Quantitative Methods III - Practice 3
Multiple Linear Regression

Prof. Lorenzo Cavallo: lorenzo.cavallo.480084@uniroma2.eu
Prof. Marianna Brunetti: marianna.brunetti@uniroma2.it

Exercise Given the table:

Regressor	**(1)**	**(2)**	**(3)**
College (X_1)	0.352	0.373	0.371
	(0.021)	(0.021)	(0.021)
Male (X_2)	0.458	0.457	0.451
	(0.021)	(0.020)	(0.020)
Age (X_3)		0.011	0.011
		(0.001)	(0.001)
North (X_4)			0.175
			(0.037)
South (X_5)			0.103
			(0.033)
East (X_6)			-0.102
			(0.043)
Intercept	12.840	12.471	12.390
	(0.018)	(0.049)	(0.057)
Summary and joint tests			
F-statistic			21.87
SER	1.026	1.023	1.020
$\bar{R}^2$	0.0710	0.0761	0.0814
n	10973	10973	10973

The table reports the results on a series of regressions estimated on more than 10,000 individuals who worked full-time for the whole year in a developing country.

The analysis concerned both quantitative and dummy variables[1]:

- AHE = logarithm of average hourly earnings (in 2007 units)
- College = dummy variable (1 if graduate, 0 otherwise)
- Male = dummy variable (1 if male, 0 if female)
- Age = age (in years)
- North = dummy variable (1 if Region = North, 0 otherwise)
- East = dummy variable (1 if Region = East, 0 otherwise)
- South = dummy variable (1 if Region = South, 0 otherwise)
- West = dummy variable (1 if Region = West, 0 otherwise)

1. For each of the three regressions, add * (level 5%) and ** (level 1%) to the table to indicate the significance of the coefficients.

2. Answer using the regression results reported in column (2)

   (a) Is age an important determinant of earnings?

   (b) Use an appropriate test statistic or confidence interval to justify your answer.

   (c) Suppose Alvo is a 30-year-old college graduate and Kal is a 40-year-old college graduate. Construct a 95% confidence interval for the difference between expected wages.

3. Using the values reported in the table:

   (a) Construct the R^2 for each of the regressions.

   (b) Show how to construct the F statistic for testing the hypothesis that $\beta_4 = \beta_5 = \beta_6 = 0$ in the regression shown in column (3).

   (c) Is the statistic significant at the 1% level?

   (d) Show how to construct the F statistic to test the hypothesis that $\beta_4 = \beta_5 = \beta_6 = 0$ in the regression shown in column (3) using the Bonferroni test.

   (e) Construct a 99% confidence interval for β_1 in the regression shown in column (3).

[1] A dummy variable is one that takes the values 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome.

Solution

1. Calculating the test statistic:

$$t_i = \frac{\beta_i}{S.E.(\beta_i)}$$

and comparing it with the critical value at 5% ($t = 1.96$) and 1% ($t = 2.58$), the following table is obtained (in which the standard errors are indicated in round brackets and test statistics in square brackets):

Regressor	**(1)**	**(2)**	**(3)**
College (X_1)	0.352^{**}	0.373^{**}	0.371^{**}
	(0.021)	(0.021)	(0.021)
	[16.762]	[17.762]	[17.667]
Male (X_2)	0.458^{**}	0.457^{**}	0.451^{**}
	(0.021)	(0.020)	(0.020)
	[21.810]	[22.850]	[12.550]
Age (X_3)		0.011^{**}	0.011^{**}
		(0.001)	(0.001)
		[11.000]	[11.000]
North (X_4)			0.175^{**}
			(0.037)
			[4.730]
South (X_5)			0.103^{**}
			(0.033)
			[3.121]
East (X_6)			-0.102^{*}
			(0.043)
			[-2.372]
Intercept	$12,840^{**}$	$12,471^{**}$	$12,390^{**}$
	(0.018)	(0.049)	(0.057)
	[692.833]	[254.510]	[217.368]

2. (a) Yes, age is an important determinant of wages.

