Economics 167, Econometrics

Computer Lab 4

1. Open *STATA*.

2. Click *File*, then *Open*.

3. Locate the *STATA* dataset file *bwages.dta* and open it.

4. The dataset is composed of seven variables taken from the European Community Household Panel. It consists of $n = 1472$ individuals, randomly sampled from the 1994 working population in Belgium. 893 of these individuals are males, 579 are females. The variables in the file are:

   - before-tax hourly wage rate, expressed in euros per hour (*wage*);
   - dummy variable for males (*male*);
   - work experience expressed in years (*exper*);
   - education level, which, here, is a discrete variable that takes the value of 1 if the individual only completed primary school, of 2 if she completed lower vocational training, of 3 if she completed intermediate level, of 4 if she has higher vocational training, and of 5 if she finished university (college).

   For our convenience, the database already includes three additional variables, the log of wages (*lnwage*), the log of experience (*lnexper*), and the log of education (*lneduc*).

5. In this lab we will explore the well-known fact that hourly wage rates of males are higher than those of females in almost all industrialized countries. As usual, we start by looking at some summary statistics. Type *summarize* in the command window:

Variable	Obs	Mean	Std. Dev.	Min	Max
wage	1472	11.05062	4.450513	2.190978	47.57552
lnwage	1472	2.334394	.3625349	.7843481	3.862319
educ	1472	3.378397	1.204522	1	5
exper	1472	17.21739	10.16667	0	47
lnexper	1472	2.690671	.7292168	0	3.871201
lneduc	1472	1.136519	.4339715	0	1.609438
male	1472	.6066576	.4886577	0	1

   We can split the sample between males and females. Type *summarize if male==1* first,

Variable	Obs	Mean	Std. Dev.	Min	Max
wage	893	11.56223	4.753789	3.467042	47.57552
lnwage	893	2.37829	.3632618	1.243302	3.862319
educ	893	3.243001	1.257386	1	5
exper	893	18.52296	10.25104	0	46
lnexper	893	2.775977	.7109817	0	3.850147
lneduc	893	1.083814	.4607368	0	1.609438
male	893	1	0	1	1

and then *summarize if male==0*,

Variable	Obs	Mean	Std. Dev.	Min	Max
wage	579	10.26154	3.808585	2.190978	33.80366
lnwage	579	2.266694	.3511076	.7843481	3.520569
educ	579	3.587219	1.086521	1	5
exper	579	15.2038	9.704987	0	47
lnexper	579	2.559102	.7379858	0	3.871201
lneduc	579	1.217806	.3752198	0	1.609438
male	579	0	0	0	0

The average wage rate for men is €11.56 per hour, whereas for women it is €10.26 per hour. The difference is about €1.30 per hour, or approximately 13%. The average education level for women is higher than the corresponding figure for men (3.58 vs 3.24). However, since the average work experience is higher for men than for women (18.52 vs 15.20), the difference in the hourly wage rate that we observe does not necessarily mean discrimination against women. Also, we are not considering, in this example, other factors that may be important for explaining the documented difference.

6. A first model to estimate the effects of gender on the hourly wage rate, correcting for differences in experience and education level, is obtainable by regressing the hourly wage rate on the male dummy variable, work experience, and level of education. That is, we estimate the equation

$$wage_i = \beta_0 + \beta_1 male_i + \beta_2 exper_i + \beta_3 educ_i + \varepsilon_i$$

by typing *regress wage male exper educ*,

Source	SS	df	MS
Model	10651.6554	3	3550.55181
Residual	18484.5373	1468	12.5916467
Total	29136.1928	1471	19.8070651

Number of obs	1472
F(3, 1468)	281.98
Prob > F	0.0000
R-squared	0.3656
Adj R-squared	0.3643
Root MSE	3.5485

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
male	1.346144	.1927364	6.98	0.000	.9680761	1.724212
exper	.1922751	.0095831	20.06	0.000	.1734771	.2110731
educ	1.98609	.0806396	24.63	0.000	1.827909	2.144271
_cons	.2136922	.386895	0.55	0.581	-.5452338	.9726183

