Exam ECON3150/4150: Introductory Econometrics.

This is an open book examination where all printed and written resources, in addition to a calculator, are allowed. If you are asked to derive something, give all intermediate steps. Do not answer questions with a "yes" or "no" only, but carefully motivate your answer.

In this exercise, we will use a data set collected from a sample of US individuals. The data are described at the end of the exercises.

Start by considering the bivariate OLS-regression that relates individual's wages to the number of years of tenure in the current employment relationship

$$\texttt{ln\_wage}_i = \alpha + \beta \texttt{tenure}_i + \epsilon_i.$$

1. (10 points) Figure 1 at the end of the exercise provides a scatter plot of `ln_wage` against `tenure`.

   (a) Explain in words how OLS finds the regression line.

   Solution: For each scatter point the OLS procedure considers the distance between the actual value of the outcome variable and the predicted value on the regression line. By changing the parameters, we can move the line closer to some points but at the same time further from others. OLS chooses the regression line that minimizes the sum of squared differences, i.e.

   $$\min_{\alpha,\beta} \sum_{i=1}^{n} \epsilon_i^2 = \min_{\alpha,\beta} \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2$$

   (b) We say that the OLS-estimator is *'unbiased'* and *'consistent'*. Explain the difference between these two concepts.

   Solution:

   i. *Unbiased* means that the expected value of the estimator is equal to the true parameter value, i.e. $E\left(\hat{\beta}\right) = \beta$.

   ii. *Consistent* means that the expected value of the estimator approaches the true parameter value, i.e. $\hat{\beta} \underset{n\to\infty}{\overset{p}{\to}} \beta$. Some estimators may be consistent but not unbiased, e.g. the IV-estimator.

   (c) The Gauss-Markov theorem tells us that the OLS-estimator is also *'efficient'*. Explain what is meant by this and what assumptions are necessary for it to be true.

Solution: *Efficient* means, in this context, that the OLS-estimator has the lowest variance of all estimators within a certain class. More precisely, the Gauss-Markov theorem tells us that if the (i) $E(u_i|X_i) = 0$, (ii) (X_i, Y_i), $i = 1, \ldots, n$ are iid, (iii) large outliers are unlikely and (iv) residuals are homoskedastic, then the OLS-estimator is the Best Linear Unbiased Estimator (BLUE).

2. (10 points) Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the mean of the variable X. Show that the OLS-estimators $\hat{\beta}$ and $\hat{\alpha}$ of β and α in the regression above are

$$\hat{\beta} = \frac{\sum_{i=1}^{n}\left(\texttt{ln\_wage}_i - \overline{\texttt{ln\_wage}}\right)\texttt{tenure}_i}{\sum_{i=1}^{n}\left(\texttt{tenure}_i - \overline{\texttt{tenure}}\right)^2}$$
$$\hat{\alpha} = \overline{\texttt{ln\_wage}} - \hat{\beta}\cdot\overline{\texttt{tenure}}$$

(Hint: Consider the first-order conditions of the OLS objective function and note that $\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\bar{X} = \bar{X}\sum_{i=1}^{n}\left(X_i - \bar{X}\right) = 0$*. You may use the simplified notation* $Y = \texttt{ln\_wage}$ *and* $X = \texttt{tenure}$ *in your derivations.)*

Solution: Let $Y = \texttt{ln\_wage}$ and $X = \texttt{tenure}$. OLS minimizes the sum of squared residuals over the parameters, i.e. in the bivariate case

$$\min_{\alpha,\beta}\sum_{i=1}^{n}\epsilon_i^2 = \min_{\alpha,\beta}\sum_{i=1}^{n}\left(Y_i - \alpha - \beta X_i\right)^2$$

Taking FOC w.r.t. the parameters, we get

$$\frac{\partial}{\partial\alpha}\sum_{i=1}^{n}\epsilon_i^2 = \sum_{i=1}^{n}2\left(Y_i - \hat{\alpha} - \hat{\beta}X_i\right)(-1) = 0$$
$$\iff 2n\hat{\alpha} = 2\sum_{i=1}^{n}Y_i - 2\hat{\beta}\sum_{i=1}^{n}X_i$$
$$\iff \hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n}Y_i - \hat{\beta}\frac{1}{n}\sum_{i=1}^{n}X_i \equiv \bar{Y} - \hat{\beta}\bar{X}$$
$$\frac{\partial}{\partial\beta}\sum_{i=1}^{n}\epsilon_i^2 = \sum_{i=1}^{n}2\left(Y_i - \hat{\alpha} - \hat{\beta}X_i\right)(-X_i)$$
$$= \sum_{i=1}^{n}2\left(y_i - \left(\bar{Y} - \hat{\beta}\bar{X}\right) - \hat{\beta}X_i\right)(-X_i) = 0$$
$$\iff 2\hat{\beta}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)X_i = 2\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)X_i$$
$$\iff \hat{\beta} = \frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)X_i}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)X_i} = \frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)X_i}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$$

