TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

Quantitative Methods III - Practice 4
Multiple Linear Regression

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

Exercise 1 The following table shows the results of 3 multiple regression models considering the average hourly wage (Y) of 7178 workers.

Regressor	**(1)**	**(2)**	**(3)**
College (X_1)	9.85*** (1.20)	9.72*** (1.18)	9.70*** (1.15)
Female (X_2)	-5.12*** (0.98)	-4.95*** (0.96)	-4.90*** (0.95)
Age (X_3)		0.55** (0.22)	0.53** (0.21)
Northeast (X_4)			0.80* (0.45)
Midwest (X_5)			-1.32** (0.52)
South (X_6)			-0.38 (0.50)
Intercept	17.50*** (2.10)	0.20 (1.95)	0.45 (1.90)
Summary statistics			
SER	11.85	11.70	11.65
R^2	0.165	0.182	0.185
n	7178	7178	7178

Specifically, the variables are:

- AHE (Y): average hourly earnings
- College (X_1): dummy variable (1 = graduated, 0 = not graduated)
- Female (X_2): dummy variable (1 = female, 0 = male)
- Age (X_3): age in years
- Northeast (X_4): dummy variable (1 if region = North-east, 0 otherwise)

- Midwest (X_5): dummy variable (1 if region = Midwest, 0 otherwise)
- South (X_6): dummy variable (1 if region = South, 0 otherwise)

1. Calculate $\bar{R}^2$ for each of the regressions.
2. **Limited to Model (1):** Do graduate workers earn on average more than graduate workers? If yes, how much more?
3. **Limited to Model (2):** Predict the wages of Marianna, a 45-year-old college graduate, and Lorenzo, a 48-year-old college graduate.
4. **A**re the model coefficients statistically significant?

   Limited to Model (3):

5. Calculate a 95% confidence interval for the effect of the variable Age on hourly earnings.
6. Suppose Francesca is a 30-year-old college graduate and Anna is a 40-year-old college graduate. Construct a 95% confidence interval for the difference between expected wages.
7. Is it possible that the wage difference is \$1 for each year of age?
8. Show how to construct the F statistic for testing the hypothesis that $\beta_4 = \beta_5 = \beta_6 = 0$.

Solutions The table presents three regression models: columns (1), (2), and (3).
Each row corresponds to a different independent variable (regressor) and coefficients are reported with standard errors in parentheses.
The three regression models are:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \quad \text{(Model 1)}$$
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \quad \text{(Model 2)}$$
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \varepsilon \quad \text{(Model 3)}$$

College (X_1): Highly positive coefficient (9.7-9.85), suggesting that having a college education is strongly associated with higher earnings.
Female (X_2): Negative and significant effect, indicating a gender wage gap.
Age (X_3): Showing a small but significant positive effect.
Region (X_4 to X_6): With mixed effects (e.g., living in the Midwest is associated with lower earnings).

1. Recalling the Adjusted-R^2 formula,

$$\bar{R}^2 = 1 - \frac{n-1}{n-k}(1 - R^2),$$

- column (1): $\bar{R}^2 = 1 - \frac{7178-1}{7178-3}(1 - 0.165) = 0.1648$
- column (2): $\bar{R}^2 = 1 - \frac{7178-1}{7178-4}(1 - 0.182) = 0.1817$

- column (3): $\bar{R}^2 = 1 - \frac{7178-1}{7178-7}(1 - 0.185) = 0.1843$

2. The coefficient of the dummy variable *College* (equal to 1 if the subject has a degree, 0 if not) is equal to 9.85; this means that graduate workers earn on average 9.85$/hour more than non-graduate workers.

3. On average, a worker earns $0.55 an hour more for each year of age.

   Expected Wage for Marianna (a 45-year-old female college graduate):
   $0.20 + (9.72 \times 1) + (-4.95 \times 1) + (0.55 \times 45) = \$29.72$ per hour.

   Expected Wage for Lorenzo (a 48-year-old male college graduate):
   $0.20 + (9.72 \times 1) + (-4.95 \times 0) + (0.55 \times 48) = \$36.32$ per hour.

   The difference is $6.6 per hour:

$$\Delta Y = \Delta AHE = AHE_{Lorenzo} - AHE_{Marianna} = \Delta Age + \Delta Female =$$
$$= (48 - 45) \times \$0.55 + (0 - 1) \times (-\$4.95) = \$1.65[X_2] + \$4.95[X_3] = \$6.6$$

4. To calculate the *t-value* of the coefficients we have to:

$$t - value = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

   In the figure we have the *t-value* and the *p-value* for each coefficient.

