ECONOMETRICS

Lecturer in charge: Tran Thu Hien, Ph.D

Textbook: Introductory Econometrics: A Modern Approach (Wooldridge J.M., 2004)

EXPECTED OUTCOME

- Know how to set up econometric models corresponding to real world economic problems
- Know how to use Stata to estimate the models
- Know how to interpret the estimated results
- Know how to use the results for policy purposes

Course Contents

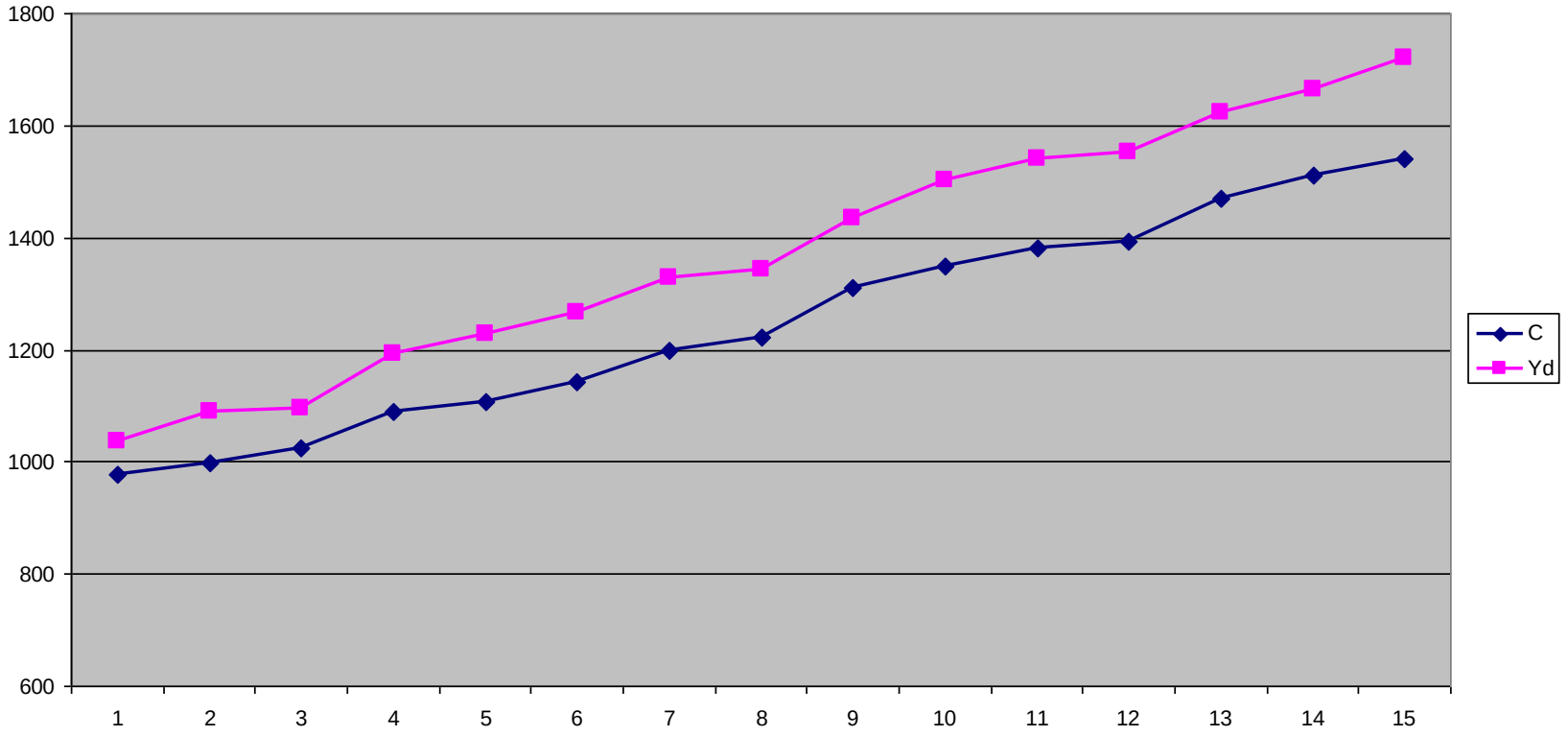
- The Nature of Econometrics and Economic Data
- The Simple Regression Model
- Multiple Regression Analysis: Estimation
- Multiple Regression Analysis: Inference
- Multiple Regression Analysis with Qualitative Information
- Specification and Data Problems: Cross-sectional Data
- Basic Regression Analysis with Time Series Data
- Basic Regression Analysis with Panel Data
- Application of Econometrics in Finance

THE NATURE OF ECONOMETRICS AND ECONOMIC DATA

OUTLINE

1. What is Econometrics?
2. Steps in Empirical Economic Analysis
3. Examples
4. Economic Data
5. Causality and the notion of “Ceteris Paribus”

