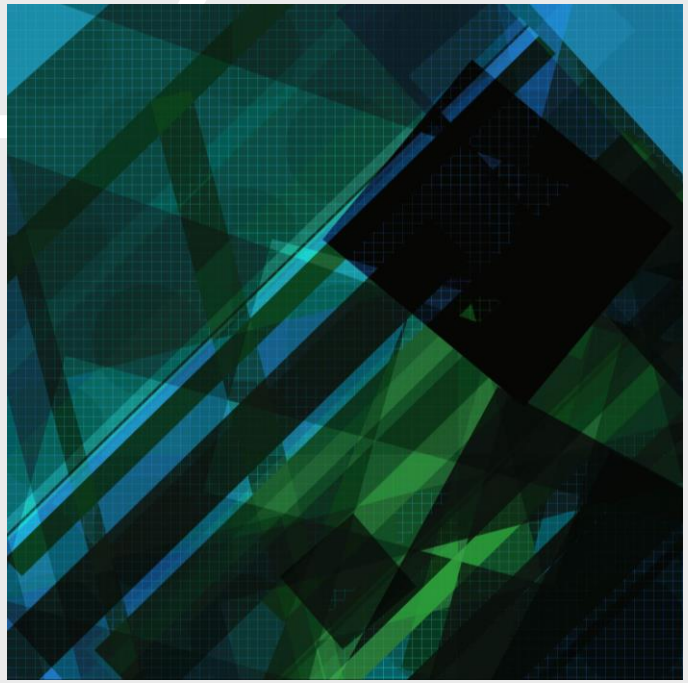
Intermediate Econometrics

IMQF & MEAE 2024/25

Aleksandra Nojković



Recommended textbooks:

- **Verbeek, M. (2017), *A Guide to Modern Econometrics*, 5th edition, John Wiley&Sons**
- **Wooldridge, J.M. (2021) *Introductory Econometrics: A Modern Approach, 7th ed.,* Cengage Learning**
- Asteriou, D. and Hall, S.H. (2021), Applied Econometrics, 4th ed., Bloomsbury Academic
- Greene, W.H. (2018), *Econometric Analysis*, 8th edition, Pearson

An introduction to linear regression

- **Two of the cornerstones** of econometrics:

- Linear regression model

- Ordinary least squares (OLS)

Simple and Multiple Regression Model
(Verbeek, Chapters 1-3)

1) Introduction to regression: the classical linear regression model (CLRM)
2) The ordinary least square (OLS) method of estimation
3) The assumptions of the CLRM
4) Properties of the OLS estimator
5) The overall goodness-of-fit
6) Hypothesis testing and confidence intervals
7) How to estimate a regressions in EViews *(Regression; Diagnostics)*

The linear regression model (simple)

- Way to examining the nature and form of the relationship among two or more variables
- Important issue: **direction of causation** between two variables
- In the case of two variables, **population regression equation** is:

$$Y=E(Y)+\varepsilon=\beta_1+\beta_2 X+\varepsilon$$

where $E(Y_i)$ denotes the average value of Y_i for given X_i ; β_1 and β_2 are unknown population parameters; ε is an **error term** or **disturbance term.**

Independent vs. Dependent Variables

- Y in the model
  - **Dependent variable**
  - Response variable
  - Explained variable
  - Predicted variable
  - Regressand

- X in the model
  - Independent variable: Meaning of ‘independent’
  - **Explanatory variable**
  - Regressor
  - Covariate variable
  - Predictor variable

Reasons why disturbance term exist

1) Aggregation of variables (to avoid having too many variables)

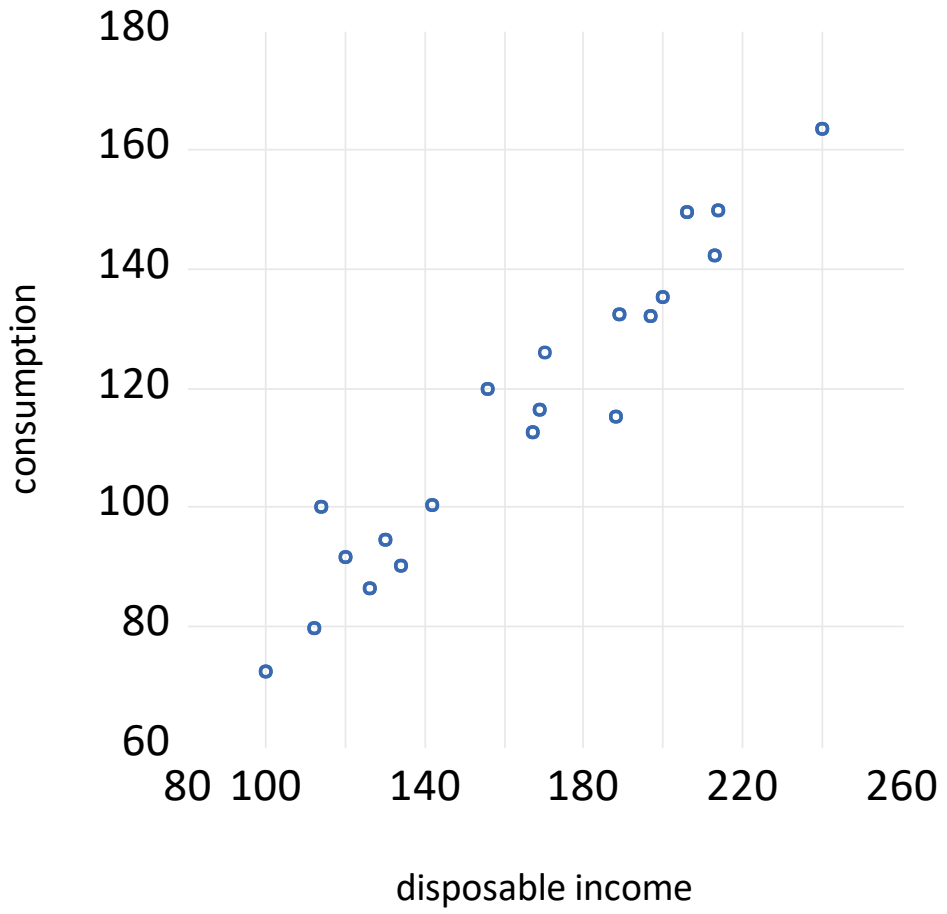
2) Omission of explanatory variables and functional misspecification

3) Unpredictability of human behavior

4) Measurement error

Sample data: Scatter plot of Y on X

(Data source: AH (2021))



Sample regression equation

- We are estimating population regression function based on **sample information** (*i*=1,2,..., N):

$$\boldsymbol{y_i = \underbrace{\beta_1 + \beta_2 x_i}_{\substack{deterministic \\ term}} + \underbrace{\varepsilon_i}_{\substack{error \\ term}}}$$

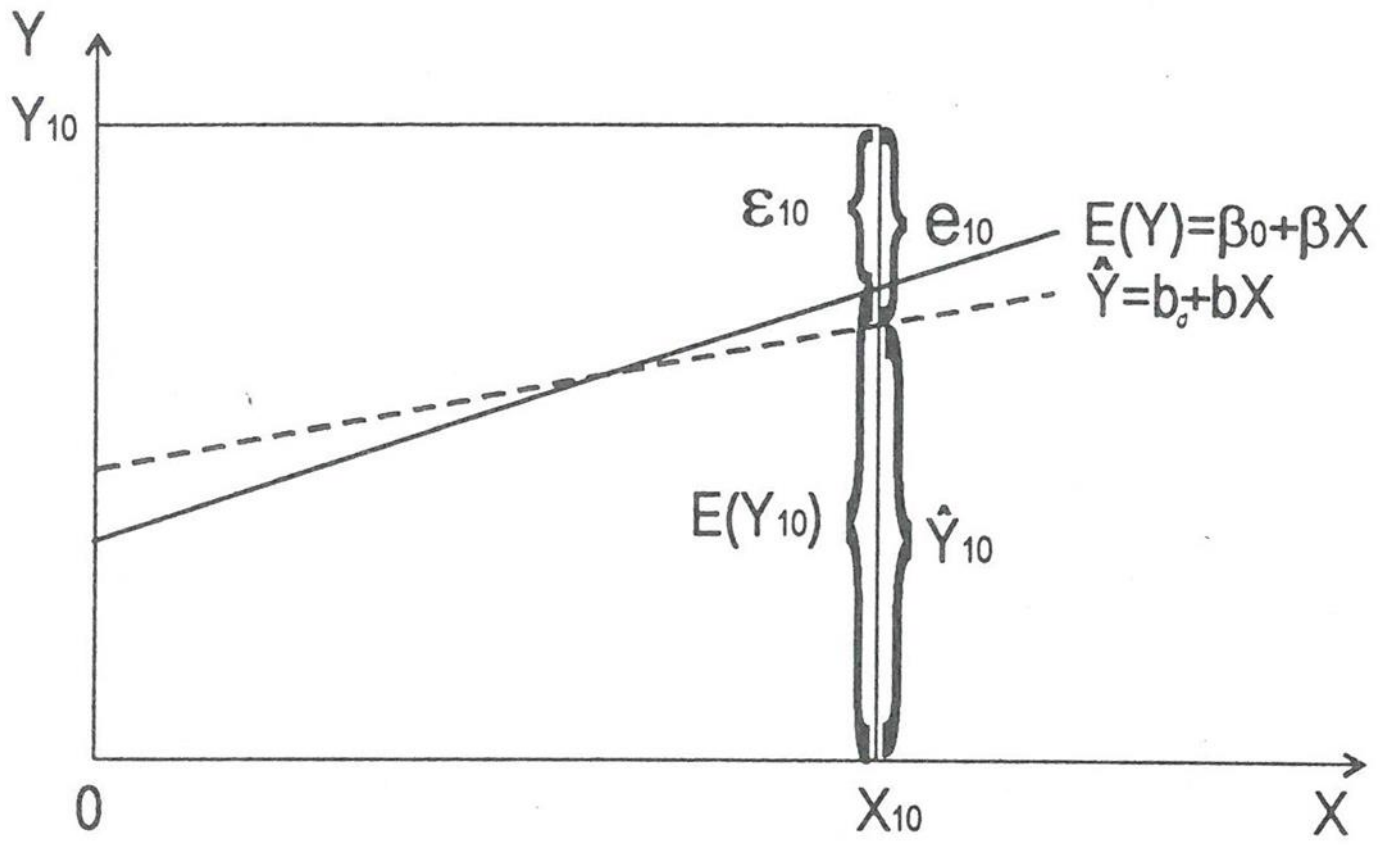
- This give us the following relationship – fitted straight line:

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i = b_1 + b_2 x_i$$

or an actual value of Y can be written as a sum of predicted value of y and residual:

$$y_i = \hat{y}_i + e_i$$

Population vs sample regression equation



Sample data: Scatter plot of Y on X (with fitted line and observation points)

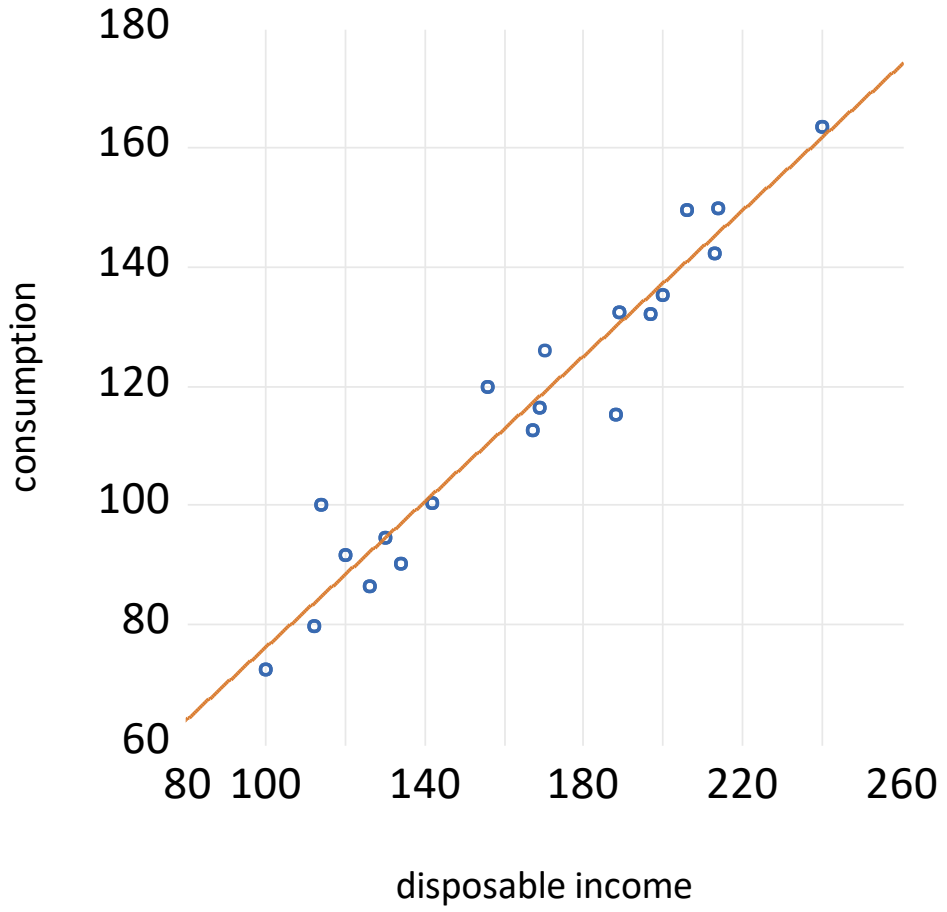


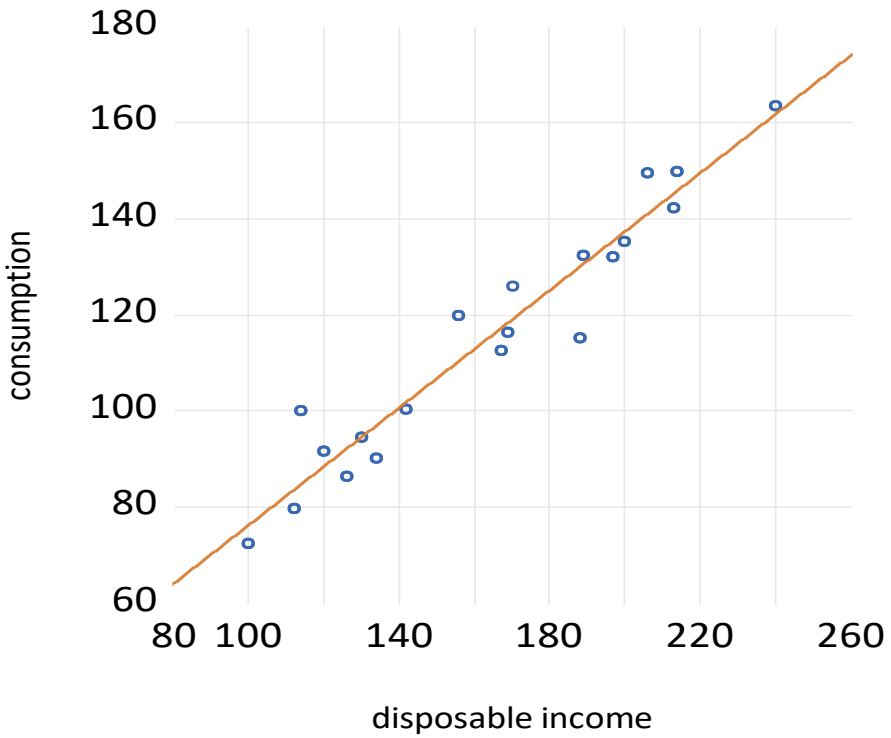
Table of OLS estimates of consumption function

Dependent Variable: Y
Method: Least Squares
Date: 12/05/21 Time: 21:37
Sample: 1 20
Included observations: 20

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	15.11641	6.565638	2.302352	0.0335
X	0.610889	0.038837	15.72951	0.0000

R-squared	0.932182	Mean dependent var	115.5160
Adjusted R-squared	0.928415	S.D. dependent var	25.71292
S.E. of regression	6.879603	Akaike info criterion	6.789639
Sum squared resid	851.9210	Schwarz criterion	6.889212
Log likelihood	-65.89639	Hannan-Quinn criter.	6.809076
F-statistic	247.4176	Durbin-Watson stat	2.283770
Prob(F-statistic)	0.000000		

