# Lecture 2
# Multiple Regression and Tests

Dr.ssa Rossella Iraci Capuccinello

2017-18

# Simple Regression Model

- The random variable of interest, y, depends on a single factor, $x_{1i}$, and this is an exogenous variable.
- The true but unknown relationship is defined as being

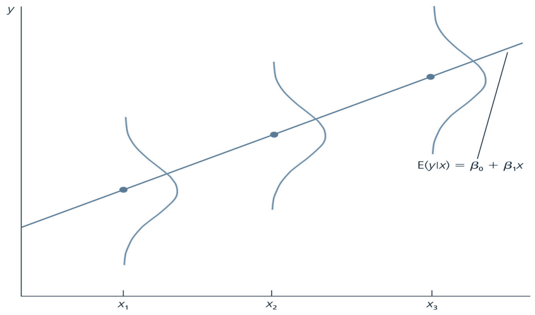
$$y_i = \beta_0 + \beta_1 x_{1i} + u_i$$

- The values of y are expected to lie on a straight line, depending on the corresponding values of x
- Their values will differ from those predicted by that line by the amount of the error term $u_i$

# Simple Regression Model Fig1

FIGURE 2.1

$E(y|x)$ as a linear function of $x$.

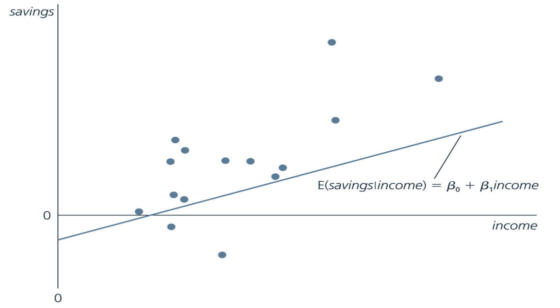



Source: Chap. 2 Woolwridge

# Simple Regression Model Fig2

**FIGURE 2.2**

**Scatterplot of savings and income for 15 families, and the population regression $E(savings|income) = \beta_0 + \beta_1 income$.**

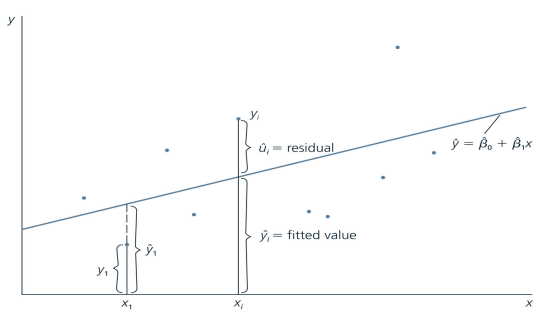



Source: Chap. 2 Woolwridge

# Simple Regression Model Fig3

FIGURE 2.4

Fitted values and residuals.

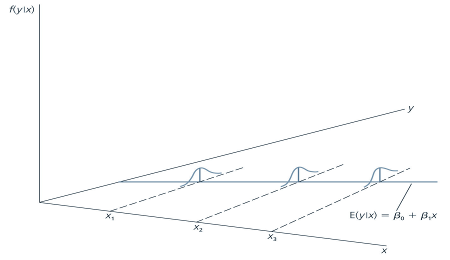



Source: Chap. 2 Woolwridge

# Simple Regression Model Fig4 - Homoskedasticity

FIGURE 2.8

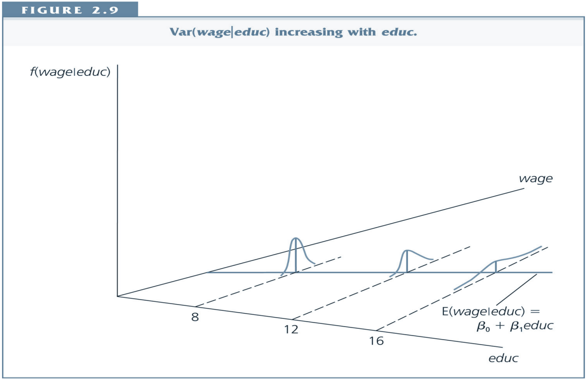
The simple regression model under homoskedasticity.




