Erasmus School of Economics

# MOOC Econometrics

## Lecture 2.2 on Multiple Regression: Representation

Christiaan Heij

Erasmus University Rotterdam

## Example

- $\log(\text{Wage})_i =$

  $\beta_1 + \beta_2 \text{Female}_i + \beta_3 \text{Age}_i + \beta_4 \text{Educ}_i + \beta_5 \text{Parttime}_i + \varepsilon_i$

| | |
|---|---|
| 'Wage': | yearly wage (index, median = 100) |
| 'Female': | gender dummy (1 for females, 0 for males) |
| 'Age': | age (in years) |
| 'Educ': | education (4 levels, from 1 for low to 4 for high) |
| 'Parttime': | part-time job dummy (1 if work on 3 or less days per week, 0 if more than 3 days per week) |



## Notation

- $y_i \;= \log(\text{Wage})_i$

  $x_{1i} = 1 \qquad x_{2i} = \text{Female}_i \qquad x_{3i} = \text{Age}_i$

  $x_{4i} = \text{Educ}_i \qquad x_{5i} = \text{Parttime}_i$

- Let $x_i$ be $(5 \times 1)$ vector with components $(x_{1i}, \ldots, x_{5i})$.

  Let $\beta$ be $(5 \times 1)$ vector with components $(\beta_1, \ldots, \beta_5)$.

- Then wage equation can be written as

  $y_i = \sum_{j=1}^{5} \beta_j x_{ji} + \varepsilon_i = x_i' \beta + \varepsilon_i$

- Symbol $'$ (prime): transposition (see Building Blocks).



## Matrix notation

- Write $y_i = x_i'\beta + \varepsilon_i$ for 500 observations in database:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{500} \end{pmatrix} = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_{500}' \end{pmatrix} \beta + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{500} \end{pmatrix}$$

- Let $y$: $(500 \times 1)$ vector with components $y_i$

  $X$: $(500 \times 5)$ matrix with rows $x_i'$

  $\varepsilon$: $(500 \times 1)$ vector with components $\varepsilon_i$

- Then wage equation for 500 observations becomes:

  $y = X\beta + \varepsilon$

## Multiple regression model

- Model with $k$ explanatory factors:

  $y_i = \beta_1 + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i = \sum_{j=1}^{k} x_{ji}\beta_j + \varepsilon_i$

  (with $x_{1i} = 1$).
- $y_i$ is dependent or explained variable,

  $x_{1i}, \ldots, x_{ki}$ are regressor variables or explanatory factors.
- First 'explanatory' factor is the constant $x_{1i} = 1$.



## Multiple regression model

- Let database contain $n$ observations for all variables.
- As before, let

  $y$: $(n \times 1)$ vector with components $y_i$

  $X$: $(n \times k)$ matrix with elements $x_{ji}$

  $\beta$: $(k \times 1)$ vector with components $\beta_j$

  $\varepsilon$: $(n \times 1)$ vector with components $\varepsilon_i$
- Then model can be written as

  $y = X\beta + \varepsilon$



## Set of linear equations

- $y = X\beta + \varepsilon$

  $\rightarrow$ $X\beta$ is 'explained' part of $y$

  $\rightarrow$ $\varepsilon$ is 'unexplained' part of $y$
- $X$ explains much of $y$ if $y \approx X\beta$ for some choice of $\beta$.
- $y = X\beta$ is set of $n$ equations in $k$ unknown parameters $\beta$.

**Test**

Let $X$ be $(n \times k)$ matrix with rank$(X) = r$.
What is the number of solutions of the equations $y = X\beta$?



## Test answers

- $y = X\beta$ where $X$ is $(n \times k)$ with rank$(X) = r$.

  $\rightarrow$ Always $r \leq k$ and $r \leq n$.

  $\rightarrow$ If $r = n = k$: $y = X\beta$ has unique solution.

  $\rightarrow$ If $r = n < k$: $y = X\beta$ has multiple solutions.

  $\rightarrow$ If $r < n$: $y = X\beta$ has (in general) no solution.
- (Nearly always) $n > k$.

  We assume $r = k < n$.

  So $y = X\beta$ has (in general) no exact solution.

## Interpretation of model coefficients

- Model: $y_i = \beta_1 + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i$.
- What happens to $y$ if $x_j$ increases by one unit while all other $x$-variables $x_h$ (with $h \neq j$) remain fixed?
- Partial effect: $\frac{\partial y}{\partial x_j} = \beta_j$ (if $x_h$ remains fixed for all $h \neq j$).
- Only possible as thought-experiment, called the 'ceteris paribus' assumption.



## Decomposition of total effect

- Total effect if factors are mutually dependent (and $x_{1i} = 1$):

$$\frac{dy}{dx_j} = \frac{\partial y}{\partial x_j} + \sum_{h=2,h\neq j}^{k} \frac{\partial y}{\partial x_h}\frac{\partial x_h}{\partial x_j} = \beta_j + \sum_{h=2,h\neq j}^{k} \beta_h \frac{\partial x_h}{\partial x_j}$$

- Indirect effects $x_j \to x_h \to y$ combined: $\sum_{h=2,h\neq j}^{k} \beta_h \frac{\partial x_h}{\partial x_j}$
- So: <u>Total effect = Partial effect + Indirect effect</u>
- Example if part-time jobs more common for higher education:

| | |
|---|---|
| Direct: | Educ $\uparrow$ $\Rightarrow$ Wage $\uparrow$ |
| Indirect: | Educ $\uparrow$ $\Rightarrow$ Parttime $\uparrow$ $\Rightarrow$ Wage $\downarrow$ |
| Total: | Sum of $\uparrow$ and $\downarrow$ effect, need effect sizes |



## Testing for model restrictions

- Factor $x_j$ in model if (relevant) effect on $y$.
- Test for single factor $j$: Test $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$.
- Test for two factors $j$ and $h$: Test $H_0 : \beta_j = \beta_h = 0$ against $H_1 : \beta_j \neq 0$ and/or $\beta_h \neq 0$.
- General: Test $H_0 : R\beta = r$ against $H_1 : R\beta \neq r$
  $\to$ $R$ is given $(g \times k)$ matrix with $\text{rank}(R) = g$
  $\to$ $r$ is given $(g \times 1)$ vector

**Test**

If $\beta_j = 0$, does this mean that $x_j$ has no effect on $y$?
Motivate your answer.



## Test answers ($\beta_j = 0 \Rightarrow x_j$ no effect on $y$?)

- <u>Yes</u>, in sense that $x_j$ has no <u>partial</u> effect

  (assuming all other explanatory factors remain fixed).
- <u>No</u>, in sense that $x_j$ may have <u>indirect</u> effect

  (via other factors $x_j \to x_h \to y$).
- Example: $\log(\text{Wage})_i = \beta_1 + \beta_2 \text{Educ}_i + \beta_3 \text{Parttime}_i + \varepsilon_i$

  If $\beta_2 = 0$ and $\beta_3 \neq 0$, then higher education still has indirect effect on wage if having part-time job is related to education level.

# TRAINING EXERCISE 2.2

- Train yourself by making the training exercise (see the website).
- After making this exercise, check your answers by studying the webcast solution (also available on the website).