Measures Covariation



Predictor: con.= 15.12 + 0.61 dis. income

$$b = \frac{Cov(x,y)}{Var(x)}$$

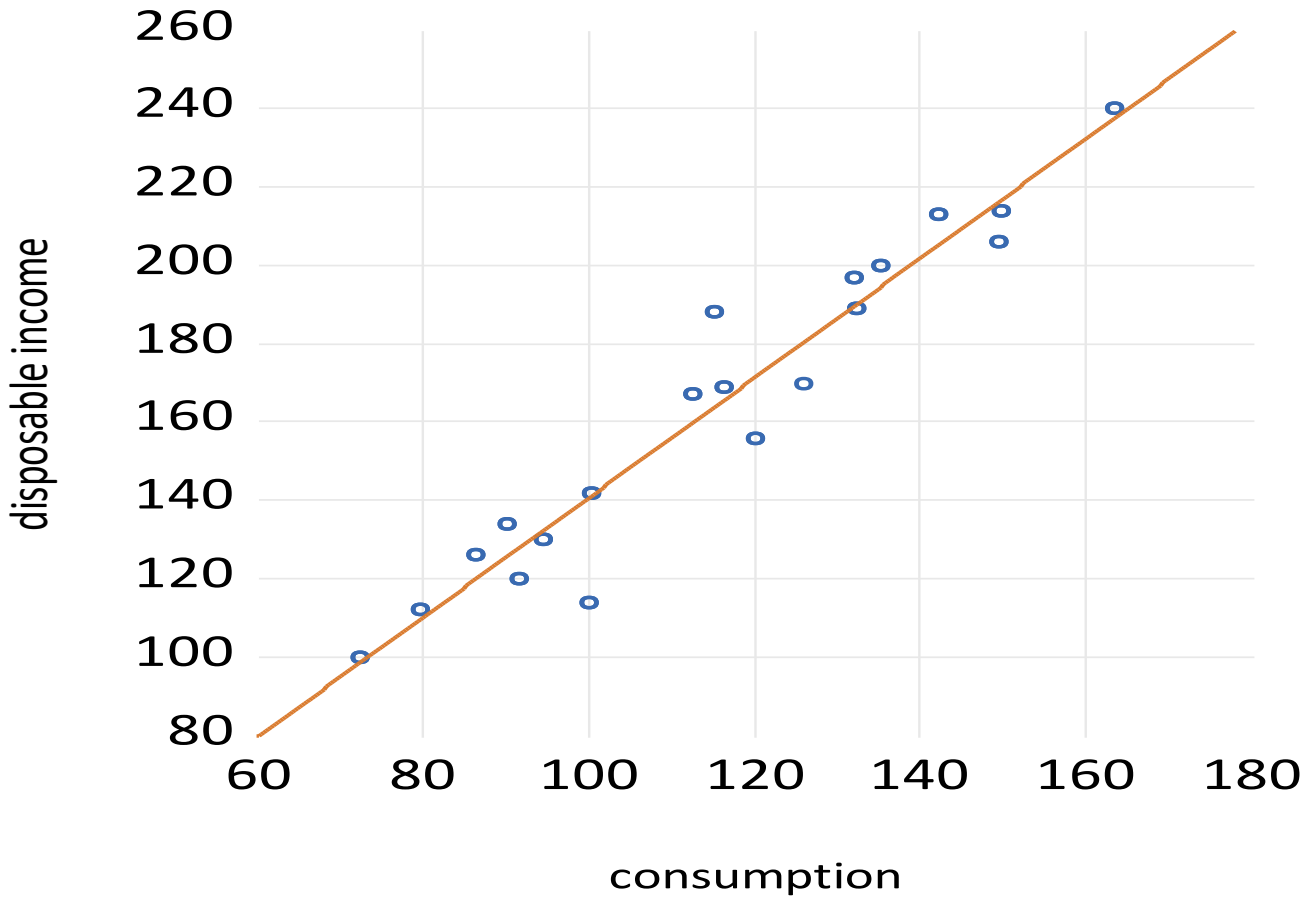
Note the numerator of b is the covariance of x and y.

If Cov(x,y) = 0, then b = 0.

Also, since the correlation is $r_{xy} = \frac{Cov(x,y)}{\sqrt{Var(x)Var(y)}} = \frac{s_{xy}}{s_x s_y}$,

$$b = \frac{s_y}{s_x} \times \text{Correlation of x and y.}$$

Causality?



Correlation = 0.96 (!)

Disp. income= -11.92 + 1.53 consumption

Ordinary Least Square (OLS)Estimator

- Method based on following criterion - chose the sample regression function in such a way that **sum of the squared residuals is as small as possible (i.e., is minimized)**

- Most popular technique in uncomplicated application of regression analysis

- The OLS estimates follow some numerical and statistical properties (such as **unbiasedness and efficiency**) - we will discuses them latter

Normal Equations

$$SSR = \sum_{i=1}^{N} (y_i - b_1 - b_2 x_i)^2$$

$$\frac{\partial SSR}{\partial b_1} = \sum_{i=1}^{N} \frac{\partial (y_i - b_1 - b_2 x_i)^2}{\partial b_1} = \sum_{i=1}^{N} 2(y_i - b_1 - b_2 x_i)(-1) = -2 \sum_{i=1}^{N} e_i = 0$$

$$\Rightarrow \frac{1}{N} \sum_{i=1}^{N} e_i = 0 \Rightarrow \bar{e} = 0$$ (The residuals sum to zero)

$$\frac{\partial SSR}{\partial b_2} = \sum_{i=1}^{N} \frac{\partial (y_i - b_1 - b_2 x_i)^2}{\partial b_2} = \sum_{i=1}^{N} 2(y_i - b_1 - b_2 x_i)(-x_i) = 0 = -2 \sum_{i=1}^{N} x_i e_i$$

$$= \Rightarrow \sum_{i=1}^{N} x_i (e_i - \bar{e}) = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})(e_i - \bar{e}) = 0$$

(The covariance of x and the residuals is zero)

- It is easy to verify that second-order conditions for a minimum are met

OLS estimators in simple CLRM

$$\widehat{\beta_2} = b_2 = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2}$$

$$\widehat{\beta_1} = b_1 = \overline{y} - b_1 \overline{x}$$

The classical linear regression model (multiple)

- Economists are usually interested in following:

$$y_i=\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK}+\varepsilon_i$$

or in matrix notation:

$$y = X\beta + \varepsilon,$$

where *y* and ε are N-dimensional vectors and *X* is of dimension N x K.

- The OLS estimator for β is given by:

$$b = (X' X)^{-1} (X'y)$$

Matrix notation

- **We define column vectors of N observations on y and the *K-1* explanatory variables (X) + const. term, and introduce the notation:**

$$\mathrm{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{12} & \cdots & x_{1K} \\ 1 & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N2} & \ldots & x_{NK} \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}$$

$$\mathrm{y} \quad = \quad \mathrm{X}\beta \quad + \quad \varepsilon$$

- **(X'X) is invertible - The assumption means that the rank of the matrix X** is ***K*. No linear dependencies => FULL COLUMN RANK of the matrix X (topic of perfect/exact multicollinearity)**

Ordinary Least Squares

By construction, OLS produces the **best linear approximation** of y from x_2 to x_K and a constant. However, **without additional assumptions**, this approximation has limited value (**we have not used** any economic or statistical theory so far) :

- the coefficients do not have an economic interpretation
- we cannot make statistical statements about these coefficients
- the approximation is valid within a given set of observations only
- the linear relationship **has no general validity outside the current set** of values (e.g., in the future or for units not in the sample)

The linear regression model

- We now start with a linear relationship between y and X_1=1 (a constant), X_2 to X_K, which we assume **to be *generally* valid**
- We write: $y_i = x_i'\beta + \varepsilon_i$
- The model is a **statistical model and has an "error term".** The error term ε_i contains all influences that are not included explicitly in the model
- The unknown coefficients β_k have a meaning and measure how we expect Y to change if X_k changes (**and all other x values remain the same)**
- As a result, OLS produces an estimator for the unknown population parameter vector β

Interpreting the linear model

- The coefficient β_k measures the expected change in y_i if x_{ik} changes by one unit (**marginal effect**), but **all other variables** in x_i **do not change**. That is,

$$\frac{\partial E\{y_i|x_i\}}{\partial x_{ik}} = \beta_k$$

- The statement that the other variables in x_i do not change is a *ceteris paribus condition*

- In a multiple regression model, single coefficients can only be interpreted **under a ceteris paribus condition**. Thus, (strictly speaking) can only be interpreted if we know which other variables are included

About ceteris paribus

- If we are interested in the relationship between y_i and x_{ik} the other variables in x_i act as **control variables**

- Sometimes, ceteris paribus is hard to maintain

- *For example, what is the impact of age upon a person's wage, keeping years of experience fixed?*

About ceteris paribus (II)

- Sometimes, ceteris paribus is impossible, for example if the model includes both age and age-squared

- *Example: model includes*

$$age_i \beta_2 + age_i^2 \beta_3$$

then the marginal effect of a changing age (ceteris paribus) is

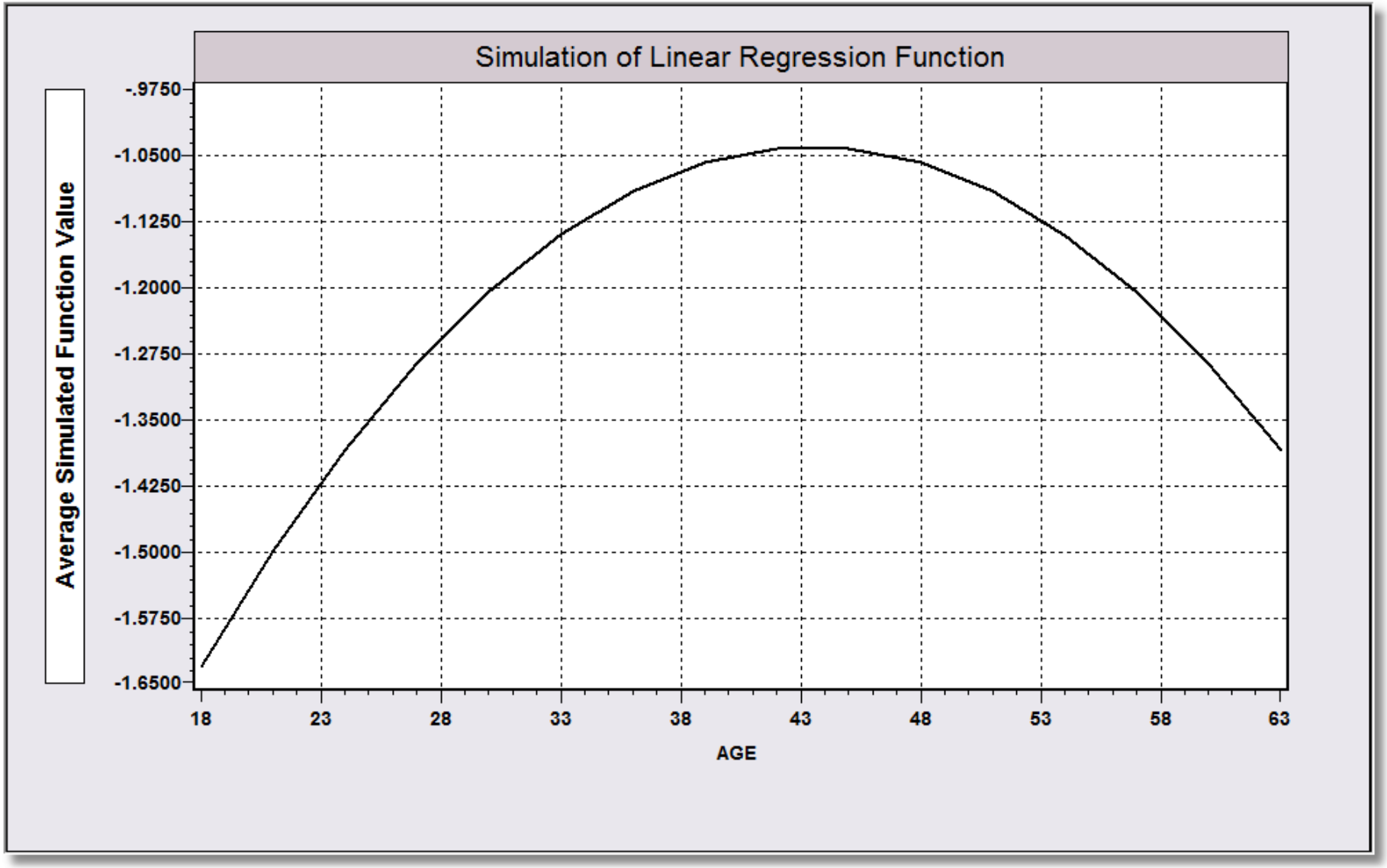
$$\frac{\partial E\{y_i | x_i\}}{\partial age_i} = \beta_2 + 2age_i \beta_3$$

Consequently, the marginal effect depends upon age

Functional Form: Quadratic

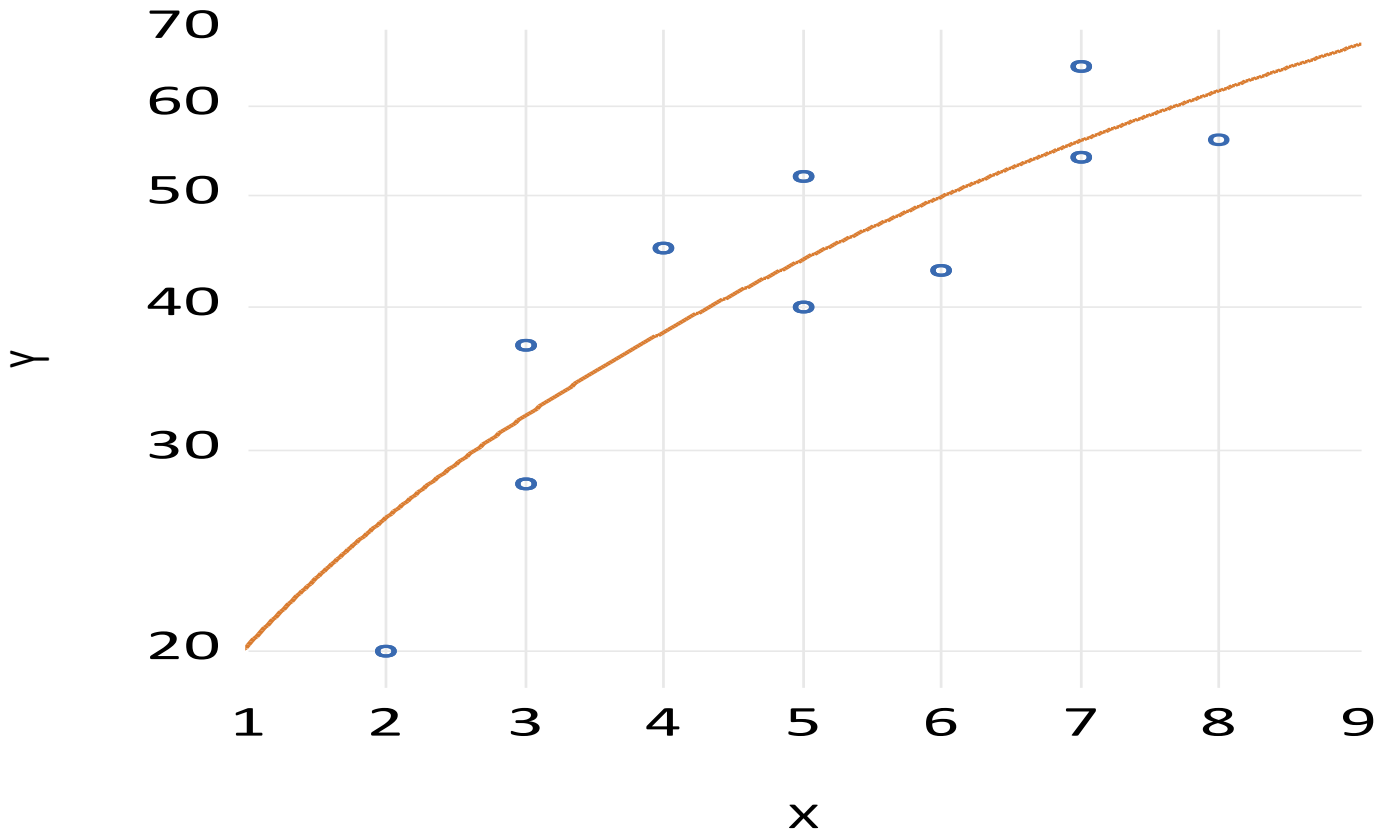
- $Y = b_1 + b_2X + b_2X^2 + e$
- $dE[Y|X]/dX = b_2 + 2b_3X$
- Diminishing marginal effect (easily seen on graph)

Functional Form: Quadratic (II)



Non-linear models

Consumption = *f* (Disp. Income)



Transformation to linear model

- Baseline model:

$$Y = \beta_0 X^{\beta} \ ,$$

by taking logarithms becomes:

$$\underbrace{\ln Y}_{Y^*} = \underbrace{\ln \beta_0}_{\beta_0^*} + \beta \underbrace{\ln X}_{X^*} \quad .$$

- Now interpretation of β is elasticity:

- $\frac{\partial lnY}{\partial lnX} = \frac{\%\ change\ Y}{\%\ change\ X} = Elasticity\ of\ Y\ with\ the\ change\ in\ X$

