wage rate for the males in the sample and estimate its standard error. Type *mean wage if male==1*. We get:

Mean estimation — Number of obs = 1725

	Mean	Std. Err.	[95% Conf.	Interval]
wage	6.313021	.0842427	6.147793	6.47825

The table also reports a 95% confidence interval for the average hourly wage rate of the males. The boundaries of the confidence interval are calculated as

$$\left[\overline{wage}_M - t_{1724;0.975} SE\left(\overline{wage}_M\right); \overline{wage}_M + t_{1724;0.975} SE\left(\overline{wage}_M\right)\right],$$

where $\overline{wage}_M$ is the sample mean of the wage rates of the males, $Prob\left(T \leq t_{1724;0.975}\right) = 0.975$, and $T \sim T_{1724}$.

9. Calculate the average hourly wage rate for the females in the sample and its standard error. Type *mean wage if male==0*:

Mean estimation — Number of obs = 1569

	Mean	Std. Err.	[95% Conf.	Interval]
wage	5.146924	.0726128	5.004495	5.289352

10. By using the *OLS* approach, estimate the model

$$wage_i = \beta_1 + \beta_2 male_i + \varepsilon_i.$$

- Click on *Statistics* $\Longrightarrow$ *Linear Models and Related* $\Longrightarrow$ *Linear Regression*. A window with five different tabs pops up. In the first tab, *Model*, we can specify the model. Choose *wage* as the dependent variable and *male* as the unique exogenous regressor. Then click *OK*. We get:

Source	SS	df	MS
Model	1117.26971	1	1117.26971
Residual	34076.9173	3292	10.351433
Total	35194.187	3293	10.6875758

Number of obs = 3294
F(1, 3292) = 107.93
Prob > F = 0.0000
R-squared = 0.0317
Adj R-squared = 0.0315
Root MSE = 3.2174

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
male	1.166097	.1122422	10.39	0.000	.9460258	1.386169
_cons	5.146924	.0812248	63.37	0.000	4.987668	5.30618

- This is a typical regression output. In the lower part of the table we can find the estimated coefficients, one for the constant term and one associated with

the regressor, *male*. What is the interpretation that should be given to the estimated coefficients of our simple model? If the dummy variable, *male*, equals zero, then the model tells us that the expected wage rate for the females is equal to $\widehat{\beta}_1 = \$5.15$. If the dummy variable equals one, the model tells us that the additional compensation for the males is $\widehat{\beta}_2 = \$1.17$. These findings suggest that the average hourly wage rate of the males is about \$6.31. Note that these are exactly the same numbers we found earlier when we computed the sample means for males and females separately. Also note that we could have obtained the same regression output by typing *regress wage male* or *reg wage male* in the *STATA* command window.

- Next to the estimates, we can see the estimated standard deviations of the *OLS* estimators – i.e., the so-called standard errors. The fourth column reports the values of the realized t statistics – i.e., the ratio between the estimated coefficients and the corresponding standard errors. The computed t statistics are used to test the null that the associated parameters are individually equal to zero. In other words, these are simple tests of statistical significance for the coefficients of interest in the linear regression model. In the fifth column we can find the corresponding p-values. Finally, the last two columns report the estimated 95% confidence intervals for the two parameters in the model. We can estimate intervals with different confidence levels by simply specifying a different confidence value in the tab *Reporting* in the estimation window. Note the dual relationship between confidence intervals and the two-sided tests we can run using the t statistics. In both cases the tests reject the null that the parameter is equal to zero at conventional levels (1%, 5%, and 10%). Both β_1 and β_2 are statistically significant: look at the size of the t statistics (pretty big) and note that the two p-values are lower than 1%, 5%, and 10%. But also note that neither confidence intervals cover zero. Given that what *STATA* automatically computes is 95% confidence intervals, a natural interpretation that can be given to these interval estimates is that 5% tests would reject the null of statistical insignificance for both parameters.

- Look at the upper panel in the table to find additional information on the regression. The *SS* column reports the sum of squares of the model, the residuals, and the dependent variable, respectively. That is, the first line is computed as $\sum \left(\widehat{wage}_i - \overline{wage}\right)^2$, the second as $\sum \widehat{\varepsilon}_i^2$, and the the third as $\sum \left(wage_i - \overline{wage}\right)^2$. The third column shows the degrees of freedom of the model, the residuals, and the dependent variable. Specifically, in the first line, the degrees of freedom equal the number of regressors that we are using in the model other than the constant term ($K-1$, where K is the number of regressors plus the constant term). In the second line, the degrees of freedom equal $N-K$. The last line is simply the sum of the first two, or $N-1$. Under *MS* we can see estimates for the mean squares of the model, the residuals, and the dependent variable divided by their corresponding degrees of freedom. Respectively, these three terms are computed as $\frac{\sum\left(\widehat{wage}_i - \overline{wage}\right)^2}{K-1}$, $\frac{\sum \widehat{\varepsilon}_i^2}{N-K}$, and $\frac{\sum (wage_i - \overline{wage})^2}{N-1}$. The second figure in the *MS* column is an unbiased estimate of the error variance.

