Erasmus
School of
Economics

# MOOC Econometrics

## Lecture 2.1 on Multiple Regression: Motivation

Christiaan Heij

Erasmus University Rotterdam

## Introduction

- Compare wage of males and females.
- They may differ, for example, in education level.
- Research Question 1: What is total gender difference in wage, including differences caused by education?
- Research Question 2: What is partial gender difference in wage, excluding differences caused by education?



## Gender difference in wage

**Test**

- For which question should education be included in the analysis?
- For which question should it be excluded?

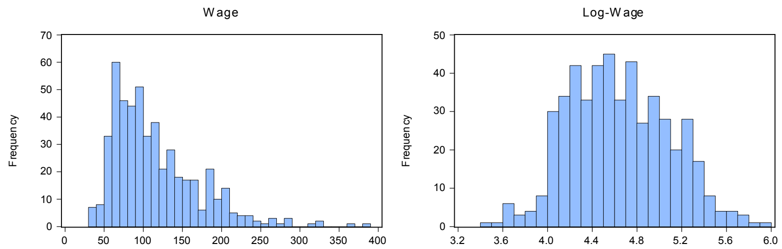
- Total gender effect including education effects:

  $\rightarrow$ Education should be excluded from model!
- Partial gender effect excluding education effects:

  $\rightarrow$ Education should be included in model!
- Coming lectures will explain the why and how.



## Wage data set

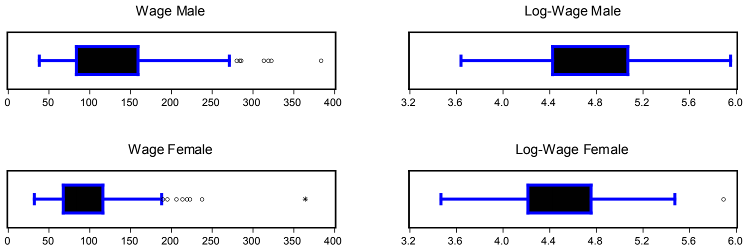



- Data set for 500 employees on wages (indexed, median $= 100$).

  $\rightarrow$ Random sample from much larger population of employees.
- Wage is much more skewed than log-wage

  ('log' denotes natural logarithm).

## Boxplots of wage and log-wage




- Females have lower wage than males.
- Research questions:
  - → How large is this difference?
  - → What are the causes of this difference?



## Simple regression

- $\log(\text{Wage}) = 4.73 - 0.25\text{Female} + e$

  $(R^2 = 0.07,\ b = -0.25,\ t_b = -6.25)$
- 'Female': gender dummy, 1 for females, 0 for males.

**Test**

What is the estimated gender difference in wage level?

- Answer: $\log(\text{Wage}_{Female}) - \log(\text{Wage}_{Male}) = -0.25$

  $\text{Wage}_{Female} = \text{Wage}_{Male} \times e^{-0.25} = \text{Wage}_{Male} \times 0.78$

  → Females earn on average 22% less than males.



## Multiple explanatory factors

- Wage depends on factors as age, education level, and part-time jobs.
- Simple regressions give:

$$\text{Age} = 40.05 - 0.11\text{Female} + e \qquad (R^2 = 0.00, t_b = -0.11)$$
$$\text{Educ} = 2.26 - 0.49\text{Female} + e \qquad (R^2 = 0.05, t_b = -5.16)$$
$$\text{Parttime} = 0.20 + 0.25\text{Female} + e \qquad (R^2 = 0.07, t_b = 6.15)$$

- Females: same age, lower education, more often part-time job.



## Gender differences in education

| | | Education level | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | Total |
| Count | Male | 108 | 77 | 72 | 59 | 316 |
| | Female | 88 | 57 | 33 | 6 | 184 |
| Percentage | Male | 34 | 24 | 23 | 19 | 100 |
| | Female | 48 | 31 | 18 | 3 | 100 |

## Gender differences in part-time jobs

| | | Part-time job | | |
|---|---|---|---|---|
| | | Yes | No | Total |
| Count | Male | 62 | 254 | 316 |
| | Female | 82 | 102 | 184 |
| Percentage | Male | $\boxed{20}$ | 80 | 100 |
| | Female | $\boxed{45}$ | 55 | 100 |



## Partial effects

- Partial effect: if all other variables remain 'fixed'.
- Research question: What is partial gender effect on wage?
- So: Gender difference in wage after correction for differences in education and part-time jobs.
- Answer obtained by multiple regression.
  → Methods: Lectures 2.2-2.4
  → Outcomes: Lecture 2.5



## TRAINING EXERCISE 2.1

- 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).

