# Multiple linear regression
# Lecture: 12

Hamid Raza
Assistant Professor in Economics
raza@business.aau.dk

Aalborg University

Statistics

# Outline

## Assumption 1

- In the population model, the relationship of the dependent variable **y** with the independent variable **x** and the error (or disturbance) $\varepsilon$ is given by:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

The model is linear in the parameters.

## Assumption 1

- In the population model, the relationship of the dependent variable **y** with the independent variable **x** and the error (or disturbance) $\varepsilon$ is given by:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

The model is linear in the parameters.

## Assumption 2

- We assume to have a random sample of size $n$. The x values are realizations of random variable X that are independent of the error terms

## Assumption 1

- In the population model, the relationship of the dependent variable **y** with the independent variable **x** and the error (or disturbance) $\varepsilon$ is given by:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

  The model is linear in the parameters.

## Assumption 2

- We assume to have a random sample of size $n$. The x values are realizations of random variable X that are independent of the error terms

## Assumption 3

- The error terms are random variables, $\varepsilon_i (i = 1, ..., n)$, which have a mean of 0 and variance $\sigma^2$. This property is called homoscedasticity, or uniform variance:

$$E(\varepsilon_i) = 0 \text{ and } E(\varepsilon_i^2) = \sigma^2 \tag{1}$$

## Assumption 4

- The random error terms, $\varepsilon$, are not correlated with one another, so that

$$E[\varepsilon_i \varepsilon_j] = 0 \tag{2}$$

## Assumption 4

- The random error terms, $\varepsilon$, are not correlated with one another, so that

$$E[\varepsilon_i \varepsilon_j] = 0 \tag{2}$$

## Assumption 5

- There is no direct linear relationship between the independent variables.

# Outline

# Multiple linear regression

## Example: the determinants of imports

# Multiple linear regression

## Example: the determinants of imports

- Economic theory argues that factors such as income of a country (GDP) and price competitiveness affect the demand for its imports. This simple argument leads to a model such as:

$$imports = \beta_0 + \beta_1 Y + \beta_2 REER + e$$

where Y is GDP, and REER is real effective exchange rate (which is a proxy of price competitiveness)

# Multiple linear regression

## Example: the determinants of imports

- Economic theory argues that factors such as income of a country (GDP) and price competitiveness affect the demand for its imports. This simple argument leads to a model such as:

$$imports = \beta_0 + \beta_1 Y + \beta_2 REER + e$$

  where Y is GDP, and REER is real effective exchange rate (which is a proxy of price competitiveness)

- In general, a multiple linear regression is given by:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ..... + \beta_k x_k + \varepsilon$$

# Multiple linear regression

- Assume a multiple linear regression model with two independent variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

# Multiple linear regression

- Assume a multiple linear regression model with two independent variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

- The intercept in a multiple linear regression is given by:

$$\beta_0 = \bar{y} - \beta_1 \bar{x_1} + \beta_2 \bar{x_2}$$

# Multiple linear regression

- Assume a multiple linear regression model with two independent variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

- The intercept in a multiple linear regression is given by:

$$\beta_0 = \bar{y} - \beta_1 \bar{x_1} + \beta_2 \bar{x_2}$$

- The estimator $\beta_1$ is now estimated as follows:

$$\beta_1 = \frac{S_y(r_{x_1 y} - r_{x_1 x_2} r_{x_2 y})}{S_{x_1}(1 - r^2_{x_1 x_2})}$$

# Multiple linear regression

- Assume a multiple linear regression model with two independent variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

- The intercept in a multiple linear regression is given by:

$$\beta_0 = \bar{y} - \beta_1 \bar{x_1} + \beta_2 \bar{x_2}$$

- The estimator $\beta_1$ is now estimated as follows:

$$\beta_1 = \frac{S_y(r_{x_1 y} - r_{x_1 x_2} r_{x_2 y})}{S_{x_1}(1 - r^2_{x_1 x_2})}$$

$r$ stands for sample correlation, $S$ stands for standard deviation

# Multiple linear regression

- Assume a multiple linear regression model with two independent variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

- The intercept in a multiple linear regression is given by:

$$\beta_0 = \bar{y} - \beta_1 \bar{x_1} + \beta_2 \bar{x_2}$$

- The estimator $\beta_1$ is now estimated as follows:

$$\beta_1 = \frac{S_y(r_{x_1 y} - r_{x_1 x_2} r_{x_2 y})}{S_{x_1}(1 - r^2_{x_1 x_2})}$$