- On the upper right-hand side of the table, see the number of observations in the sample (first line). The second line shows an F statistic for the test

$$\begin{cases} H_0: \text{ the slope coefficients are jointly equal to zero} \\ H_1: \text{ not } H_0 \end{cases}.$$

- The two numbers in parenthesis are the degrees of freedom of the F distribution according to which the F statistic for this test is distributed. The first figure corresponds to the number of restrictions we are imposing in the model (in this case, just one, since we only have one slope coefficient). The second figure equals $N - K$. Remember the form of the F statistic! In this case we reject the null, since the value of the realized test statistic is large. We can get the same piece of information by observing the corresponding p-value in the next line.
- Then we can see the values for the R^2 and the adjusted R^2 of the model. Finally, find the root mean squared error of the residuals, calculated as $\sqrt{\frac{\sum \hat{\varepsilon}_i^2}{N-K}}$.

11. Create another dummy variable that equals one when the individual is female and zero when the individual is male. Type *gen female=0* and then *replace female=1 if male==0*. Estimate the model

$$wage_i = \beta_1 + \beta_2 male_i + \beta_3 female_i + \varepsilon_i$$

by typing *reg wage male female*. We get

Source	SS	df	MS		
Model	1117.26971	1	1117.26971	Number of obs	= 3294
Residual	34076.9173	3292	10.351433	F(1, 3292)	= 107.93
				Prob > F	= 0.0000
Total	35194.187	3293	10.6875758	R-squared	= 0.0317
				Adj R-squared	= 0.0315
				Root MSE	= 3.2174

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
male	1.166097	.1122422	10.39	0.000	.9460258	1.386169
female	(dropped)					
_cons	5.146924	.0812248	63.37	0.000	4.987668	5.30618

- One of the two dummy variables has been dropped automatically by the software. The two dummy variables, by construction, sum up to a vector (a list) of ones, which is the vector already included in the set of regressors when we consider an intercept term in the model specification. In other words, we have a situation of perfect multicollinearity, known as dummy variable trap. Technically, this means that the matrix $(X'X)$ in the OLS estimator formula cannot be inverted, given that its rank is not full. As we know, in this case, not all parameters in the model can be estimated. The solution is to drop one of the dummy variables, or to include both but remove the constant term. Type *reg wage male female, noco*:

Source	SS	df	MS		
Model	110312.663	2	55156.3313	Number of obs	= 3294
Residual	34076.9173	3292	10.351433	F(2, 3292)	= 5328.38
				Prob > F	= 0.0000
Total	144389.58	3294	43.8341166	R-squared	= 0.7640
				Adj R-squared	= 0.7638
				Root MSE	= 3.2174

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
male	6.313021	.077465	81.50	0.000	6.161137	6.464906
female	5.146924	.0812248	63.37	0.000	4.987668	5.30618

The two estimated coefficients represent the two average wage rates for the males and females in the sample.

12. Now run the model

$$wage_i = \beta_1 + \beta_2 male_i + \beta_3 school_i + \beta_4 exper_i + \varepsilon_i.$$

The coefficient β_2 measures the difference in expected wage between a male and a female with the same level of schooling and experience. In general, the coefficients in a multiple regression model can be interpreted only under a *ceteris paribus* condition.

- Click again on *Statistics* $\Longrightarrow$ *Linear Models and Related* $\Longrightarrow$ *Linear Regression* and choose *male*, *school*, and *exper* as regressors in the model. Alternatively, type *regress wage male school exper* in the command window. We get:

Source	SS	df	MS
Model	4666.31659	3	1555.43886
Residual	30527.8705	3290	9.27898798
Total	35194.187	3293	10.6875758

Number of obs	3294
F(3, 3290)	167.63
Prob > F	0.0000
R-squared	0.1326
Adj R-squared	0.1318
Root MSE	3.0461

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
male	1.344369	.1076759	12.49	0.000	1.13325	1.555487
school	.6387977	.0327958	19.48	0.000	.5744954	.7031
exper	.1248255	.0237628	5.25	0.000	.0782342	.1714167
_cons	-3.380018	.4649765	-7.27	0.000	-4.291691	-2.468346

- Suppose we want to run the test

$$\begin{cases} H_0: \ \beta_2 = \beta_3 = \beta_4 = 0 \\ H_1: \ \text{not } H_0 \end{cases}.$$

To do so, just look at the F statistic automatically produced by STATA and reported in the regression output. This time, we are imposing three restrictions and, in fact, the degrees of freedom are $J = 3$ and $N - K = 3290$. The null hypothesis is rejected, since the value of the test statistic is big and the associated p-value is virtually zero (so, smaller than any conventional test level).