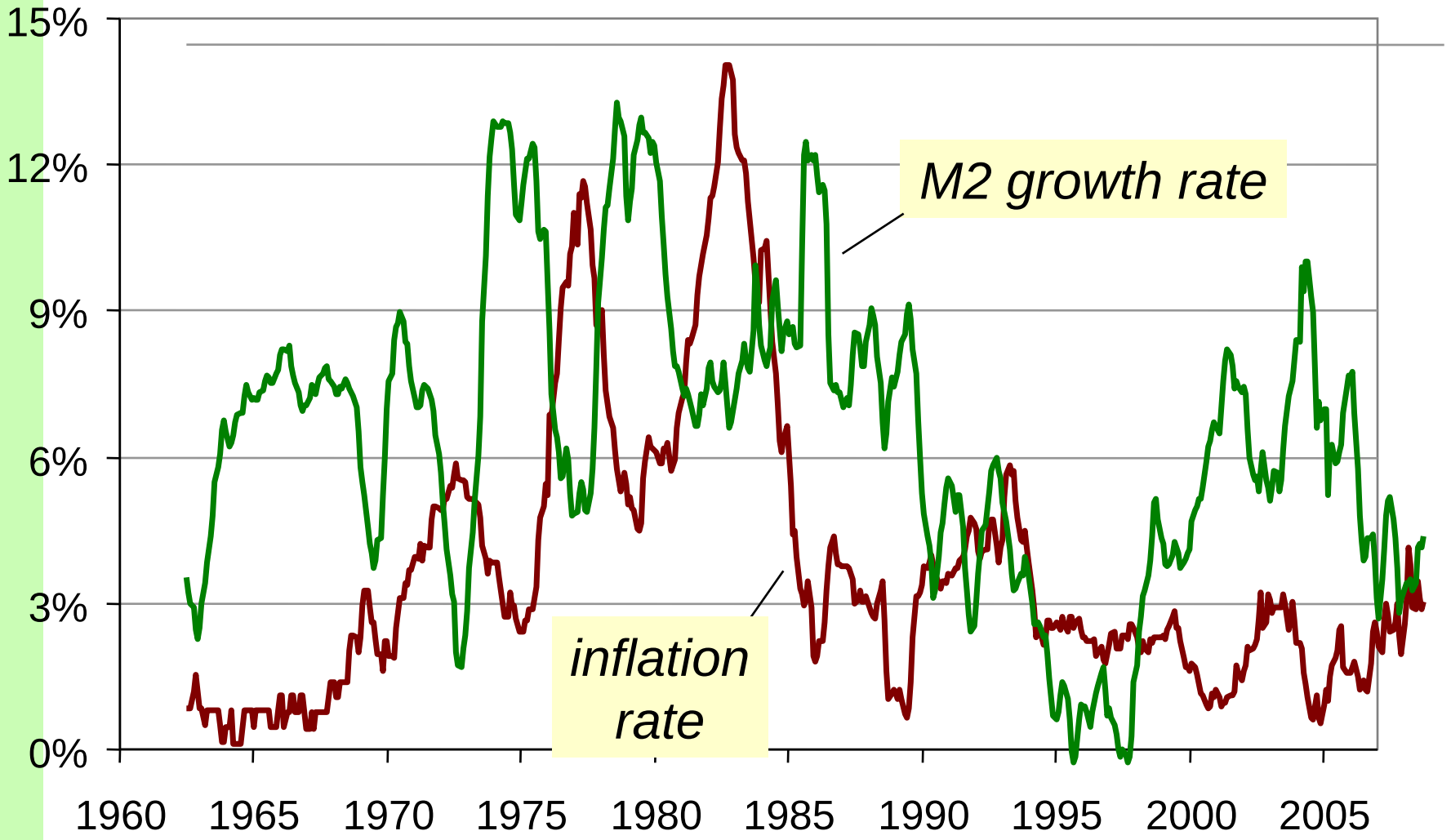
INTRODUCTION – look at the picture



INTRODUCTION

- Example1:
  - Statisticians say: income and expenditure go in the same direction
  - Economists say: an increase in income will raise expenditure, other things being equal
  - Econometricians say: 1usd increase in income will result in 0.70 cent increase in expenditure, other things being equal

INTRODUCTION – look at the picture



1. WHAT IS ECONOMETRICS?

- Combination of **statistical methods**, **economics** and **data** to answer empirical questions in economics.

- There are many different **types of empirical questions** in economics. Some examples:

- **Forecasting:**

  Use current and past economic data to predict future values of variables such as inflation, GDP, stock prices, etc.

- **Testing economic theories:**

  - Test of the Becker's economic model of criminal behavior
  - Test of the Capital Asset Pricing Model (CAPM)

1. WHAT IS ECONOMETRICS?

- **Estimation of economic relationships:**
  - Demand and supply equations;
  - Production functions;
  - Wage equations, etc.

- **Evaluating government policies:**
  - Employment effects of an increase in the minimum wage;
  - Effects of monetary policy on inflation.

- **Evaluating business policies:**
  - Estimate the impact of job training on worker productivity;
  - Compare profits under two pricing policies.

1. WHAT IS ECONOMETRICS?

- Econometrics is relevant in virtually **every branch of applied economics**: finance, labor, health, industrial, macro, development, international, trade, marketing, strategy, etc.

- There are **two important features which distinguish Econometrics from other applications of statistics**:

1. **Economic data is non-experimental data**. We cannot simply classify individuals or firms in an *experimental group* and a *control group*. Individuals are typically free to self-select themselves in a group (e.g., education, occupation, product market, etc).

2. **Economic models** (either simple or sophisticated) are key to interpret the statistical results in econometric applications.

2. STEPS IN EMPIRICAL ECONOMIC ANALYSIS

- The **research process** in applied econometrics **is not simply linear, but it has "loops".**

- Keeping this in mind, it is useful to describe the different steps of the research process in econometrics:

1. Formulation of the **question(s)** of interest.
2. Construction of an **economic model**
3. Specification of the **econometric model**
4. **Hypotheses** postulated
5. Collection of **data**
6. Estimation, validation, hypotheses testing, prediction.

3. EXAMPLE: Job Training and Worker Productivity

Step 1: Empirical question(s)

- Suppose that the government wants to evaluate the effectiveness of a publicly-funded job training program.
- Regulators : this is important for the decision to continue investments and organization of these the training program.
- Society: this is a solution to increase labor productivity, and subsequently economic development.
- Initial question: What is the impact of the job training program on worker productivity?

3. EXAMPLE: Job Training and Worker Productivity

Step 2: Economic Model

- Verify economic factors that affect worker productivity: education, experience, training, etc.

- Operationalize the variable of interest: Workers are paid commensurate with their productivity

$$wage = f(education, experience, training, ability)$$

3. EXAMPLE: Job Training and Worker Productivity

Step 3: Econometric Model

- An Econometric Model is an economic model where we take into account what is observable and not to the researcher.

- A researcher's decision of which economic model to estimate depends critically on what is observable.

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 training + \mu$$

wage: hourly wage
educ: years of formal education
exper: year of workforce experience
training: weeks spent in job training

- The β's are parameters to estimate.
- *u* represents unobservable inputs, e.g., ability.

3. EXAMPLE: Job Training and Worker Productivity

Step 4: Hypothesis

Other things kept constant, the workers who spent more weeks in job training have higher wage than those who spent less weeks in job training.

Step 5: Collection of data

- Data on hourly wage, years of education, years of experience, and weeks spent in job training of every worker in the sample are collected
- Data on control factors: family background, age, gender, etc.

3. EXAMPLE: Job Training and Worker Productivity

Step 4: Estimation, validation, hypotheses testing, prediction

- The parameters β are estimated by a relevant econometrics method. After estimation, we have to make specification tests in order to validate some of the specification assumptions that we have made for estimation.

- The results of these tests may imply a re-specification and re-estimation of the model.

- Once we have a validated model, we can interpret the results from an economic point of view, make tests, and predictions.

4. Economic Data

- Different types of datasets have their own issues, advantages and limitations.

- Some econometric methods may be valid for some types of data but not for others.

- We typically distinguish three types of datasets:
  1. Cross-Sectional Data
  2. Time Series Data
  3. Panel Data or Longitudinal Data

4. Economic Data

Cross-Sectional Data

- A cross-sectional dataset is a sample of individuals, or households, or firms, or cities, or states, or countries, …, taken at a given point in time.

- We often assume that these data have been obtained by **random sampling**.

- Sometimes we do not have a random sample: sample selection problem; spatial correlation.

4. Economic Data

- **Example of cross-section**

obsno	*wage*	*educ*	*exper*	*female*	*married*
1	3,10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
499	11.56	16	5	0	1
500	3.50	14	5	1	0

4. Economic Data

Time Series Data

- A time series dataset consists of observations on a variable or several variables over several periods of time (days, weeks, months, years).
- A key feature of time series data is that, typically, observations are correlated across time. We do not have a random sample.
- This time correlation introduces very important issues in the estimation and testing of econometric models using time series data.
- Seasonality is other common feature in many weekly, monthly or quarterly time series data.