Source: Chap. 2 Woolwridge

The errors are considered drawn from a fixed distribution, with a mean of zero and a constant variance of $\sigma^2$

# Simple Regression Model Fig5 - Heteroskedasticity




Source: Chap. 2 Woolwridge

# Multiple Regression

- The random variable of interest, y, depends upon a number of different factors, $x_{1i}, x_{2i}, \ldots, x_{ki}$, and these are exogenous variables.
- The true but unknown relationship is defined as being

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \ldots x_{ki} + u_i \qquad i = 1, \ldots, n$$

# CLRM assumptions

- the Classical Linear Regression Model (CLRM) assumptions are:
  1. $x_{ij}$ $j = 1, \ldots k$ **are non−stochastic**
  2. $E(u_i|x_1, x_2, \ldots, x_k) = 0$ (Exogeneity $\rightarrow$ regressors are uncorrelated with the errors)
  3. $Var(u_i|x_1, x_2, \ldots, x_k) = \sigma^2$ (error variance constant (**homoscedasticity**), points distributed around true regression line with a constant spread)
  4. $cov(u_i, u_j|x_1, x_2, \ldots, x_k) = 0$ (errors serially uncorrelated over observations)
  5. $(u_i|x_1, x_2, \ldots, x_k) \sim N(0, \sigma^2)$ $\rightarrow$ Normality

# Running Multiple Regression

Simple regressions are easy:

- Type reg followed by
  1. Dependent variable : y
  2. Independent variables : $x_1, x_2, \ldots, x_k$

# Simplest specification

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$
$$\frac{\partial y}{\partial x_1} = \beta_1$$

- change in y for a unit increase in $x_1$

# Scaled dependent variable

$$y/\alpha = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$
$$\frac{\partial y}{\partial x_1} = \frac{\beta_1}{\alpha}$$
$$\frac{\partial y}{\partial x_2} = \frac{\beta_2}{\alpha}$$

- Interpretation coefficient:
  - each new coefficient and s.e. will be the corresponding old coefficient and s.e. multiplied by the scalar $\alpha$
  - t statistics are identical.

# Standardized regressors

$$zy = \delta_0 + \delta_1 zx_1 + \delta_2 zx_2 + \epsilon$$

$$\frac{\partial zy}{\partial zx_1} = \delta_1 = \frac{\sigma_1}{\sigma_y}\beta_1$$

$$where \quad zy = \frac{y - \overline{y}}{\sigma_y} \quad zx = \frac{x - \overline{x_j}}{\sigma_j}$$

- if $x_1$ increases by 1 s.d. then y changes by $\delta_1$ standard deviations.
- to generate a sdtzed variable: egen zvarname=std(varname)
- Interpretation: 1 s.d. increase in $x_1$ decreases $y$ by $\delta_1 s.d.$

# Log forms

$$\ln(y) = \alpha + \beta_1 \ln(x_1) + \beta_2 x_2 + \beta_3 (1/x_3) + \beta_4 x_3 + \beta_5 x_4 + \beta_6 x_4^2 + u$$

- $$\frac{\partial \ln(y)}{\partial \ln(x_1)} = \beta_1$$
  % change in y for a 1% increase in $x_1$, elasticity of y w.r.t. $x_1$
- $$\frac{\partial \ln(y)}{\partial x_2} = \beta_2$$
  change in $\ln(y)$ for a unit increase in $x_2$; when $\beta_2$ multiplied by 100, this is the percentage change in y (also called semi−elasticity of y w.r.t. $x_2$)

## Examples

reg logearn age ages yearsed deg_all

1. $\ln(y) = logearn$ and $x_2 = s$ years of education. Suppose $\beta_2 = 0.054$ says that each year of education increases wages by a constant percentage, 5.4%.

$$\%\Delta wage \approx (100 \cdot \beta_2)\Delta x_2$$

2. The coefficient of *deg_all* (0.5613) says that having a degree or a higher qualification increases wages by 56.13% relative to those individuals with lower or no qualifications, holding other factors fixed.

