MASTER
Economics & Psychology

Universitè Paris 1 Panthèon-Sorbonne
Universitè Paris Descartes

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.

   Sol. First compute the marginal probabilities of X, recalling the formula $Pr(X = x_i) = \sum_j Pr(X = x_i, Y = y_j)$, which is the sum by row, and get $Pr(X = 0) = 0.25$, $Pr(X = 1) = 0.5$ and $Pr(X = 2) = 0.25$. Then $E(X) = 0 \times 0.25 + 1 \times 0.5 + 2 \times 0.25 = 1$

2. Are X and Y independent? Is their correlation equal to zero?

   Sol. X and Y are independent if $Pr(Y = y_j) = Pr(Y = y_j|X = x_i)$ for all i and j, which is that the conditional probability of Y is equal to the marginal (if this holds the same is true for X). This can be easily checked by recalling that $Pr(Y = y_j|X = x_i) = \frac{Pr(Y=y_j,X=x_i)}{Pr(X=x_i)}$ which combined with the equation above gives $Pr(X = x_i, Y = y_j) = Pr(Y = y_j)Pr(X = x_i)$. Thus, using the joint probability table above we easily see that the two variables are not independent, for example $Pr(X = 0)Pr(Y = 0) = 0.2 > Pr(X = 0, Y = 0) = 0.15 \times 0.6 = 0.25$. Still, we have to check that the correlation is not zero, because we know from class that independence implies cov and corr =0 but not viceversa, then even if they are dependent we might have zero correlation, indeed:

   $Cov(X, Y) = E[(X - \overline{X})(Y - \overline{Y}))]$, where $E(Y) = 0 \times 0.7 + 1 \times 0.3 = 0.3$. Then

   $$\begin{aligned} Cov(X, Y) =& 0.2(0 - 1)(0 - 0.3) + 0.3(1 - 1)(0 - 0.3) + 0.2(0 - 2)(0 - 0.3)+ \\ &+ 0.05(0 - 1)(1 - 0.3) + 0.05(1 - 1)(1 - 0.3) + 0.05(0 - 2)(1 - 0.3) = 0 \end{aligned}$$

   and also the $Corr(X, Y) = 0$.

3. Compute the expected value of Y among the people with no children, and its variance.

   Sol. This means to compute $E(Y|X = 0) = \sum_j y_j Pr(Y_j|X = 0)$ and $Var(Y|X = 0) = \sum_j (y_j - E[Y|X = 0]])^2 Pr(Y_j|X = 0)$. Then, first you need to compute the conditional probability of Y given X=0: $Pr(Y = 0|X = 0) = \frac{0.2}{0.25} = 0.8$, $Pr(Y = 0|X = 0) = \frac{0.05}{0.25} = 0.2$. Then, $E[Y|X = 0] = 0 * 0.8 + 1 * 0.2 = 0.2$ which is lower than the unconditional expectation, which is given that a person has no children we expect a smaller share of married people. $Var(Y|X = 0) = 0.8(0 - 0.2)^2 + 0.2(1 - 0.2)^2 = 0.16$

4. What is the probability to find a person with no children among those not married?

   Sol.

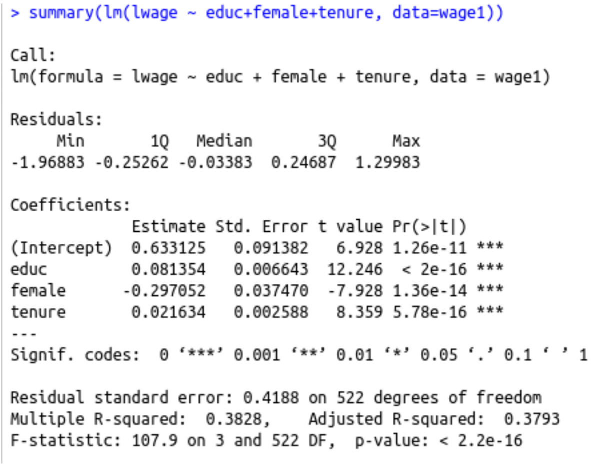
   $$Pr[X = 0|Y = 1] = \frac{Pr(X = 0, Y = 1)}{Pr(Y = 1)} = \frac{0.05}{0.30} = 0.1\overline{6}$$

Econometrics - Reading coefficients and testing

You have the following regression:

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

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.