4. Economic Data

- **Example of time series dataset:**

obsno	*year*	*month*	*exrate*	*irate*	
1	1990	1	1.32	7.35	
2	1990	2	1.30	7.30	
3	1990	3	1.29	7.32	
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	
191	2005	11	1.11	4.26	
192	2005	12	1.10	4.31	

4. Economic Data

Panel Data or Longitudinal Data

- In panel data we have a group of individuals (or households, firms, countries, …) who are observed at several points in time. **That is, we have time series data for each individual in the sample.**

- The key feature of panel data that distinguishes them from **pooled cross sections** is that the same individuals are followed over a given period of time.

- Using panel data we can control for time-invariant unobserved characteristics of individuals, firms, countries, …

4. Economic Data

- **Example of panel dataset: 150 cities over 2 years**

obsno	*city*	*Year*	*murders*	*population*	*police*
1	1	1999	5	350,000	440
2	1	2000	8	359,200	471
3	2	1999	2	64.300	75
4	2	2000	1	65,100	75
.	.	.	.	.	.
.	.	.	.	.	.
299	150	1999	25	543,000	520
300	150	2000	32	546,200	493

5. Causality and the notion of "Ceteris Paribus"

- Most empirical questions in economics are associated to the identification of **CAUSAL EFFECTS**.

- The notion of **ceteris paribus** (i.e., "other factors being equal") plays an important role in the analysis of causality.

- What we need to identify causal effects is **to hold constant all the relevant factors which are not independent of the causal variable under study.**

Basic Statistical Concepts: A Review

- E(X)
  - If X is discrete: $x_1,..,x_n$:

$$E(X) = \frac{x_1 + x_2 + .. + x_n}{n} := \mu$$

  - E(aX + bY) = aE(X) + bE(Y)
- Variance

$$var(X) = E[(X - E(X))^2]$$

$$var(X) = \frac{(x_1 - \mu)^2 + .. + (x_n - \mu)^2}{n}$$

- Standard deviation:

$$\sigma = \sqrt{var(X)}$$

- cov(X, Y)= E[(X-E(X))(Y-E(Y)]

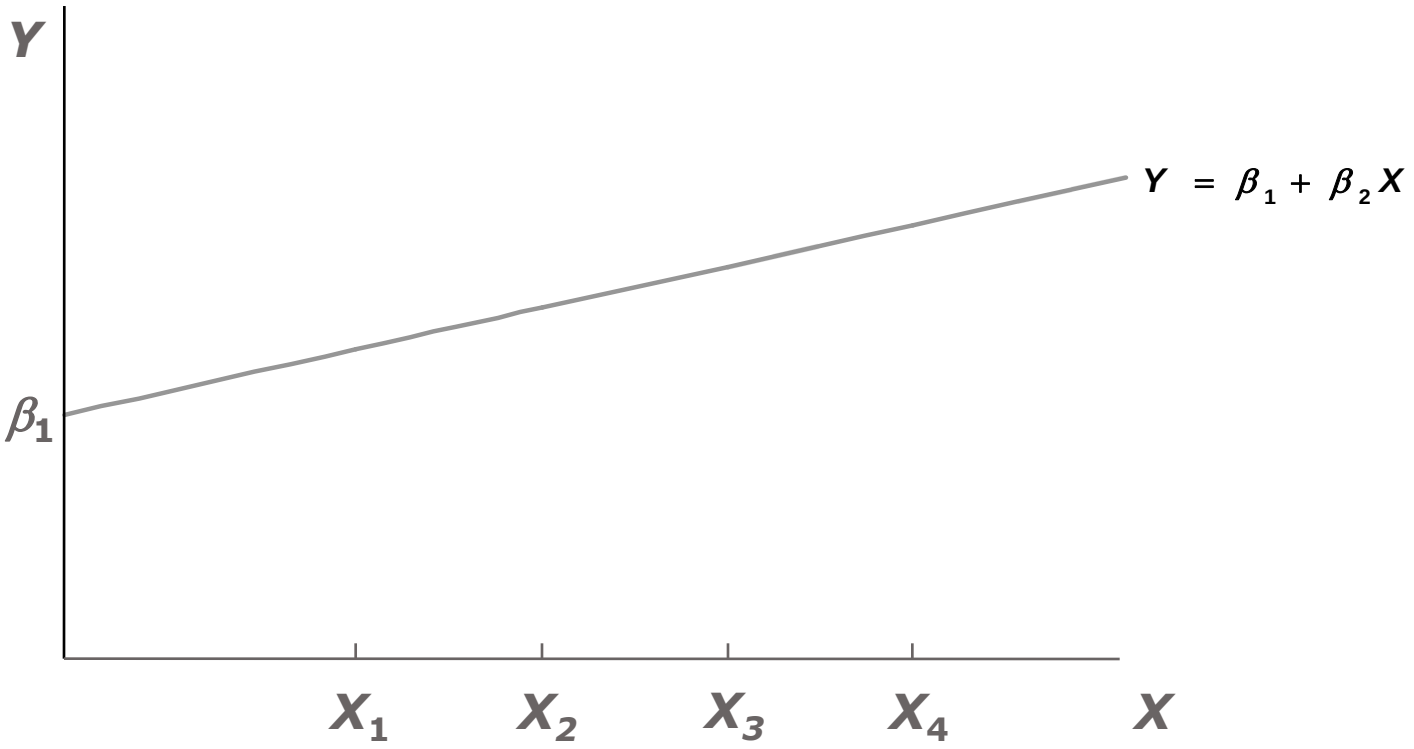
$$cov(X,Y) = \frac{(x_1 - \mu_X)(y_1 - \mu_Y) + .. + (x_n - \mu_X)(y_n - \mu_Y)}{n} = \frac{\sum x_i y_i}{n} - \mu_X \mu_Y$$

- Correlation coefficient:

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

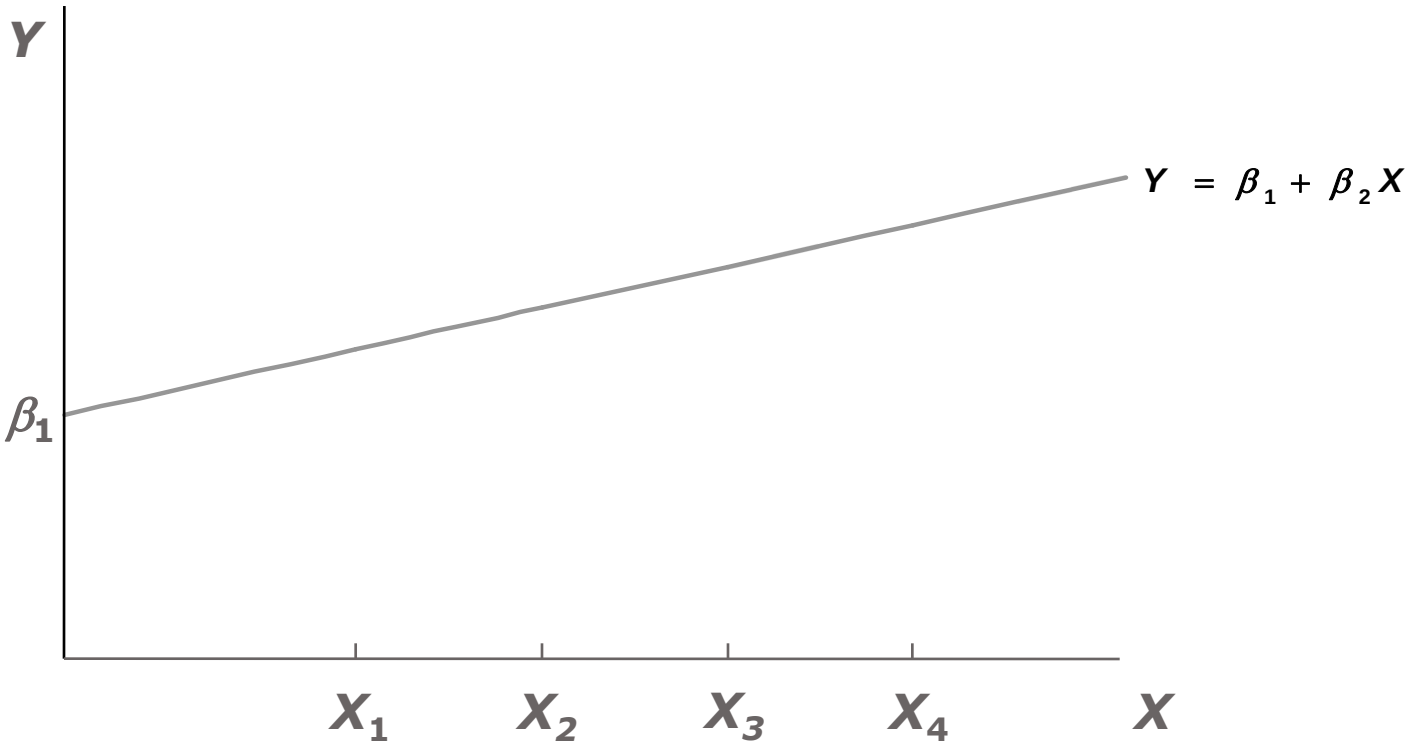
THE SIMPLE REGRESSION MODEL

SIMPLE LINEAR REGRESSION MODEL



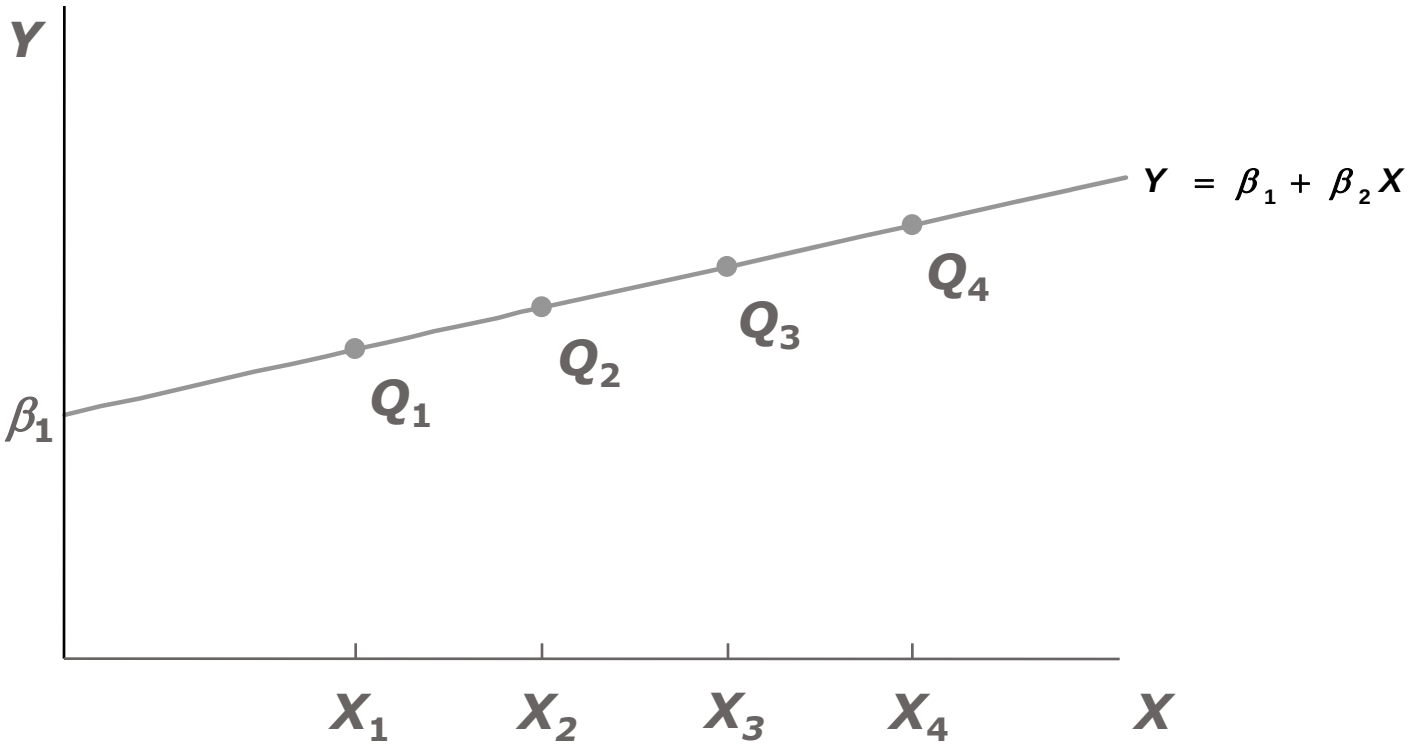
Suppose that a variable *Y* is a linear function of another variable *X*, with unknown parameters β_1 and β_2 that we wish to estimate.