OLS estimator and OLS estimates

- An estimator is a **random variable:**

- because the sample is randomly drawn from a larger population

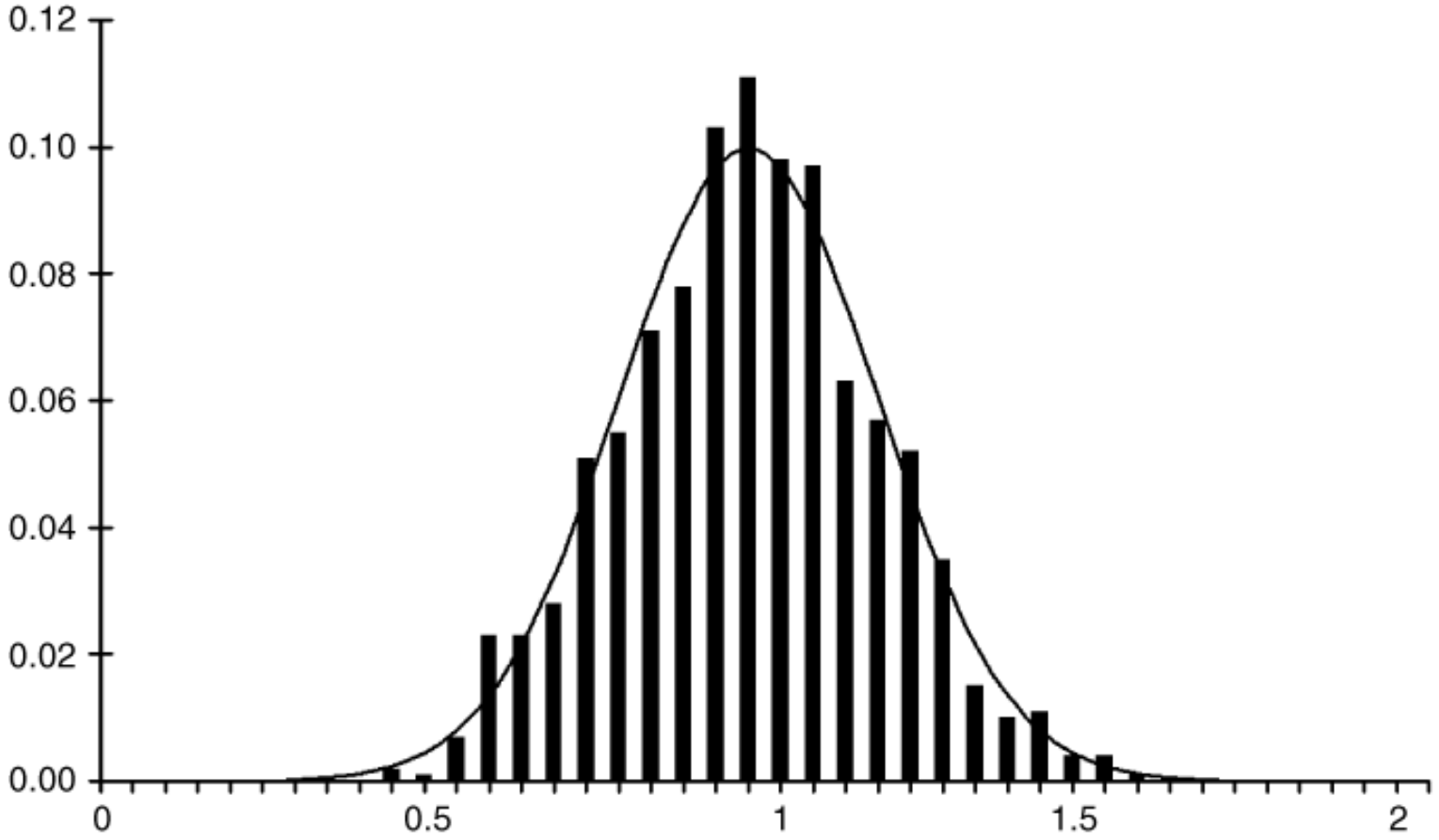
- because the data are generated by some random process (each ε_i is random drawing from the population distribution, independent from the other error terms)

- A **new sample means a new estimates** (**new set of N observations** (y_i, x_i))

- When we consider the different estimates for many different samples, we obtain **the sampling distribution of the OLS estimator** (see Figure in the next slide)

- We evaluate the "quality" of the OLS estimator (and a given OLS estimate) **by the properties of the sampling distribution**

Figure of sampling distribution: histogram and normal density



OLS estimator and OLS estimates (II)

- The estimator is a vector of random variables
- The estimate is a vector of numbers
- While given sample only produces **a single estimate**, we evaluate a **properties of the underlaying estimator**

b is a statistic

- Random because it is a sum of the ε's
- It has a distribution, like any sample statistic

Error terms and residuals

- Note that we write:

$$y_i = x_i'\beta + \varepsilon_i$$

and

$$y_i = x_i'b + e_i$$

- We call ε_i the **error term** and e_i **the residual**

- The error term is unobserved, **the residual is constructed** (after the estimation) using the estimate b

Error terms and residuals (II)

- By **virtue of the first order conditions** of OLS, the residual is mean zero and uncorrelated with x_i:

1) $$\sum_{i=1}^{N} e_i = 0$$

2) $$\sum_{i=1}^{N} x_i e_i = 0$$

- This **does not necessarily hold for the error term**

Is OLS a good estimator?

- The answer to this question depends upon the assumptions we are willing to make
- The most standard and most convenient ones are given by the **Gauss-Markov assumptions** (assumption made about ε_i and x_i)
- Note that these assumptions are very strong and **often not satisfied**
- Under the Gauss-Markov assumptions, the OLS estimator has nice properties
- Later, we shall discuss how essential the Gauss-Markov assumptions are and how they **can be relaxed**

The Gauss-Markov assumptions

(A1) Error terms have mean zero: $E\{\varepsilon_i\}=0$

(A2) *All* error terms are independent of *all* X variables (exogeneity):

$\{\varepsilon_1 ,\dots \varepsilon_N\}$ is independent of $\{x_1,\dots x_N\}$

(A3) All error terms have the same variance (homoskedasticity):

$$V\{\varepsilon_i\} = \sigma^2$$

(A4) The error terms are mutually uncorrelated (no autocorrelation):

$$cov\{\varepsilon_i,\varepsilon_j\} = 0, \quad i \neq j$$

- **These assumptions imply** that $E\{ y_i \mid x_i \} = x_i'\beta$
- Under (A2) we can treat the **explanatory variables as fixed** (deterministic)

Estimator properties

- *Under assumptions (A1) - (A4):*

1. The OLS estimator **is unbiased**. That is, $E\{b\} = \beta$

- Under assumptions (A1), A(3) and (A4):

2. Error terms are uncorrelated drawings from a distribution with **expectations zero** and **constant variances** σ^2

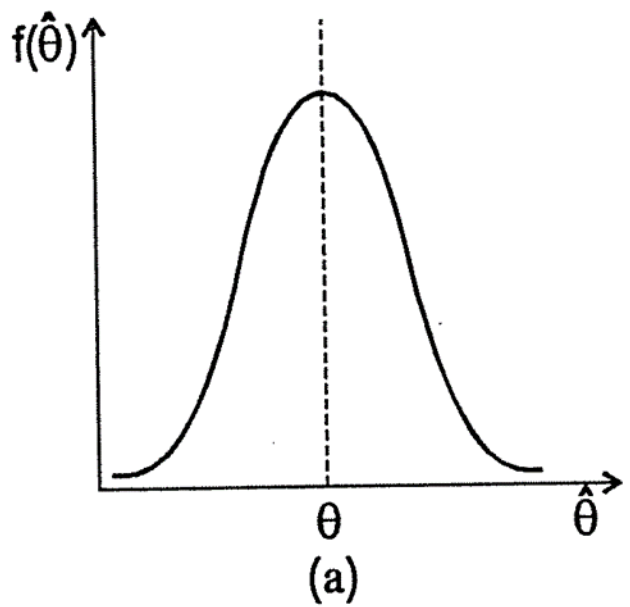
- *Under assumptions (A1), (A2), (A3) and (A4):*

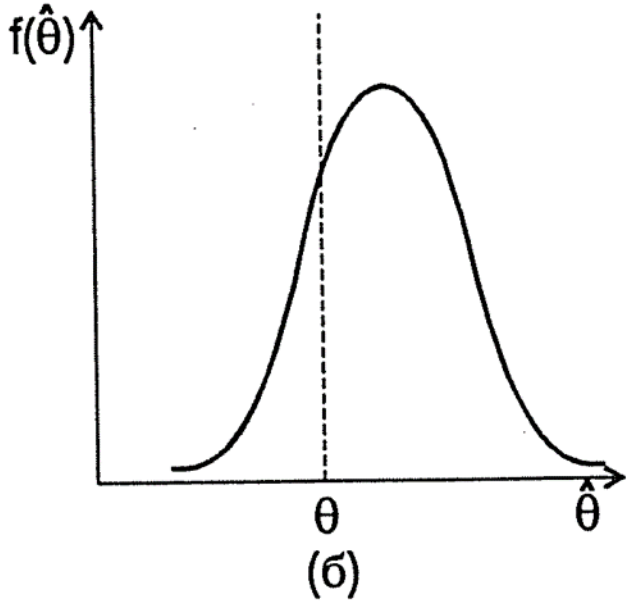
3. **The variance of the OLS** estimator is given by:

$$V\{b\} = \sigma^2(\Sigma_i x_i x_i')^{-1}$$

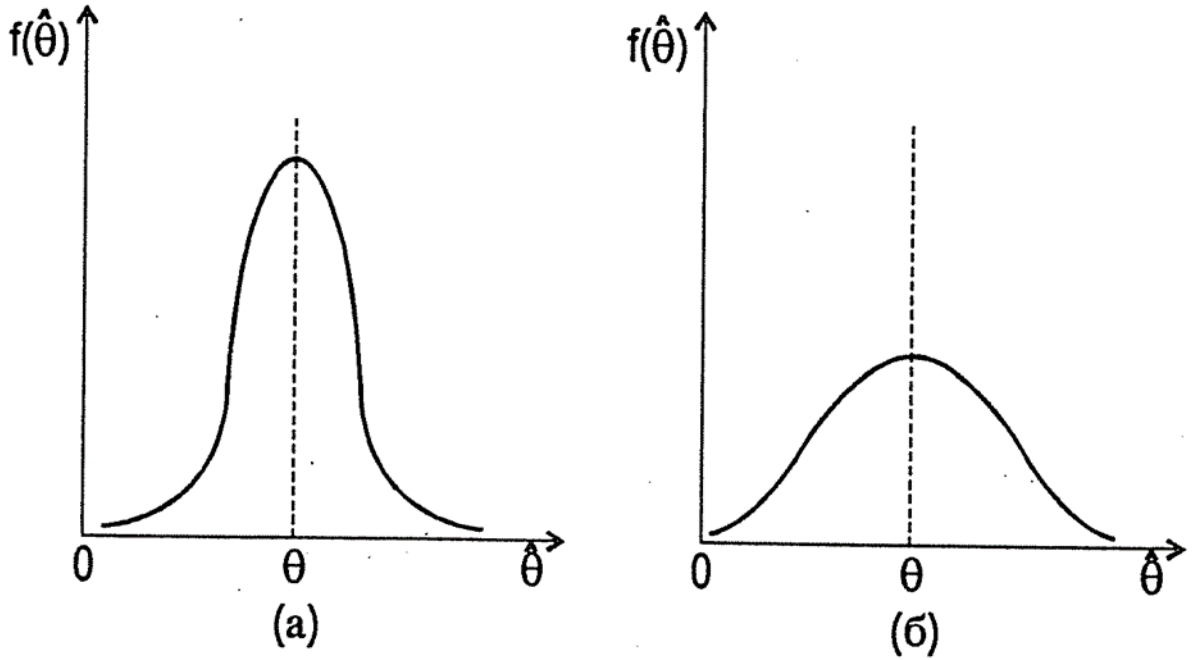
4. The OLS estimator is BLUE: **best linear unbiased estimator** for β

Unbiased and biased estimator

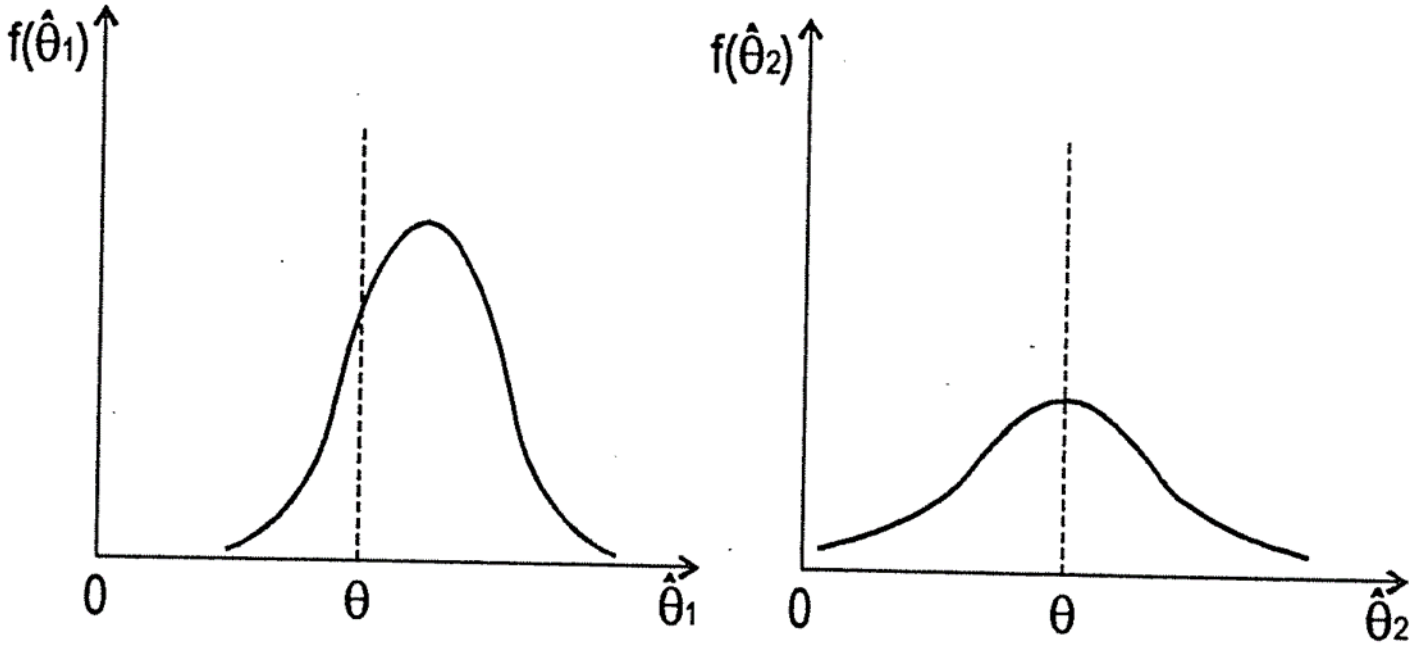




Relatively efficient and inefficient estimator



Mean Square Error criteria
(MSE = Bias^2 + Variance)



Estimator properties (II)

- We **estimate the variance of the error term** σ^2 by the sampling variance of the residuals
- However, because K parameters were chosen to minimize the residual sum of squares, we employ a **degrees of freedom correction**:

$$s^2 = (N-K)^{-1} \Sigma_i e_i^2$$

- Under assumptions (A1)-(A4), s^2 **is unbiased** for σ^2
- We estimate the variance (covariance matrix) of b by :

$$\hat{V}\{b\} = s^2(\Sigma_i x_i x_i')^{-1}$$

- The square root of the k^{th} diagonal element is the **standard error** of b_k

Estimator properties (III)

- **A convenient fifth assumption** is that all error terms have a normal distribution. We specify:

(A5): $\varepsilon_i \sim NID(0, \sigma^2)$

which is shorthand for: all ε_i are *independent* drawings from a *normal* distribution with mean 0 and variance σ^2. ("normally and independently distributed")

- **(A5) replaces (A1)+(A3)+(A4)**

- *Under assumptions* ***(A2) + (A5):***

4. The OLS estimator **b has a normal distribution** with mean β and covariance matrix $V\{b\} = \sigma^2(\Sigma_i x_i x_i')^{-1}$

Example: individual wages

- Consider a sample of N=526 individuals (252 females; W(2021)). We observe wage rates (per hour), gender, experience and years of schooling.