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: Total effect = Partial effect + Indirect effect
- 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 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 $y$?)

- Yes, in sense that $x_j$ has no partial effect

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

Erasmus School of Economics

# MOOC Econometrics

## Lecture 2.3 on Multiple Regression: Estimation

Christiaan Heij

Erasmus University Rotterdam

## OLS criterion

- Model: $y = X\beta + \varepsilon$
- Dimensions: $y$ $(n \times 1)$, $X$ $(n \times k)$: observed data
  $\beta$ $(k \times 1)$, $\varepsilon$ $(n \times 1)$: unobserved
- Objective:
  $\rightarrow$ Estimate $\beta$ by $(k \times 1)$ vector $b$ so that $Xb$ is 'close' to $y$.



## OLS criterion

- We assume that $(n \times k)$ matrix $X$ has $\text{rank}(X) = k$.

**Test**

Prove that $\#(\text{parameters}) = k \leq n = \#(\text{observations})$.

- Answer: $X$ is $(n \times k)$ matrix, hence $k = \text{rank}(X) \leq n$.
- Wish: small vector of residuals $y - Xb = e = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$.
- Least squares criterion ('ordinary least squares', OLS):
  $\rightarrow$ minimize $S(b) = e'e = \sum_{i=1}^{n} e_i^2$.



## OLS estimation

- $$\begin{aligned} S(b) = e'e &= (y - Xb)'(y - Xb) \\ &= y'y - y'Xb - b'X'y + b'X'Xb \\ &= y'y - 2b'X'y + b'X'Xb \end{aligned}$$

**Test**

We used $y'Xb = b'X'y$. Prove this result.

- Answer: $y'Xb$ is $(1 \times 1)$, so $y'Xb = (y'Xb)' = b'X'y$.
- Facts of matrix derivatives (see Building Blocks):
  $\frac{\partial b'a}{\partial b} = a$
  $\frac{\partial b'Ab}{\partial b} = (A + A')b$

## OLS estimation

- First order conditions for $S(b) = y'y - 2b'X'y + b'X'Xb$:

  $$\frac{\partial S}{\partial b} = -2X'y + (X'X + X'X)b = -2X'y + 2X'Xb = 0.$$

- So: $X'Xb = X'y$.

**Test**

Prove that $\text{rank}(X) = k$ implies that $X'X$ is invertible.

- Answer: $X'X$ is $(k \times k)$ matrix, and

  $X'Xa = 0 \Rightarrow a'X'Xa = (Xa)'Xa = 0 \Rightarrow Xa = 0 \Rightarrow a = 0.$
  Last step follows from $\text{rank}(X) = k$.

- So: $\boxed{b = (X'X)^{-1}X'y}$



## Geometric aspects

- $y$ is $(n \times 1)$, $X$ is $(n \times k)$

- Define $H = X(X'X)^{-1}X'$

  $$M = I - H = I - X(X'X)^{-1}X'$$

**Test**

Show that $M' = M$, $M^2 = M$, $MX = 0$, $MH = 0$.

- Answer: Direct calculations.
  Use $(X'X)^{-1}$ symmetric and $(X'X)^{-1}X'X = I$.

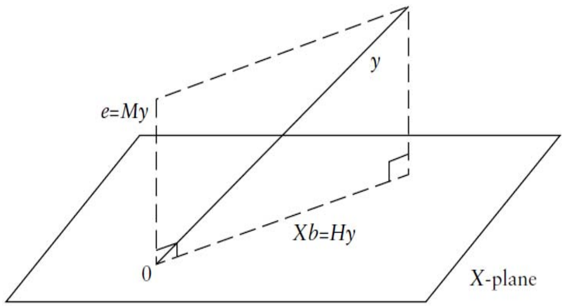
- Fitted values: $\hat{y} = Xb = X(X'X)^{-1}X'y = Hy$.

  Residuals: $e = y - Xb = y - Hy = My$.