# Log and quadratics forms

$$\ln(y) = \alpha + \beta_1 \ln(x_1) + \beta_2 x_2 + \beta_3 x_3 + \beta_4 (1/x_3) + \beta_5 x_4 + \beta_6 x_4^2 + u$$

$$\frac{\partial \ln(y)}{\partial x_3} = \beta_3 - \frac{\beta_4}{x_3^2}$$

proportionate change in y for a unit increase in $x_3$

$$\frac{\partial \ln(y)}{\partial x_4} = \beta_5 + 2\beta_6 x_4$$

proportionate change in y for a unit increase in $x_4$

- if $\widehat{\beta_5} > 0$ and $\widehat{\beta_6} < 0 \Longrightarrow$ quadratic relationship between x and y, diminishing effects of x on y.
- e.g. tot effect at age 35 $\Longrightarrow 0.1239 - 2 * 0.00140 * 35 = 0.0255$
- tot effect at age 60 $\Longrightarrow 0.1239 - 2 * 0.00140 * 60 = -0.04470$

# t-test

$$H_o : \beta_1 = a_1 \quad \Longleftrightarrow \beta_{age} = 0$$
$$H_1 : \beta_1 \neq a_1 \quad \Longleftrightarrow \beta_{age} \neq 0$$

$$t = \frac{\widehat{\beta}_1 - a_1}{s.e.(\widehat{\beta}_1)} \sim t_{\alpha/2, dof=n-k-1} = \frac{0.1239212 - 0}{0.0029524} \sim t_{0.025, 13724-22-1}$$

- Recalling that the degrees of freedom ($dof$) are the difference: *number of observations minus number of estimated parameters*
- Rejection rule: if $|t| > t^c \Longrightarrow 41.97 > 1.96 \Longrightarrow$ *reject* $H_o$

# t-test in Stata

- The t-stat appears in the regression output.
- You can perform the test manually
  test age=0
- but it shows an F-test
- knowing that $t^2_{n-k-1} = F_{1,n-k-1}$ the results are identical

$$F(1, 13701) = 1761.69$$

$$di\ sqrt(1761.69) \Longrightarrow 41.97$$

$$di\ invttail(13702, 0.025) \Longrightarrow 1.96$$

# p-value

Stata shows also the p-value: the largest significance level at which the null hypothesis would not be rejected, given the observed t. Generally, one **rejects** the null hypothesis if the **p-value is smaller** than or equal to the **significance** level.

$$P(T > t_{observed}|H_o) = p$$
$$P(|T| > |t|\,|H_o) = 2 * P(T > |t|) = p$$
$$in\ Stata\ ttail(n, t)$$
$$di\ 2 * ttail(13724, 41.97) \Longrightarrow 0.000$$

# t-test - other commands

- testparm dresid2 dresid3, equal
  test whether the coeff are equal
- test age=5
- testnl To test non-linear constraints

# Confidence Interval

From

$$\frac{\widehat{\beta}_i - \beta_i}{se(\widehat{\beta}_i)} \sim t_{\alpha/2,n-k-1}$$

simple manipulations leads to $(1-\alpha)$% CI for unknown $\beta_i$:

$$\widehat{\beta}_i \pm t^c_{\alpha/2,n-k-1} \cdot se(\widehat{\beta}_i)$$

where $t^c_{\alpha/2,n-k-1}$ is $(1-\alpha/2)^{th}$ percentile in $t_{\alpha/2,n-k-1}$ distribution.
In our example, 95% CI for *deg_all*:

$$\beta_{-i} = 0.5612611 - 1.96 * 0.0152126 = 0.53144$$
$$\beta_i = 0.5612611 + 1.96 * 0.0152126 = 0.59107.$$

It is a good practice, when running models to check the CI for the same parameter estimated.

# $R^2$

Defining

- $SST = \sum_{i=1}^{n}(y_i - \overline{y})^2$ total sum of squares
- $SSE = \sum_{i=1}^{n}(\widehat{y_i} - \overline{y})^2$ explained sum of squares
- $RSS = \sum_{i=1}^{n} \widehat{u_i}^2$ **residual sum of squares**

Knowing that

$$SST = SSE + RSS$$

The $R^2$ is defined to be

$$R^2 = \frac{SSE}{SST} = 1 - \frac{RSS}{SST}$$

# $R^2$ interpretation

- It is the proportion of the sample variation in $y_i$ explained by the OLS line.
- It never decreases and increases when an additional regressor is added to a regression.
- In our example, $R^2 = 0.2171$ means that all the independent variables together explain about 21.71% of the variation of log wages for our sample of workers.