- **First model:** explain wage from a female dummy (= 1 if female, 0 if male). That is:

$$wage_i = \beta_1 + \beta_2\, female_i + \varepsilon_i$$

- The interpretation is: the expected wage of a person, given his or her gender is $\beta_1 + \beta_2\, female_i$

- That is, the expected wage of an arbitrary female is $\beta_1 + \beta_2$, for an **arbitrary** male it is β_1

Dummy Variable

- D = 0 in one case and 1 in the other
- $Y = b_1 + b_2D + e$ (change in level – constant term)
- When D = 0, $E[Y|X] = b_1$
- When D = 1, $E[Y|X] = b_1 + b_2D$

Set of Dummy Variables

- Usually, Z = Type = 1,2,…,K (e.g., in wage1.wf1: west, south, north-central and east US)

- $Y = b_1 + b_2 X + d_1$ if Type=1
  $+ d_2$ if Type=2
  …
  $+ d_K$ if Type=K

Set of Dummy Variables (II)

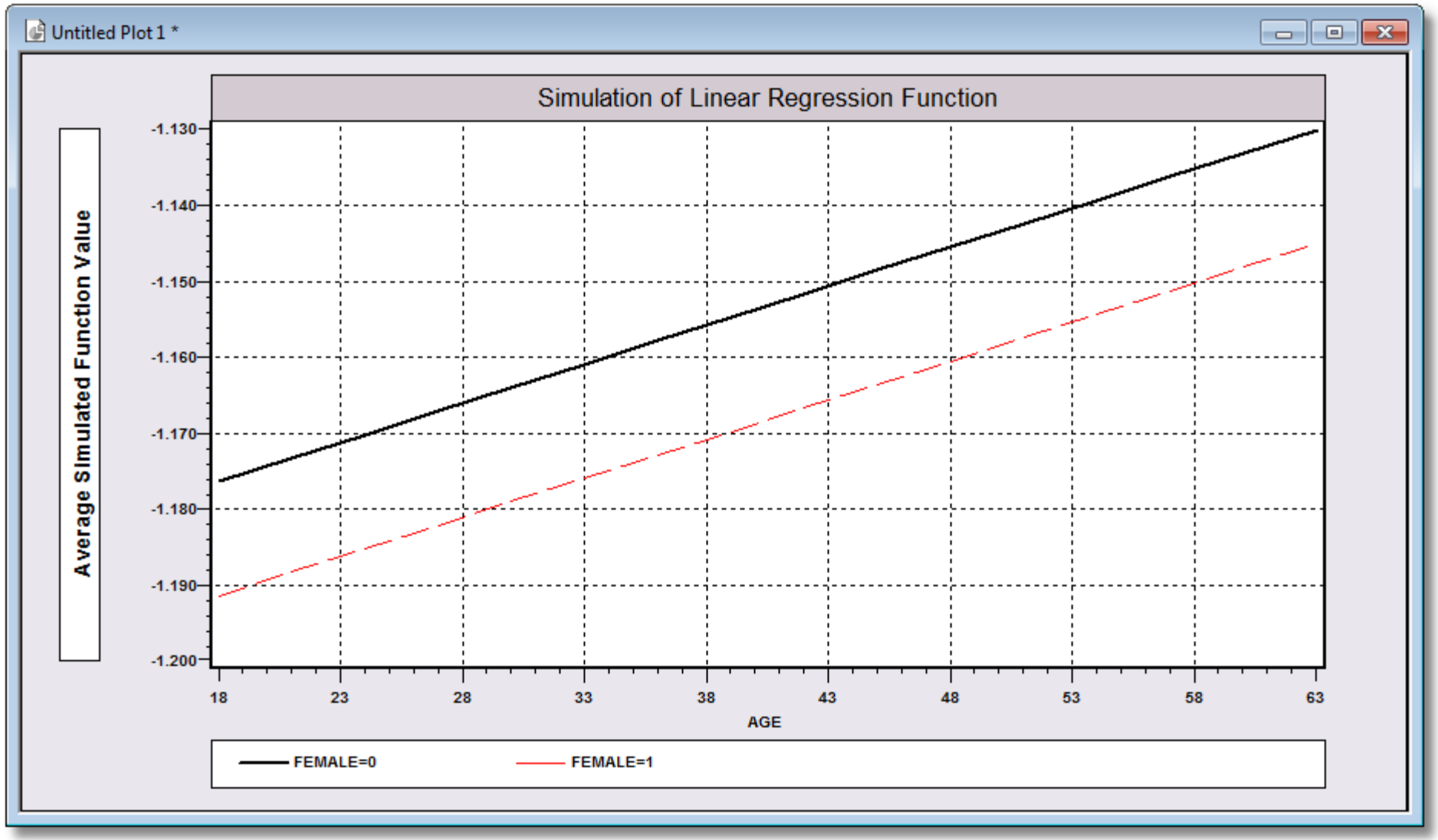
- Complete set of dummy variables divides the sample into groups
- Fit the regression with “group” effects
- **Need to drop one** (anyone) of the variables to compute the regression. (Avoid the “dummy variable trap”)
- **Interaction effect** between two dummy variables (e.g., females employed in services) or dummy and numerical variables (change in marginal effect)

Table of OLS estimates wage equation (I)

Dependent Variable: LWAGE
Method: Least Squares
Date: 12/05/21 Time: 20:19
Sample: 1 526
Included observations: 526

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.813570	0.029814	60.83028	0.0000
FEMALE	-0.397217	0.043073	-9.221915	0.0000

R-squared	0.139635	Mean dependent var	1.623268
Adjusted R-squared	0.137993	S.D. dependent var	0.531538
S.E. of regression	0.493503	Akaike info criterion	1.429220
Sum squared resid	127.6177	Schwarz criterion	1.445438
Log likelihood	-373.8848	Hannan-Quinn criter.	1.435570
F-statistic	85.04372	Durbin-Watson stat	1.825492
Prob(F-statistic)	0.000000		


Untitled Plot 1 *
Simulation of Linear Regression Function
Average Simulated Function Value
-1.130
-1.140
-1.150
-1.160
-1.170
-1.180
-1.190
-1.200
18
23
28
33
38
43
48
53
58
63
AGE
FEMALE=0
FEMALE=1

Regression Arithmetic

$$y_i = \hat{y}_i + e$$

$$y_i - \bar{y} = \left(\hat{y}_i - \bar{y}_i\right) + e_i$$

A few algebra steps later...

$$\Sigma_{i=1}^{N}\left(y_i - \bar{y}\right)^2 = \Sigma_{i=1}^{N}\left(\hat{y}_i - \bar{y}\right)^2 + \Sigma_{i=1}^{N} e_i^2$$

TOTAL LARGE?? small??

TOTAL = Regression + Residual

This is the analysis of (the) variance (of y); ANOVA

Fit of the Equation to the Data

The original question about the model fit to the data :

$$\Sigma_{i=1}^{N}\left(y_i - \bar{y}\right)^2 = \Sigma_{i=1}^{N}\left(\hat{y}_i - \bar{y}\right)^2 + \Sigma_{i=1}^{N} e_i^2$$

TOTAL LARGE?? small??

TOTAL = Regression + Residual

$$\frac{\text{TOTAL SS}}{\text{TOTAL SS}} = \frac{\text{Regression SS}}{\text{TOTAL SS}} + \frac{\text{Residual SS}}{\text{TOTAL SS}}$$

1 = Proportion Explained + Proportion Unexplained

Analysis of Variance Table

Source	Degrees of Freedom	Sum of Squares	Mean Square	
Regression	1	$\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$	$\frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{1}$	
Residual	N-2	$\sum_{i=1}^{N} e_i^2$	$\frac{\sum_{i=1}^{N} e_i^2}{N-2}$	
Total	N-1	$\sum_{i=1}^{N}(y_i - \bar{y})^2$	$\frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-1}$	

Explained Variation

- The proportion of variation “explained” by the regression is called R-squared (R^2)
- It is also called the **Coefficient of Determination**
- (It is the square of something – to be shown later)

ANOVA Table

$$R^2 = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

Source	Degrees of Freedom	Sum of Squares	Mean Square	
Regression	1	$\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$	$\frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{1}$	
Residual	N-2	$\sum_{i=1}^{N} e_i^2$	$\frac{\sum_{i=1}^{N} e_i^2}{N-2}$	
Total	N-1	$\sum_{i=1}^{N}(y_i - \bar{y})^2$	$\frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-1}$	

Goodness-of-fit

- The quality *of the linear approximation* offered by the model can be measured by the R^2

- The R^2 indicates the proportion of the variance in y that can be explained by the linear combination of x variables

- In formula:

$$R^2 = \frac{\hat{V}\{\hat{y}_i\}}{\hat{V}\{y_i\}} = \frac{1/(N-1)\sum_{i=1}^{N}(\hat{y}_i - \overline{y})^2}{1/(N-1)\sum_{i=1}^{N}(y_i - \overline{y})^2},$$

- If the model contains an intercept (as usual), it holds that

$$\hat{V}\{y_i\} = \hat{V}\{\hat{y}_i\} + \hat{V}\{e_i\},$$

Goodness-of-fit (II)

- Accordingly, we can also write

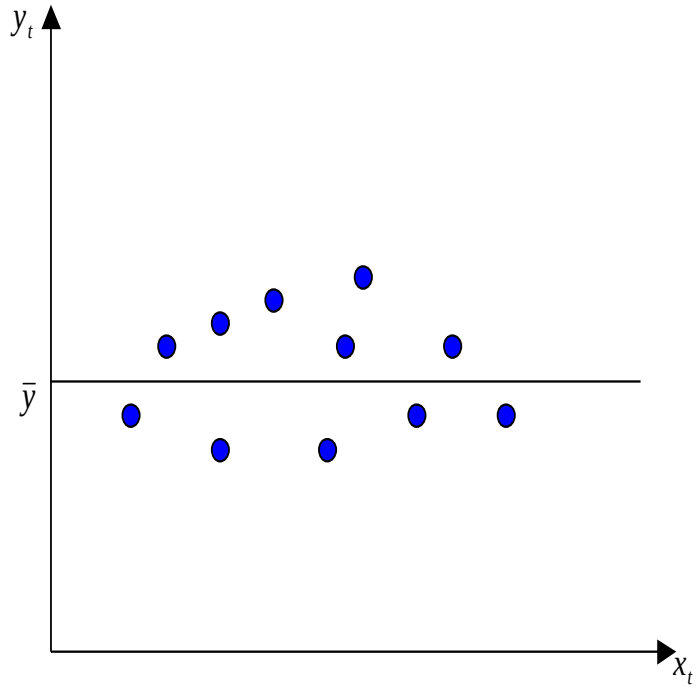
$$R^2 = 1 - \frac{\hat{V}\{e_i\}}{\hat{V}\{y_i\}} = 1 - \frac{1/(N-1)\sum_{i=1}^{N} e_i^2}{1/(N-1)\sum_{i=1}^{N}(y_i - \bar{y})^2}.$$

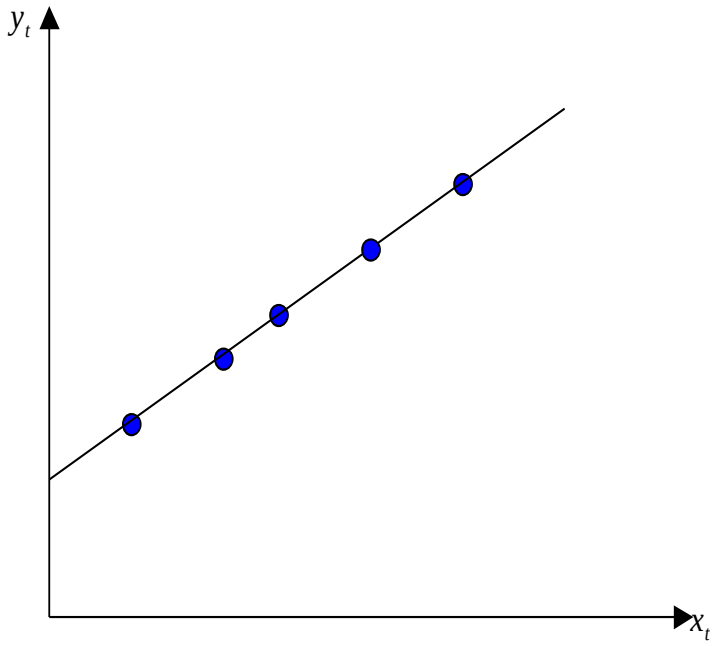
- If the model does *not* contain an intercept, these two expressions are *not* equivalent. (Statistical software may use either definition)

- It is also possible to define the R^2 as the squared correlation coefficient between observed and fitted values of y:

$$R^2 = \mathrm{corr}^2\{y_i, \hat{y}_i\}$$

$R^2 = 0$ and $R^2 = 1$





Correlation Coefficient

$$r_{xy} = \text{Correlation}(x,y)$$

$$= \frac{\text{Sample Cov}[x,y]}{[\text{Sample Standard deviation (x)}]\,[\text{Sample standard deviation (y)}]}$$

$$= \frac{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2}\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i-\bar{y})^2}}$$

$$-1 \le r_{xy} \le 1$$

R-Squared is r_{xy}^2 (in simple model)

- R-squared is the square of the correlation between y_i and the predicted y_i which is $b_1 + b_2x_i$
- The correlation between y_i and $(b_1+b_2x_i)$ is the same as the correlation between y_i and x_i.
- Therefore,….
- A regression with a high R^2 predicts y_i well

Goodness-of-fit (II)

- In general: $0 \leq R^2 \leq 1$.
- There is no general rule to say that an **R^2 is high or low**. This **depends upon the particular context**
- R^2s of 0 or 1 **are suspicious** (close to 1 - "nonsense regressions" in time series analysis)
- R^2s cannot be compared if y is different
- R^2 will never decrease if a variable is added. Therefore, we define **adjusted R^2** as:

$$\overline{R}^2 = 1 - \frac{1/(N-K)\sum_{i=1}^{N} e_i^2}{1/(N-1)\sum_{i=1}^{N}(y_i - \overline{y})^2}.$$

(has a penalty for larger *K*)

Notes About Adjusted R^2

(1) Adjusted R^2 is denoted $\bar{R}^2$. $\bar{R}^2$ is less than R^2.

(2) $\bar{R}^2$ is not the square of $\bar{R}$. It is not the square of anything. Adjusted R squared is just a name, not a formula.

(3) Adjusting R^2 makes no sense when there is only one variable in the model. You should pay no attention to $\bar{R}^2$ when K = 1.

(4) $\bar{R}^2$ can be less than zero! See point (2).

Criteria for model selection

- Maximum of adjusted R^2
- Minimum of S.E. of regression (s): $s^2 = (1 - \bar{R}^2)\frac{\sum y_i^2}{N-1}$
- Minimum of information criterion function (IC):

$$IC(K) = \ln(s^2) + g(K/N)$$

- AIC – Akaike IC (g=2)
- SIC – Schwarz Bayesian IC (g=ln(n))
- HQC – Hannan and Quin IC (g=2lnln(n))

- Finite Prediction Error (FPE)