- $e$ and $\hat{y}$ orthogonal: $e'\hat{y} = (My)'Hy = y'M'Hy = 0$.



## Relation between $y$, $X$, $b$, and $e$




The '$X$-plane' is $k$-dimensional subspace spanned by columns of $X$, that is, set of vectors $Xa$ with $a$ arbitrary $(k \times 1)$ vector.



## Estimation of error variance $\sigma^2$

- $\sigma^2 = E(\varepsilon_i^2)$

  Estimate unknown $\varepsilon = y - X\beta$ by residuals $e = y - Xb$.

- Sample variance of residuals: $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(e_i - \bar{e})^2$.

**Test**

Check that the $(n \times 1)$ vector of residuals $e$ satisfies k linear restrictions, so that $e$ has $(n-k)$ 'degrees of freedom'.

- Answer: $\text{rank}(X) = k$, and $X'e = X'(y - Xb) = X'y - X'Xb = 0$.

- OLS estimator: $s^2 = \frac{1}{n-k}e'e = \frac{1}{n-k}\sum_{i=1}^{n} e_i^2$

- Unbiased under standard assumptions (see next lecture).

## R-squared ($R^2$)

- Definition: $R^2 = \left(\text{cor}(y, \hat{y})\right)^2 = \frac{\left(\sum(y_i - \overline{y})(\hat{y}_i - \overline{\hat{y}})\right)^2}{\sum(y_i - \overline{y})^2 \sum(\hat{y}_i - \overline{\hat{y}})^2}$,

  where 'cor' is correlation coefficient and $\hat{y} = Xb$.

- Higher $R^2$ means better fit of $Xb$ to observed $y$.

- If model contains constant term ($x_{1i} = 1$ for all $i = 1, \ldots n$):
  $R^2 = 1 - \frac{e'e}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$.



## TRAINING EXERCISE 2.3

- 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).

Erasmus School of Economics

# MOOC Econometrics

## Lecture 2.4.1 on Multiple Regression: Evaluation - Statistical Properties

Christiaan Heij

Erasmus University Rotterdam

## Six DGP assumptions

- A1 Linear model: $y = X\beta + \varepsilon$.
- A2 Fixed regressors: $X$ non-random.
- A3 Random error terms with mean zero: $E(\varepsilon) = 0$.
- A4 Homoskedastic error terms: $E(\varepsilon_i^2) = \sigma^2$ for all $i = 1, \ldots, n$.
- A5 Uncorrelated error terms: $E(\varepsilon_i \varepsilon_j) = 0$ for all $i \neq j$.
- A6 Parameters $\beta$ and $\sigma^2$ are fixed and unknown.

**Test**

Prove that A4 and A5 imply that $E(\varepsilon\varepsilon') = \sigma^2 I$.

- Answer: Direct calculation of variance-covariance matrix.



## A1,2,3,6: $b$ unbiased, $E(b) = \beta$

Under A1, A2, A3, and A6, OLS is unbiased: $E(b) = \beta$.

**Test**

Express OLS estimator $b$ in terms of $\varepsilon$.

- Answer: $b = (X'X)^{-1}X'y \overset{(A1)}{=} (X'X)^{-1}X'(X\beta + \varepsilon)$
  $= \beta + (X'X)^{-1}X'\varepsilon$.
- $E(b) \overset{(A6)}{=} \beta + E((X'X)^{-1}X'\varepsilon) \overset{(A2)}{=} \beta + (X'X)^{-1}X'E(\varepsilon)$
  $\overset{(A3)}{=} \beta + (X'X)^{-1}X'0 = \beta$.



## A1-A6: $\text{var}(b) = \sigma^2(X'X)^{-1}$

- Seen before: $b = \beta + (X'X)^{-1}X'\varepsilon$.
- $$\begin{aligned}\text{var}(b) = \; & E\Big((b - Eb)(b - Eb)'\Big) \overset{(A1,2,3,6)}{=} \\ & E\Big((b-\beta)(b-\beta)'\Big) = \\ & E\Big((X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}\Big) \overset{(A2)}{=} \\ & (X'X)^{-1}X' \; E(\varepsilon\varepsilon') \; X(X'X)^{-1} \overset{(A4,5)}{=} \\ & (X'X)^{-1}X' \; \sigma^2 I \; X(X'X)^{-1} = \\ & \sigma^2(X'X)^{-1}X'X(X'X)^{-1} = \sigma^2(X'X)^{-1}.\end{aligned}$$
- Let $a_{jh}$ be $(j,h)$-th element of $(k \times k)$ matrix $(X'X)^{-1}$, then $\text{var}(b_j) = \sigma^2 a_{jj}$ and $\text{cov}(b_j, b_h) = \sigma^2 a_{jh}$.