$r$ stands for sample correlation, $S$ stands for standard deviation

- The estimator $\beta_2$ is estimated as follows:

$$\beta_2 = \frac{S_y(r_{x_2 y} - r_{x_1 x_2} r_{x_1 y})}{S_{x_2}(1 - r^2_{x_1 x_2})}$$

# Multiple linear regression

- Assume a multiple linear regression model with two independent variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

- The intercept in a multiple linear regression is given by:

$$\beta_0 = \bar{y} - \beta_1 \bar{x_1} + \beta_2 \bar{x_2}$$

- The estimator $\beta_1$ is now estimated as follows:

$$\beta_1 = \frac{S_y(r_{x_1 y} - r_{x_1 x_2} r_{x_2 y})}{S_{x_1}(1 - r^2_{x_1 x_2})}$$

  $r$ stands for sample correlation, $S$ stands for standard deviation

- The estimator $\beta_2$ is estimated as follows:

$$\beta_2 = \frac{S_y(r_{x_2 y} - r_{x_1 x_2} r_{x_1 y})}{S_{x_2}(1 - r^2_{x_1 x_2})}$$

  $r$ stands for sample correlation, $S$ stands for standard deviation

## Multiple linear regression

- The error term $\varepsilon$ in a multiple regression is given by:

$$\varepsilon = y - (\beta_0 + \beta_1 x_1 + \beta_2 x_2)$$

## Multiple linear regression

- The error term $\varepsilon$ in a multiple regression is given by:

$$\varepsilon = y - (\beta_0 + \beta_1 x_1 + \beta_2 x_2)$$

The notation $\varepsilon$ is usually reserved for population. The residuals obtained from a sample are usually denoted by $e$

## Multiple linear regression

- The error term $\varepsilon$ in a multiple regression is given by:

$$\varepsilon = y - (\beta_0 + \beta_1 x_1 + \beta_2 x_2)$$

  The notation $\varepsilon$ is usually reserved for population. The residuals obtained from a sample are usually denoted by $e$

- OLS will estimate $\beta_1$ and $\beta_2$ in such a way that it will minimise the sum of squared residuals:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y - (\beta_0 + \beta_1 x_1 + \beta_2 x_2))^2$$

  The minimisation process involves taking partial derivatives (which we will do in Econometrics I).

# Outline

# Interpretation of regression estimators

## Interpretation of regression estimators

- Assume a regression model:

$$wage = \beta_0 + \beta_1 educ + e \tag{3}$$

# Interpretation of regression estimators

- Assume a regression model:

$$wage = \beta_0 + \beta_1 educ + e \tag{3}$$

where wage is hourly earnings and educ is the no. of years of education

## Interpretation of regression estimators

- Assume a regression model:
$$wage = \beta_0 + \beta_1 educ + e \tag{3}$$
where wage is hourly earnings and educ is the no. of years of education
- **Interpretation**: How do we interpret the above model? In other words, what is the effect of a change in education on wages?

## Interpretation of regression estimators

- Assume a regression model:

$$wage = \beta_0 + \beta_1 educ + e \tag{3}$$

  where wage is hourly earnings and educ is the no. of years of education

- **Interpretation**: How do we interpret the above model? In other words, what is the effect of a change in education on wages? To interpret the effect of a change in education on wages, we can write:

$$\Delta(wage) = \beta_1 \Delta educ + \Delta e \tag{4}$$

## Interpretation of regression estimators

- Assume a regression model:
$$wage = \beta_0 + \beta_1 educ + e \tag{3}$$
where wage is hourly earnings and educ is the no. of years of education
- **Interpretation**: How do we interpret the above model? In other words, what is the effect of a change in education on wages? To interpret the effect of a change in education on wages, we can write:
$$\Delta(wage) = \beta_1 \Delta educ + \Delta e \tag{4}$$
Assuming $\Delta e = 0$, we get:

## Interpretation of regression estimators

- Assume a regression model:

$$wage = \beta_0 + \beta_1 educ + e \tag{3}$$

  where wage is hourly earnings and educ is the no. of years of education

- **Interpretation**: How do we interpret the above model? In other words, what is the effect of a change in education on wages? To interpret the effect of a change in education on wages, we can write:

$$\Delta(wage) = \beta_1 \Delta educ + \Delta e \tag{4}$$

  Assuming $\Delta e = 0$, we get:

$$\Delta(wage) = \beta_1 \Delta educ \tag{5}$$

## Interpretation of regression estimators

- Assume a regression model:

$$wage = \beta_0 + \beta_1 educ + e \tag{3}$$

where wage is hourly earnings and educ is the no. of years of education

- **Interpretation**: How do we interpret the above model? In other words, what is the effect of a change in education on wages? To interpret the effect of a change in education on wages, we can write:

$$\Delta(wage) = \beta_1 \Delta educ + \Delta e \tag{4}$$

Assuming $\Delta e = 0$, we get:

$$\Delta(wage) = \beta_1 \Delta educ \tag{5}$$

- Now to interpret the above, first we need to know the units of education and wages.

# Interpretation of regression estimators

- Assume a regression model:
$$wage = \beta_0 + \beta_1 educ + e \tag{3}$$
where wage is hourly earnings and educ is the no. of years of education
- **Interpretation**: How do we interpret the above model? In other words, what is the effect of a change in education on wages? To interpret the effect of a change in education on wages, we can write:
$$\Delta(wage) = \beta_1 \Delta educ + \Delta e \tag{4}$$
Assuming $\Delta e = 0$, we get:
$$\Delta(wage) = \beta_1 \Delta educ \tag{5}$$
- Now to interpret the above, first we need to know the units of education and wages. For example, if this model investigates the effect of years of education on hourly wage. Then, we will interpret the model as:

## Interpretation of regression estimators

- Assume a regression model:

$$wage = \beta_0 + \beta_1 educ + e \tag{3}$$

  where wage is hourly earnings and educ is the no. of years of education

- **Interpretation**: How do we interpret the above model? In other words, what is the effect of a change in education on wages? To interpret the effect of a change in education on wages, we can write:

$$\Delta(wage) = \beta_1 \Delta educ + \Delta e \tag{4}$$

  Assuming $\Delta e = 0$, we get:

$$\Delta(wage) = \beta_1 \Delta educ \tag{5}$$

- Now to interpret the above, first we need to know the units of education and wages. For example, if this model investigates the effect of years of education on hourly wage. Then, we will interpret the model as: *an increase in one year of education increases hourly wage by* $\beta_1$.
- Note: if we take the natural log of the variables, the interpretations will change which we will cover in Econometrics I

# Example: A simple regression model

```
library(foreign)
data1<-read.dta("http://fmwww.bc.edu/ec-p/data/wooldridge/wage1.dta")
```

# Example: A simple regression model

```
library(foreign)
data1<-read.dta("http://fmwww.bc.edu/ec-p/data/wooldridge/wage1.dta")
```

Extract the relevant variables from the above dataset

```
wage = data1$wage
educ=data1$educ
slr=lm(wage~educ)
slr

##
## Call:
## lm(formula = wage ~ educ)
##
## Coefficients:
## (Intercept)         educ
##      -0.905        0.541
```

In the above example, one year of additional education is associated with a 0.5 dollar increase in hourly wage.

# Multuple linear regression (cont)

## Interpretation

- Assume a multiple regression model with two independent variables ($x_1$ and $x_2$):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e \tag{6}$$

# Multuple linear regression (cont)

## Interpretation

- Assume a multiple regression model with two independent variables ($x_1$ and $x_2$):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e \tag{6}$$

- $\beta_0$ is the predicted value of y, when $x_1$ and $x_2$ are assumed zero.

# Multuple linear regression (cont)

## Interpretation

- Assume a multiple regression model with two independent variables ($x_1$ and $x_2$):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e \tag{6}$$

- $\beta_0$ is the predicted value of y, when $x_1$ and $x_2$ are assumed zero.
- We can interpret $\beta_1$ as follows:

$$\Delta y = \beta_1 \Delta x_1 \tag{7}$$

# Multuple linear regression (cont)

## Interpretation

- Assume a multiple regression model with two independent variables ($x_1$ and $x_2$):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e \tag{6}$$

- $\beta_0$ is the predicted value of y, when $x_1$ and $x_2$ are assumed zero.
- We can interpret $\beta_1$ as follows:

$$\Delta y = \beta_1 \Delta x_1 \tag{7}$$

  That is, $\beta_1$ captures the change in y explained by a change in $x_1$, when $x_2$ is held fixed.