D. Hendry (*Economica, 47*, 1980): Econometrics-Alchemy or Science?

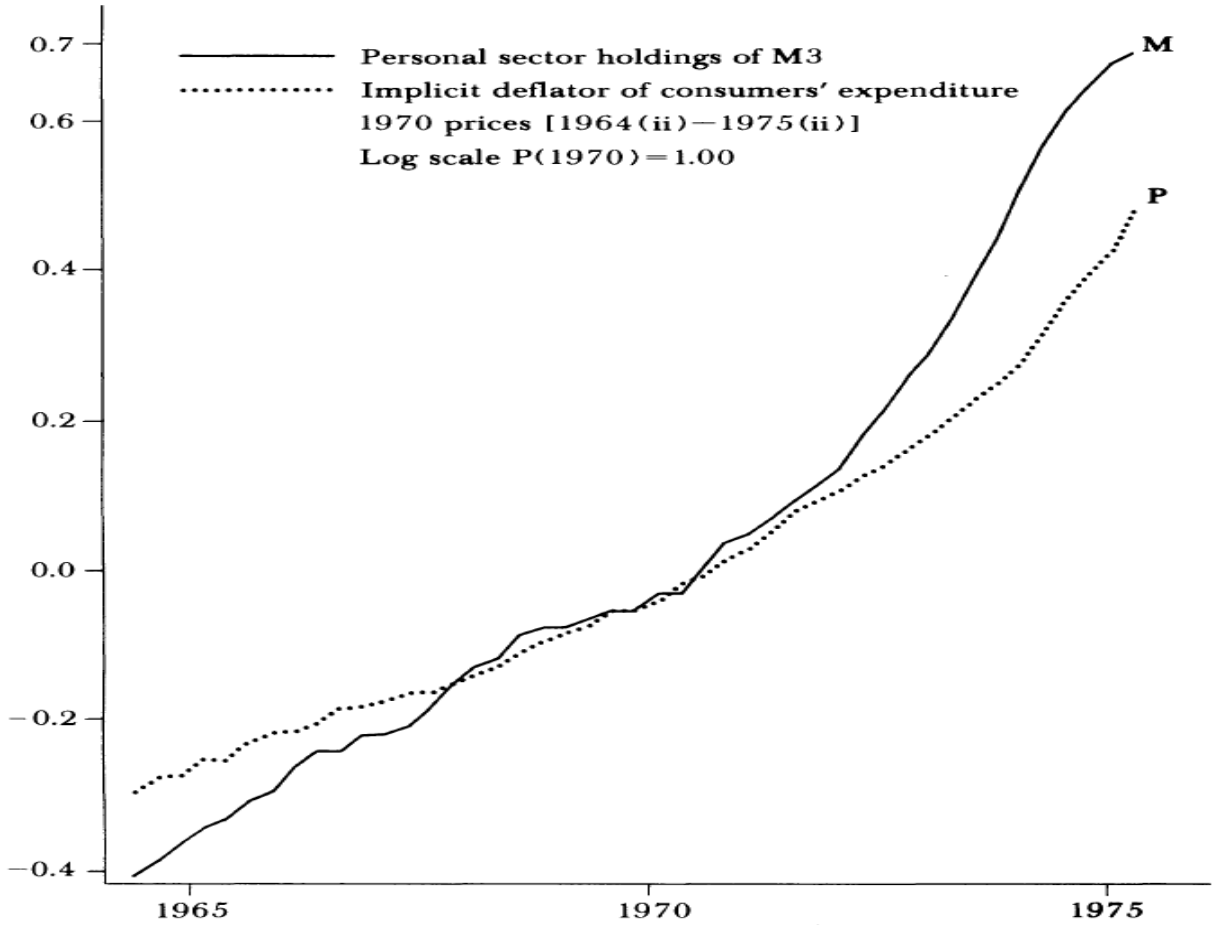


FIGURE 1

Nonsense regression

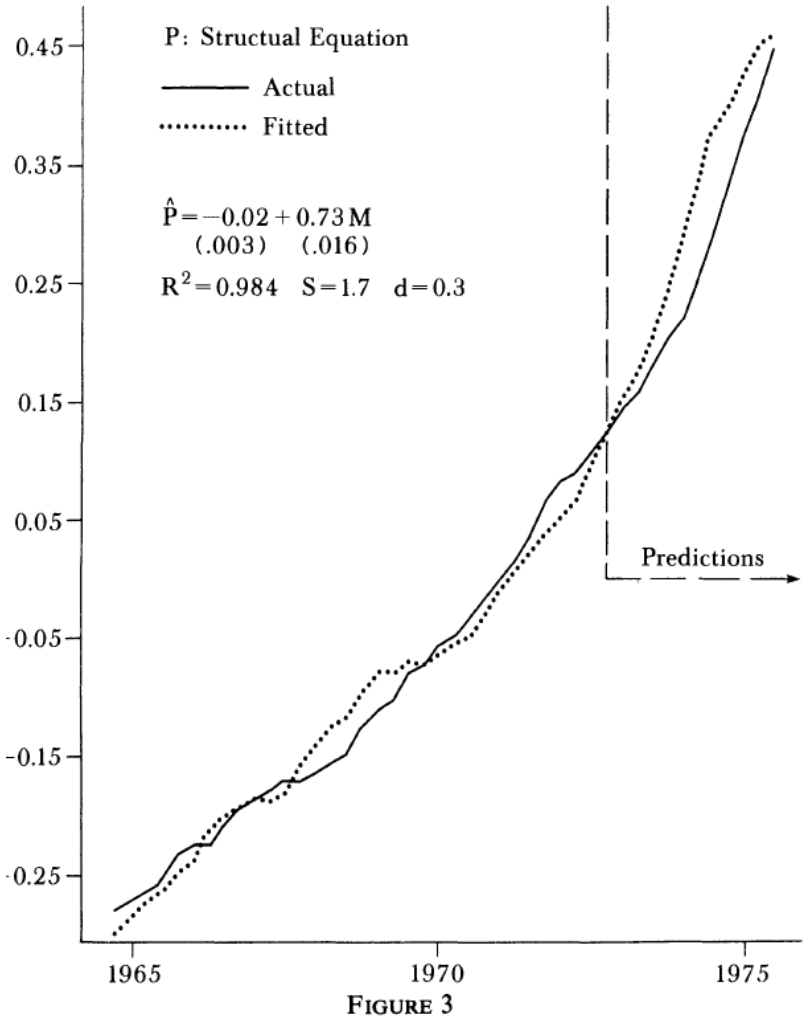


FIGURE 3

Nonsense regression (II)

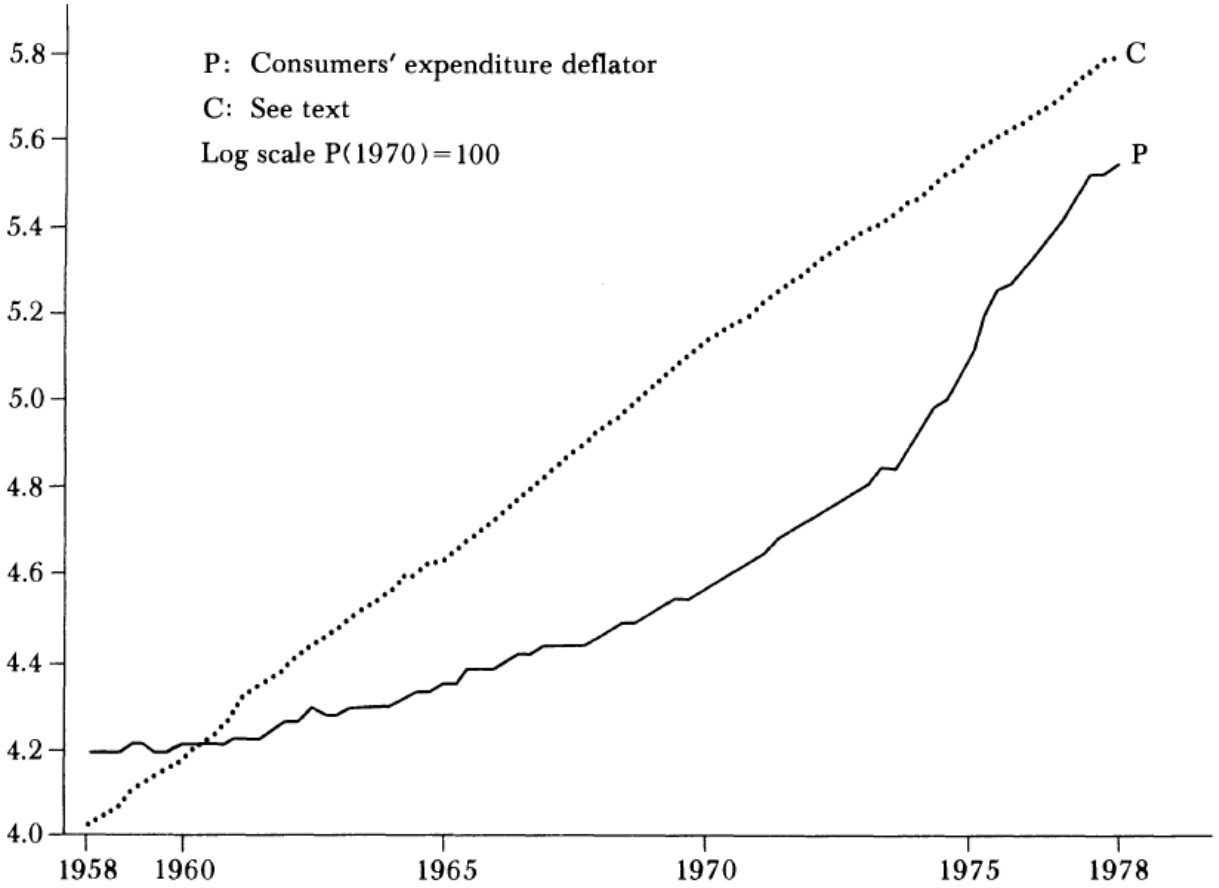


FIGURE 5

Nonsense regression(III);
C is simply cumulative rainfall in the UK!

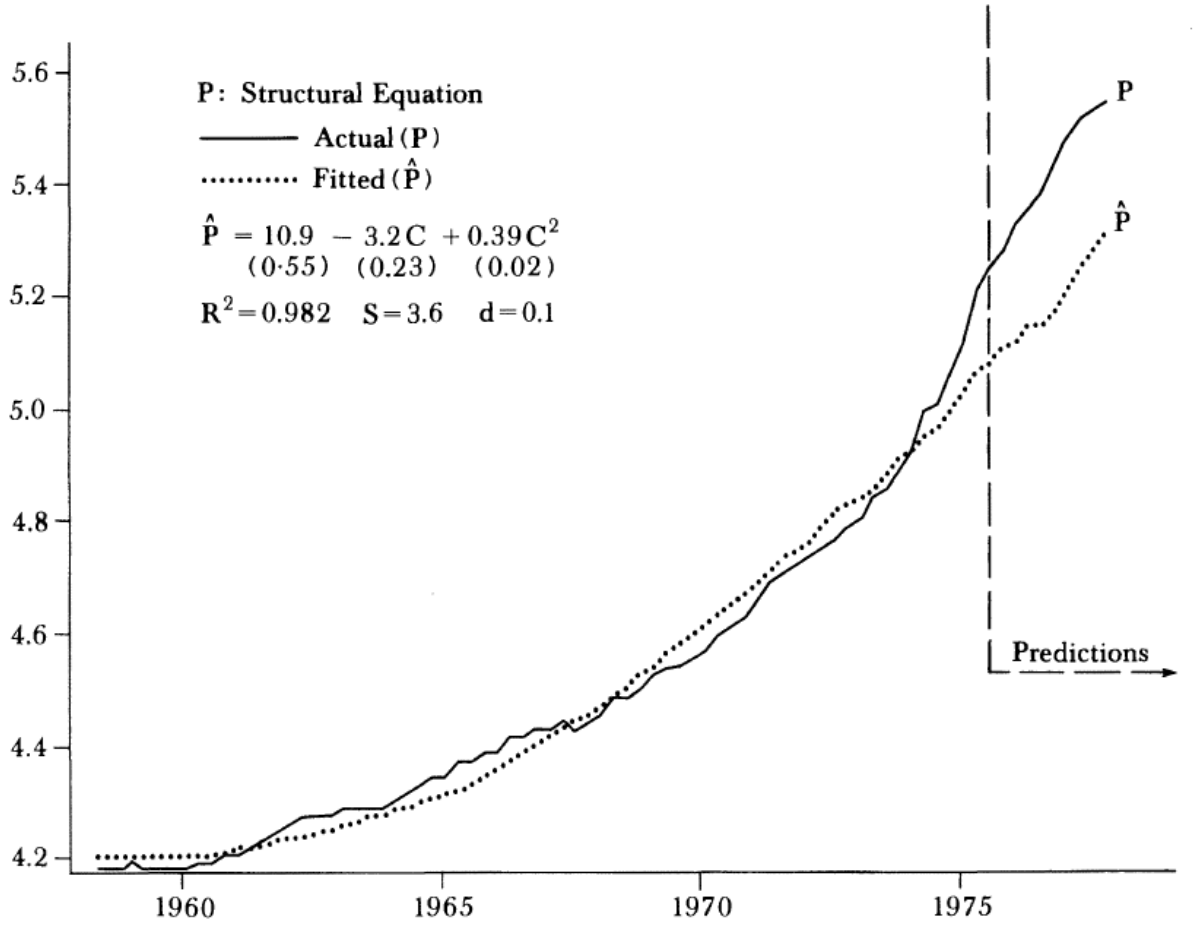


FIGURE 7

Table of OLS estimates wage equation (I)

Dependent Variable: LWAGE
Method: Least Squares
Date: 12/05/21 Time: 20:19
Sample: 1 526
Included observations: 526

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.813570	0.029814	60.83028	0.0000
FEMALE	-0.397217	0.043073	-9.221915	0.0000

R-squared	0.139635	Mean dependent var	1.623268
Adjusted R-squared	0.137993	S.D. dependent var	0.531538
S.E. of regression	0.493503	Akaike info criterion	1.429220
Sum squared resid	127.6177	Schwarz criterion	1.445438
Log likelihood	-373.8848	Hannan-Quinn criter.	1.435570
F-statistic	85.04372	Durbin-Watson stat	1.825492
Prob(F-statistic)	0.000000		

Table of OLS estimates wage equation (II)

Dependent Variable: LWAGE
Method: Least Squares
Date: 11/05/23 Time: 18:00
Sample: 1 526
Included observations: 526

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.826269	0.094054	8.785044	0.0000
EDUC	0.077203	0.007047	10.95525	0.0000
FEMALE	-0.360865	0.039024	-9.247156	0.0000

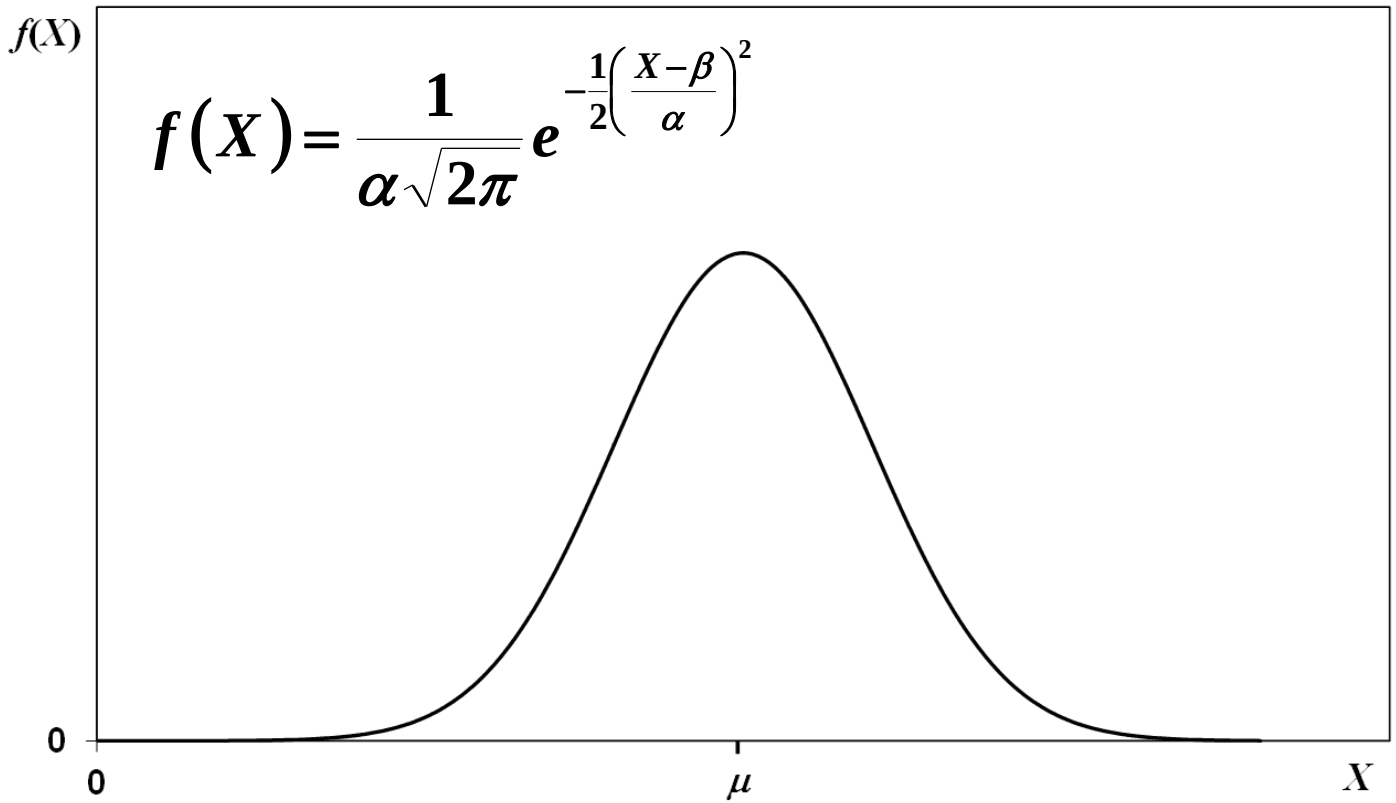
R-squared	0.300220	Mean dependent var	1.623268
Adjusted R-squared	0.297544	S.D. dependent var	0.531538
S.E. of regression	0.445496	Akaike info criterion	1.226432
Sum squared resid	103.7982	Schwarz criterion	1.250758
Log likelihood	-319.5515	Hannan-Quinn criter.	1.235957
F-statistic	112.1887	Durbin-Watson stat	1.813768
Prob(F-statistic)	0.000000		

Relevant theoretical distributions for our course

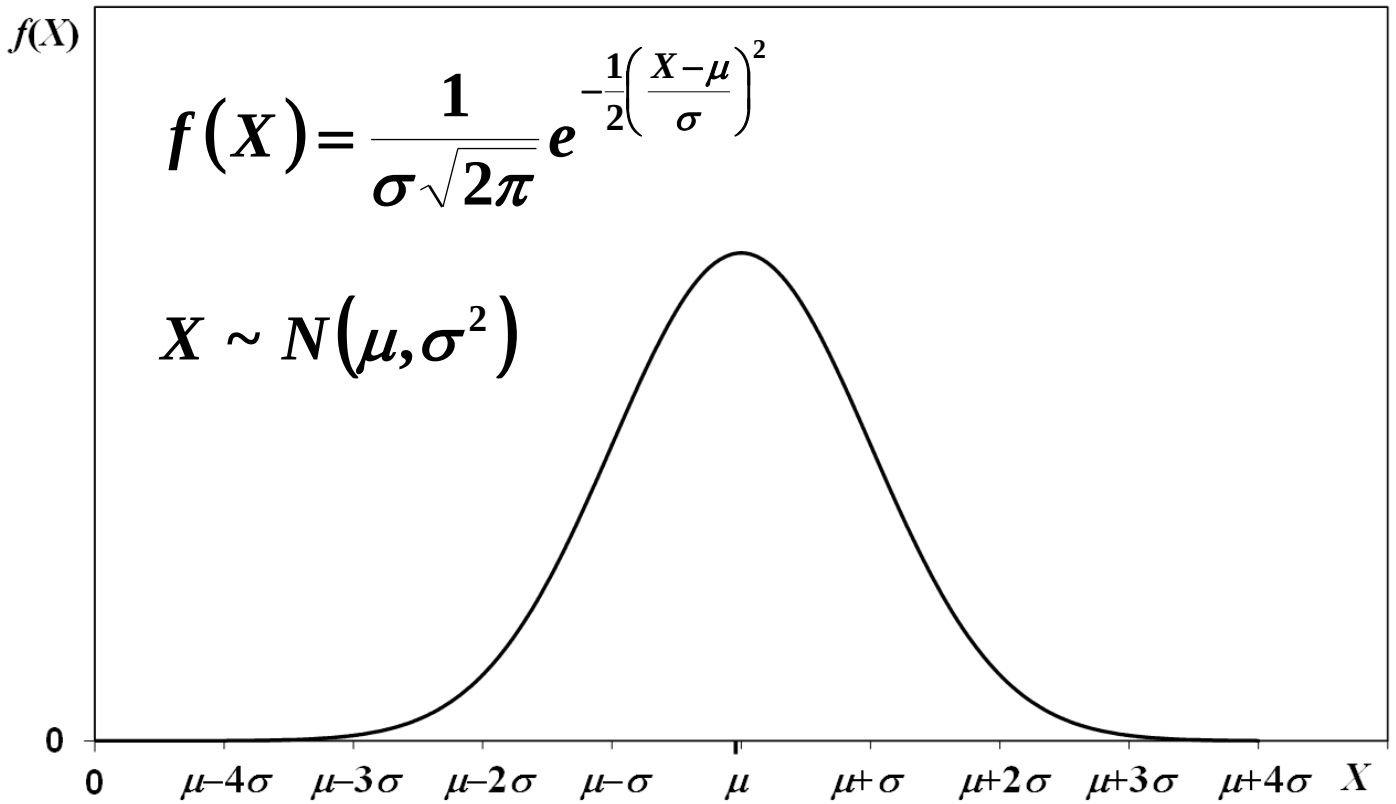
- There are **only four distributions**, all of them continuous, that are going to be of importance to us:

1) Normal distribution
2) *t*-distribution
3) *F* - distribution
4) Chi-squared (χ^2) distribution

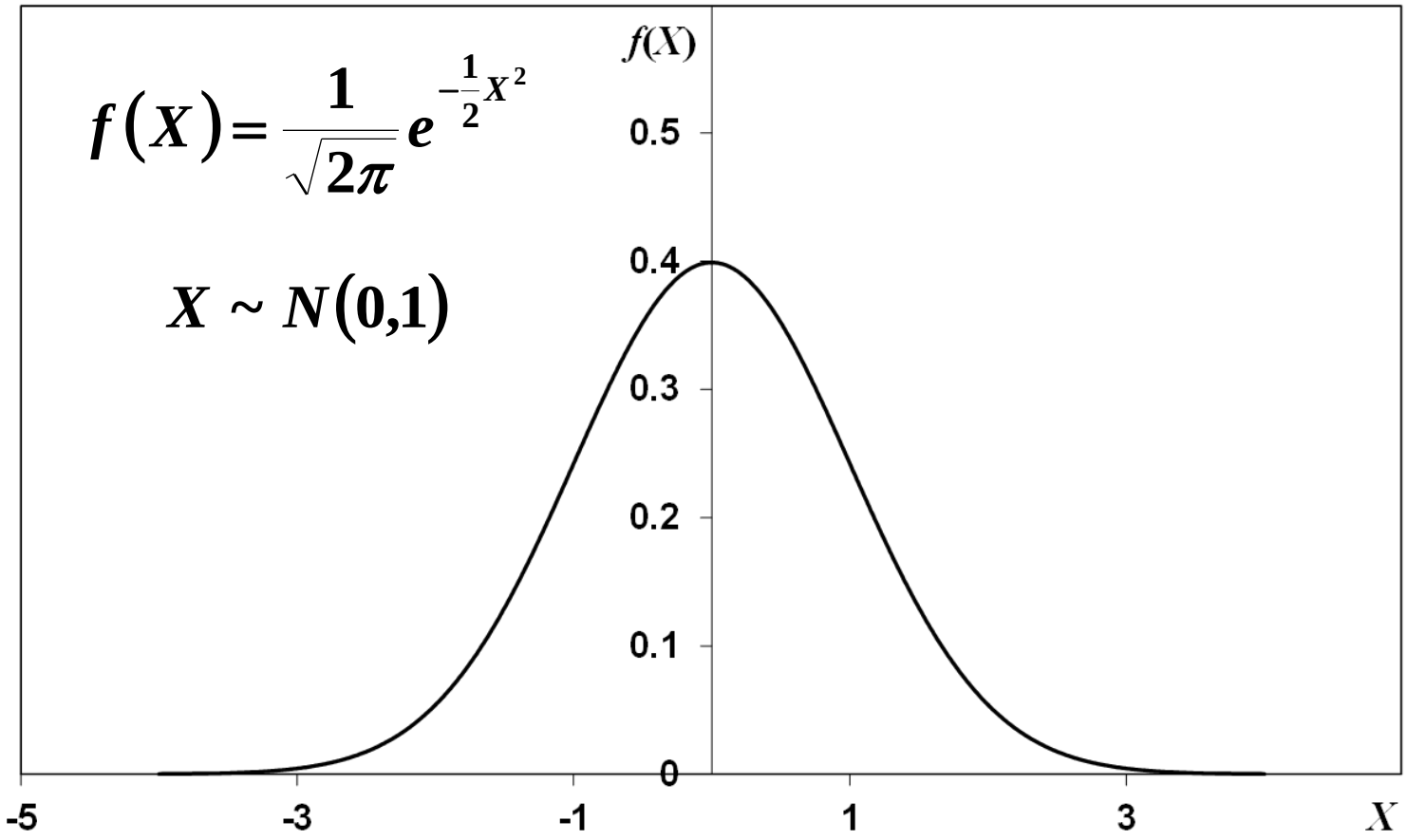
Normal distribution



Normal distribution



Standard normal distribution



An important special case is the standard normal distribution, where μ = 0 and σ = 1. This is shown in the figure.

χ^2-distribution

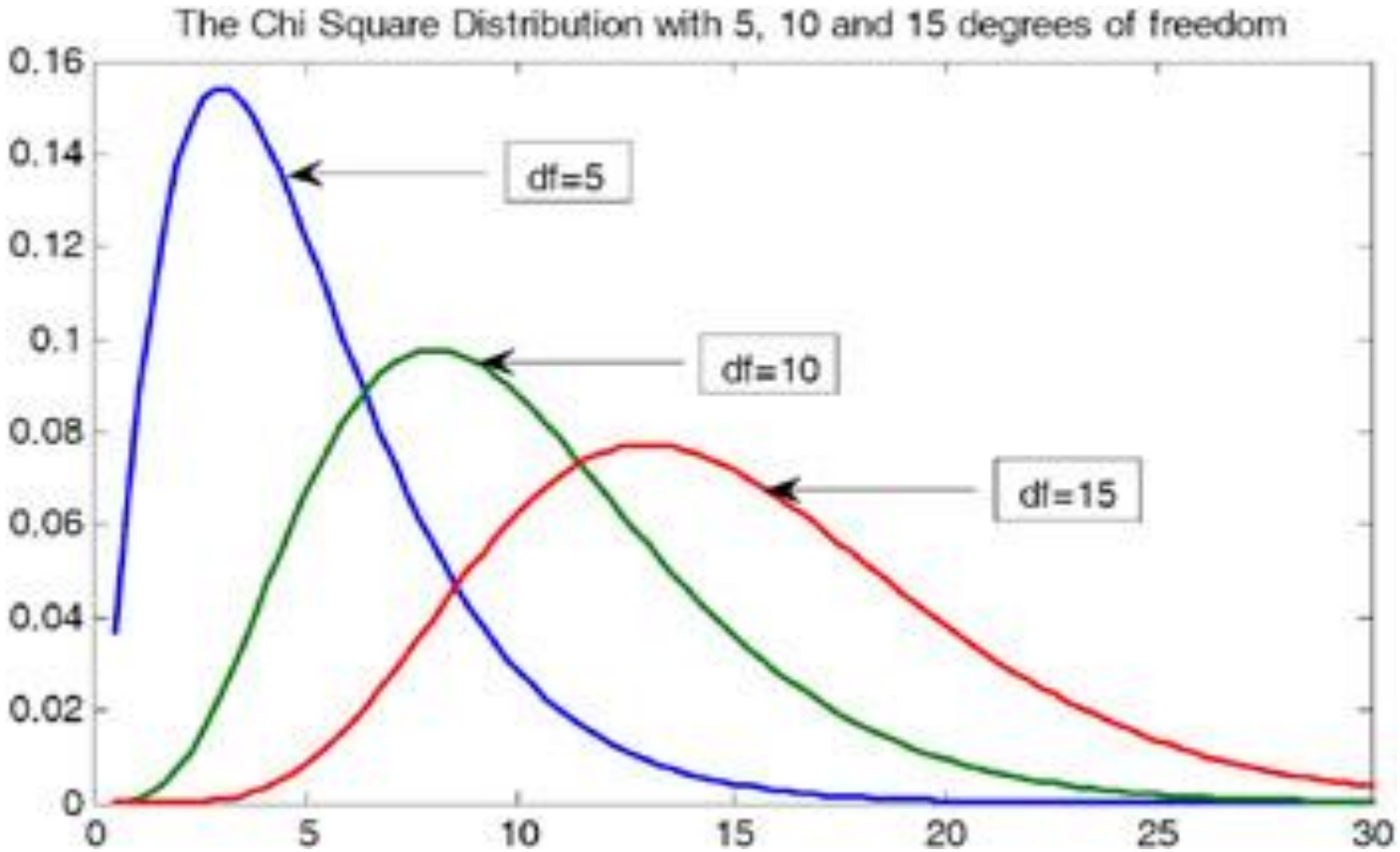
- First, we define **Chi-squared distribution** as follows. If Z_1, Z_2, …., Z_n is a set of independent standard normal variables, it hold that:

$$Z_1^2+Z_2^2+\ldots+Z_n^2=\sum_{i=1}^{n} Z_i^2,$$

has a Chi-squared distribution with n degrees of freedom.

- We denote $\sum_{i=1}^{n} Z_i^2 \sim\chi_n^2$, and distribution is with expected value equal to n and variance equal to 2n

Chi-square (χ^2) Distribution



Student *t* - distribution

- Next, we consider ***t* - distribution** (or Student distribution).

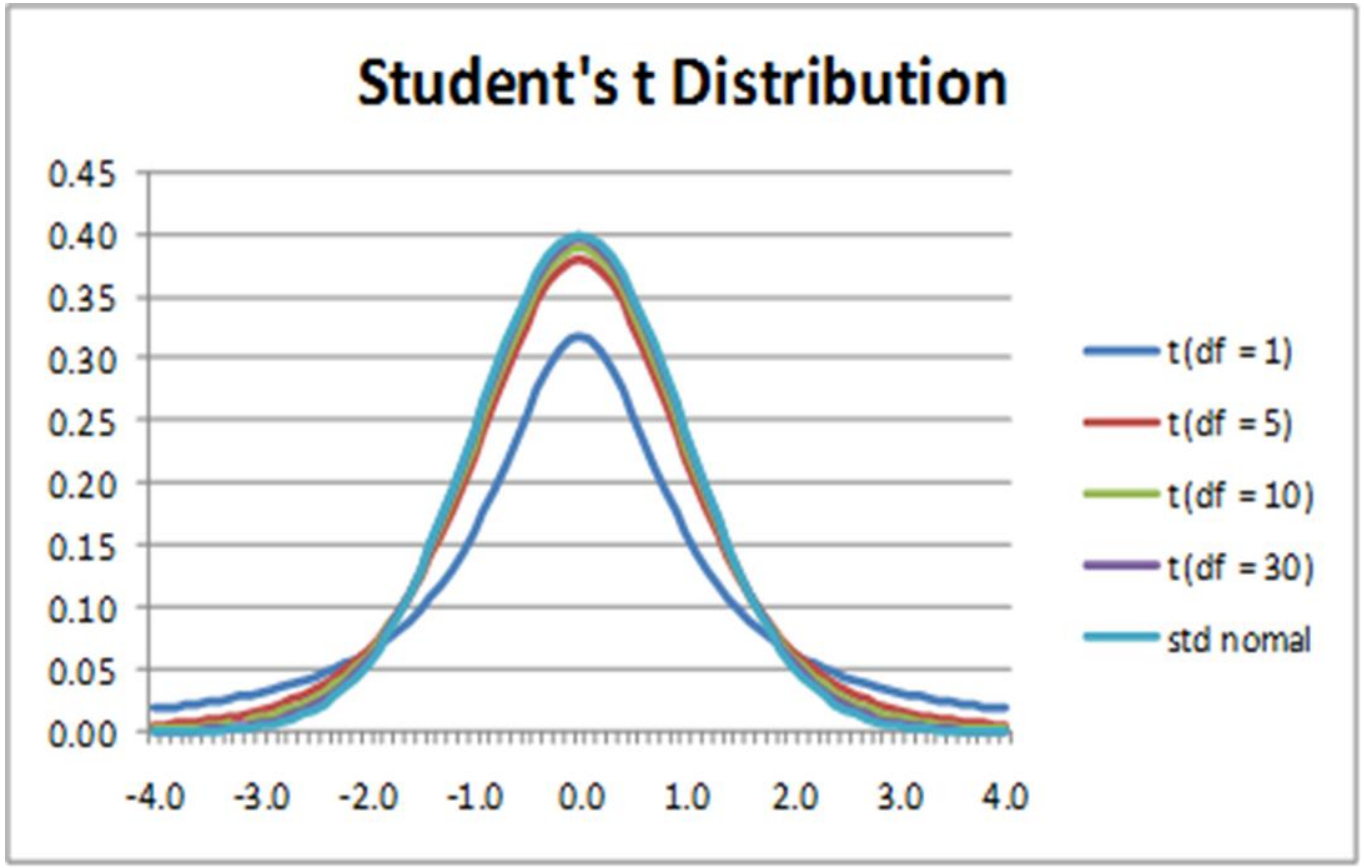
- If Z_0 has a standard normal distribution, $Z_0 \sim N(0, 1)$, and $\sum_{i=1}^{n} Z_i^2 \sim \chi_n^2$, and the distributions are independent, the ratio:

$$t = \frac{Z_0}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} Z_i^2}},$$

has a t- distribution with n degrees of freedom. Like standard normal, t distribution is symmetric around zero , but has fatter tails, particularly for small n.

- t-distribution has expected value of 0, and variance of n/n-2.

t - distribution



F – distribution

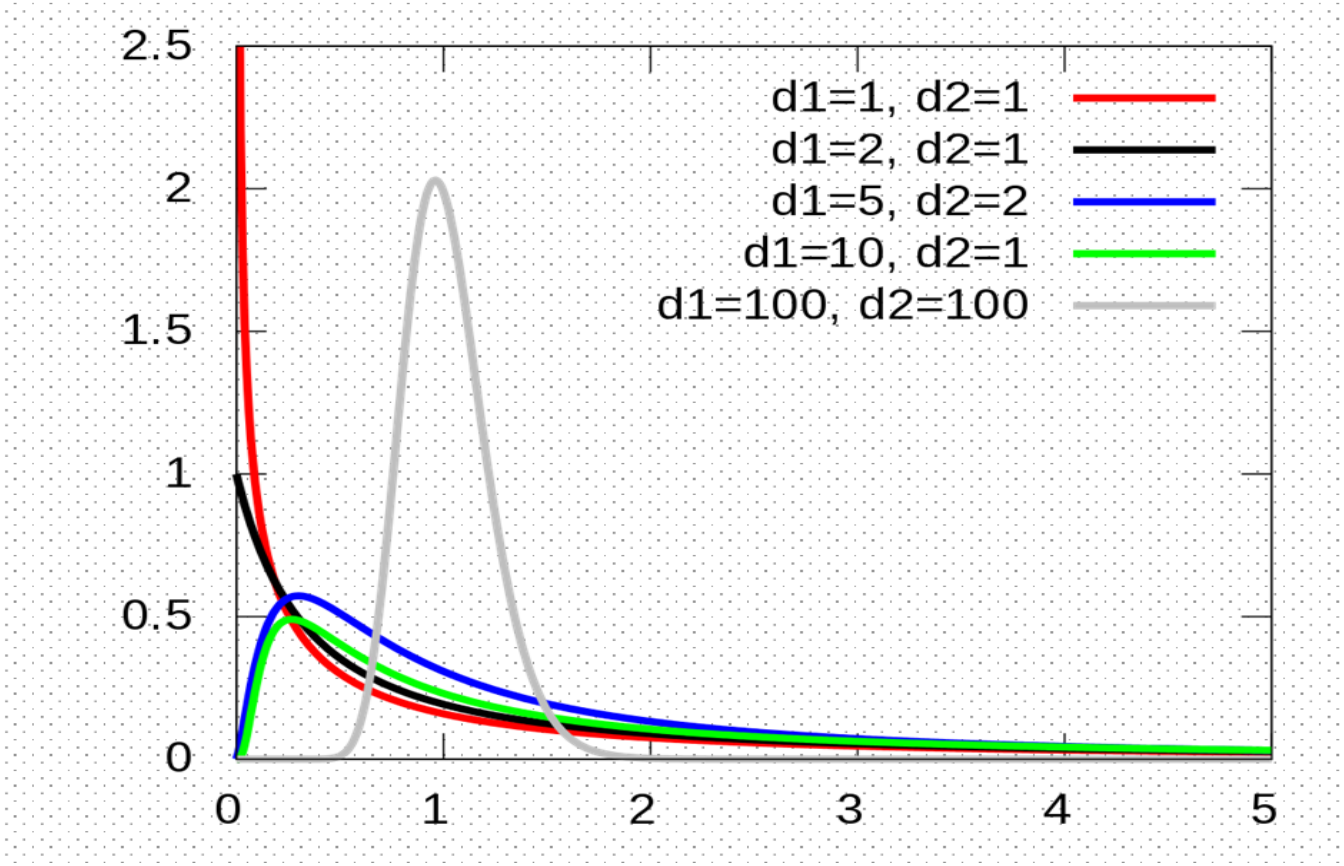
- If U $\sim \chi^2_m$ and V$\sim\chi^2_n$, and U and V are independent, it follows that ratio:

$$F = \frac{U/m}{V/n}$$

has **F distribution** with *m* (d_1) and *n* (d_2) degrees of freedom in the numerator and denominator respectively (we denote F_n^m).