## OLS estimator of $\sigma^2$

Under A1-A6, $s^2 = e'e/(n-k)$ is unbiased: $E(s^2) = \sigma^2$.

- Idea of proof: (a) Express $e$ in $\varepsilon$.
  (b) Compute $E(ee')$.
  (c) Use 'trace trick' to get $E(e'e)$.
- (a) Previous lecture: $e = My$ where $M = I - X(X'X)^{-1}X'$ with $M' = M = M^2$ and $MX = 0$.

  Then $e = My \overset{(A1)}{=} M(X\beta + \varepsilon) = MX\beta + M\varepsilon = M\varepsilon$.
- (b) $E(ee') = E(M\varepsilon\varepsilon'M') \overset{(A2)}{=} ME(\varepsilon\varepsilon')M \overset{(A4,5)}{=} M\sigma^2 IM = \sigma^2 M$.
- (c) 'Trace trick': $E(e'e) = \text{trace}(E(ee')) = \sigma^2\text{trace}(M) = (n-k)\sigma^2$.



## Details of 'trace trick' (optional)

- $\text{trace}(AB) = \text{trace}(BA)$, where 'trace' is sum of diagonal elements of square matrix (see Building Blocks).
- Trace trick:

$$
\begin{aligned}
E(e'e) = E(\textstyle\sum_{i=1}^n e_i^2) &= E(\text{trace}(ee')) = \text{trace}(E(ee')) \\
&= \text{trace}(\sigma^2 M) = \sigma^2\text{trace}(I_n - X(X'X)^{-1}X') \\
&= \sigma^2\text{trace}(I_n) - \sigma^2\text{trace}(X(X'X)^{-1}X') \\
&= n\sigma^2 - \sigma^2\text{trace}((X'X)^{-1}X'X) \\
&= n\sigma^2 - \sigma^2\text{trace}(I_k) = (n-k)\sigma^2.
\end{aligned}
$$

- As $E(e'e) = (n-k)\sigma^2$, it follows that $E(s^2) = \sigma^2$.



## Efficiency of OLS

- A1-A6: OLS $b$ is Best Linear Unbiased Estimator (BLUE).
- This is the so-called Gauss-Markov theorem.
- If $\hat{\beta} = Ay$ is linear estimator, $A$ non-random $(k \times n)$ matrix,

  and if $\hat{\beta}$ is unbiased, $E(\hat{\beta}) = \beta$,

  then $\text{var}(\hat{\beta}) - \text{var}(b)$ is positive semi-definite (PSD).

  (see Building Blocks for PSD)
- As $b$ has smallest variance of all linear unbiased estimators, OLS is efficient (in this class).



## TRAINING EXERCISE 2.4.1

- 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).

Erasmus School of Economics

# MOOC Econometrics

## Lecture 2.4.2 on Multiple Regression: Evaluation - Statistical Tests

Christiaan Heij

Erasmus University Rotterdam

## Test for a single restriction: $t$-test

- Under assumptions A1-A6:

  $E(b) = \beta$ and $\mathrm{var}(b) = \sigma^2 (X'X)^{-1}$.

- A7: $\varepsilon$ is normally distributed.

**Test**

Check that A1-A7 imply $b \sim N(\beta, \sigma^2 (X'X)^{-1})$.

- Answer: $b = (X'X)^{-1} X'(X\beta + \varepsilon) = \beta + (X'X)^{-1} X'\varepsilon$

  is linear function of $\varepsilon \sim N(0, \sigma^2 I)$.



## $t$-test

- Test for relevance of single explanatory factor $j$:

  Test $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$.

- A1-A7: $b_j \sim N(\beta_j, \sigma^2 a_{jj})$, $a_{jj}$ is element $(j,j)$ of $(X'X)^{-1}$.