# Multuple linear regression (cont)

## Interpretation

- Assume a multiple regression model with two independent variables ($x_1$ and $x_2$):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e \tag{6}$$

- $\beta_0$ is the predicted value of y, when $x_1$ and $x_2$ are assumed zero.
- We can interpret $\beta_1$ as follows:

$$\Delta y = \beta_1 \Delta x_1 \tag{7}$$

  That is, $\beta_1$ captures the change in y explained by a change in $x_1$, when $x_2$ is held fixed.
- We can interpret $\beta_2$ as follows:

$$\Delta y = \beta_2 \Delta x_2 \tag{8}$$

# Multuple linear regression (cont)

## Interpretation

- Assume a multiple regression model with two independent variables ($x_1$ and $x_2$):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e \tag{6}$$

- $\beta_0$ is the predicted value of y, when $x_1$ and $x_2$ are assumed zero.
- We can interpret $\beta_1$ as follows:

$$\Delta y = \beta_1 \Delta x_1 \tag{7}$$

  That is, $\beta_1$ captures the change in y explained by a change in $x_1$, when $x_2$ is held fixed.
- We can interpret $\beta_2$ as follows:

$$\Delta y = \beta_2 \Delta x_2 \tag{8}$$

  That is, $\beta_2$ captures the change in y explained by a change in $x_2$, when $x_1$ is held fixed.

# Multuple linear regression (cont)

## Interpretation

- Assume a multiple regression model with two independent variables ($x_1$ and $x_2$):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e \tag{6}$$

- $\beta_0$ is the predicted value of y, when $x_1$ and $x_2$ are assumed zero.
- We can interpret $\beta_1$ as follows:

$$\Delta y = \beta_1 \Delta x_1 \tag{7}$$

  That is, $\beta_1$ captures the change in y explained by a change in $x_1$, when $x_2$ is held fixed.
- We can interpret $\beta_2$ as follows:

$$\Delta y = \beta_2 \Delta x_2 \tag{8}$$

  That is, $\beta_2$ captures the change in y explained by a change in $x_2$, when $x_1$ is held fixed.
- $\beta_1$ and $\beta_2$ are the **partial effects** of $x_1$ and $x_2$, respectively.

# Example: Multiple linear regression

# Example: Multiple linear regression

```
library(foreign)
data1<-read.dta("http://fmwww.bc.edu/ec-p/data/wooldridge/wage1.dta")
```

# Example: Multiple linear regression

```
library(foreign)
data1<-read.dta("http://fmwww.bc.edu/ec-p/data/wooldridge/wage1.dta")
```

Extract the relevant variables from the above dataset

```
wage = data1$wage
educ=data1$educ
exper=data1$exper
mlr=lm(wage~educ+exper)
mlr


##
## Call:
## lm(formula = wage ~ educ + exper)
##
## Coefficients:
## (Intercept)         educ        exper
##     -3.3905       0.6443       0.0701
```

# Outline

# Hypothesis testing

## Formulating hypothesis:

- Hypothesis testing in multiple linear regression is the same as in simple linear regression (See the previous lecture)

# Hypothesis testing

## Formulating hypothesis:

- Hypothesis testing in multiple linear regression is the same as in simple linear regression (See the previous lecture)
- We usually test the null hypothesis, $H_0 : \beta_j = 0$, against the alternative hypothesis $H_1 : \beta_j \neq 0$

# Hypothesis testing

## Formulating hypothesis:

- Hypothesis testing in multiple linear regression is the same as in simple linear regression (See the previous lecture)
- We usually test the null hypothesis, $H_0 : \beta_j = 0$, against the alternative hypothesis $H_1 : \beta_j \neq 0$
- We calculate our t statistics as follows

$$t = (\hat{\beta}_j - \beta_j)/se(\hat{\beta}_j),$$

# Hypothesis testing

## Formulating hypothesis:

- Hypothesis testing in multiple linear regression is the same as in simple linear regression (See the previous lecture)
- We usually test the null hypothesis, $H_0 : \beta_j = 0$, against the alternative hypothesis $H_1 : \beta_j \neq 0$
- We calculate our t statistics as follows

$$t = (\hat{\beta}_j - \beta_j)/se(\hat{\beta}_j),$$

- We can obtain our p values like before

# Hypothesis testing

## Formulating hypothesis:

- Hypothesis testing in multiple linear regression is the same as in simple linear regression (See the previous lecture)
- We usually test the null hypothesis, $H_0 : \beta_j = 0$, against the alternative hypothesis $H_1 : \beta_j \neq 0$
- We calculate our t statistics as follows

$$t = (\hat{\beta}_j - \beta_j)/se(\hat{\beta}_j),$$

- We can obtain our p values like before
- R usually performs all these calculations for you. You can see the details of your model by typing: `summary(mlr)`. Note mlr is the name of my model.

