Econometrics 2
Central European University

Assignment 2
Due Friday 28 February

Access the data on Moodle: nls.dta
Data: The data are drawn from the NLS Young Men Cohort. We have a sample of 14-24-year-old men for whom the data contains information about family background, living arrangements, and region of residence in 1966. A follow-up survey in 1976 provides information on educational attainment and earnings for the same individuals (for details see Card 1993).

The observational unit of the data is the individual. The data set has 3010 observations.
The key variables for this exercise are:
lwage76 – natural log of hourly wage in 1976
age76 – age of the person in 1976
black – indicator variable equal to 1 if the person is black, zero otherwise
ed76 – the individual's completed years of schooling in 1976
daded – father's completed years of schooling; set equal to sample mean if missing
nodaded – father's education imputed
momed – mother's completed years of schooling; set equal to sample mean if missing
nomomed – mother's education imputed
famed – father and mother educational class 1-9
momdad14 – lived with both parents at age of 14
sinmom14 – lived with single mother at age of 14
step14 – lived with a step parent at age of 14
nearc4 – grew up near a 4-year college
nearc2 – grew up near a 2-year college
kww – Knowledge of the world of work test score
south66 – lived in the South in 1966
smsa66 – lived in a metropolitan area in 1966
reg661-reg669 – dummies for region of residence in 1966
south76 – lived in the South in 1976
smsa76 – lived in a metropolitan area in 1976

Question 1: Omitted variables bias and proxy variables:

a. Suppose you are interested in understanding how education affects wages. Which is the specification you are going to run? What are the variables in this dataset that you are going to use for this?

   Lwage76 = a + b1*ed76 + b2*x + e

   Specification:

1) Lwage76 = a + b1*ed76

2) Lwage76 = a + b1*ed76 + b2*black + b3*daded + b4*momed + b5* kww

3) Lwage76 = a + b1*ed76 + b2*black + b3*daded + b4*momed + b5* kww + b6* reg66`i' + b7*smsa76

For the first specification we try to measure the effect of education on wages without considering other variables. In the second specification, we add variables standing for an individual background that can be correlated with one person's wages and/or education level (race, parents' level of education, and test score (ability)). In the third specification more relevant variables on living places are added (region fixed effects and whether they live in a metropolitan area) which potentially have effects on the dependent variable.

VARIABLES	(1)	(2)	(3)
Education level (1976)	0.0521***	0.0199***	0.0161***
	(0.00287)	(0.00336)	(0.00330)
Black (1 = yes)		-0.133***	-0.110***
		(0.0195)	(0.0202)
Father Education		-0.00384	-0.00613**
		(0.00275)	(0.00269)
Mother Education		0.00465	0.00366
		(0.00303)	(0.00296)
KWW score		0.0160***	0.0150***
		(0.00104)	(0.00102)
Metropolitan area (1976) (1 = yes)			0.139***
			(0.0165)
Constant	5.571***	5.481***	5.432***
	(0.0388)	(0.0453)	(0.0555)
Observations	3,010	2,963	2,963
R-squared	0.099	0.210	0.252
RegionFE	NO	NO	YES

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

b. Explain how omitting unobservable information about "ability" could bias the OLS estimate of the return of education on wages. What is your expectation of the direction of the bias? Explain how you could use a proxy variables approach to reduce the omitted variables bias.

The unobservable information about "ability" is a potential determinant of the dependent variable (wages), which means its regression coefficient is significantly different than 0, and possibly correlated with an individual's education level, which means their covariance is not equal to 0. Therefore, not including the variable ability can lead to a biased estimate of the regression coefficient of education variable, indicating not the direct effect of education on wages but rather the sum of the indirect effect of education on wages and the relationship between education and ability. Since the correlation of education with ability is probably

positive, and the effect of ability on wage is very likely to be positive, then omitting ability generates a positive bias on the parameter of education.
In an attempt to reduce OVB, we can use a proxy variable, which is correlated and measurable to serve the place of an individual's ability which cannot be measured directly. Such a proxy variable could be IQ or GPA. We can assume these variables capture well ability, which would make the coefficient of education less biased or at least consistent.

c. Generate a measure for potential experience, as exper=age76-ed76-6, and estimate two regression models: First, regress log-wage on education, experience, experience squared, black, and region indicators. Second, use the same regression models and add family background characteristics, like mother's and father's education as proxy variables for unobservable ability. Interpret the difference in the coefficient on education in both specifications.