   Sol. Recall the formula seen in class and plug in $\Delta educ = 1$

   $$\frac{\Delta lwage}{\Delta educ} = \hat{\beta}_1$$

   $$100\frac{\Delta wage}{wage}\% = (e^{\hat{\beta}_1} - 1)100\% = 8.475\%$$

2. Compute what is the expected percentage change in WAGE associated to the fact of being a male wrt to the case of being a female.

   For a male, $FEMALE_i = 0$ and then the estimated wage is $W\hat{A}GE^{male} = e^{\hat{\beta}_0 + \hat{\beta}_1 EDUC_i + \hat{\beta}_3 TENURE_i}$, for a female $FEMALE_i = 1$ and then $W\hat{A}GE^{fem} = e^{\hat{\beta}_0 + \hat{\beta}_1 EDUC_i + \hat{\beta}_2 + \hat{\beta}_3 TENURE_i}$. Then

   $$100\frac{W\hat{A}GE^{mal} - W\hat{A}GE^{fem}}{W\hat{A}GE^{fem}}\% = 100(e^{-\beta_2} - 1)\% = 100(1 - 1.3458)\% = 34.5\%.$$

3. Under the assumption that n is large, construct a 56.2% confidence interval on β_3. Provide a precise interpretation of this confidence interval.

   Sol. The confidence interval around β_3 reads $\hat{\beta}_3 - t_{28.1\%} \times SE(\hat{\beta}_3) < \beta_3 < \hat{\beta}_3 + t_{28.1\%} \times SE(\hat{\beta}_3)$, where $\hat{\beta}_3 = 0.0216$, $SE(\hat{\beta}_1) = 0.002588$ and $t_{21.9\%} = 0.781$, from $Pr(t < t_{21.9\%}) = 0.5 + 0.562/2 = 0.781$. Hence,

   $$0.01958 < \beta_3 < 0.02367$$

   With a 56.2% confidence, holding EDUC and FEMALE constant, a one year increase in TENURE is associated with a change in wage between 1.96% and 2.4%.

4. 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.

   Sol. Testing this hypothesis implies setting the following null and alternative:

   $$H_0 : \beta_2 = 0 \qquad\qquad H_1 : \beta_2 < 0 \ \ .$$

The test statistics under H_0 reads

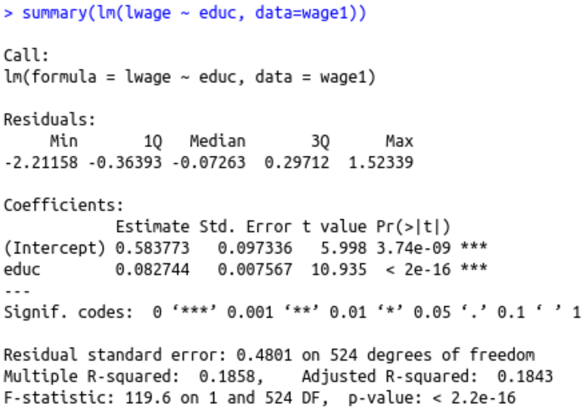
$$t = \frac{\hat{\beta}_2 - 0}{SE(\hat{\beta}_2)}$$

and it is asymptotically distributed as N(0,1). We have

$$t^{act} = \frac{-0.297 - 0}{0.0374} = -7.94$$

and $t_{22.96\%} = 0.74$, note that since this is a left tail test we test if $t^{act} < -t_{22.96\%}$, and since this is the case, we reject H_0 meaning that being a female does have a negative significant (at 74%) effect on log(WAGE), keeping the other variable constant.