   Regression Results (t-values in brackets; statistical significance indicated by: $^{***}: p < 0.01$ - $^{**}: p < 0.05$ - $^{*}: p < 0.10$):

Regressor	**(1)**	**(2)**	**(3)**
College (X_1)	9.85*** (1.20) [8.21]	9.72*** (1.18) [8.24]	9.70*** (1.15) [8.43]
Female (X_2)	-5.12*** (0.98) [-5.22]	-4.95*** (0.96) [-5.16]	-4.90*** (0.95) [-5.16]
Age (X_3)		0.55** (0.22) [2.50]	0.53** (0.21) [2.52]
Northeast (X_4)			0.80* (0.45) [1.78]
Midwest (X_5)			-1.32** (0.52) [-2.54]
South (X_6)			-0.38 (0.50) [-0.76]
Intercept	17.50*** (2.10) [8.33]	0.20 (1.95) [0.10]	0.45 (1.90) [0.24]
Summary statistics			
SER	11.85	11.70	11.65
R^2	0.165	0.182	0.185
n	7178	7178	7178

   The t-values help assess the statistical significance of each coefficient. Generally, we compare them against critical values from the standard normal distribution:

   - For $\alpha = 0.10$: $|t| > 1.645$

- For $\alpha = 0.05$: $|t| > 1.96$
- For $\alpha = 0.01$: $|t| > 2.576$

The coefficients of the 3 Models are:

- **College (X_1)**: High t-values (8.21, 8.24 and 8.43) indicate a very strong and highly significant positive effect of having a college education on the dependent variable.
- **Female (X_2)**: Negative t-values (-5.22 and -5.16), all far below -2.576 ($\alpha = 0.01$), confirm a highly significant gender effect, likely indicating a gender wage gap.
- **Age (X_3)**: t-values (2.50-2.52) are above 1.96 ($\alpha = 0.05$), meaning age has a statistically significant positive effect at the 5% level.
- **Northeast (X_4)**: The t-value (1.78) is slightly below 1.96 ($\alpha = 0.05$) but above 1.645 ($\alpha = 0.1$), suggesting a weak significance at the 10% level.
- **Midwest (X_5)**: The t-value (-2.54) is close to -2.576, implying a significant negative effect at the 5% level.
- **South (X_6)**: The t-value (-0.76) is well above -1.645, indicating no statistical significance.
- **Intercept**: High t-value in Model 1 (8.33), but nearly zero in Models 2 and 3, suggesting omitted variable bias in the simplest model.

5. The 95% confidence interval for the *Age* (X_3) coefficient in Model 3 is calculated as:

$$(1-\alpha)\% \cdot CI(\beta_i) = \hat{\beta}_i \pm z_{\alpha/2} \cdot SE(\hat{\beta}_i)$$

where:

- $\hat{\beta}_3 = 0.53$
- $SE(\hat{\beta}_3) = 0.21$
- $(1-\alpha)\% = 95\% \rightarrow \alpha/2 = 2.5\%$
- $z_{\alpha/2} = 1.96$ (for 95% confidence level)

$$CI(\hat{\beta}_3) = 0.53 \pm (1.96 \times 0.21) = [\$0.12; \$0.94]$$

6. Both Anna and Francesca are women and graduates. The difference between the two is age.

   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:

$$\Delta Age \times [\hat{\beta}_3 \pm z_{0.025} \times \mathrm{SE}(\hat{\beta}_3)] = 10 \times [0.53 \pm 1.96 \times 0.21] = [\$1.2; \$9.4]$$

7. We test the null hypothesis that the coefficient for Age is equal to 1:

$$t = \frac{\hat{\beta}_3 - \beta_0}{SE(\hat{\beta}_3)} = \frac{0.53 - 1}{0.21} = \frac{-0.47}{0.21} = -2.24$$

Since $|t| = 2.24$ exceeds the critical value of 1.96 for a 95% confidence level, we **reject** the hypothesis that the impact of age is exactly \$1 per year.

We could have reached the same conclusions by looking at the 95% confidence interval, which does not include the \$1 value.

8. $H_0 : \beta_4 = \beta_5 = \beta_6 = 0 \qquad H_1 : \beta_i \neq 0$, 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 Model (3), while for the restricted we are going to consider the R^2 of Model (2).

Considering the number of restrictions $q = 3$ and $n = 7178$, it follows that

$$F = \frac{(0.185 - 0.165)/3}{(1 - 0.185)/(7178 - 7)} = 58.66$$

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.

