,

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. Derive the *OLS* estimators in the multiple regression model $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$, where $\mathbf{y}$ is vector $n \times 1$ of the dependent variable, $\mathbf{X}$ is the $n \times k$ matrix of the regressors, β is the coefficient vector $k \times 1$ and $\mathbf{u}$ is the vector $n \times 1$ of errors.

2. The following table shows the results of 3 multiple regression models considering the average hourly wage of 7178 workers.

Regressor	**(1)**	**(2)**	**(3)**
College (X_1)	10.47	10.44	10.42
Female (X_2)	-4.69	-4.56	-4.57
Age (X_3)		0.61	0.61
Northeast (X_4)			0.74
Midwest (X_5)			-1.54
South (X_6)			-0.44
Intercept	18.15	0.11	0.33
Summary statistics			
SER	12.15	12.03	12.01
R^2	0.165	0.182	0.185
n	7178	7178	7178

Specifically, the variables are:

- AHE: average hourly wage

- College: dummy variable (1 = graduated, 0 = not graduated)
- Female: dummy variable (1 = female, 0 = male)
- Age: age in years
- Northeast: dummy variable (1 if region = North-east, 0 otherwise)
- Midwest: dummy variable (1 if region = Midwest, 0 otherwise)
- South: dummy variable (1 if region = South, 0 otherwise)
- West: dummy variable (1 if region = West, 0 otherwise)

2.a) Calculate $\bar{R}^2$ for each of the regressions.

Limited to column 1:

2.b) Do graduate workers earn on average more than graduate workers? If yes, how much more?

2.c) Do men earn more than women on average? How much more?

Limited to column 2:

2.d) Is age an important determinant of wages? Predict the wages of a 29-year-old university graduate and a 34-year-old university graduate.

Limited to column 3:

2.e) Are there important regional differences? Why is the West variable excluded from the regression?

2.f) Calculate the expected difference in the salaries of a 28-year-old college graduate from the South and a 28-year-old college graduate from the Midwest.

Solutions

1. Given the multiple linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \qquad \text{with} \qquad \mathbf{u} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta},$$

where $\mathbf{y}$ is vector $n \times 1$ of the dependent variable, $\mathbf{X}$ is the $n \times k$ matrix of predictors, $\boldsymbol{\beta}$ is the vector $k \times 1$ of the coefficients and $\mathbf{u}$ is the vector $n \times 1$ of the residuals.

Knowing that the Least Squares Method aims to minimize the sum of the squared residuals, the *OLS* estimators is obtained by minimizing

$$\begin{aligned}
\mathbf{u}'\mathbf{u} =& (\mathrm{y} - \mathrm{X}\boldsymbol{\beta})'(\mathrm{y} - \mathrm{X}\boldsymbol{\beta}) \\
=& \mathrm{y}'\mathrm{y} - \boldsymbol{\beta}'\mathrm{X}'\mathrm{y} - \mathrm{y}'\mathrm{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathrm{X}'\mathrm{X}\boldsymbol{\beta} \\
=& \mathrm{y}'\mathrm{y} - 2\boldsymbol{\beta}'\mathrm{X}'\mathrm{y} + \boldsymbol{\beta}'\mathrm{X}'\mathrm{X}\boldsymbol{\beta}
\end{aligned}$$

To minimize $\mathbf{u}'\mathbf{u}$ we have to set the derivative with respect to $\boldsymbol{\beta}$ equal to zero. Consequentially,

$$\frac{\delta \mathbf{u}'\mathbf{u}}{\delta \boldsymbol{\beta}} = -\,2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = 0$$

$$2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = 2\mathbf{X}'\mathbf{y} \rightarrow \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

2.a) 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.b) The coefficient of the dummy variable *College* (equal to 1 if the subject has a degree, 0 if not) is equal to 10.47; this means that graduate workers earn on average 10.47$/hour more than non-graduate workers.

2.c) Following the same reasoning we can state that, given the coefficient of the dummy variable *Female*, women earn on average 4.69$/hour less than men.

2.d) On average, a worker earns \$0.61 an hour more for each year of age.

Expected Wage for a 29-year-old female college graduate:
$0.11 + 10.44 \times 1 - 4.56 \times 1 + 0.61 \times 29 = \$23.68$ per hour.

Expected Wage for a 34-year-old female college graduate:
$0.11 + 10.44 \times 1 - 4.56 \times 1 + 0.61 \times 34 = \$26.73$ per hour.

The difference is \$3.05 per hour ($= (34 - 29) \times 0.61$).

2.e) The regional differences are highlighted by the difference in the coefficients of the dummy variables *Northeast*, *Midwest* and *South*.

Taking into consideration the *West* as a basic mode, it results that:

- workers in the *Northeast* earn an average of \$0.74 hour more than workers in the *West*;
- workers in the *Midwest* earn an average of \$1.54 an hour less than workers in the *West*;
- workers in the *South* earn an average of \$0.44 per hour less than workers in the *West*.

The variable *West* is excluded from the regression to avoid the so-called "*liquidity trap*". In fact, if this variable were also included in the regression we would have perfect collinearity between the predictors: in this case, the constant would turn out to be a linear combination of the dummies *Northeast*, *Midwest*, *South* and *West* (the sum of which would be exactly equal to 1).

2.f) The expected difference in wages of a 28-year-old graduate from the *South* and a 28-year old graduate from the *Midwest* is: $-0.44 - (-1.54) = \$1.1$ per hour