All slope coefficients in this regression are statistically significant. The R^2 is low, but not that low, if consider the fact that we are using only three explanatory variables on the right-hand side of the regression. The estimated model describes the expected wage rate given gender, experience, and education level. *Ceteris paribus*, the effect of gender is virtually identical to the average wage differential that we computed above – i.e., the relevant slope coefficient is €1.35. In other words, adjusting for differences in education and experience does not change the expected wage differential between males and females. Moreover, this difference is statistically significant. As expected, the effects of experience and education on the wage rate are, *ceteris paribus*, positive. One extra year of work experience is expected to increase the hourly wage rate by €0.19. The wage differential between two people of the same gender and with the same number of years of work experience, but with two adjacent education levels, is expected to be €1.99.

7. It could be argued that experience affects a person's wage non-linearly. For example, we may think that, after so many years of work, the effect of an additional year of

work experience on one's wage may become increasingly smaller. To model this idea, we may want to run the regression

$$wage_i = \beta_0 + \beta_1 male_i + \beta_2 exper_i + \beta_3 exper_i^2 + \beta_4 educ_i + \varepsilon_i.$$

To do so in *STATA*, we first need to create the variable $exper^2$: type *gen exper2=exper^2*. Then we can estimate the equation above by typing *regress wage male exper exper2 educ*,

Source	SS	df	MS	Number of obs =	1472
Model	11023.4381	4	2755.85953	F(4, 1467) =	223.20
Residual	18112.7546	1467	12.3467993	Prob > F =	0.0000
				R-squared =	0.3783
				Adj R-squared =	0.3766
Total	29136.1928	1471	19.8070651	Root MSE =	3.5138

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
male	1.333693	.1908668	6.99	0.000	.9592925	1.708094
exper	.3579993	.0316566	11.31	0.000	.2959024	.4200963
exper2	-.0043692	.0007962	-5.49	0.000	-.005931	-.0028073
educ	1.988127	.0798526	24.90	0.000	1.831489	2.144764
_cons	-.8924851	.4329127	-2.06	0.039	-1.741679	-.0432912

All slope coefficients are statistically significant, this time even the constant term. The estimated coefficient associated with the gender dummy is virtually identical to the one estimated in the previous regression, the one without the quadratic regressor. Furthermore, the estimated coefficient associated with the education level has not changed almost at all. The R^2 and the adjusted R^2 have both increased a little. Since the t statistic associated with the coefficient of the new quadratic regressor that we just added is high, we can conclude that this model probably performs significantly better that the first one at explaining the variance of the dependent variable. Given the presence of both experience and its square in the equation, we cannot interpret their coefficients in isolation, though. One way to describe the effect of experience is to say that the expected wage differential associated with a marginal increase of experience is, *ceteris paribus*, given by

$$\frac{\partial E\left(wage_i | male_i, exper_i, educ_i\right)}{\partial exper_i} = \beta_2 + 2\beta_3 exper_i,$$