   Using the test statistic, we get:

$$t = \frac{0.011}{0.001} = 11$$

The test statistic is greater than 2.58 and this implies that the age coefficient is statistically significant at the 1% level.

(b) The (1-α)% confidence interval of a coefficient β_i is:

$$(1-\alpha)\%CI(\beta_i) = [\beta_i \pm z_{1-\alpha/2} \times \text{SE}(\beta_i)]$$

In this case the 95% confidence interval of β_3 is:

$$95\%CI(\beta_3) = [\beta_3 \pm z_{0.025} \times \text{SE}(\beta_3)] = [0.011 \pm 1.96 \times 0.001] = [0.009; 0.013]$$

(c) To construct the confidence interval we need to consider the age difference of the two subjects ($\Delta Age = 40 - 30 = 10$ years).
The confidence interval will be:

$$\Delta Age \times [\beta_3 \pm z_{0.025} \times \text{SE}(\beta_3)] = 10 \times [0.011 \pm 1.96 \times 0.001] = [0.0904; 0.1296]$$

As a percentage it is a difference in expected wages between 9.04% and 12.96%.

3. (a) From the formula of the Adjusted-R^2 ($\bar{R}^2$), we can get R^2. In particular, knowing that

$$\bar{R}^2 = 1 - \frac{n-1}{n-k} \times (1 - R^2) = 1 - \frac{n-1}{n-k} \times \frac{RSS}{TSS}$$
$$\rightarrow R^2 = 1 - \frac{n-k}{n-1}(1 - \bar{R}^2)$$

.

It follows that:

- column (1) = $R^2 = 1 - \frac{10973-3}{10973-1}(1 - 0.0710) = 0.0712$
- column (2) = $R^2 = 1 - \frac{10973-4}{10973-1}(1 - 0.0761) = 0.0764$
- column (3) = $R^2 = 1 - \frac{10973-7}{10973-1}(1 - 0.0814) = 0.0819$

(b) $H_0 : \beta_4 = \beta_5 = \beta_6 = 0 \qquad H_1 : \beta_i \neq 0$, for $i = 4, 5, 6$.

Knowing that:

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

we can test the null hypothesis starting from the R^2 of the regressions. For the non-restricted, the R^2 is that of column (3), while for the restricted we are going to consider the R^2 of column (2).
Considering the number of restrictions $q = 3$ and $n = 10973$, it follows that

$$F = \frac{0.0819 - 0.0764/3}{(1 - 0.0819)/(10973 - 7)} = 21.898$$

.

(c) The critical value at 1% $F_{3,\infty} = 3.78$, is less than the F statistic, so we reject the null hypothesis at 1% significance level.

(d) The Bonferroni test allows testing joint hypotheses on q coefficients starting from the t statistics relating to individual hypotheses but correcting the critical value as follows:

$$c = z_{1-\frac{\alpha/q}{2}}$$

H_0 is rejected if at least one of the individual t-ratios is, in absolute value, greater than the critical value c, adjusted as above.

In our case, H_0 is a joint hypothesis on 3 coefficients, therefore $q = 3$. Setting $\alpha = 1\%$, we have:

$$c = z_{1-\frac{\alpha/q}{2}} = z_{1-\frac{0.01/3}{2}} = z_{0.9983} = 2.93$$

The t-statistics of the single coefficients subject to the null hypothesis are $t_4 = 4.730$, $t_5 = 3.121$ and $t_6 = -2.372$. Since $|t_6| < c$ but $|t_5| > c$ and $|t_4| > c$, the null hypothesis is rejected.

(e) $[\beta_1 \pm 2.58 \times \mathrm{SE}(\beta_1)] = [0.371 \pm 2.58 \times 0.021] = [0.3168; 0.4252]$