  If $H_0 : \beta_j = 0$ holds, then $z_j = \frac{b_j - \beta_j}{\sigma\sqrt{a_{jj}}} = \frac{b_j}{\sigma\sqrt{a_{jj}}} \sim N(0,1)$.

- Replace unknown $\sigma$ by $s$, square root of $s^2 = e'e/(n-k)$.

  Test statistic: $t_j = \frac{b_j}{s\sqrt{a_{jj}}} = \boxed{\frac{b_j}{\mathrm{SE}(b_j)}}$, with $\mathrm{SE}(b_j) = s\sqrt{a_{jj}}$.

- A1-A7: $t_j \sim t(n-k)$ (close to normal unless $n-k$ small).



## Test for multiple restrictions: $F$-test

- Test for multiple linear restrictions:

  Test $H_0 : R\beta = r$ against $H_1 : R\beta \neq r$.

  $\rightarrow$ $R$ is given $(g \times k)$ matrix with $\mathrm{rank}(R) = g$
  $\rightarrow$ $r$ is given $(g \times 1)$ vector

- A1-A7 imply $b \sim N(\beta, \sigma^2 (X'X)^{-1})$.

**Test**

Under $H_0$: $Rb \sim N(m, \sigma^2 V)$. Compute $m$ and $\sigma^2 V$.

- Answer: $m = E(Rb) = RE(b) = R\beta = r$.

  $$\sigma^2 V = \mathrm{var}(Rb) = R\mathrm{var}(b)R' = \sigma^2 R(X'X)^{-1}R'.$$

## $F$-test

- Then $(1/\sigma)(Rb - r) \sim N(0, V)$.
- Facts: $(1/\sigma^2)(Rb - r)'V^{-1}(Rb - r) \sim \chi^2(g)$.

$$F = (1/s^2)(Rb - r)'V^{-1}(Rb - r)/g \sim F(g, n - k).$$

- $F$ can be computed from residual sums of squares:

$$\boxed{F = \frac{(e_0'e_0 - e_1'e_1)/g}{e_1'e_1/(n-k)}}$$

$\rightarrow$ $e_0'e_0$: sum of squared residuals of restricted model ($H_0$)

$\rightarrow$ $e_1'e_1$: sum of squared residuals of unrestricted model ($H_1$)



## Test for removing a set of explanatory factors

- Restricted model: remove set of $g$ explanatory factors.
- Re-order $k$ factors so that last $g$ are removed:

Re-order $X = (X_1 \; X_2)$, $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$, and $b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$

$X_2$: last $g$ columns of $X$ (factors removed in restricted model)

$\beta_2$: last $g$ elements of $\beta$

$b_2$: last $g$ elements of $b$

- Then $y = X_1\beta_1 + X_2\beta_2 + \varepsilon = X_1b_1 + X_2b_2 + e$.



## $F$-test

- $y = X_1\beta_1 + X_2\beta_2 + \varepsilon$.
- Test $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$.
- If $H_0$ holds, then $F = \frac{(e_0'e_0 - e_1'e_1)/g}{e_1'e_1/(n-k)} \sim F(g, n - k)$

$\rightarrow$ $e_0'e_0$: sum of squared residuals of restricted model

(OLS in model $y = X_1\beta_1 + \varepsilon$)

$\rightarrow$ $e_1'e_1$: sum of squared residuals of unrestricted model

(OLS in model $y = X_1\beta_1 + X_2\beta_2 + \varepsilon$)



## TRAINING EXERCISE 2.4.2

- 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).

Erasmus
School of
Economics

# MOOC Econometrics

## Lecture 2.5 on Multiple Regression: Application

Christiaan Heij

Erasmus University Rotterdam

## Wage equation

- Wage data of Lecture 2.1 with model of Lecture 2.2.
- Model: $\log(\text{Wage})_i = \beta_1 + \beta_2 \text{Female}_i$
$$+\beta_3 \text{Age}_i + \beta_4 \text{Educ}_i + \beta_5 \text{Parttime}_i + \varepsilon_i$$
- OLS gives: $\log(\text{Wage})_i = 3.05 - 0.04\text{Female}_i$
$$+0.03\text{Age}_i + 0.23\text{Educ}_i - 0.37\text{Parttime}_i + e_i$$
- $R^2 = 0.704$ and $s = 0.245$.
- Data are random sample from population of employees.