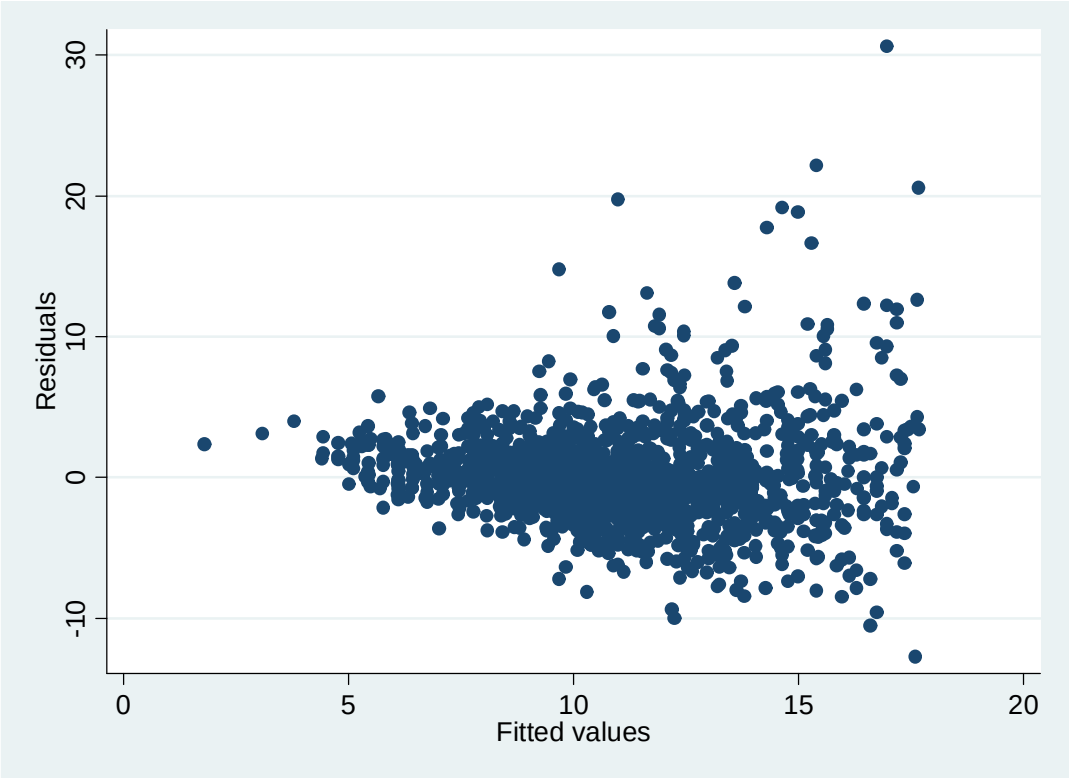
which we estimate to be $0.3580 - 2 \times 0.0044 \times exper_i$ – i.e., the effect of experience changes with its level. For example, if we want to estimate, *ceteris paribus*, the effect on wage of a marginal increase in the work experience for a person that has already 10 years of experience, we get

$$0.3580 - 2 \times 0.0044 \times 10 = €0.27.$$

8. Before we continue with our statistical analysis, it is important to analyze to what extent the assumptions regarding the error terms are satisfied in this example. For the standard errors and the statistical tests to be valid, we need to rule out the possibility that problems of autocorrelation and heteroskedasticity are present. We saw in class that we can prove unbiasedness and consistency of the OLS estimator even if we violate the assumption of spherical errors (homoskedasticity plus non-autocorrelation of the error term). But, if we want to correctly compute t statistics and F statistics, we really need to make sure that the error terms are homoskedastic and non-autocorrelated. Given that there is no natural ordering in the data and individuals are randomly sampled, autocorrelation is probably not an issue. Autocorrelation is

usually a problem in time-series application. When we a have a cross-section of individuals, as in this case, we can almost always rule out the presence of serial correlation in the data. On the other hand, heteroskedasticity could be problematic. In class we will analyze the problem of heteroskedasticity in depth and we will study how to formally test for non-constant variance in the error term. However, an informal way to look at this issue is to consider a scatter plot of the residuals of the regression against the fitted dependent variable.

To obtain this plot in *STATA*, type *rvfplot*:



If there is no heteroskedasticity, we should expect the dispersion of the residuals not to vary with different levels of the fitted values. From the graph above, however, the dispersion of the residuals gets higher as we move to the right. This finding casts serious doubts about the assumption of homoskedasticity. If homoskedasticity is violated, all the inference we made so far (the t statistics produced by *STATA*, for example) is invalid and we should try to find a solution to the problem. Inference is invalid because the standard errors are calculated by *STATA* under the assumption of homoskedasticity. They would be different if we took heteroskedasticity into account. We will see all this in class. It is important to notice at this point that *STATA* did not warn us about this issue.