3. (10 points) Using the output below, calculate the OLS-estimate of β and α. *(Hint: Note that* $\sum_{i=1}^{n} X_i = n \cdot \bar{X}$*.)*

```
. gen tenure_sq = tenure*tenure

. gen ln_wage_sq = ln_wage*ln_wage

. gen tenure_lwage = tenure*ln_wage

. sum ln_wage tenure tenure_sq ln_wage_sq tenure_lwage
```

Variable	Obs	Mean	Std. Dev.	Min	Max
ln_wage	2,231	1.872672	.573017	.0049396	3.707372
tenure	2,231	5.97785	5.510331	0	25.91667
tenure_sq	2,231	66.08483	102.5389	0	671.6736
ln_wage_sq	2,231	3.835101	2.34625	.0000244	13.74461
tenure_lwage	2,231	12.1415	12.62416	0	75.7233

```
. reg ln_wage tenure
```

Source	SS	df	MS		
				Number of obs	= 2,231
				F(1, 2229)	= 220.51
Model	65.9164523	1	65.9164523	Prob > F	= 0.0000
Residual	666.300616	2,229	.29892356	R-squared	= 0.0900
				Adj R-squared	= 0.0896
Total	732.217068	2,230	.328348461	Root MSE	= .54674

ln_wage	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
tenure	[redacted]	.0021011	[redacted]	0.000	[redacted]	[redacted]
_cons		.0170805		0.000		

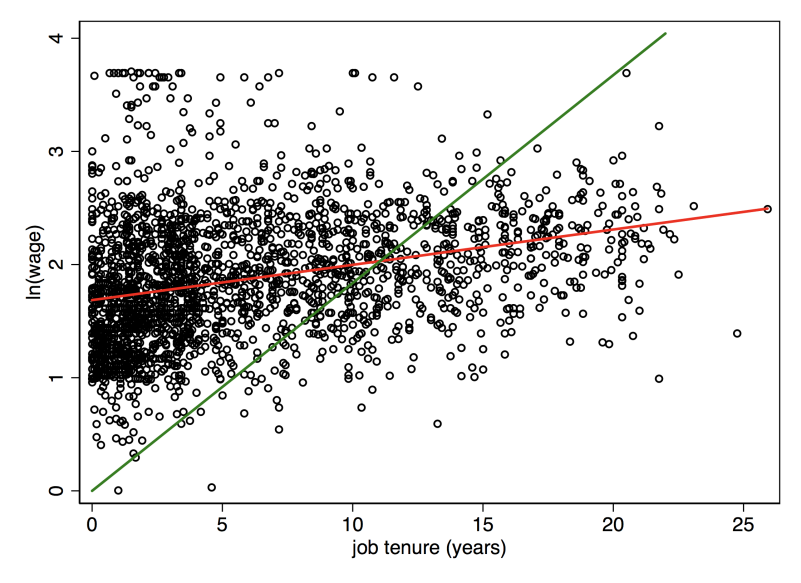
Solution: Using the expression for the OLS-estimators above and the simplified notation, we can expand to

$$\begin{aligned}
\hat{\beta} &= \frac{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right) X_i}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2} \\
&= \frac{\sum_{i=1}^{n} Y_i X_i - \bar{Y} \sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} X_i^2 - 2\bar{X} \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} \bar{X}^2} \\
&= \frac{n \cdot \overline{Y_i X_i} - \bar{Y} \cdot n \cdot \bar{X}}{n \cdot \overline{X^2} - 2 \cdot \bar{X} \cdot n \cdot \bar{X} + n \cdot \bar{X}^2} \\
&= \frac{\overline{Y_i X_i} - \bar{Y} \cdot \bar{X}}{\overline{X^2} - \bar{X}^2} \\
&= \frac{12.142 - 1.873 \cdot 5.978}{66.085 - 5.978^2} = 0.0311 \\
\hat{\alpha} &= \bar{Y} - \hat{\beta}\bar{X} = 1.873 - 0.0311 \cdot 5.978 = 1.687
\end{aligned}$$

4. (5 points) Say that we omit the constant term from the OLS regression.

   (a) Make a rough sketch of the scatter plot in Figure 1 and include an approximation to the OLS fitted line through the scatter plot with and without a constant term.