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
```

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

Sol. In this question you are asked to test if in the original model

$$H_0 : \beta_2 = \beta_3 = 0 \qquad\qquad H_1 : \beta_2 \neq 0 \text{ and/or } \beta_3 \neq 0 \ .$$

This means that you have to run an F test with two restrictions on a homoskedastic sample. The F test statistic reads

$$F = \frac{(R^2_{unrestricted} - R^2_{restricted})/q}{\frac{(1 - R^2_{unrestricted})}{(n - (k_{unrestricted} + 1))}} \ .$$

where in this case $q = 2$ and $n - k - 1 = 526 - (3 + 1) = 522$ and hence where $F \sim F_{2,\infty}$. The critical value of F at 10% can be read on the statistical table and is 2.30. The corresponding rejection rule then is

$$\text{Reject } H_0 \text{ if } F^{act} > F_c \quad \Rightarrow \quad F^{act} > 2.3 \ ,$$

and since $F^{act} = \frac{(0.3828 - 0.1858)/2}{(1-0.3828)/(522)} = 83.31$ we reject the null that β_2 and β_3 are jointly equal to 0.

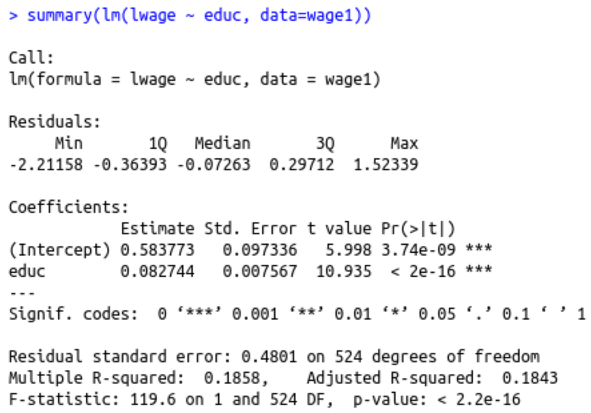
Omitted Variable Bias

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

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

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



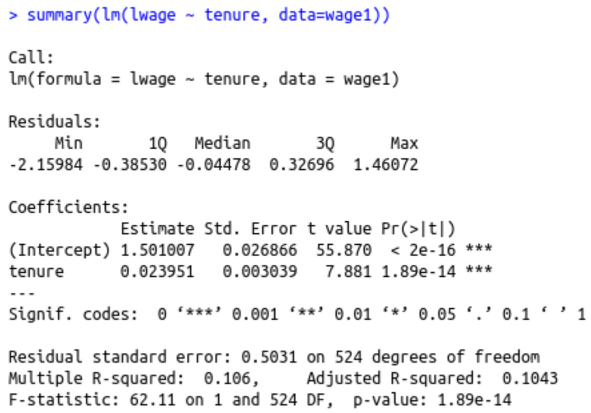
```
> summary(lm(lwage ~ educ, data=wage1))

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
```



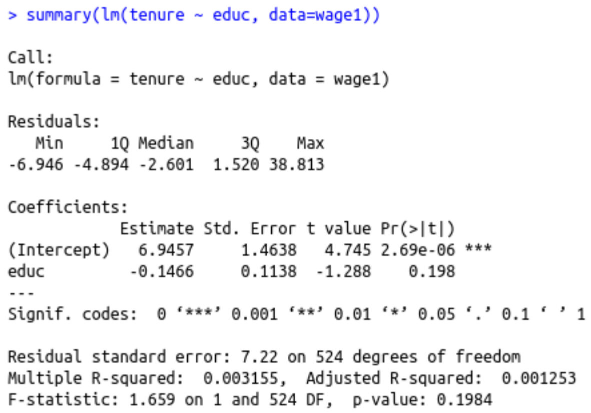
```
> summary(lm(lwage ~ tenure, data=wage1))