SIMPLE LINEAR REGRESSION MODEL



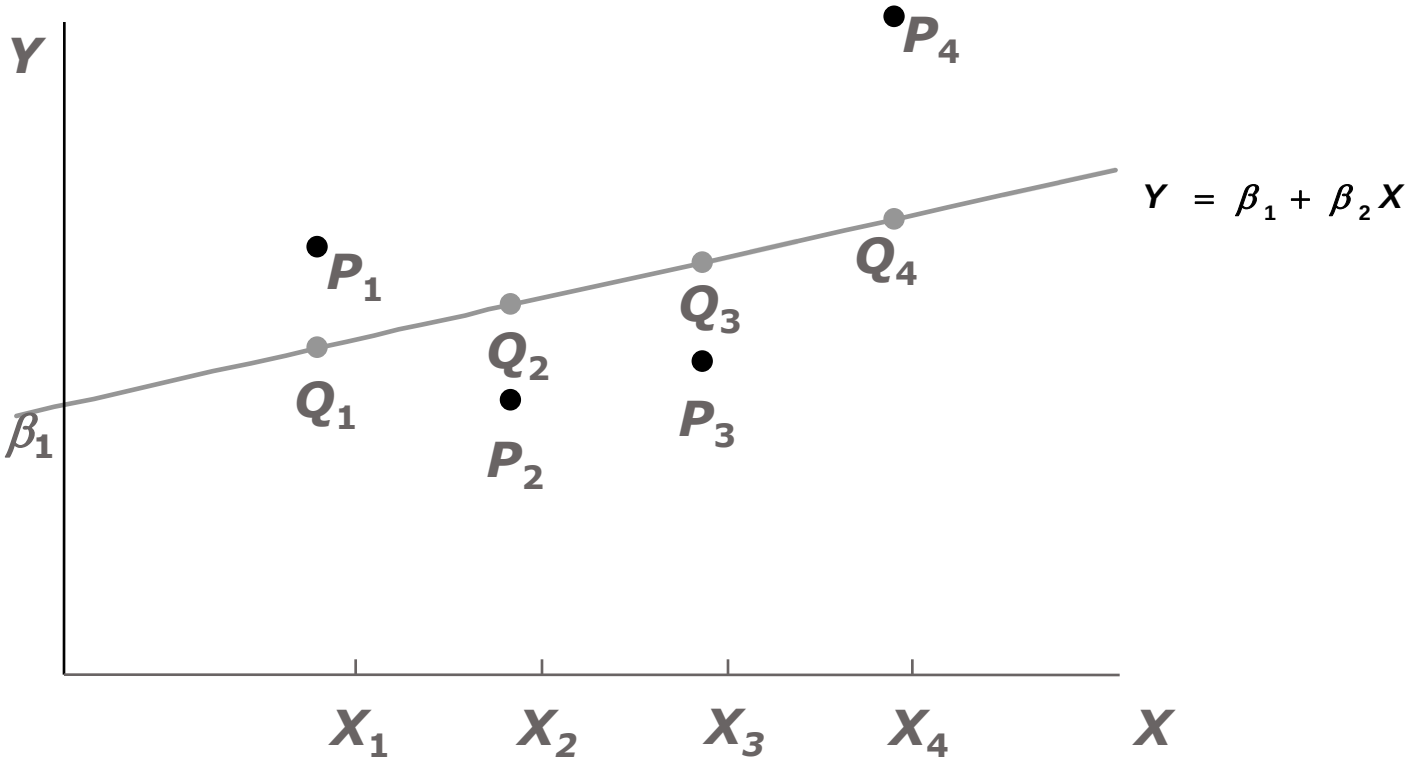
Suppose that we have a sample of 4 observations with *X* values as shown.

SIMPLE LINEAR REGRESSION MODEL



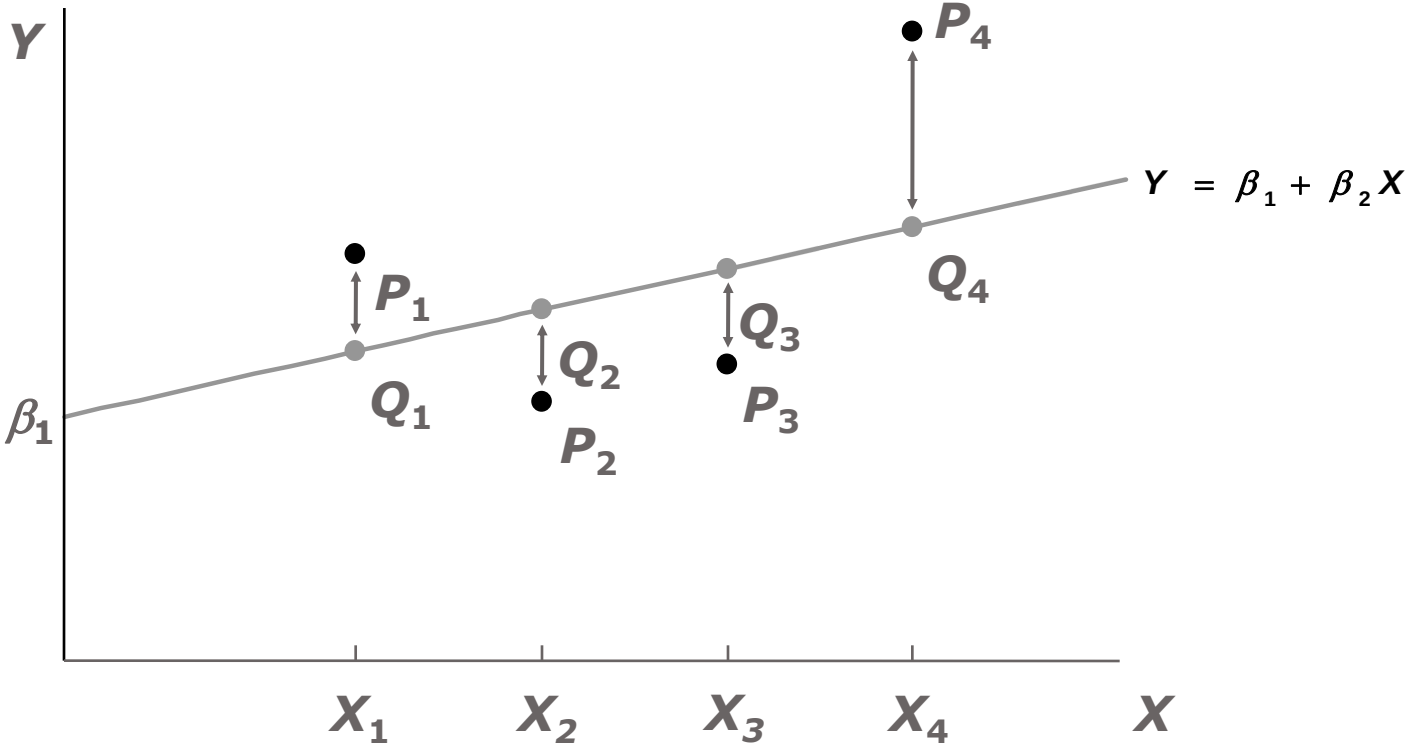
If the relationship were an exact one, the observations would lie on a straight line and we would have no trouble obtaining accurate estimates of β_1 and β_2.