Solution: See red and green lines below.



   (b) Consider that in the true model, $\alpha = 0$, but that you included the constant term in the regression. How would you expect this to affect your estimates?

Solution: Omitting the constant term may be appropriate when we have strong theoretical reasons to believe that the regression line should pass through the origin. It is hard to think of an example, but wage earnings and hours could be one, since wage earnings is a product of hours. Another could be a first difference model where we do not want to allow for a trend/drift. Including the constant term is usually a good idea: Omitting it wrongly will bias all of your estimates, since you force the line through the origin. Including it wrongly (i.e. when it is truly zero) is likely to do little harm, since you are still allowing for $\alpha = 0$ in the model. However, including an unnecessary parameter will make the estimation less efficient, since it uses one degree of freedom.

5. (5 points)

   (a) Explain what we mean by heteroskedasticity.

Solution: Heteroskedasticity means that the conditional variance of the error term is not constant, i.e. that the variance of the dependent variable is a function of the included covariates.

   (b) Considering Figure 1, would you be concerned about heteroskedasticity in this case? Why or why not?

Solution: In the scatter plot, the data look heteroskedastic, since the variance of Y is decreasing in X.

(c) How would heteroskedasticity affect the OLS-estimates of our model? How would you account for this in your estimation?

Solution: Heteroskedasticity affects the efficiency of the estimates, meaning that the standard errors will not be correct. It does not affect the estimated parameters directly. We can guard against wrong SEs by using robust standard errors (or by correctly specifying the functional form of the variance and using WLS).

6. We may be concerned that the bivariate model is inappropriate. Consider the extended multivariate regression

$$\texttt{ln\_wage}_\texttt{i} = \alpha + \beta \texttt{tenure}_\texttt{i} + \gamma_1 \texttt{BlueCollar}_i + \gamma_2 \texttt{WhiteCollar}_i + \epsilon_i$$

where `BlueCollar` and `WhiteCollar` are mutually exclusive dummy variables equal to one if the individual's occupation is classified as blue collar and white collar, respectively. The estimation output from this regression is included below.

```
. gen WhiteCollar = (occupation == 1)

. gen Managerial = (occupation == 2)

. gen BlueCollar = (occupation == 3)

. reg ln_wage tenure BlueCollar WhiteCollar
```

Source	SS	df	MS	Number of obs	=	2,231
				F(3, 2227)	=	127.78
Model	107.530402	3	35.8434674	Prob > F	=	0.0000
Residual	624.686666	2,227	.280505912	R-squared	=	0.1469
				Adj R-squared	=	0.1457
Total	732.217068	2,230	.328348461	Root MSE	=	.52963

ln_wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
tenure	.0299802	.0020387	14.71	0.000	.0259822	.0339782
BlueCollar	-.1886312	.0260202	-7.25	0.000	-.2396575	-.1376048
WhiteCollar	-.3471735	.0288962	-12.01	0.000	-.4038399	-.2905072
_cons	1.852852	.0225853	82.04	0.000	1.808561	1.897142

(a) (5 points) Give an interpretation of the estimate on `tenure`.

Solution: One year increase in tenure is associated with an increase in expected earnings of about 3 percent.

(b) (5 points) Calculate a 90 percent confidence interval for β. Give an interpretation of this interval.

Solution: With repeated sampling from the full distribution, the probability that the true parameter will be contained in the CI is 90 percent.

$$\begin{aligned}\hat{\beta} \pm t_{\frac{0.1}{2},df} \cdot \hat{se}\left(\hat{\beta}\right) &= 0.0299 \pm t_{0.05,2201-4} \cdot 0.0019 \\ &= 0.0299 \pm 1.645 \cdot 0.0019 \\ &= (0.0268, 0.0330)\end{aligned}$$

(c) (5 points) Test the hypothesis $H_0 : \beta = 0.05$ on the 5 percent-level.

Solution: Use the test statistic

$$\frac{\hat{\beta} - \beta^{H_0}}{\hat{se}\left(\hat{\beta}\right)} = \frac{0.0299 - 0.05}{0.0019} = -10.58$$

which is much higher than the critical value from the t-table of 1.96. We can therefore reject the null hypothesis.