  OLS results depend on these data, hence also random.



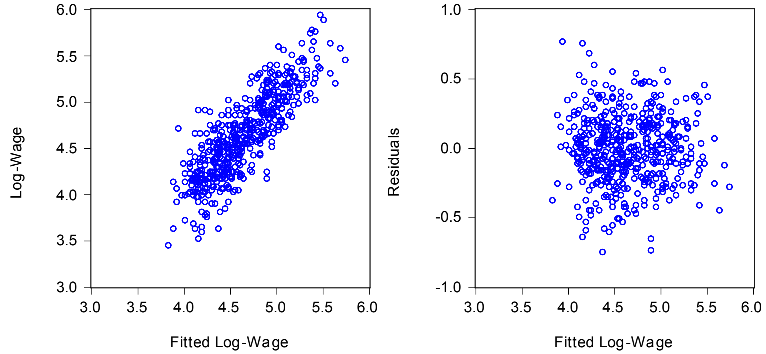
## Regression outcomes

Dependent variable: log(Wage)
Sample size: 500

| | Coefficient $b_j$ | Standard error $SE(b_j)$ | t-Statistic $t_j$ | p-value $H_0 : \beta_j = 0$ |
|---|---|---|---|---|
| Constant | 3.053 | 0.055 | 55.168 | 0.000 |
| Female | -0.041 | 0.025 | -1.663 | 0.097 |
| Age | 0.031 | 0.001 | 24.041 | 0.000 |
| Educ | 0.233 | 0.011 | 21.874 | 0.000 |
| Parttime | -0.365 | 0.032 | -11.576 | 0.000 |
| R-squared | 0.704 | | | |
| SE of regression | 0.245 | | | |



## Two scatter diagrams




- Left diagram: Actual log-wage against fitted log-wage.
- Right diagram: Residuals against fitted log-wage.

## Regression outcomes

- Age, Education, and Parttime are significant (p-values 0.000).

  Female is not significant at 5% level (p-value 0.097).

- Interpretation in terms of average wage effects:

  Extra year of age: $e^{0.031} - 1 = 3\%$

  Extra level of education: $e^{0.233} - 1 = 26\%$

  Part-time job: $e^{-0.365} - 1 = -31\%$

- After controlling for age, education, and part-time job effects, the (partial) gender effect of -4% for females is not significant.

- Lecture 2.1: Significant gender effect of -25% for females: total effect, including education and part-time jobs.



## Model with absolute (instead of relative) effects

- If explained variable is log-wage, parameters

$$\beta_j = \partial \log(\text{Wage})/\partial x_j = (\partial \text{Wage}/\partial x_j)/\text{Wage}$$

  measure relative wage effects of each factor.

- If explained variable is wage (instead of log-wage), parameters $\beta_j = \partial \text{Wage}/\partial x_j$ measure wage level effects.

- OLS in this model gives:

$$\text{Wage}_i = -77.87 - 2.12\text{Female}_i$$
$$+3.62\text{Age}_i + 29.47\text{Educ}_i - 43.10\text{Parttime}_i + e_i.$$

- $R^2 = 0.681$ and $s = 31.276$.



## Wage or log-wage?

- Age, Education, and Parttime are significant (p-values 0.000).

  Female is not significant at 5% level (p-value 0.501).

**Test**

Why can we not choose between the two models (with log-wage and wage) on the basis of $R^2$ and $s$?

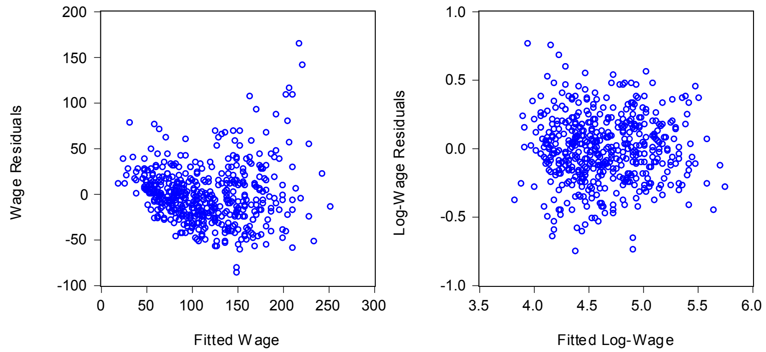
- Answer: $R^2$ and $s$ are based on sum of squares of $y$ and $e = y - Xb$, and $y$ differs in the two models.