One way (not the only one, nor always the right one) to eliminate or reduce the heteroskedasticity problem is to change the functional form of the regression model and use log wages, log education, and log experience rather than the corresponding variables in levels. Note that it is not possible to apply the log transformation on the gender dummy variable, since the natural logarithm is not defined at zero. We can run the regression

$$log\left(wage_i\right) = \beta_0 + \beta_1 male_i + \beta_2 log\left(exper_i\right) + \beta_3 \left[log\left(exper_i\right)\right]^2 + \beta_4 log\left(educ_i\right) + \varepsilon_i$$

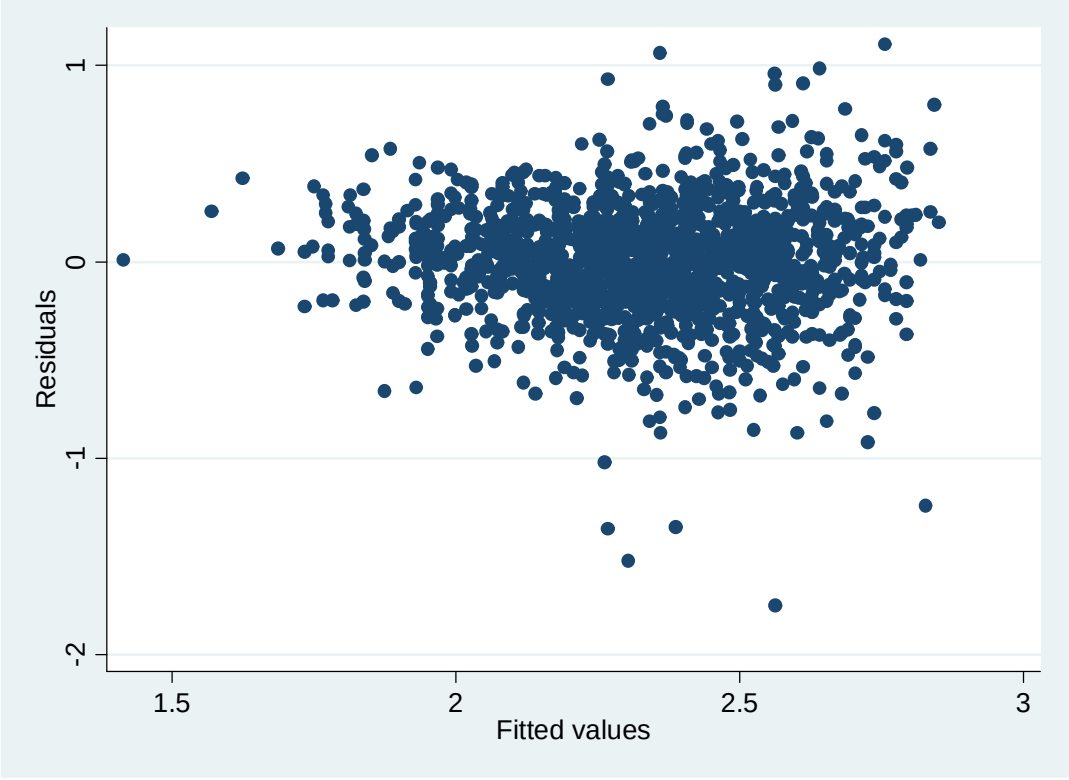
by first creating the variable *[log(exper)]*^{2} (*gen lnexper2=lnexper^2*) and then typing *regress lnwage male lnexper lnexper2 lneduc*,

Source	SS	df	MS
Model	73.1312577	4	18.2828144
Residual	120.204562	1467	.081939033
Total	193.33582	1471	.131431557

Number of obs	1472
F(4, 1467)	223.13
Prob > F	0.0000
R-squared	0.3783
Adj R-squared	0.3766
Root MSE	.28625

lnwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
male	.1179433	.0155711	7.57	0.000	.0873993	.1484874
lnexper	.1098205	.0543838	2.02	0.044	.0031421	.2164988
lnexper2	.0260073	.0114762	2.27	0.024	.0034958	.0485188
lneduc	.4421763	.0181921	24.31	0.000	.4064911	.4778616
_cons	1.262706	.0663418	19.03	0.000	1.132571	1.39284