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
```



```
> 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
```

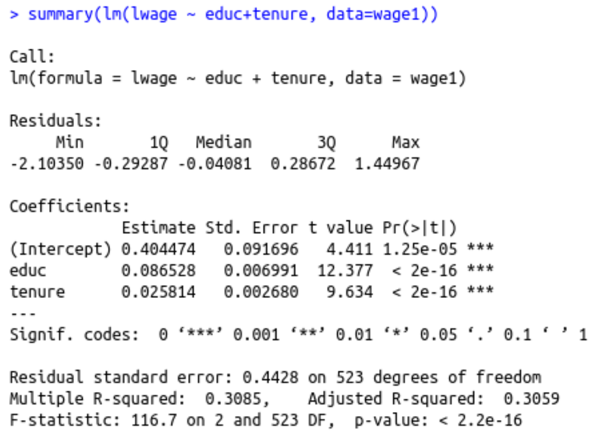
1. Explain why the OLS estimator of b_1 fails to produce an unbiased estimation of the true value of this parameter. Which of the OLS assumption is likely to be violated? Explain the meaning of this assumption.

   Sol. The ZCM assumption is likely to be violated here inducing an OVB. This is because in the error term u_i we have $TENURE_i$ (nb this is because it correlates significantly with $log(WAGE_i)$, in simple words it has an effect on the wage of a person) which correlates significantly with $EDUC_i$, this makes $E[u|EDUC] \neq E[u] = 0$. [CAUTION: please note that the relation between tenure and educ is instead not significant, and therefore we should not be worried of breaking the ZCM assumption here. Please, proceed in the exercise as it was the case since to correct the exercise I need a bit of time; Thanks to point this out!]

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 (and assume well

specified) model:

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



```
> 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
```

Sol. The sign of the OVB depends on the sign of the correlation between TENURE and EDUC and between WAGE and TENURE. Since they are negative and positive, respectively, then $b_1 = 0.08$ is likely to be a downward biased estimate of β_1. In particular, recall the equation for the OVB you have in the slide reads:

$$\hat{b_1} = \hat{\beta}_1 + \hat{\beta}_2 \times \frac{Cov(tenure_i, educ_i)}{Var(educ_i)}$$
$$= \hat{\beta}_1 + \hat{\beta}_2 \times \hat{c}_1 = 0.0865 + 0.0258 \times (-0.1466) = 0.0827$$

The bias for the data at hand is $\hat{b_1} - \hat{\beta_1} = 0.0827 - 0.0865 = -0.0038$, which is a negative bias as expected.

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.

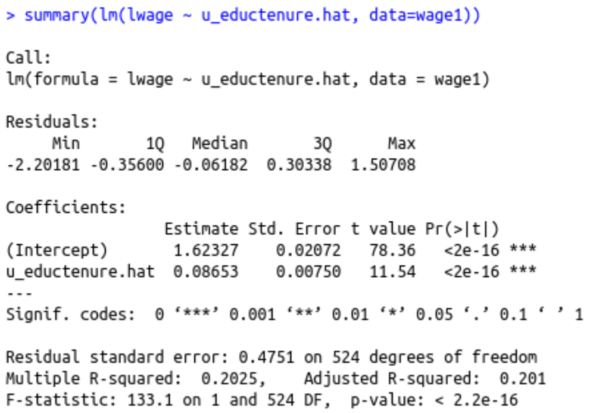
   Sol. We could use the following two stages approach, first we estimate:

   $$EDUC_i = d_0 + d_1 TENURE_i + \varepsilon_i$$

   And we call $\hat{\varepsilon_i}$= *u_eductenure.hat*, second we estimate

   $$lwage_i = \delta_0 + \delta_1 \hat{\varepsilon_i} + n_i$$

   Indeed, as shown in the following results δ_1=β_1=0.08653