(d) (5 points) Assuming that residuals are homoskedastic, test the hypothesis $\gamma_1 = \gamma_2 = 0$. *(Hint: Output from the previous regression will be necessary for this test.)*

Solution: To test multiple hypotheses we use an F-test. The bivariate regression above gives the model under H0, i.e.

$$\begin{aligned}F &= \frac{\left(R^2_{unrestricted} - R^2_{restricted}\right)/q}{\left(1 - R^2_{restricted}\right)/\left(n - k_{unrestricted} - 1\right)} \\ &= \frac{\left(0.1465 - 0.0903\right)/2}{\left(1 - 0.0903\right)/\left(2231 - 4\right)} = 74.27\end{aligned}$$

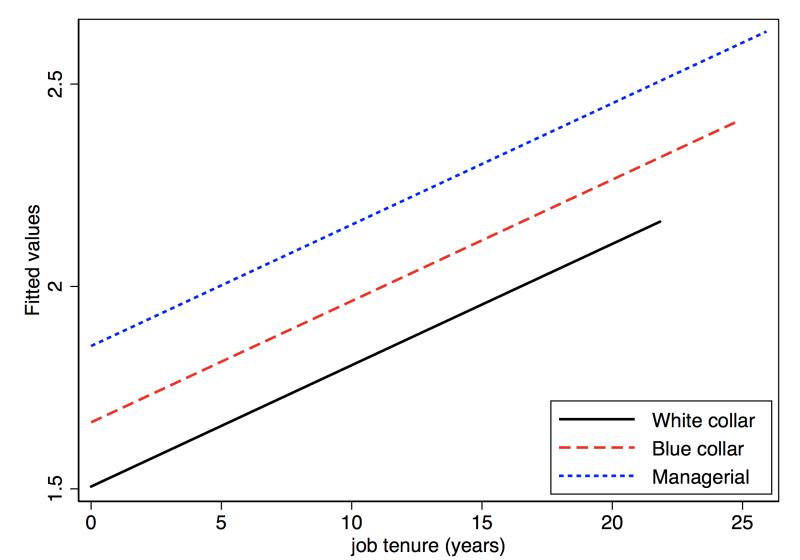
or alternatively using the SSR

$$\begin{aligned}F &= \frac{\left(SSR_{restricted} - SSR_{unrestricted}\right)/q}{SSR_{unrestricted}/\left(n - k_{unrestricted} - 1\right)} \\ &= \frac{\left(666.3006 - 624.6867\right)/2}{624.6867/\left(2231 - 4\right)} = 74.18\end{aligned}$$

Which in both cases is much higher than the critical value from the F-table.

(e) (5 points) Draw a sketch of the estimated regression lines for the three occupation groups: Managers, blue collar workers and white collar workers. Give an interpretation of γ_1 and γ_2. Discuss whether these estimates seem reasonable or if you believe that there may be some important omitted variables or that the functional form may be misspecified.

Solution: See graph below. γ_1 and γ_2 give an estimate of the difference in wage level between blue collar and white collar workers, respectively, compared to managers (the omitted category). The estimates suggest that blue collar workers have about 19 percent higher wages than managers, while white collar workers have about 15 percent lower wages than managers.
It seems unreasonable that wages of blue collar workers are higher than wages of managers. We may be suspicious that the regression model is misspecified. For instance, maybe the wage profile of the three groups is different, such that the coefficient on tenure should be allowed to vary.



(f) (5 points) Suppose that you included the variable `Managerial` *instead of* the variable `BlueCollar`? What would be the values of the coefficients in this regression?

Solution: $\hat{\beta}$ is unchanged, while the estimated coefficient on the dummy variables change to be relative to blue collar rather than managers. More precisely, the coefficient on `Managerial` will equal $-\hat{\gamma}_1 = 0.1886$, since the comparison is the same but reversed. The coefficient on `WhiteCollar` will equal $\hat{\gamma}_2 - \hat{\gamma}_1 = -0.3472 + 0.1886 = -0.1586$, since the difference between `WhiteCollar` and `BlueCollar` equals `WhiteCollar` − `Managerial` − (`BlueCollar` − `Managerial`).

(g) (5 points) What would happen to your estimates if you included the variable `Managerial` *in addition to* the variables `BlueCollar` and `WhiteCollar`?

Solution: The three dummies are perfectly collinear and cannot be estimated in the same regression. STATA will drop one of them by default.

7. (5 points each) Discuss whether each of the following statements is correct or not. Note that these *do not* relate to the regression model we studied above.

(a) With municipality fixed effects in the regression, we cannot include the distance from the municipality to the closest city in our regression model.

Solution: True. The fixed effects will account for all time-invariant differences between municipalities. Since the distance to the closest city does not change over time, this variable will be perfectly collinear with the fixed effect.

(b) The causal effect of a treatment D_i is given by the difference between the observed outcome of the treated ($D_i = 1$) and the observed outcome of the untreated ($D_i = 0$) .