Since the dependent (or endogenous) variable is different, the R^2 of the new model cannot be compared to the R^2's of the previous two. The interpretation of the estimated coefficients is also different from before. This time, the coefficient associated with the gender dummy describes the expected percent difference in the hourly wage rate between males and females. From the descriptive statistics we analyzed at the beginning, we know that the average difference is 13%. These estimates show that the expected differential is, *ceteris paribus*, 11.8% – i.e., after controlling for education and work experience. Consider again the issue of heteroskedasticity. If the problem is solved, then we can safely look at the t statistics in the regression output. Type *rvfplot* again:



There seem to be some traces of heteroskedasticity, still. But the problem looks less serious than before. Therefore, we can probably assume that the assumption of homoskedasticity holds this time. If we do so, we should also assume that the standard errors and the routinely computed t statistics and F statistics are appropriate.

The coefficients for log experience and its square are somewhat hard to interpret. The elasticity of the hourly wage rate with respect to work experience can be computed as

$$\frac{\partial E\left[log\left(wage_i\right)|male_i, log\left(exper_i\right), log\left(educ_i\right)\right]}{\partial log\left(exper_i\right)} = \beta_2 + 2\beta_3 log\left(exper_i\right),$$

which, in this case, is estimated to be $0.1098 + 2 \times 0.0260 \times log\left(exper_i\right)$. It is surprising to see that the elasticity is increasing with experience. However, the two coefficients associated with log experience and its square are significant at the 5% level, but not at the 1% level. Given the high number of individuals in the dataset, it could be more appropriate to consider 1%-level tests to assess statistical significance. To figure whether we should keep these two regressors in the model or not, we could run a joint test of significance on β_2 and β_3,

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

Then type *test(lnexper=lnexper2=0)*,

```
( 1)  lnexper - lnexper2 = 0
( 2)  lnexper = 0

       F(  2,  1467) =  234.09
            Prob > F =    0.0000
```

Based on the outcome of the F test, we reject the null and we may decide to keep both the variables in the specification or, maybe, just one of them. Before modifying the model, have a quick look at the *AIC* and *BIC*. Type *estat ic*:

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	1472	-594.6353	-244.8633	5	499.7267	526.1985

Note: N=Obs used in calculating BIC; see **[R] BIC note**

If we drop the squared term and estimate the model

$$log\left(wage_i\right) = \beta_0 + \beta_1 male_i + \beta_2 log\left(exper_i\right) + \beta_3 log\left(educ_i\right) + \varepsilon_i$$

by typing *regress lnwage male lnexper lneduc*, we have

Source	SS	df	MS
Model	72.7104488	3	24.2368163
Residual	120.625371	1468	.082169871
Total	193.33582	1471	.131431557

```
Number of obs =    1472
F(  3,  1468) =  294.96
Prob > F      =  0.0000
R-squared     =  0.3761
Adj R-squared =  0.3748
Root MSE      =  .28665
```

lnwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
male	.1200777	.0155645	7.71	0.000	.0895467	.1506087
lnexper	.2306474	.0107336	21.49	0.000	.2095925	.2517023
lneduc	.4366162	.0180512	24.19	0.000	.4012072	.4720252
_cons	1.14473	.0411808	27.80	0.000	1.06395	1.225509

To calculate *AIC* and *BIC* of this model, type *estat ic*:

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	1472	-594.6353	-247.4354	4	502.8708	524.0483

Note: N=Obs used in calculating BIC; see **[R] BIC note**

To conclude, the F test suggests that we should keep log experience and its square in the model specification. Comparing the R^2's and the adjusted R^2's of the model with the squared term and the model without it is not very useful. The difference we observe in the values of the R^2's and the adjusted R^2's is very small. The AIC is smaller in the case of the model with the squared term, but the BIC is smaller in the other model. What to choose, then? Probably, we should keep the squared term in the model specification even if we may have problems with the economic interpretation of the estimated coefficients. As you could see, though, very often, different tests or criteria may lead us to different conclusions.