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4)

Residuals:
    Min      1Q  Median      3Q     Max
-373215  -35779   -2688   32137  745175

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8934.7626 17205.4495
X1             15.2476     1.7871
X2             -0.3141     1.1286
X3            -12.9828     2.1917
X4              4.0356     2.1431
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 136200 on 102 degrees of freedom
Multiple R-squared:  0.7525,    Adjusted R-squared:
F-statistic: 77.54 on 4 and 102 DF,  p-value: < 2.2e-16
```

Figure 1

Exercise 2 Given the **Figure 1**, calculate:

1. The adjusted R^2.
2. Analyze the goodness of fit of the regression model.
3. The significance of the regression coefficients.
4. Test the hypothesis that variables $X3$ and $X4$ are jointly not significant, considering that the regression model with independent variables $X1$ and $X2$ has a coefficient of determination equal to 0.68.
5. Is there an alternative method to test the joint significance of the two coefficients? If so, apply it.

Solutions

1. The formula of the R^2_{adj} is:

$$R^2_{\text{adj}} = 1 - \frac{n-1}{n-k} \cdot (1 - R^2)$$

Given:

- $R^2 = 0.7525$ (coefficient of determination),
- $n = 107$ (number of observations),
- $k = 5$ (number of coefficients),

we compute:

$$R^2_{\text{adj}} = 1 - \frac{107-1}{107-5} \cdot (1 - 0.7525) = 1 - \frac{106}{102} \cdot (0.2475) = 0.7428$$

Thus, the adjusted R^2 value is 0.7428.

2. From the **Figure 1**:

- The goodness-of-fit is calculated by the **Adjusted** R^2 that is equal 0.7428, meaning that 74.28% of the variance in Y is explained by the model, accounting for the number of predictors.
- The **Residual Standard Error (***SER***)** is equal to 136.200, which provides an estimate of the standard deviation of residuals.

```
R 4.3.0 · ~/

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4)

Residuals:
    Min      1Q  Median      3Q     Max
-373215  -35779   -2688   32137  745175

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8934.7626 17205.4495  -0.519   0.6047
X1             15.2476     1.7871   8.532 1.41e-13 ***
X2             -0.3141     1.1286  -0.278   0.7813
X3            -12.9828     2.1917  -5.924 4.29e-08 ***
X4              4.0356     2.1431   1.883   0.0625 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 136200 on 102 degrees of freedom
Multiple R-squared:  0.7525,    Adjusted R-squared:  0.7428
F-statistic: 77.54 on 4 and 102 DF,  p-value: < 2.2e-16
```

3. To calculate the *t-value* of the coefficients we have to:

$$t - value = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

In the figure we have the *t-value* and the *p-value* for each coefficient.

The significance of each coefficient is assessed using the *p-value*:

- **X1** is highly significant ($p = 1.41 \times 10^{-13}$), meaning it has a strong relationship with Y.
- **X2** is not significant ($p = 0.7813$), implying it does not contribute much to explaining Y.
- **X3** is highly significant ($p = 4.29 \times 10^{-8}$), indicating a strong relationship with Y.
- **X4** is borderline significant at the 10% level ($p = 0.0625$), suggesting a weak association.
- The **intercept** is also not significant ($p = 0.6047$).

4. We test whether $X3$ and $X4$ jointly add explanatory power using the **F-test for joint significance**.

$$H_0 : \beta_3 = \beta_4 = 0 \quad \text{(both coefficients are zero)}$$
$$H_1 : \text{At least one of } \beta_3 \text{ or } \beta_4 \text{ is different from zero.}$$

The **F-statistic** is given by:

$$F = \frac{(R^2_{\text{full}} - R^2_{\text{restricted}})/q}{(1 - R^2_{\text{full}})/(n-k)} \tag{1}$$

where:

- $R^2_{\text{full}} = 0.7525$ (model with all predictors)
- $R^2_{\text{restricted}}$ is the R^2 of the model without $X3$ and $X4$. Let us assume it is 0.68.
- $q = 2$ (number of restricted variables)
- $n = 107$ (sample size)
- $k = 5$ (total number of estimated parameters, including the intercept)

Substituting the values:

$$\begin{aligned} F &= \frac{(0.7525 - 0.68)/2}{(1 - 0.7525)/(107 - 5)} \\ &= \frac{0.0725/2}{0.2475/102} \\ &= \frac{0.03625}{0.002426} \\ &= 14.95 \end{aligned}$$

Since this F-statistic is large, we **reject** H_0, meaning that $X3$ and $X4$ are jointly significant.

5. Bonferroni correction adjusts for multiple comparisons. With a family-wise error rate of $\alpha = 0.05$ and two hypotheses ($X3$ and $X4$), the adjusted significance threshold is:

$$c = \frac{\alpha}{2} = \frac{0.05}{2} = 0.025 \tag{2}$$

- $p(X3) = 4.29 \times 10^{-8}$ is still highly significant ($p < 0.001$).
- $p(X4) = 0.0625$ is **not significant** under Bonferroni correction.