- To run the test

$$\begin{cases} H_0: \ \beta_3 = \beta_4 = 0 \\ H_1: \ \text{not } H_0 \end{cases},$$

estimate the corresponding restricted model (i.e., the same model as above after imposing the restrictions in the null) and then compute the F statistic as

$$F = \frac{(RSS_R - RSS_U)/J}{RSS_U/(N-K)},$$

where $J = 2$ is the number of restrictions. The restricted model is the one already estimated with only one regressor. Take the residual sum of squares from the two regression outputs and calculate:

$$\begin{aligned} F &= \frac{(RSS_R - RSS_U)/J}{RSS_U/(N-K)} \\ &= \frac{(34076.92 - 30527.87)/2}{30527.87/(3294-4)} \\ &= 191.24 \end{aligned}$$

Calculate the same F statistic using the R^2's of the two regressions:

$$\begin{aligned} F &= \frac{\left(R_U^2 - R_R^2\right)/J}{\left(1 - R_U^2\right)/(N-K)} \\ &= \frac{(0.1326 - 0.0317)/2}{(1 - 0.1326)/(3294 - 4)} \\ &= 191.35. \end{aligned}$$

Do you want to calculate the critical values for a 10%, or 5%, or 1% test? Type *mata* and then *invF(2,3290,0.90)*, or *invF(2,3290,0.95)*, or *invF(2,3290,0.99)*:

```
: invF(2,3290,0.90)
  2.304197364

: invF(2,3290,0.95)
  2.998461715

: invF(2,3290,0.99)
  4.611622282
```

The realized value of the test statistic is greater than the conventional critical values. As such, we reject the null hypothesis that $\beta_3 = \beta_4 = 0$. The second model seems to do a better job at explaining the dependent variable. Type *end* to exit *MATA*.

- Test the restrictions

$$\begin{cases} H_0: \ \beta_3 = \beta_4 = 0.30 \\ H_1: \ \text{not } H_0 \end{cases}.$$

Click on *Statistics* $\Longrightarrow$ *Postestimation* $\Longrightarrow$ *Tests* $\Longrightarrow$ *Test Linear Hypotheses*. In the right panel of the new window select *Linear expressions are equal* and then, right below, type *school=exper=0.30*. Click *OK*:

```
( 1)  school - exper = 0
( 2)  school = .3

       F(  2,  3290) =   97.50
            Prob > F =    0.0000
```

The test rejects the null once again. Note that we could have run the same test by directly typing *test (school=exper=0.30)* in the command line.

- Run a Wald test for the multiple linear restrictions

$$\begin{cases} H_0: \ \beta_2 + \beta_3 = 2 \ \wedge \ \beta_3 + \beta_4 = 0.75 \\ H_1: \ \text{not } H_0 \end{cases}.$$

Click on *Statistics* $\Longrightarrow$ *Postestimation* $\Longrightarrow$ *Tests* $\Longrightarrow$ *Test Linear Hypotheses*. In the right panel of the new window select *Linear expressions are equal* and then, below, type *male+school=2*. Click *Specification 2* under *Specifications*, on the upper left-hand side of the window, and type *school+exper=0.75* right below. Click *OK*:

```
( 1)  male + school = 2
( 2)  school + exper = .75

       F(  2,  3290) =    0.08
            Prob > F =    0.9277
```

The same test could have been run by directly typing *test (male+school=2) (school+exper=0.75)* in the *STATA* command line. This time we do not reject the null. The value of the test statistic is low and the p-value is high, much bigger

than the conventional levels of 10%, 5%, and 1%. Note that *STATA* runs Wald tests using F statistics. So, we must be sure that the Gauss-Markov assumptions really hold to reliably interpret the results of this test. Or, we can just notice that the sample size is big enough and that the F distribution we are using for this test is asymptotically approaching a chi-squared distribution with 2 degrees of freedom.

Finally, note the following. Given that the sample size seems to be rather large, we do not need to analyze the residuals to make sure that the t statistics and the F statistics we have been using have the right distributions. We can just rely on the weak law of large numbers and on the central limit theorem to be sure of the consistency and asymptotic normality of the OLS estimator. Things would have been different in small samples. In a small-sample case, before even looking at the t statistics and the F statistics, we should have checked the non-autocorrelation property and normality of the regression residuals.