SIMPLE LINEAR REGRESSION MODEL



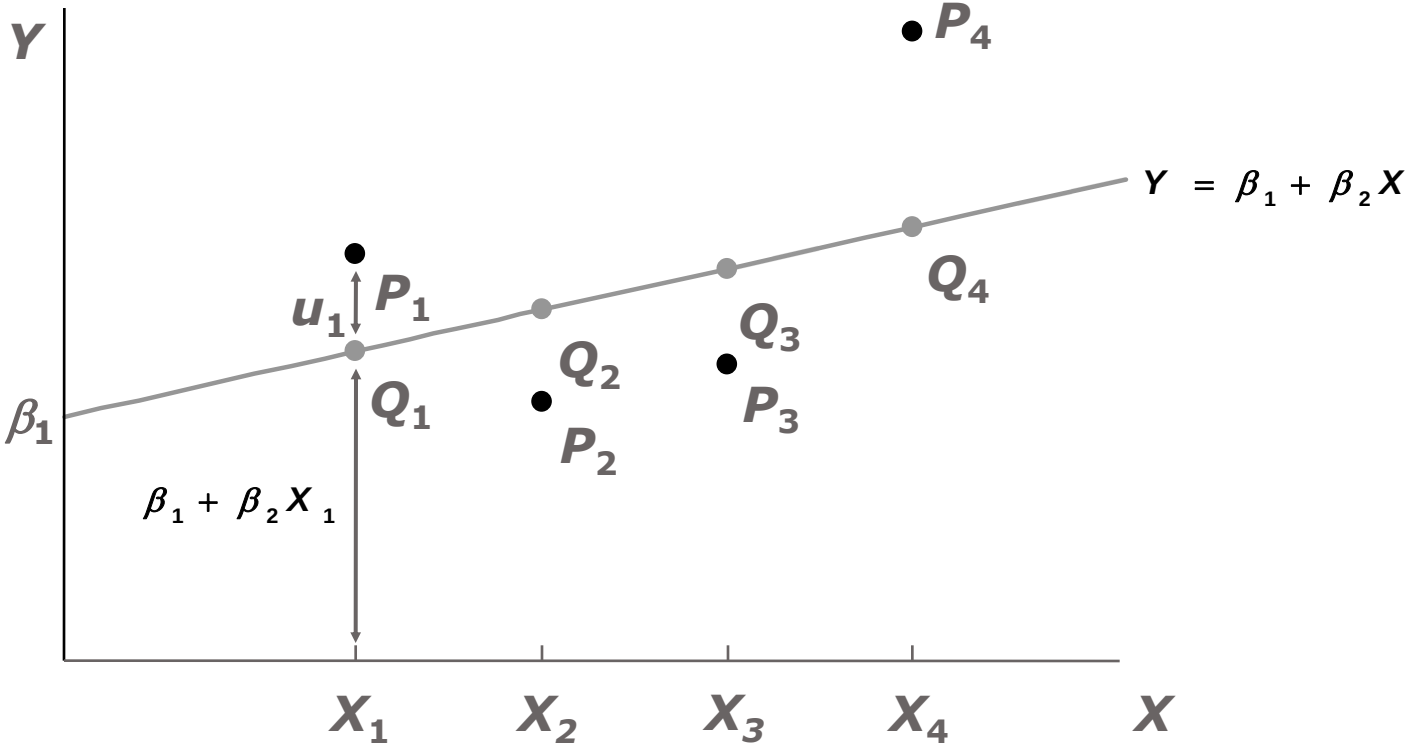
In practice, most economic relationships are not exact and the actual values of *Y* are different from those corresponding to the straight line.

SIMPLE LINEAR REGRESSION MODEL



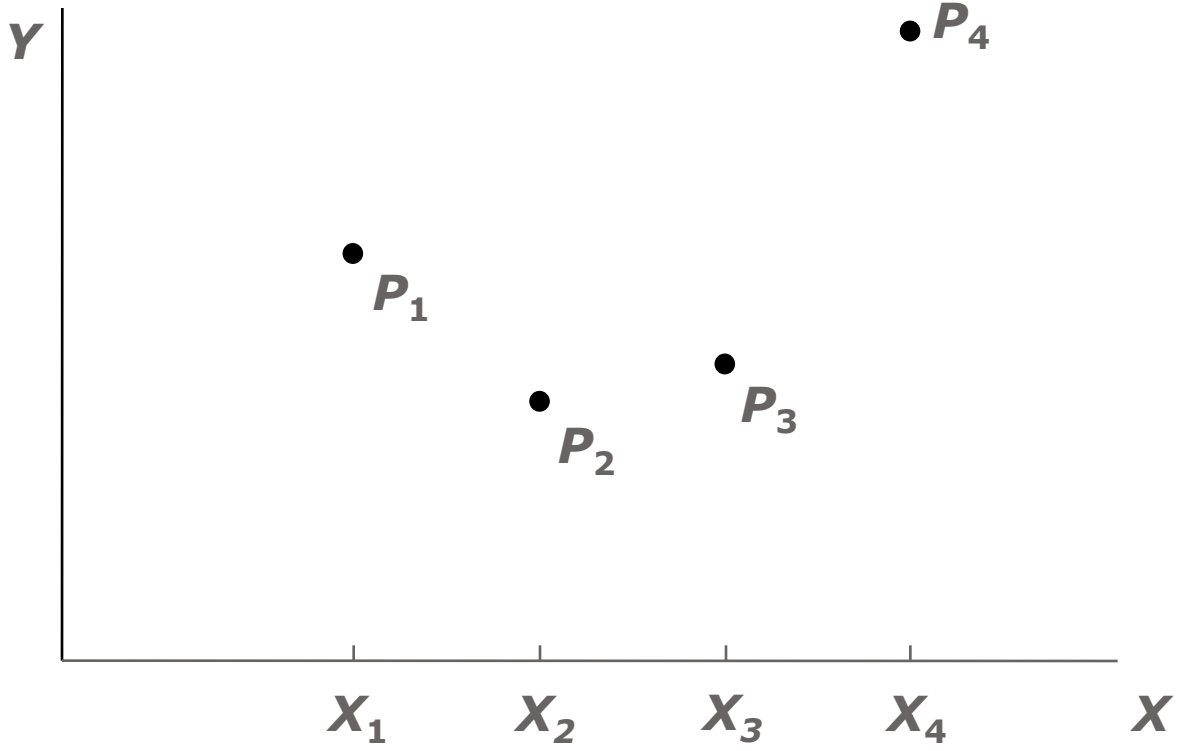
To allow for such divergences, we will write the model as $Y = \beta_1 + \beta_2 X + u$, where *u* is a disturbance term.