# Confidence interval

## Formulating hypothesis:

- We calculate our confidence intervals as follows

$$\hat{\beta}_j \pm t_{n-k-1,\frac{\alpha}{2}} se(\hat{\beta}_j)$$

# Confidence interval

## Formulating hypothesis:

- We calculate our confidence intervals as follows

$$\hat{\beta}_j \pm t_{n-k-1,\frac{\alpha}{2}} se(\hat{\beta}_j)$$

- You can calculate this in R by typing;

# Confidence interval

## Formulating hypothesis:

- We calculate our confidence intervals as follows

$$\hat{\beta}_j \pm t_{n-k-1,\frac{\alpha}{2}} se(\hat{\beta}_j)$$

- You can calculate this in R by typing;
`confint(model,level=0.95)`

# Confidence interval

## Formulating hypothesis:

- We calculate our confidence intervals as follows

$$\hat{\beta}_j \pm t_{n-k-1,\frac{\alpha}{2}} se(\hat{\beta}_j)$$

- You can calculate this in R by typing;
  `confint(model,level=0.95)`
  hvor model er modellen.

# Outline

## Exercise:

- Exam assignment no. 5 is uploaded on moodle

## Exercise:

- Exam assignment no. 5 is uploaded on moodle
- This assignment is related to multiple linear regression

## Exercise:

- Exam assignment no. 5 is uploaded on moodle
- This assignment is related to multiple linear regression
- You should work with your group mates on this exercise (unless you strongly prefer to work alone)

## Exercise:

- Exam assignment no. 5 is uploaded on moodle
- This assignment is related to multiple linear regression
- You should work with your group mates on this exercise (unless you strongly prefer to work alone)
- First, carefully look at the model equations. Before performing any statistical analysis, discuss with your group whether the relationships in these equations should be positive or negative, and why?

## Exercise:

- Exam assignment no. 5 is uploaded on moodle
- This assignment is related to multiple linear regression
- You should work with your group mates on this exercise (unless you strongly prefer to work alone)
- First, carefully look at the model equations. Before performing any statistical analysis, discuss with your group whether the relationships in these equations should be positive or negative, and why?
- Then, import the data for this assignment and run the regression models

## Exercise:

- Exam assignment no. 5 is uploaded on moodle
- This assignment is related to multiple linear regression
- You should work with your group mates on this exercise (unless you strongly prefer to work alone)
- First, carefully look at the model equations. Before performing any statistical analysis, discuss with your group whether the relationships in these equations should be positive or negative, and why?
- Then, import the data for this assignment and run the regression models
- Carefully, think about the estimates and discuss what are these estimates showing? How will you interpret these results

## Exercise:

- Exam assignment no. 5 is uploaded on moodle
- This assignment is related to multiple linear regression
- You should work with your group mates on this exercise (unless you strongly prefer to work alone)
- First, carefully look at the model equations. Before performing any statistical analysis, discuss with your group whether the relationships in these equations should be positive or negative, and why?
- Then, import the data for this assignment and run the regression models
- Carefully, think about the estimates and discuss what are these estimates showing? How will you interpret these results
- And finally, you have to test whether these estimates are statistically significant or not?

## Exercise:

- Exam assignment no. 5 is uploaded on moodle
- This assignment is related to multiple linear regression
- You should work with your group mates on this exercise (unless you strongly prefer to work alone)
- First, carefully look at the model equations. Before performing any statistical analysis, discuss with your group whether the relationships in these equations should be positive or negative, and why?
- Then, import the data for this assignment and run the regression models
- Carefully, think about the estimates and discuss what are these estimates showing? How will you interpret these results
- And finally, you have to test whether these estimates are statistically significant or not?
- Calculate the confidence interval of your estimators