```
> 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
```

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

Sol. The information provided gives us $\sum_i (lwage_i - \overline{lwage})^2 = 148.3$ and $\sum_i \hat{u}_i^2 = 102.56$, which are SST and SSR, respectively. Then,

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{102.56}{148.3298} = 0.30855$$

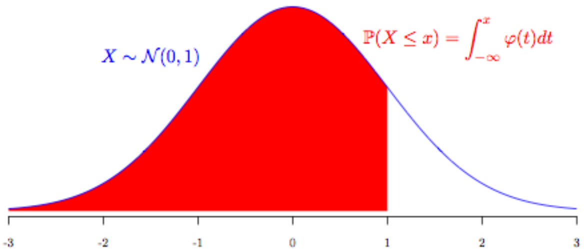
We can compute the $\overline{R}^2$ as

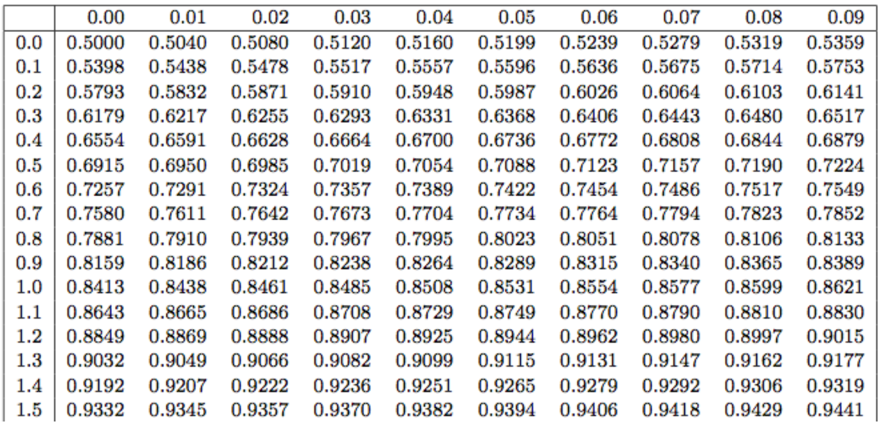
$$\overline{R}^2 = 1 - \frac{n-1}{n-(k+1)} \frac{SSR}{SST} = 1 - \frac{n-1}{n-(k+1)}(1 - R^2)$$

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.

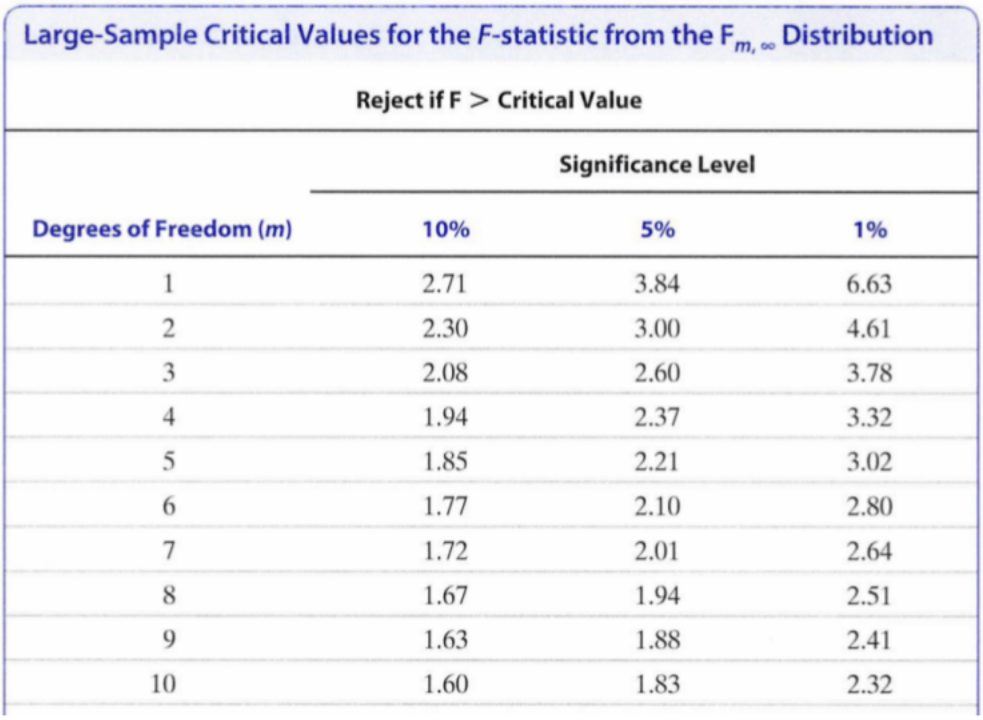
$$\overline{R}^2 = 1 - \frac{526-1}{526-(2+1)}(1 - 0.30855) = 0.3059$$

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





	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