Exercise Class - Review

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

Overview of Statistics
You have the following table representing the joint distribution of X, number of children category variable and Y, representing the martial status:

	Married (Y=0)	Not Married (Y=1)
No children (X=0)	0.2	0.05
Few children (X=1)	0.3	0.2
Many children (X=2)	0.2	0.05

1. Compute the expected value of X.
2. Are X and Y independent? Is their correlation equal to zero?
3. Compute the expected share of married people among the people with no children, and its variance.
4. What is the probability to find a person with no children among those not married?

Econometrics - Reading coefficients and testing
You have the following regression:

$$log(WAGE_i) = \beta_0 + \beta_1 EDUC_i + \beta_2 FEMALE_i + \beta_3 TENURE_i + u_i \tag{1}$$

where EDUC is a variable capturing the number of years of completed education, TENURE the number of years of work experience in the same enterprise and FEMALE is a dummy variable equal 1 for female workers.Assume that u is an homoskedastic error term and that the standard OLS assumptions (HP1-HP4) hold. Estimating this regression model with OLS over a sample of Italian workers we obtain:

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

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

Residuals:
     Min       1Q   Median       3Q      Max
-1.96883 -0.25262 -0.03383  0.24687  1.29983

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.633125   0.091382   6.928 1.26e-11 ***
educ         0.081354   0.006643  12.246  < 2e-16 ***
female      -0.297052   0.037470  -7.928 1.36e-14 ***
tenure       0.021634   0.002588   8.359 5.78e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4188 on 522 degrees of freedom
Multiple R-squared:  0.3828,    Adjusted R-squared:  0.3793
F-statistic: 107.9 on 3 and 522 DF,  p-value: < 2.2e-16
```

1. Compute what is the expected percentage change in WAGE associated to an additional year of education.
2. Under the assumption that n is large, construct a 56.2% confidence interval on β_3. Provide a precise interpretation of this confidence interval.

3. Formulate and run a test for the hypothesis that Italian women earns less than Italian men. Test this hypothesis using the 22.96% significance level. Briefly comment the result and the level of significance of the test.

   Dropping FEMALE and TENURE from the regression model, but using the same sample of observations, you get:

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

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

Residuals:
     Min       1Q   Median       3Q      Max
-2.21158 -0.36393 -0.07263  0.29712  1.52339

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.583773   0.097336   5.998 3.74e-09 ***
educ        0.082744   0.007567  10.935  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4801 on 524 degrees of freedom
Multiple R-squared:  0.1858,    Adjusted R-squared:  0.1843
F-statistic: 119.6 on 1 and 524 DF,  p-value: < 2.2e-16
```

4. Are FEMALE and TENURE jointly insignificant in the original equation at the 10% significance level?

Omitted Variable Bias

Pampilio Piratta is deciding if it is worthwhile to work an extra year at his enterprise or to go back to study for a master. With this aim he is exploring the relation between wage, education and tenure. Sadly Pampilio Piratta's research efforts are limited by the fact that he knows how to estimate linear regression models only if they contain one single regressor. Then he estimates the following three models:

i. $log(WAGE_i) = b_0 + b_1 EDUC_i + u_i$


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

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

Residuals:
     Min       1Q   Median       3Q      Max
-2.21158 -0.36393 -0.07263  0.29712  1.52339

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.583773   0.097336   5.998 3.74e-09 ***
educ        0.082744   0.007567  10.935  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4801 on 524 degrees of freedom
Multiple R-squared:  0.1858,    Adjusted R-squared:  0.1843
F-statistic: 119.6 on 1 and 524 DF,  p-value: < 2.2e-16
```

ii. $log(WAGE_i) = a_0 + a_1 TENURE_i + w_i$

```
> 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
```

iii. $TENURE_I = c_0 + c_1 EDUC_i + v_i$

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

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

Residuals:
   Min     1Q Median     3Q    Max
-6.946 -4.894 -2.601  1.520 38.813

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.9457     1.4638   4.745 2.69e-06 ***
educ         -0.1466     0.1138  -1.288    0.198
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.22 on 524 degrees of freedom
Multiple R-squared:  0.003155,  Adjusted R-squared:  0.001253
F-statistic: 1.659 on 1 and 524 DF,  p-value: 0.1984
```

1. Explain why the OLS estimator of b_1 fails to produce an unbiased estimation of the true value of this parameter. Which of the OLS assumption is likely to be violated? Explain the meaning of this assumption.

2. Based on the results obtained above discuss, using the OVB formula asses whether b_1 is likely to be upward or downward biased. Then compute the value of the bias for the data at hand through the same formula and the results below from estimating the long model:

$$log(WAGE_i) = \beta_0 + \beta_1 EDUC_i + \beta_2 TENURE_i + \epsilon_i \tag{2}$$

3. Would it be possible for Pampilios Piratta to obtain an estimate of b_1 not affected by this OVB but without estimating a linear model with both EDUC and TENURE? Check your answer with the results provided.

4. Compute the R^2 and $\overline{R^2}$ of the model in (2) knowing that $\sum_i (lwage_i - \overline{lwage})^2 = 148.3$ and that $\sum_i \hat{u}_i^2 = 102.56$.

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

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

Residuals:
     Min       1Q   Median       3Q      Max
-2.10350 -0.29287 -0.04081  0.28672  1.44967

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.404474   0.091696   4.411 1.25e-05 ***
educ        0.086528   0.006991  12.377  < 2e-16 ***
tenure      0.025814   0.002680   9.634  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4428 on 523 degrees of freedom
Multiple R-squared:  0.3085,    Adjusted R-squared:  0.3059
F-statistic: 116.7 on 2 and 523 DF,  p-value: < 2.2e-16
```

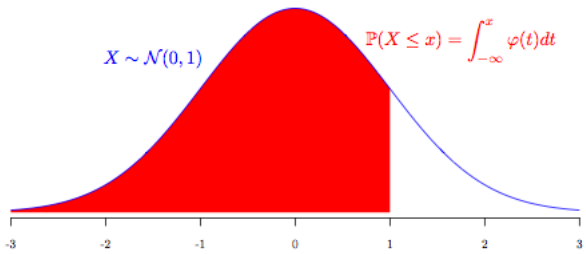
```
> summary(lm(lwage ~ u_eductenure.hat, data=wage1))

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

Residuals:
     Min       1Q   Median       3Q      Max
-2.20181 -0.35600 -0.06182  0.30338  1.50708

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.62327    0.02072   78.36   <2e-16 ***
u_eductenure.hat  0.08653    0.00750   11.54   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4751 on 524 degrees of freedom
Multiple R-squared:  0.2025,    Adjusted R-squared:  0.201
F-statistic: 133.1 on 1 and 524 DF,  p-value: < 2.2e-16
```



	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441

Large-Sample Critical Values for the *F*-statistic from the $F_{m,\infty}$ Distribution

Reject if F > Critical Value

	Significance Level		
Degrees of Freedom (*m*)	**10%**	**5%**	**1%**
1	2.71	3.84	6.63
2	2.30	3.00	4.61
3	2.08	2.60	3.78
4	1.94	2.37	3.32
5	1.85	2.21	3.02
6	1.77	2.10	2.80
7	1.72	2.01	2.64
8	1.67	1.94	2.51
9	1.63	1.88	2.41
10	1.60	1.83	2.32