Exercise Class 2 - OLS

Instructor: Irene Iodice
Email: irene.iodice@malix.univ-paris1.fr

In this class we solve together the first 2 questions of last year mid-term exam that Professor Secchi sent you. I would also like to review an additional exercise on Omitted Variable Bias, that you find in the following. The data used come from the same database used in class from Wooldrige, import this in R as
library(foreign)
wage2 <- read.dta("http://fmwww.bc.edu/ec-p/data/wooldridge/wage2.dta")

Doing a bit of exercises with the program is worth: try to rerun the code that is reported before the results!

Ex.1: OVB
We want to test the association between wages and tenure (how many years has the person worked by that enterprise), and our 'omitted' variable will be gender . Suppose our population model is:

$$log(wage)_i = \beta_0 + \beta_1 tenure_i + \beta_2 female_i + u_i \tag{1}$$

1. *Through the use of a correlation matrix identifies the sign of the bias in γ_1 of the following model.*

$$log(wage)_i = \gamma_0 + \gamma_1 tenure_i + e_i \tag{2}$$

   A simple correlation matrix already helps us in detecting the direction of the bias when we estimate a SRM wrt MRM.

```
> df = data.frame(wage1$lwage,wage1$female,wage1$tenure)
> cor(df)
             wage1.lwage wage1.female wage1.tenure
wage1.lwage    1.0000000   -0.3736774    0.3255380
wage1.female  -0.3736774    1.0000000   -0.1979103
wage1.tenure   0.3255380   -0.1979103    1.0000000
```

   The correlation matrix above tells us that $\gamma_1 > \beta_1$.

2. *Through the estimation of the two models compute the bias*

```
> summary(lm(lwage ~ tenure+female, data=wage1))

Call:
lm(formula = lwage ~ tenure + female, data = wage1)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.00085 -0.28200 -0.06232  0.31200  1.57325 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.688842   0.034368  49.141  < 2e-16 ***
tenure       0.019265   0.002925   6.585 1.11e-10 ***
female      -0.342132   0.042267  -8.095 4.06e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4747 on 523 degrees of freedom
Multiple R-squared:  0.2055,	Adjusted R-squared:  0.2025 
F-statistic: 67.64 on 2 and 523 DF,  p-value: < 2.2e-16
```

   The bias is equal to $\gamma_1 - \beta_1 = 0.0239521 - 0.019265 = 0.0046871$.

```
> summary(lm(lwage ~ tenure, data=wage1))

Call:
lm(formula = lwage ~ tenure, data = wage1)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.15984 -0.38530 -0.04478  0.32696  1.46072 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.501007   0.026866  55.870  < 2e-16 ***
tenure      0.023951   0.003039   7.881 1.89e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5031 on 524 degrees of freedom
Multiple R-squared:  0.106,	Adjusted R-squared:  0.1043 
F-statistic: 62.11 on 1 and 524 DF,  p-value: 1.89e-14
```

3. *What is the parameter missing to compute the bias on γ_1 through the OVB formula? Which regression do we have to run to find its value?*
The regression we have to estimate is:

$$female_i = \alpha_0 + \alpha_1 tenure_i + v_i \tag{3}$$

this is because note that the equation for the OVB you have in the slide reads:

$$\begin{aligned} E[\hat{\gamma_1}] &= \beta_1 + \beta_2 \frac{Cov(tenure_i, female_i)}{Var(tenure_i)} \\ &= \beta_1 + \beta_2 \alpha_1 \\ &= 0.019265 + (-0.342132)(-0.013698) = 0.019265 \end{aligned}$$

```
> summary(lm(female ~ tenure, data=wage1))

Call:
lm(formula = female ~ tenure, data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.5490 -0.5079 -0.1929  0.4750  0.9167 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.549011   0.026201  20.954  < 2e-16 ***
tenure      -0.013698   0.002964  -4.622  4.8e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4906 on 524 degrees of freedom
Multiple R-squared:  0.03917,	Adjusted R-squared:  0.03733 
F-statistic: 21.36 on 1 and 524 DF,  p-value: 4.796e-06
```

4. *How could we get an unbiased estimate of β_1 without estimating a MRM?*

We could use the following two stages approach, first we estimate:

$$tenure_i = \sigma_0 + \sigma_1 female + w_i \tag{4}$$

And we call $\hat{w}_i$= $u\_tenurefemale.hat$, second we estimate

$$lwage_i = \delta_0 + \delta_1 \hat{w}_i + \epsilon_i \tag{5}$$

Indeed, as shown in the following results δ_1=β_1=0.019265

```
> u_tenurefemale.hat<-resid(lm(tenure ~ female, data=wage1))
> summary(lm(lwage ~ u_tenurefemale.hat, data=wage1))

Call:
lm(formula = lwage ~ u_tenurefemale.hat, data = wage1)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.20777 -0.39993 -0.04825  0.33359  1.49600 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)        1.623268   0.022421  72.399  < 2e-16 ***
u_tenurefemale.hat 0.019265   0.003169   6.079 2.33e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5142 on 524 degrees of freedom
Multiple R-squared:  0.06587,	Adjusted R-squared:  0.06409 
F-statistic: 36.95 on 1 and 524 DF,  p-value: 2.333e-09
```

5. *Compute the R^2 for model (1) with the info displayed below and then compute $\overline{R}^2$.*

```
> sum( (wage1$lwage - mean(wage1$lwage) )^2 )
[1] 148.3298
> sum( ( u_lwagetenurefemale.hat )^2 )
[1] 117.8466
```

The information provided gives us $\sum_i (lwage_i - \overline{lwage})^2 = 148.3298$ and $\sum_i \hat{u}_i^2 = 117.8466$, which are SST and SSR, respectively. Then,

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{117.8466}{148.3298} = 0.2055$$

We can compute the $\overline{R}^2$ as

$$\overline{R}^2 = 1 - \frac{n-1}{n-(k+1)} \frac{SSR}{SST} = 1 - \frac{n-1}{n-(k+1)}(1 - R^2)$$

where k is the total number of explanatory variables in the model (not including the constant term), and n is the sample size, which is 526, you can compute this from the info on df.

$$\overline{R}^2 = 1 - \frac{526-1}{526-(2+1)}(1 - 0.2055) = 0.2025$$

You can compare the results obtained with those displayed in the results from R presented before.