VARIABLES	(1)	(2)
Education level (1976)	0.0790***	0.0764***
	(0.00355)	(0.00371)
Experience	0.0875***	0.0880***
	(0.00677)	(0.00678)
Experience squared	-0.00244***	-0.00245***
	(0.000323)	(0.000323)
Black (1 = yes)	-0.183***	-0.175***
	(0.0186)	(0.0189)
Mother Education		0.00654**
		(0.00292)
Father Education		0.000632
		(0.00266)
Constant	4.599***	4.552***
	(0.0809)	(0.0833)
Observations	3,010	3,010
R-squared	0.268	0.270
RegionFE	YES	YES

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

When we add the measure on parents' education, the coefficient of education falls from .079 to .076. This is the variation in wages that could be attributed by the variation in parents' education that is not included in the first specification, so the coefficient estimate of education level in the first specification contains the indirect effect of parents' education on wages as well. This change could be interpreted that omitting these variables generates a bias, even though the change seen in the education coefficient is very small, and besides that, the coefficient on father's education is not significant. Since we can assume that the bias of omitting ability is big (as we will see when we add kww test scores), there must be better proxies for that. Therefore, parents' education is not the best proxy for ability.

d. Now, regress education on the rest of the variables included in the specification and predict the residuals from this regression, call them res_ed. Regress log wages on the residuals

obtained above (res_ed). What is the coefficient on the res variable? Is this coefficient similar to the one you got in 1c? Why do you think this is the case? (Hint: Regression Anatomy)

VARIABLES	(1)
Residuals	0.0764***
	(0.00410)
Constant	6.262***
	(0.00766)
Observations	3,010
R-squared	0.103

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

We can see from the table above that two coefficients are equal. It is because according to the regression anatomy theorem, the coefficient of variable x in a multiple regression is equivalent to the coefficient of a bivariate model using the residual from a regression of that variable regressed on all the other variables. In other words, the residuals obtained from regressing education on the rest of the variables is the part of the variation in education which is not attributed to the variations in all the other variables. By regressing wages on the residuals, we have the same coefficient as the coefficient of education when we regress wages on education and other variables (the part of the variation in wages that is explained by the variation in education and also the part of the variation in education that is attributed to the variation in wages in a multiple regression of wages on all the variables). As the results are similar, we could say that we have good control variables in our model.

Let's show this result

First, we have our original regression equation (for simplification we do not show the individuals subscript)

$$Y = B_0 + B_1 x_1 + \cdots + B_k x_k + e \qquad (1)$$

We know the formula of the beta coefficients in the OLS estimation

$$\beta = \frac{Cov(Y, x)}{Var\,(x)} \qquad (2)$$

Now, we need to regress the regressor of interest x_k (education) against all other regressor on the
Specification

$$x_k = B_0 + B_1 x_1 + \cdots + e \qquad (3)$$

We store the residuals of this regression. Let's call these residuals $\widetilde{x_k}$
Now, we need to regress $Y \quad (wages)$ on the residuals $(\widetilde{x_k})$

$$Y \quad = \gamma_0 + \gamma\widetilde{x_k} + u \qquad (4)$$

Now we need to show that we will get B_k from this regression
Substituting on (2)

$$\beta \quad = \frac{Cov\,(Y \quad = B_0 + B_1 x_1 + \cdots + B_k x_k + e \quad , \widetilde{x_k})}{Var\,(\widetilde{x_k})}$$

The covariance of the residuals of the regression (3) with the regressors of (1) must be zero.
Also, $Cov(B_k x_k, \widetilde{x_k}) = B_k\, Var\,(\widetilde{x_k})$
Therefore,

$$\beta \quad = \frac{B_k\, Var\,(\widetilde{x_k})}{Var\,(\widetilde{x_k})} = B_k$$

Therefore, in our regression, we get the coefficient of the education regressor, as we wanted to show.

e. Now add the kww test score as a proxy for ability. How does the regression coefficient on education change in comparison to the models estimated in (a)? Check the correlation coefficient between the kww test score and education.

VARIABLES	(1)
Education level (1976)	0.0567***
	(0.00442)
exper	0.0717***
	(0.00705)
exper2	-0.00206***
	(0.000329)
Black (1 = yes)	-0.123***
	(0.0199)
Mother Education	0.00555*
	(0.00291)
Father Education	-0.000837
	(0.00266)
KWW score	0.00882***
	(0.00113)
Constant	4.680***
	(0.0783)
Observations	2,963
R-squared	0.280
RegionFE	YES

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Correlation: 0.49

There is a significant drop in the education coefficient, meaning that there is an omitted variable bias when we do not control for ability. The correlation is 0.49, and, therefore, it is not a surprise that omitting this variable causes a positive bias in the education coefficient.

f. Summarizing the results from b) and c), what do proxy variables tell you about the severity of the ability bias problem in this application?
Without the proxy variables indicating an individual's ability, the effect of education on wages is overestimated, as we can see that the coefficient estimate decreased after we add two proxy variables, however, as discussed above, parents' education is not a very good proxy. When we add kww test score, it might be a better proxy for ability as the change in the coefficient on education is much bigger (the coefficient estimate is less biased).

g. Export your main results in a table called Exercise1 with appropriate variable labels and footnotes. (please check the do file)

Question 2: Measurement Error:

In your folder, you will find a file named measurement_error.do.

a. Comment on what the code does linking theory seen in class with the code simulations. What are your main conclusions?