- Graphical check of regression assumptions:

  scatter diagram of residuals against fitted values.



## Scatter diagrams of residuals against fitted values




- Left for wage: nonlinear and heteroskedastic

- Right for log-wage: no indication violation regression assumptions

## Testing for constant education effects

- Allow that education effect varies per education level:

$$\log(\text{Wage})_i = \beta_1 + \beta_2\text{Female}_i + \beta_3\text{Age}_i$$
$$+\beta_4\text{DE2}_i + \beta_5\text{DE3}_i + \beta_6\text{DE4}_i + \beta_7\text{Parttime}_i + \varepsilon_i$$

- $\text{DE2}_i = 1$ if employee $i$ has education level 2
  $\text{DE2}_i = 0$ if employee $i$ has education level 1, 3, or 4

  (similar definitions for DE3 and DE4)

- Effect of education is constant if $\beta_5 = 2\beta_4$ and $\beta_6 = 3\beta_4$.

- Test $H_0 : \beta_5 = 2\beta_4$ and $\beta_6 = 3\beta_4$ against $H_1$: $H_0$ not true.



## Regression outcomes

Dependent variable: log(Wage)
Sample size: 500

| | Coefficient $b_j$ | Standard error $SE(b_j)$ | t-Statistic $t_j$ | p-value $H_0 : \beta_j = 0$ |
|---|---|---|---|---|
| Constant | 3.318 | 0.051 | 64.554 | 0.000 |
| Female | -0.031 | 0.024 | -1.267 | 0.206 |
| Age | 0.030 | 0.001 | 24.269 | 0.000 |
| DE2 | 0.171 | 0.027 | 6.308 | 0.000 |
| DE3 | 0.380 | 0.029 | 12.996 | 0.000 |
| DE4 | 0.767 | 0.035 | 21.610 | 0.000 |
| Parttime | -0.366 | 0.031 | -11.813 | 0.000 |
| R-squared | 0.716 | | | |
| SE of regression | 0.241 | | | |



## Outcomes

- OLS in unrestricted model (under $H_1$) gives:

$$\log(\text{Wage})_i = 3.32 - 0.03\text{Female}_i + 0.03\text{Age}_i$$
$$+0.17\text{DE2}_i + 0.38\text{DE3}_i + 0.77\text{DE4}_i - 0.37\text{Parttime}_i + e_i.$$

- $R^2 = 0.716$ and $s = 0.241$.

- All factors are significant, except for 'Female' (p-value 0.206).

- Test for constant education effects:

  Test $H_0 : \beta_5 = 2\beta_4, \beta_6 = 3\beta_4$ against $H_1$: $H_0$ not true.

**Test**

Compute the $F$-test, using $R_1^2 = 0.716$ and $R_0^2 = 0.704$.



## Computation of F-test

- $R_1^2 = 0.716$ and $R_0^2 = 0.704$

- $g = 2$, $n = 500$, $k = 7$ (under $H_1$), $n - k = 500 - 7 = 493$

$$F = \frac{(R_1^2 - R_0^2)/g}{(1 - R_1^2)/(n-k)} = \frac{(0.716 - 0.704)/2}{(1 - 0.716)/493} = 10.4$$

- 5% critical value of $F(2, 493)$ is 3.0.

  As $F = 10.4 > 3.0$, $H_0$ is rejected (at 5% level).

- Conclusion: Wage effect of one extra level of education differs significantly across education levels.

## Wage effect of extra education

- Coefficients of education dummies:

  level 2: 0.171

  level 3: 0.380

  level 4: 0.767

- Wage increase for higher education level:

  $1 \rightarrow 2$: $e^{0.171} - 1 = 0.19 = 19\%$

  $2 \rightarrow 3$: $e^{(0.380-0.171)} - 1 = e^{0.209} - 1 = 23\%$

  $3 \rightarrow 4$: $e^{(0.767-0.380)} - 1 = e^{0.387} - 1 = 47\%$

- Effect much larger for highest education level.



## TRAINING EXERCISE 2.5

- 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).