- The F distribution is thus the ratio of two independent Chi-squared distributed variables, divided by their respective degrees of freedom.

F - distribution



Tests based on the OLS estimator

- Often, economic theory implies certain restrictions upon our coefficients. For example, $\beta_k = 0$
- We can check whether our estimates deviate "significantly" from these restrictions by means of a statistical test
- If they do, we will reject the null hypothesis that these restrictions are true
- To perform a test, we need **a *test statistic***. A test statistic is something we can compute from our sample and has a known distribution if the null hypothesis is true

Tests involving one parameter

- The most common test is the ***t*-test.** It can be used to test a single restriction

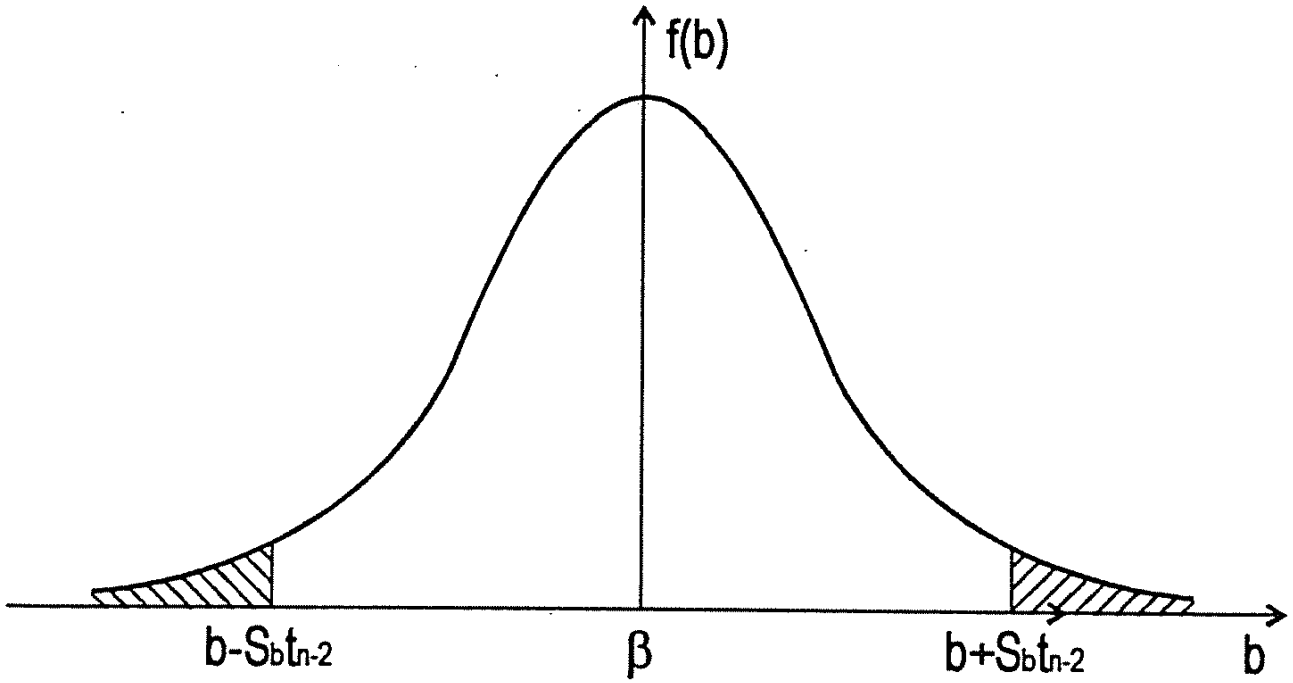
- Suppose the null hypothesis is $\beta_k = q$ for some given value q.

- Consider **the test statistic:** $t = (b_k - q) / se(b_k)$.

- If the null hypothesis is true, and under the Gauss-Markov assumptions (A1)-(A4) + normality (A5), *t* has a *t*-distribution with N-K degrees of freedom

- We will reject the null hypothesis if the absolute value of *t* is "too large"

Confidence interval for β



Tests involving one parameter (II)

- We consider values "too large" if they are *unlikely* to come from a *t*-distribution

- If we want to test with 95% confidence, we reject the null hypothesis if the absolute value of ***t* is larger than (approximately) 2**

- The ratio $t = b_k / se(b_k)$ is the ***t-value*** (or *t*-ratio) and is routinely supplied by any regression package

- It can be used to test the hypothesis that the true coefficient β_k is equal to 0

Tests involving one parameter (III)

- If assumption **(A5) does not hold**, but the other assumptions (A1)-(A4) hold, **the *t*-distribution only holds approximately**

- We can also state that under (A1)-(A4), it holds that

$$t = (b_k - q) / se(b_k)$$

(under the null hypothesis that $\beta_k = q$) has *approximately* a standard normal distribution, denoted N(0,1)

- The approximation error becomes smaller if the sample size N becomes larger. We refer to this as **asymptotic theory** as N goes to infinity ($N \rightarrow \infty$)

Tests involving more parameters

- Suppose we want to test whether J coefficients are *jointly* equal to zero.

- The easiest way to obtain a test statistic for this **is to estimate the model twice**:

1) once without the restrictions,

2) once with the restrictions imposed, i.e., with omitting the corresponding x variables.

- Let the R^2s of the two models be given by R^2_1 and R^2_0 , respectively. Note that $R^2_1 \geq R^2_0$

Tests involving more parameters (II)

- The restrictions are unlikely to be valid if the difference between the two R^2s is "large"

- A test statistic can be computed as:

$$F = \frac{(R_1^2 - R_0^2)/J}{(1 - R_1^2)/(N - K)}$$

- Under the null hypothesis (and assumptions (A1)-(A5)), *F* has an *F*-distribution with *J* and *N*-*K* degrees of freedom

- We reject if *F* is too large

- For example, with *N*-*K*=60 and *J*=3, we reject if *F* > 2.76 (95% confidence)

Tests involving more parameters (III)

- Suppose that relevant hypothesis are:

$$H_0: R^2 = 0 \Leftrightarrow H_0: \beta_2 = \ldots = \beta_K = 0$$

$$H_1: H_0 \; is \; not \; true \Leftrightarrow H_1: R^2 \neq 0$$

- Consider **the test statistic:**

$$F_{N-K}^{K-1} = \frac{R^2/(K-1)}{(1-R^2)/(N-K)}$$

- We will reject the null hypothesis If the value of *F* is "too large" ($F > F_{N-K}^{K-1}$; 95% confidence)

Table of OLS estimates wage equation (I)

Dependent Variable: LWAGE
Method: Least Squares
Date: 12/05/21 Time: 20:19
Sample: 1 526
Included observations: 526

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.813570	0.029814	60.83028	0.0000
FEMALE	-0.397217	0.043073	-9.221915	0.0000

R-squared	0.139635	Mean dependent var	1.623268
Adjusted R-squared	0.137993	S.D. dependent var	0.531538
S.E. of regression	0.493503	Akaike info criterion	1.429220
Sum squared resid	127.6177	Schwarz criterion	1.445438
Log likelihood	-373.8848	Hannan-Quinn criter.	1.435570
F-statistic	85.04372	Durbin-Watson stat	1.825492
Prob(F-statistic)	0.000000		

Do females earn less than males?

- We would like to test the null hypothesis H_0: $\beta_2=0$
- Our test statistic is:

$$t_2 = (b_2 - 0)/se(b_2) = -0.3972/0.043 = |-9.22|$$

- Since this is much larger than 2, we reject the null hypothesis that the average wage rate (in the population) is identical for males and females
- Note that $R^2 = 0.1396$, so that the simple model explains about 14% of the differences in individual wages
- Data source: http://fmwww.bc.edu/ecp/data/wooldridge/datasets.list.html

Extending the model

- Why?
- Wage differentials between males and females may be explainable by other factors (e.g., education or experience).
- Consider the more general model:

$$wage_i = \beta_1 + \beta_2\, female_i + \beta_3\, educ_i + \beta_4\, exper_i + \varepsilon_i$$

- Now, β_2 measures the difference in expected wage between a male and a female *with the same years of schooling (educ) and experience*
- The latter statement is a *ceteris paribus condition*

Table of OLS estimates wage equation (III)

Dependent Variable: LWAGE
Method: Least Squares
Date: 12/05/21 Time: 21:44
Sample: 1 526
Included observations: 526

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.480836	0.105016	4.578678	0.0000
FEMALE	-0.343597	0.037667	-9.122002	0.0000
EDUC	0.091290	0.007123	12.81591	0.0000
EXPER	0.009414	0.001449	6.495556	0.0000

R-squared	0.352552	Mean dependent var	1.623268
Adjusted R-squared	0.348831	S.D. dependent var	0.531538
S.E. of regression	0.428925	Akaike info criterion	1.152506
Sum squared resid	96.03584	Schwarz criterion	1.184942
Log likelihood	-299.1092	Hannan-Quinn criter.	1.165206
F-statistic	94.74734	Durbin-Watson stat	1.786623
Prob(F-statistic)	0.000000		

The RESET test

- A simple test on the functional form of the model. It is based on a simply auxiliary regression
- RESET = regression equation specification error test (Ramsey, 1969).
- Construct the fitted value from the model and test whether nonlinear functions of it help explaining y_i. Auxiliary regression:

$$y_i = x_i'\beta + \alpha_2 \hat{y}_i^2 + \alpha_3 \hat{y}_i^3 + \cdots + \alpha_Q \hat{y}_i^Q + v_i$$

where $\hat{y}_i = x_i'b$ (fitted value). Often Q=2

RESET test = *F*-test on Q-1 restrictions (α's are 0)

(Note: *auxiliary regression* is for testing purposes only.)

The RESET test- wage equation (II)

Unrestricted Test Equation:
Dependent Variable: LWAGE
Method: Least Squares
Date: 12/07/21 Time: 10:58
Sample: 1 526
Included observations: 526

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.794642	0.168049	4.728641	0.0000
EDUC	-0.258297	0.135101	-1.911882	0.0564
EXPER	-0.026999	0.014126	-1.911335	0.0565
FEMALE	0.998531	0.519108	1.923550	0.0550
FITTED^2	2.128346	1.002903	2.122186	0.0343
FITTED^3	-0.370887	0.219477	-1.689868	0.0917

R-squared	0.370324	Mean dependent var	1.623268
Adjusted R-squared	0.364270	S.D. dependent var	0.531538
S.E. of regression	0.423810	Akaike info criterion	1.132277
Sum squared resid	93.39963	Schwarz criterion	1.180931
Log likelihood	-291.7888	Hannan-Quinn criter.	1.151327
F-statistic	61.16441	Durbin-Watson stat	1.790640
Prob(F-statistic)	0.000000		

Table of OLS estimates wage equation (IV)

Dependent Variable: LWAGE
Method: Least Squares
Date: 12/07/21 Time: 11:02
Sample: 1 526
Included observations: 526

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.390483	0.102210	3.820413	0.0001
EDUC	0.084136	0.006957	12.09407	0.0000
EXPER	0.038910	0.004824	8.066682	0.0000
FEMALE	-0.337187	0.036321	-9.283424	0.0000
EXPERSQ	-0.000686	0.000107	-6.388842	0.0000

R-squared	0.399590	Mean dependent var	1.623268
Adjusted R-squared	0.394981	S.D. dependent var	0.531538
S.E. of regression	0.413446	Akaike info criterion	1.080882
Sum squared resid	89.05862	Schwarz criterion	1.121427
Log likelihood	-279.2720	Hannan-Quinn criter.	1.096757
F-statistic	86.68521	Durbin-Watson stat	1.775544
Prob(F-statistic)	0.000000		

The RESET test- wage equation (III)

Unrestricted Test Equation:
Dependent Variable: LWAGE
Method: Least Squares
Date: 12/07/21 Time: 11:04
Sample: 1 526
Included observations: 526

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.057290	0.214927	4.919302	0.0000
EDUC	-0.179859	0.140267	-1.282264	0.2003
EXPER	-0.083726	0.065923	-1.270056	0.2046
FEMALE	0.744732	0.573537	1.298490	0.1947
EXPERSQ	0.001471	0.001165	1.261825	0.2076
FITTED^2	1.584461	1.093791	1.448596	0.1481
FITTED^3	-0.239763	0.230758	-1.039025	0.2993

R-squared	0.416802	Mean dependent var	1.623268
Adjusted R-squared	0.410060	S.D. dependent var	0.531538
S.E. of regression	0.408262	Akaike info criterion	1.059401
Sum squared resid	86.50561	Schwarz criterion	1.116164
Log likelihood	-271.6225	Hannan-Quinn criter.	1.081626
F-statistic	61.82012	Durbin-Watson stat	1.768058
Prob(F-statistic)	0.000000		

The RESET test

- Thus, we cannot reject the current specification. However, this does not necessarily mean that other variables are irrelevant (i.e., have no impact on the house prices)

- In fact, we **may want to include other characteristics** too

Statistical measures used to describe distribution: skewness and kurtosis

- Skewness is **a measure of the asymmetry of a distribution**. A distribution is asymmetrical when its left and right side are not mirror images. A distribution can have right (or positive), left (or negative), or zero skewness, α_3:N(0, 6/n)

- Kurtosis is **a measure of "taildness" of a probability distribution**. Whereas skewness differentiates extreme values in one versus the other tail, kurtosis measures **extreme values in either tail**, α_4:N(3, 24/n)

- Kurtosis is a measure of whether the data are **heavy-tailed (fat-tailed)** or **light-tailed (thin-tailed)** relative to a normal distribution

Testing for Normality: Jarque-Bera (JB) test

- Test statistic:

$$JB = z_3^2 + z_4^2 = \frac{T}{6}\left[\hat{\alpha}_3{}^2 + \frac{(\hat{\alpha}_4 - 3)^2}{4}\right] : \chi_2^2$$

- Relevant hypothesis are:

H_0: sample data matching a normal distribution $(\alpha_3 = 0 \ and \ \alpha_4 = 3).$

H_1: H_0 *is not true* $(\alpha_3 \neq 0 \ and/or \ \alpha_4 \neq 3).$

- We will reject the null hypothesis If the value of *JB* is “too large” (the chi-squared approximation for the JB statistic's distribution is only used for large sample sizes).
- *Note: In Verbeek textbook - Ch.6 pp. 202-203*

Ceteris Paribus in Wage example

- Log of wages:

  lwage = f(gender /female, education and experience)

- Returns to education:

β_3 = one additional year of schooling increase wage for 9.1% when **holding experience and gender constant**

- How do you do that in the real world?
  - The “percentage changes”
  - How to change years of schooling and hold two other explanatory variables constant?

