# Multiple Linear Regression

Part 4

**EE325**

**Introductory Econometrics**

Revision Aug 2020

# List of the topics to cover

## Multiple linear regression

- Concept: adding more independent variables into our model.
- Result of the OLS, properties and assumptions.
- The Coefficient of Determination $R^2$ (not the same as $r^2$).

## Hypothesis testing

- Concept: Individual test and joint test
- Restricted and unrestricted model
- Testing overall significance and equality of the coefficients
- Test for structural change

## Model with 2 or more independent variables

G. 188

If we add more independent variable(s) into our model to increase fitness to the model, the specification of the stochastic form becomes

• $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{u}_i$

We are not going to use the notation $X_{1i}$ to make our estimators number corresponded with the variable names. Hence the SRF becomes

• $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}$

Now $\hat{\beta}_2$ and $\hat{\beta}_3$ are known as '**partial slope coefficients**' since when we consider

• $\hat{\beta}_2$ represents the change in $\hat{Y}_i$ when $X_{2i}$ increases for 1 unit, holding everything else constant.

• $\hat{\beta}_3$ represents the change in $\hat{Y}_i$ when $X_{3i}$ increases for 1 unit, holding everything else constant.

If we add more independent variables into this model, the specification becomes

• $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \hat{\beta}_k X_{ki} + \hat{u}_i$

Let's focus on 2 independent variables model first.

**Assumptions**

(1) Linear in parameters.

(2) All $X_i$ are independent of the error term $u_i$.

(3) Zero mean of error term $E(u_i|X_{2i}, X_{3i}) = 0$.

(4) Homoscedasticity $Var(u_i) = \sigma^2$.

(5) No autocorrelation $cov(u_i, u_j) = 0$.

*(6)* $n > k$ and $X_i$ must not all be the same.

(7) No specification bias.

(8) No exact collinearity between $X_i$.

The 8th assumption becomes more significant in this model. Imagine that

• $X_{2i} = 2X_{3i}$

we then can turn this model into

• $Y_i = \hat{\beta}_1 + \hat{\beta}_2 2X_{3i} + \hat{\beta}_3 X_{3i} + \hat{u}_i$ and then $\hat{\beta}_2 = \hat{\beta}_3$ so

• $Y_i = \hat{\beta}_1 + (\hat{\beta}_2 + 2\hat{\beta}_3) X_{3i} + \hat{u}_i$

which means that either $X_{2i}$ or $X_{3i}$ does not add any more information to this model.

## (1) OLS estimators

G. 192

We follow the same procedure as when we did, minimizing the term $\sum \hat{u}_i^2$

- $\min_{\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3} \sum \hat{u}_i^2 = \sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i}\right)^2$

Skipping all the prove because we utilize the same logic of minimization of a function with calculus, we get

- $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}_2 - \hat{\beta}_3 \bar{X}_3$
- $\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\sum x_{3i}^2) - (\sum y_i x_{3i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2}$
- $\hat{\beta}_3 = \frac{(\sum y_i x_{3i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2}$

**Notations**

- $y_i = Y_i - \bar{Y}$
- $x_{2i} = X_{2i} - \bar{X}$
- $x_{3i} = X_{3i} - \bar{X}$

**Variances of OLS estimators**

- $Var(\hat{\beta}_1) = \sigma^2 [\frac{1}{n} + \frac{\bar{X}_2^2 \sum x_{3i}^2 + \bar{X}_3^2 \sum x_{2i}^2 - 2\bar{X}_2^2 \bar{X}_3^2 \sum x_{2i} x_{3i}}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2}]$
- $Var(\hat{\beta}_2) = \sigma^2 [\frac{\sum x_{3i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2}]$
- $Var(\hat{\beta}_3) = \sigma^2 [\frac{\sum x_{2i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2}]$

As for the estimator of $\sigma^2$, we have $\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-3}$.
Furthermore, we can flip these formula a bit as equivalently as

- $Var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)}$
- $Var(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2 (1 - r_{23}^2)}$

where $r_{23}$ is the coefficient of correlation between $X_2$ and $X_3$.

All the OLS estimator properties from simple linear regression **still apply**.

## (2) Coefficient of Determination

G. 196

According to the concept of measuring $r^2$

- $r^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = R^2$

where $TSS$ is the total sum of squares,

$ESS$ is the estimated sum of squares and

$RSS$ is the residual sum of squares.