# F-test of multiple restrictions

$$H_o : \beta_1 = \beta_1^0, \beta_2 = \beta_2^0 \dots \beta_q = \beta_q^0$$
$$H_1 : \beta_j \neq \beta_j^0, \; j = 1, \dots, q$$

The null constitutes **q restrictions** $\Longrightarrow$ **multiple (or joint) hypothesis test**.

# Unrestricted and Restricted models

The **unrestricted model** has k independent variables + the intercept

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u \tag{1}$$

Suppose that the restriction is that q of the k variables (for simplicity the last q) have zero coefficients, then

$$H_o : \beta_{k-q+1} = 0, \ldots, \beta_k = 0$$

and imposing these restriction in (1) we obtain the **restricted model**

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{k-q} x_{k-q} + u \tag{2}$$

# Perform the F-test

1. Run regression 1 and get $RSS_{ur}$
2. Run regression 2 and get $RSS_r$
3. Compute the F statistic

$$F = \frac{(RSS_r - RSS_{ur})/q}{RSS_{ur}/dof} \sim F^{\alpha}_{q,dof}$$

- **q** $=$ numerator degrees of freedom $=$ $dof_r - dof_{ur} = (n-(k-q+1)) - (n-(k+1)) =$ number of restrictions under $H_o$, i.e. the number of equality signs in $H_o$
- **dof**$=$ denominator degrees of freedom $= dof_{ur} = n-k-1$

# Rejection Rule

- If $F > F^c \implies$ *reject* $H_0$
  then $x_{k-q+1}, \ldots, x_k$ are **jointly statistical significant**.
- If $H_o$ is not rejected, then the variables are **jointly insignificant**.

In this context the p-value is defined as

$$p = P(\mathcal{F} > F)$$

where $\mathcal{F}$ is an F random variable with $(q, n-k-1)$ degrees of freedom, and F is the actual value of the test statistic.
A small p-value is evidence against $H_o$.

# F-test in Stata

$H_o$ : nonwhite = female = married = numdep = 0 $\Longrightarrow q = 4$

- use wage1.dta
- **Unrestricted model**
  reg lwage educ exper tenure nonwhite female married numdep
  $\Longrightarrow k = 7$
- test nonwhite female married numdep

$H_o$ : female =-0.3 , married =0.15, numdep = 0 $\Longrightarrow q = 3$

- testnl (_b[*female*] = −0.3) ( _b[*married*] = 0.15) (_b[*numdep*] = 0)

# Example - manually

$H_o$ : nonwhite = female = married = numdep = 0 $\Longrightarrow q = 4$

- **Unrestricted model**
  reg lwage educ exper tenure nonwhite female married numdep
  $\Longrightarrow k = 7$
- take note of $RSS_{UR}$
- **Restricted model**
  reg lwage educ exper tenure $\Longrightarrow k - q = 3$
- take note of $RSS_R$
- compute $F = \frac{(RSS_r - RSS_{ur})/q}{RSS_{ur}/(n-k-1)} \sim F^{\alpha}_{q,(n-k-1)}$
- find in the Tables F critical value or use Stata command
  di $invF(q, n-k-1, \alpha)$

# F-test of overall significance

$$H_o : \beta_1 = \beta_2 = \ldots = \beta_q = 0 \quad H_1 : Any\, \beta_j \neq 0 \; j = 1, \ldots, q$$

$$\mathbf{UR} : y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

$$\mathbf{R} : y = \beta_0 + u$$

$$F = \frac{(RSS_r - RSS_{ur})/k}{RSS_{ur}/(n-k-1)} \sim F^{\alpha}_{k,(n-k-1)} \quad or$$

$$F = \frac{R^2_{ur}/k}{(1 - R^2_{ur})/(n-k-1)}$$

The F statistic with the $R^2$ is valid only for testing joint exclusion of **all** regressors.

# Saving typing - **macro**

For defining lists of vars ( globally).

- **global**
  1. type **global** followed by the groupname followed by the vars that you want to group together
  2. in a regression type reg depvar followed by $groupname
- Pay careful attention to the sign $ *in the global*.

# Example - overall significance

The F test of overall significance is reported automatically in Stata output.
You can also perform it either computing the F statistic or by using that Stata command test. For example, using a macro

- global indvars educ exper tenure nonwhite female married numdep
- reg lwage $indvars
- test $indvars