Size, power and *p*-values

- **Type I error**: we reject the null hypothesis, while it is actually true

- The probability of **a type I error** (**the size α of the test**) is directly controllable by the researcher by choosing the **confidence level** (e.g., a confidence level of 95% corresponds with a size of 5%)

- **Type II error:** we do not reject the null hypothesis while it is false (alternative is true)

- The **reverse probability**, that is, the probability of rejecting the null when it is false, is known the **power of a test.** We would like the power of a test to be high

Size, power and *p*-values (II)

- By **reducing the size of a test** to e.g., 1%, **the probability of rejecting the null hypothesis will decrease**, even if it is false

- Thus, a lower probability of a type I error will imply a higher probability of a type II error (**There is a trade off between the two error types**)

- In general, larger samples imply better power properties

- Accordingly, in large samples we may prefer to work with a size of 1% rather than the "standard" 5%

Size, power and *p*-values (III)

- Note that we say:

"We reject the null hypothesis" (at the 95% confidence level) or

"We do not reject the null hypothesis"

- We typically **do not say**:

"We accept the null hypothesis"

- **Why?**

Two mutually exclusive hypotheses (e.g., $\beta_2 = 0$ and $\beta_2 = 0.01$) may not be rejected by the data, but it is silly to *accept* both hypotheses.(Sometimes, tests are just not very powerful)

p-values

- **Final probability that plays a role in statistical test**

- The *p*-value denotes the **marginal significance level** for which the null hypothesis is rejected

- If a *p*-value **is smaller than the size α** (e.g., 0.05) we reject the null hypothesis

- Many modern software packages provide *p*-values with their tests. This allows you to perform the test without checking tables of critical values. It also allows you to perform the test without understanding what is going on

- Note that a ***p*-value of 0.08** indicates that the **null is rejected at the 10% level** but not at the 5% level

Asymptotic properties of OLS

- If some of the assumptions (A1) to (A5) are violated, the properties of the OLS estimator may differ from those reported above
- In many cases, the exact properties are unknown, and we employ asymptotic theory
- *Asymptotic theory* refers to the question what happens if, hypothetically, the sample size grows infinitely large. In formula: $N \to \infty$
- We use this to *approximate* the properties of our estimator in a given sample (in reality, sample sizes rarely grow)

Asymptotic properties of OLS (II)

- Under assumptions (A1)-(A4) it holds that b is a **consistent estimator for β** or "b converges in probability to β":

$$\text{plim } b = \beta,$$

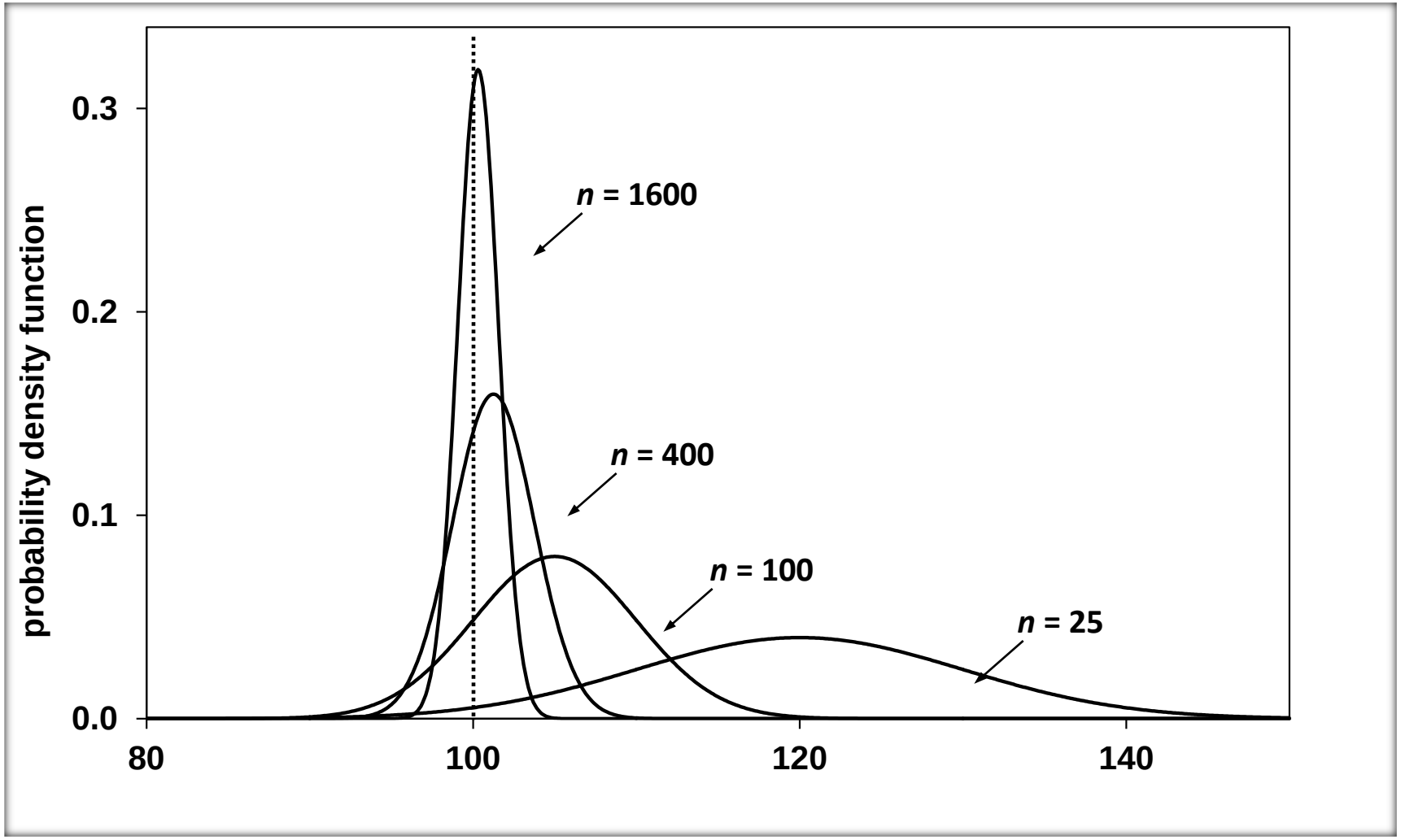
provided some regularity condition (A6) is satisfied (asymptotically there is no multicollinearity)

- This says that: if N grows, the probability that b differs from β **becomes smaller and smaller**

- Actually, **consistency of b already holds** if

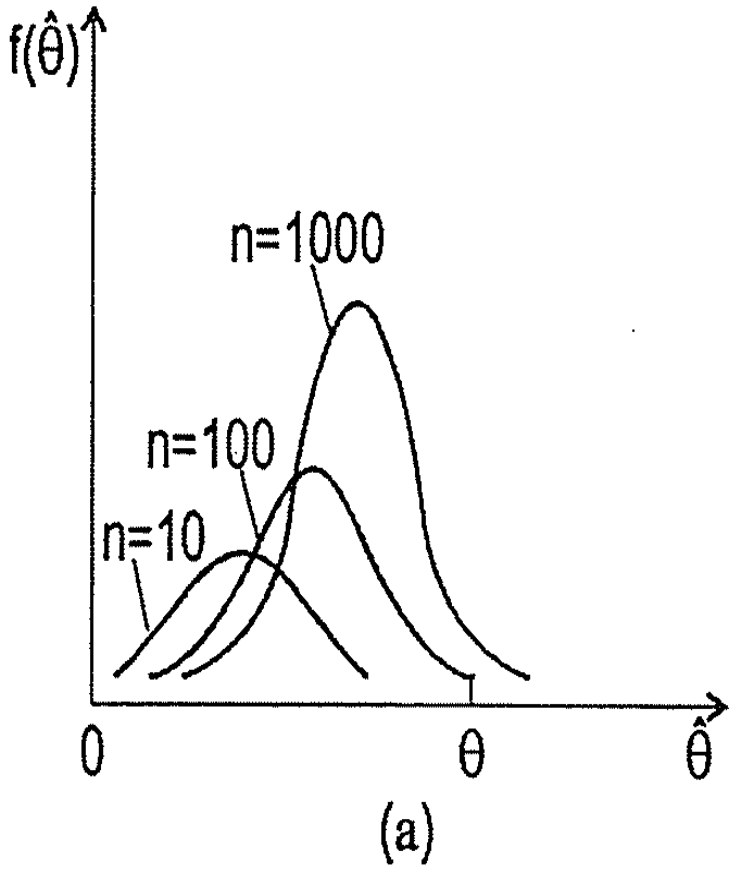
$$E\{\varepsilon_i \, x_i\}=0 \qquad \text{(A7)}$$

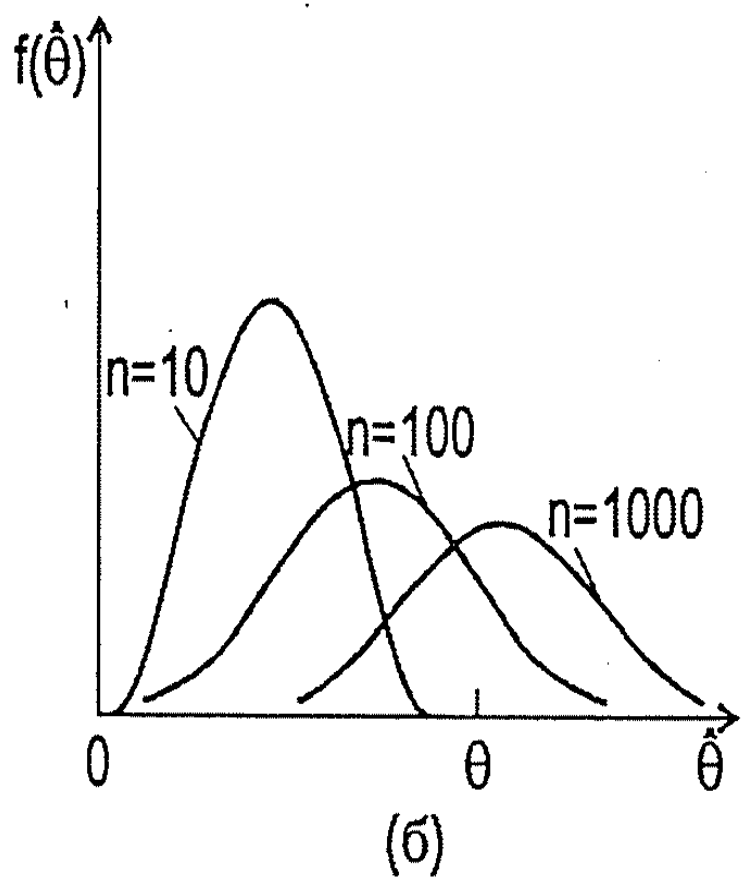
(no correlation between errors and regressors)



This is an example where the bias disappears altogether as the sample size tends to infinity. Estimator that is consistent despite being biased in finite sample.

Inconsistent estimator (plim$\left(\hat{\theta}\right) \neq \theta$)





Asymptotic properties of OLS (III)

- Further, under assumptions (A1)-(A4) it holds that b is asymptotically normal

- This means that in finite samples, b has approximately a normal distribution, where the **approximation is better if N is large**

- So, we have (approximately):

$$b \overset{a}{\sim} \mathcal{N}\left(\beta, s^2 \left(\sum_{i=1}^{N} x_i x_i'\right)^{-1}\right).$$

(Same result as before, except for “approximately”)

Asymptotic properties of OLS (IV)

- Fortunately, it is possible **to relax assumption (A2)** to x_i and ε_i **are independent** (A8), without affecting the distributional result

  (**does not rule out the dependence between x_i and ε_i for $i \neq j$**)

- Note that (A8) implies (A7): $E\{\varepsilon_i x_i\}=0$

- Thus under (A1), (A8), (A3) [homoskedasticity] and (A4) [no serial correlation], the OLS estimator is:

  - **Consistent;**
  - Asymptotically normal;
  - Routinely computed standard errors are (approximately) correct;

Data problems: Multicollinearity

- In general, there is nothing wrong with including variables in your model that are correlated, for example
  - experience and schooling,
  - age and experience,
  - inflation rate and nominal interest rate.

- However, when correlations are high, it becomes hard to identify the *individual* impact of each of the variables

- Multicollinearity is used to describe the situation when an **exact or approximate linear relationship exists** between the explanatory variables

Multicollinearity

- The signs of multicollinearity are:
- High standard errors (low t-values)
- Strange signs or magnitudes of coefficients
- Reasonable (or good) R^2 or F-statistics

- Multicollinearity has little impact on forecasts/fitted values.

- The variance of b_k is inflated if x_k can be approximated by the other explanatory variables; see

$$V\{b_k\} = \frac{\sigma^2}{1 - R_k^2} \frac{1}{N} \left[\frac{1}{N} \sum_{i=1}^{N} (x_{ik} - \bar{x}_k)^2 \right]^{-1}, \quad k = 2, \ldots, K$$

Exact multicollinearity

- **Exact multicollinearity** arises when an exact linear relationship exists between the explanatory variables. For example:

$$exper = age - school - 6$$

$$male = 1 - female$$

Note: Alternative parameterizations

- In case of exact multicollinearity, the OLS estimator *cannot* be computed. This is because the matrix $\Sigma_i x_i x_i'$ is not invertible

- The natural solution is to drop one explanatory variable (or more than one, if necessary). Some programs (e.g., Stata) do this automatically, other programs (e.g., Eviews) give an error message. [“near collinear matrix”]

Variance Inflation and Multicollinearity

- When variables are **highly but not perfectly correlated**, least squares is difficult to compute accurately (problem of that matrix (x'x) **is close** to being not invertible)
- Variances of least squares slopes become very large
- Variance inflation factors: For each x_k, $VIF(k) = 1/[1 - R^2(k)]$ where $R^2(k)$ is the R^2 in the regression of x_k on all the other x variables in the data matrix
- Values over 10(5) are considered as "high" (rule of thumb)

R_K^2 and corresponding values for VIF_K

R_K^2	0	0.5	0.8	0.9	0.95	0.975	0.99	0.995	0.999
VIFK	1	2	5	10	20	40	100	200	1000

Data problems: Outliers

- In calculating the OLS estimator, some observations may have a disproportional impact

- An **outlier** is an observation that deviates markedly from the rest of the sample. *Could be due to mistakes or problems in the data*

- An outlier becomes an influential observation if it has a substantial impact on the estimated regression line. (Large residuals are penalized more than proportionally.) See above Figure 2.3 ($\beta_1 = 3\ and\ \beta_2 = 1; outlier\colon\ x = 6, y = 0.5 \rightarrow$ slope coefficient drops from **0.94 to 0.52**, **R^2 from 0.94 to 0.18**)
- Approaches: investigate sensitivity of results, "test" for the presence of outliers, use robust estimation methods (e.g., LAD)

Impact of outliers

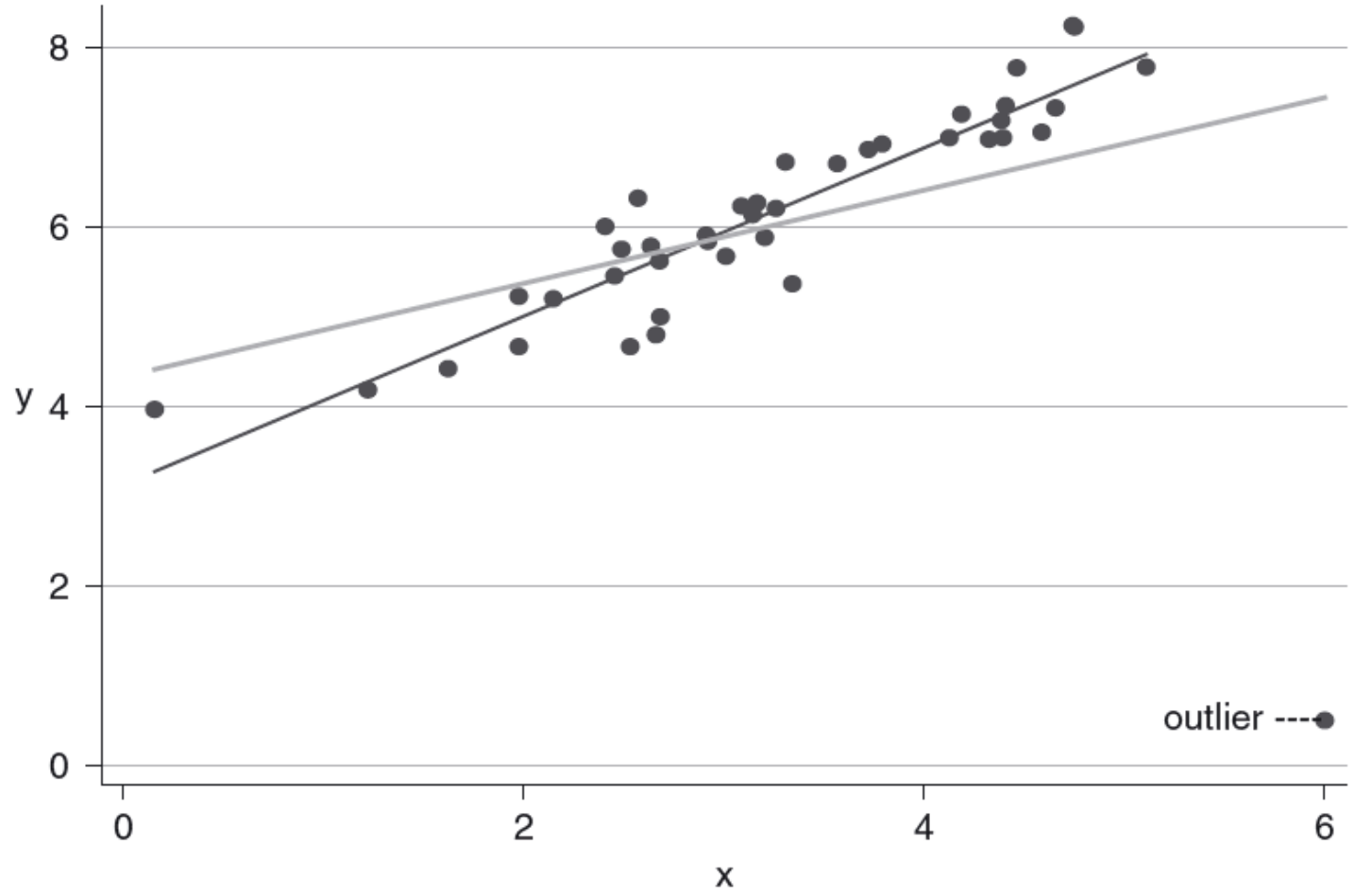


Figure 2.3 The impact of estimating with and without an outlying observation.

Least Absolute Deviations (LAD)

- Estimation method less sensitive to outliers
- $Minimize\ \sum_{i=1}^{N}|y_i - (\beta_1 + \beta_2 x_i)|$
- Solution is obtained by linear programming (there is no closed−form solution to minimizing above)
- Special case of a **so-cold quantile regressions** (available in EViews and Stata)

Least Squares vs. LAD