VARIABLES	(1) No measurement error	(2) Measurement error in wages	(3) Measurement error in education	(4) Measurement error in education 2	(5) Measurement error in wages 2
Education level (1976)	0.0521***	0.0604***			0.0656***
	(0.00287)	(0.0137)			(0.0213)
education			0.0337***	0.0185***	
			(0.00234)	(0.00188)	
Constant	5.571***	5.447***	5.815***	6.020***	5.374***
	(0.0388)	(0.186)	(0.0320)	(0.0259)	(0.288)
Observations	3,010	3,010	3,010	3,010	3,010
R-squared	0.099	0.006	0.065	0.031	0.003

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

We can see that measurement error in the dependent variable does not lead to a biased coefficient of the independent variable. The noise variable has zero mean and is uncorrelated with the explanatory variable (education), so the measurement error in wages is statistically independent of education. Thus the OLS estimator from our model is unbiased and consistent. On the other hand, the measurement error in the dependent variable results in a larger error variance than when no error occurs, which results in a larger variance of the coefficient estimate on education.
Measurement error in the explanatory variable is considered a much more important problem as it can lead to attenuation bias and the coefficient on education will be biased towards 0, as we see in the table above. The classical errors-in-variables assumption is that the

measurement error is uncorrelated with the unobserved explanatory variable, then the observed explanatory variable and the error term must be correlated, thus, the regression of the dependent variable on the observed variable gives a biased and inconsistent estimator.

b. What happens if instead of using normally distributed measurement errors you use uniformly distributed ones? What about measurement errors with a positive mean? Provide answers using simulations.

- Using uniformly distributed errors:

VARIABLES	(1) No measurement error	(2) Measurement error in wages	(3) Measurement error in education	(4) Measurement error in education 2	(5) Measurement error in wages 2
Education level (1976)	0.0521***	0.0529***			0.0388***
	(0.00287)	(0.0101)			(0.0144)
Education (with ME)			0.0414***	0.0367***	
			(0.00256)	(0.00244)	
Constant	5.571***	5.599***	5.712***	5.773***	5.900***
	(0.0388)	(0.137)	(0.0349)	(0.0334)	(0.206)
Observations	3,010	3,010	3,010	3,010	1,637
R-squared	0.099	0.009	0.080	0.070	0.004

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Here we present the simulation results using noise with uniform distribution (runiform(...) with zero mean and 2 variance and runiform(...) with zero mean and 4 variance. We usually assume a zero mean value for error distributions (measurement errors are tightly constrained around zero) when estimating measurement uncertainty, so the uncertainty in their values is small in comparison with errors that are widely spread. In other words, the uncertainty in the error is synonymously the spread in an error distribution, which is quantified by the standard deviation of the distribution. The central limit theorem demonstrates that, even though the individual constituent errors or deviations may not be normally distributed, the combined error or deviation is approximately so. If we use uniformly distributed errors instead of normally distributed one, the non-normality in the residuals and heteroscedasticity means that the amount of error in our model is not consistent across the full range of our observed data. Regarding the explanatory variables, this means that the amount of predictive ability they have (i.e., as calculated in their coefficient estimates) is not the same across the full range of the dependent variable. Thus, the explanatory variables technically mean differently across the range of the dependent variable. Hence, the coefficient estimate obtained from the model with uniformly distributed error is inconsistent and uninterpretable.

- Using normally distributed errors with a positive mean:

VARIABLES	(1) No measurement error	(2) Measurement error in wages	(3) Measurement error in education	(4) Measurement error in education 2	(5) Measurement error in wages 2
Education level (1976)	0.0521***	0.0488***			0.0661***
	(0.00287)	(0.0141)			(0.0214)
education			0.0327***	0.0207***	
			(0.00234)	(0.00186)	
Constant	5.571***	6.587***	5.797***	5.966***	6.307***
	(0.0388)	(0.191)	(0.0342)	(0.0276)	(0.290)
Observations	3,010	3,010	3,010	3,010	3,010
R-squared	0.099	0.004	0.061	0.040	0.003

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Here we present the simulations using noise with normal distribution but with positive mean (rnormal(1,2) and rnormal (1,4)). When the assumption on the zero mean of the measurement error does not hold, we can expect a large uncertainty in our model, and the average error term across our data set would not be zero (positive in this case). Our model uses the observed independent variables; therefore, the coefficient estimate of interest would be biased (closer to 0).

Question 3: Causality and R2

Regress wages on education and education on wages. What is the R2 in each case? What do you conclude?

VARIABLES	(1) Log wage (1976)	(2) Education level (1976)
Education level (1976)	0.0521***	
	(0.00287)	
Log wage (1976)		1.895***
		(0.104)
Constant	5.571***	1.395**
	(0.0388)	(0.655)
Observations	3,010	3,010
R-squared	0.099	0.099

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

R-squared of the regression of education on wages and wages on education is identical. Both equal 0.099, which means approximately 10% of the variation in log wages is attributed to the variation in education and vice versa. Not only are they identical numerically (as we can see from the table above)

but they are also identical conceptually. Both show the degree of variation in one variable that is explained by the variation in the other variable.