The $TSS$ is easy to understand, which is the sum of deviation from the mean or

- $TSS = \sum(Y_i - \bar{Y})^2 = \sum y_i^2$

while the $RSS$ is the part which cannot be estimated by the independent variables or

- $RSS = \sum \hat{u}_i^2 = \sum y_i^2 - \hat{\beta}_2 \sum y_i x_{2i} - \hat{\beta}_3 \sum y_i x_{3i}$

(see proof on the right-hand side) and the $ESS$ is the explained part or

- $ESS = \sum \hat{y}_i^2 = \hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i}$

Therefore, from

- $R^2 = \frac{ESS}{TSS} = \frac{\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i}}{\sum y_i^2}$ or
- $R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum y_i^2 - \hat{\beta}_2 \sum y_i x_{2i} - \hat{\beta}_3 \sum y_i x_{3i}}{\sum y_i^2}$

**Proof**

- From $\hat{u}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i}$

  $= y_i - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i}$

- Now $\sum \hat{u}_i^2 = \sum(\hat{u}_i \hat{u}_i) = \sum \hat{u}_i (y_i - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i})$

  $= \sum \hat{u}_i y_i = \sum y_i \hat{u}_i = \sum y_i (y_i - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i})$

  $= \sum y_i^2 - \hat{\beta}_2 \sum y_i x_{2i} - \hat{\beta}_3 \sum y_i x_{3i}$

## (3) Adjusted Coefficient of Determination

G. 201

When we add more and more independent variables into the model, the coefficient of determination is likely to increase (at least it will not decrease) from decreasing $\sum \hat{u}_i^2$.

Recall that when we define $R^2$ as

- $R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum \hat{u}_i^2}{\sum y_i^2}$

Now we define **adjusted $R^2$**, denoted as

- $\bar{R}^2 = 1 - \frac{\sum \hat{u}_i^2/(n-k)}{\sum y_i^2/(n-1)} = 1 - (1 - R^2)\frac{n-1}{n-k}$

The word **adjusted** means that the $R^2$ is adjusted for the degrees of freedom associated with the sums of the squares entering the specification. Note that the d.f. of

- $\sum y_i^2$ is $n - 1$ and
- $\sum \hat{u}_i^2$ is $n - k$.

**Comparison between $R^2$ and $\bar{R}^2$**

(1) $0 \leq R^2 \leq 1$ but $\bar{R}^2$ can be negative (taken as 0).

(2) As $k$ increases, $R^2$ is increasing but $\bar{R}^2$ may not.

(3) $\bar{R}^2 < R^2$

## (4) Examples



### (1) The Cobb-Douglas Production Function

Usual form of the Cobb-Douglas function is

- $Y = AK^{\alpha}L^{\beta}$

where $Y$ is the value of output of an economy,

$A$ is total factor productivity (TFP or sometimes simplified as production technology),

$K$ is number of capital input,

$L$ in number of labor input,

$\alpha$ and $\beta$ is the output elasticity.

The stochastic form can be expressed as

- $Y_i = AK_i^{\alpha}L_i^{\beta}e^{u_i}$

Taking natural logarithm to enable linear estimation yields

- $\ln Y_i = \ln A + \alpha \ln K_i + \beta \ln L_i + u_i$

where $A$, $\alpha$ and $\beta$ are the estimators.

### Example of result from the US

The data obtained from manufacturing sector of all states in the US, represented for each observation $i$. The results of linear regression is as follows.

- $\widehat{\ln Y_i} = 3.8876 + 0.5213 \ln K_i + 0.4683 \ln L_i$

$t = (9.8115) \quad (5.3803) \qquad (4.7342) \quad d.f. = 48$

$R^2 = 0.9642 \qquad \bar{R}^2 = 0.9627$

We can test each estimator for its significance or we can also jointly test $\alpha + \beta$ to check returns to scale such as

- $\mathrm{H}_0$: $\alpha + \beta = 1$ or constant returns to scale
- $\mathrm{H}_a$: otherwise.

## (4) Examples



### (2) Polynomial models

There are multiple economic models incorporating polynomial form, such as total cost and marginal cost, an example here is the effect of age to wage in the quadratic form. Given that wage is a function of age as follows (in the stochastic form).

- $w_i = \hat{\beta}_1 + \hat{\beta}_2 age_i + \hat{\beta}_3 age_i^2 + u_i$

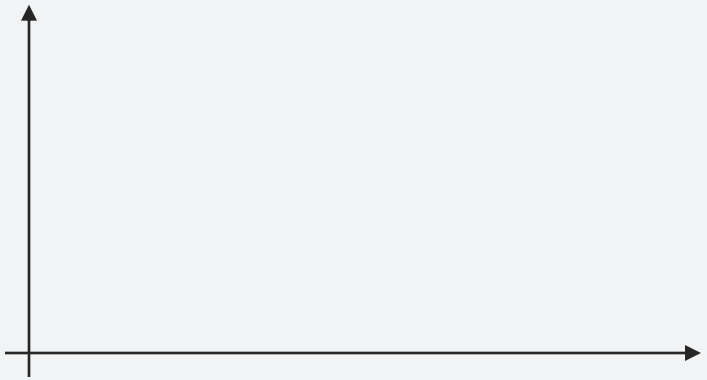
where $w_i$ is the value of output of an economy,

$age_i$ is straightforward.

$age_i^2$ is the squares of age and

$\hat{\beta}_1, \hat{\beta}_2$ and $\hat{\beta}_3$ are the estimators.

*Fitted curve of wage and age.*



### Example of result

Given that

- $\widehat{w_i} = 3.73 \quad 0.298 age_i \quad 0.0061 age_i^2$

Can you guess the sign of these estimators?