SIMPLE LINEAR REGRESSION MODEL



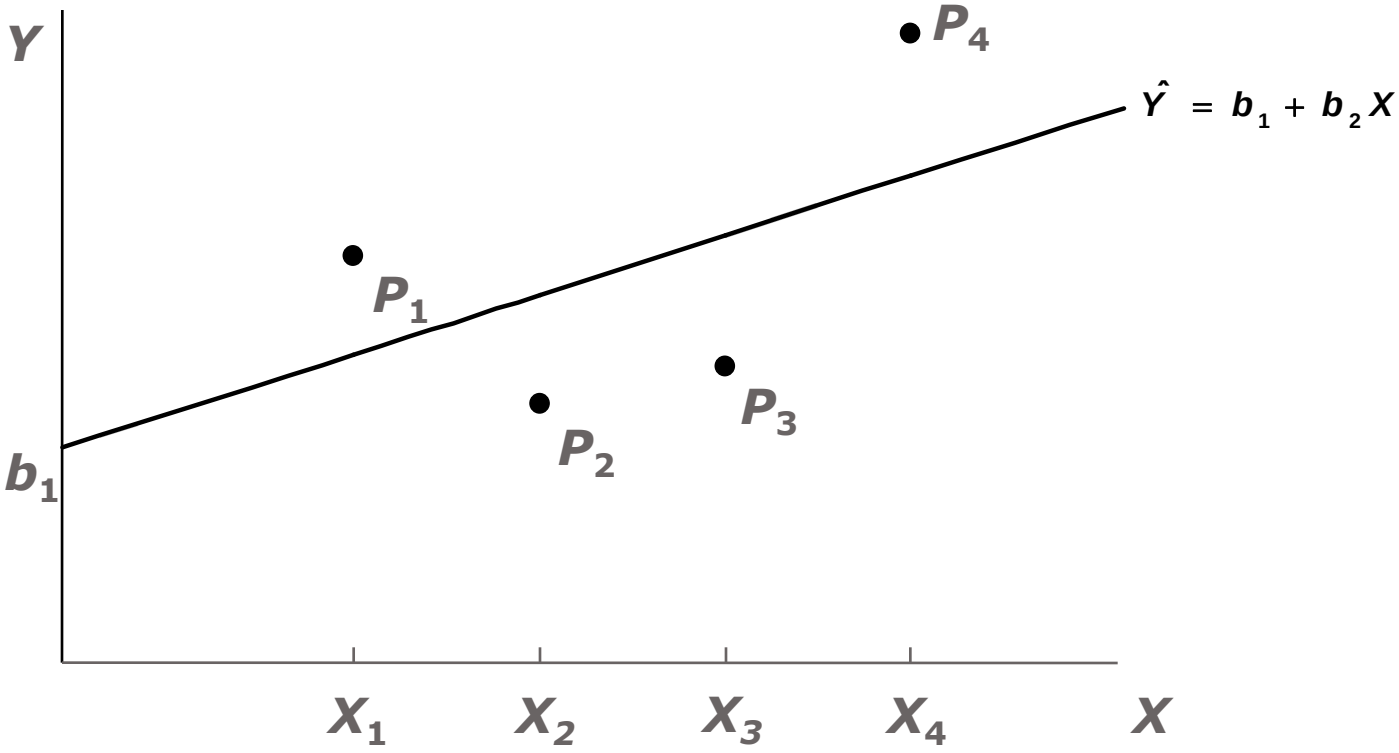
Each value of *Y* thus has a nonrandom component, $\beta_1 + \beta_2 X$, and a random component, *u*. The first observation has been decomposed into these two components.

SIMPLE LINEAR REGRESSION MODEL



In practice we can see only the *P* points.

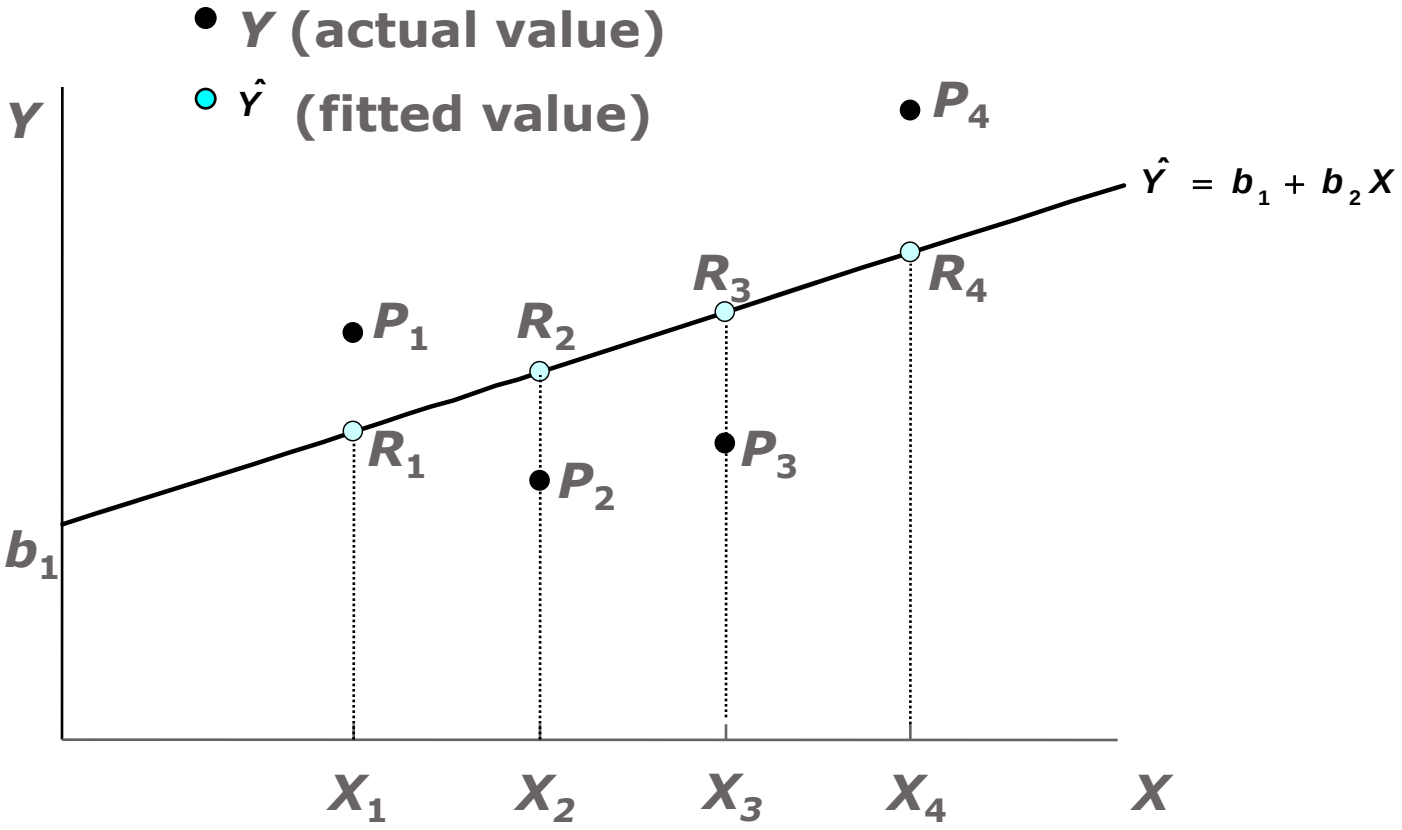
SIMPLE LINEAR REGRESSION MODEL



Obviously, we can use the *P* points to draw a line which is an approximation to the line $Y = \beta_1 + \beta_2 X$.