Solution: False. The causal effect is the difference between the potential outcomes, not the observed outcomes. For instance, the ATT is the difference between the observed outcome of the treated and the counterfactual outcome that they would have experienced if they had not been treated. Using the observed outcomes directly ignores the selection into treatment and will therefore be susceptible to OMV bias.

(c) In a study of the impact of education on wages, the education of parents is a good instrument for the education of their child.

Solution: False. The education of parents is likely to be strongly relevant, but neither excludable nor as good as random. It is likely not excludable because parents education may also explain other important covariates, like the gpa performance conditional on an education. It is likely not as good as random because it correlates with e.g. the child's innate ability or network which will also be important determinants of wages.

(d) Excluding a covariate that explains the outcome will cause estimates on all included covariates to be biased.

Solution: False. Excluding a covariate will cause estimates on other covariates to be biased only insofar as there is a correlation between the excluded and included covariates. (A candidate that answers true but clearly conveys an understanding of this point should receive full points.)

Additional material

```
. describe

Contains data from /Applications/Stata/ado/base/n/nlsw88.dta
  obs:         2,231                          NLSW, 1988 extract
 vars:             4                          1 May 2016 22:52
 size:        35,696                          (_dta has notes)
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
wage            float   %9.0g                 hourly wage
tenure          float   %9.0g                 job tenure (years)
ln_wage         float   %9.0g                 ln(wage)
occupation      long    %12.0g     occupation
-------------------------------------------------------------------------------
Sorted by:
     Note: Dataset has changed since last saved.

. label list _all
occupation:
           1 Blue collar
           2 Managerial
           3 White collar

. sum

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        wage |      2,231    7.792448    5.764505   1.004952   40.74659
      tenure |      2,231     5.97785    5.510331          0   25.91667
     ln_wage |      2,231    1.872672     .573017   .0049396   3.707372
  occupation |      2,231    2.120574    .7851096          1          3
```

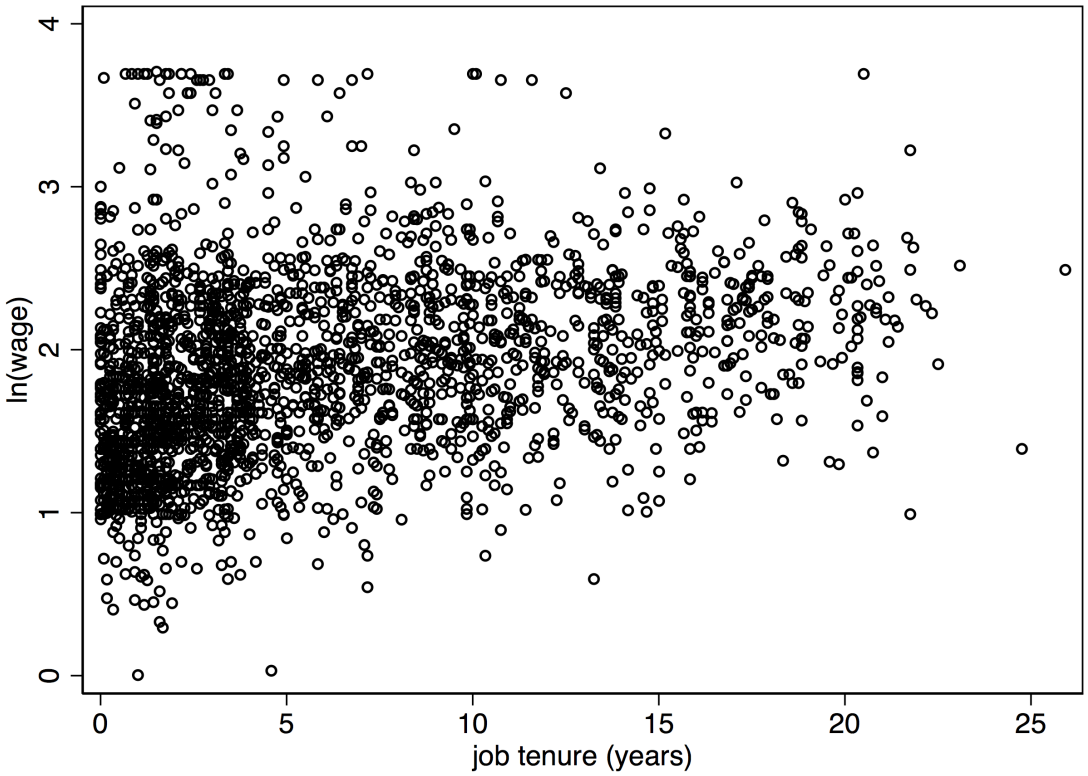


Figure 1: Log hourly wages and tenure