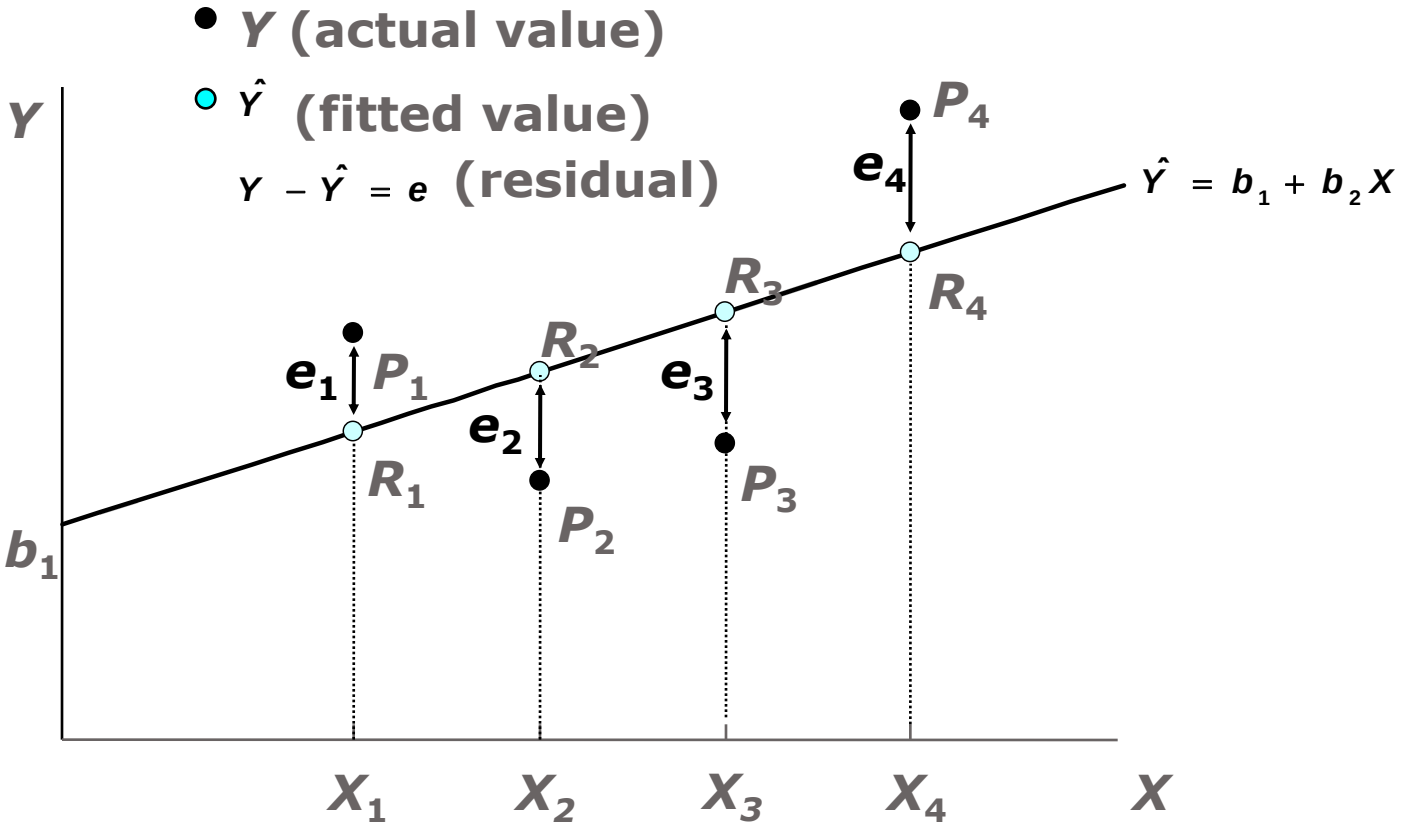
If we write this line $\hat{Y} = b_1 + b_2 X$, b_1 is an estimate of β_1 and b_2 is an estimate of β_2.

SIMPLE LINEAR REGRESSION MODEL



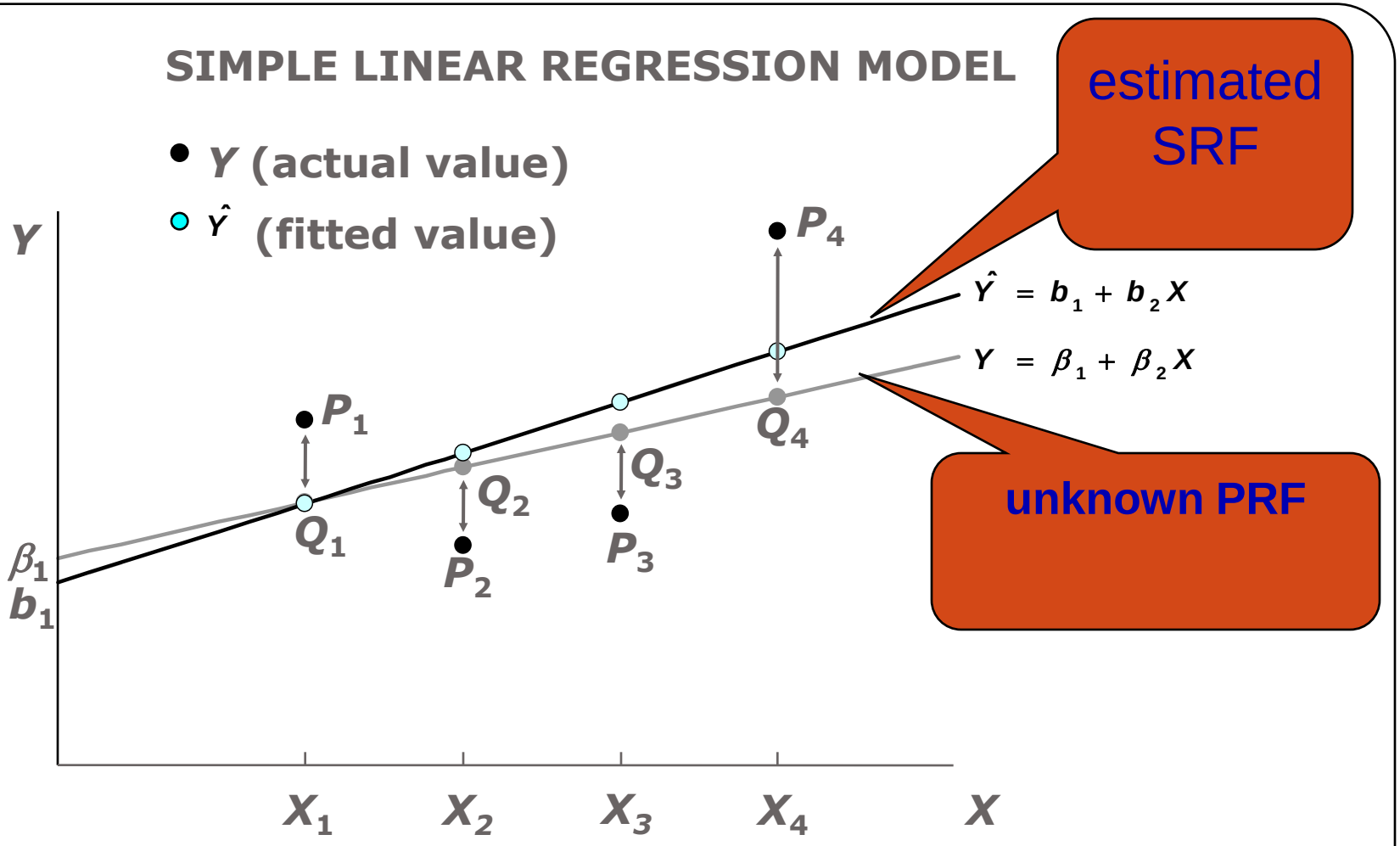
The line is called the fitted model and the values of *Y* predicted by it are called the fitted values of *Y*. They are given by the heights of the *R* points.

SIMPLE LINEAR REGRESSION MODEL



The discrepancies between the actual and fitted values of *Y* are known as the residuals.

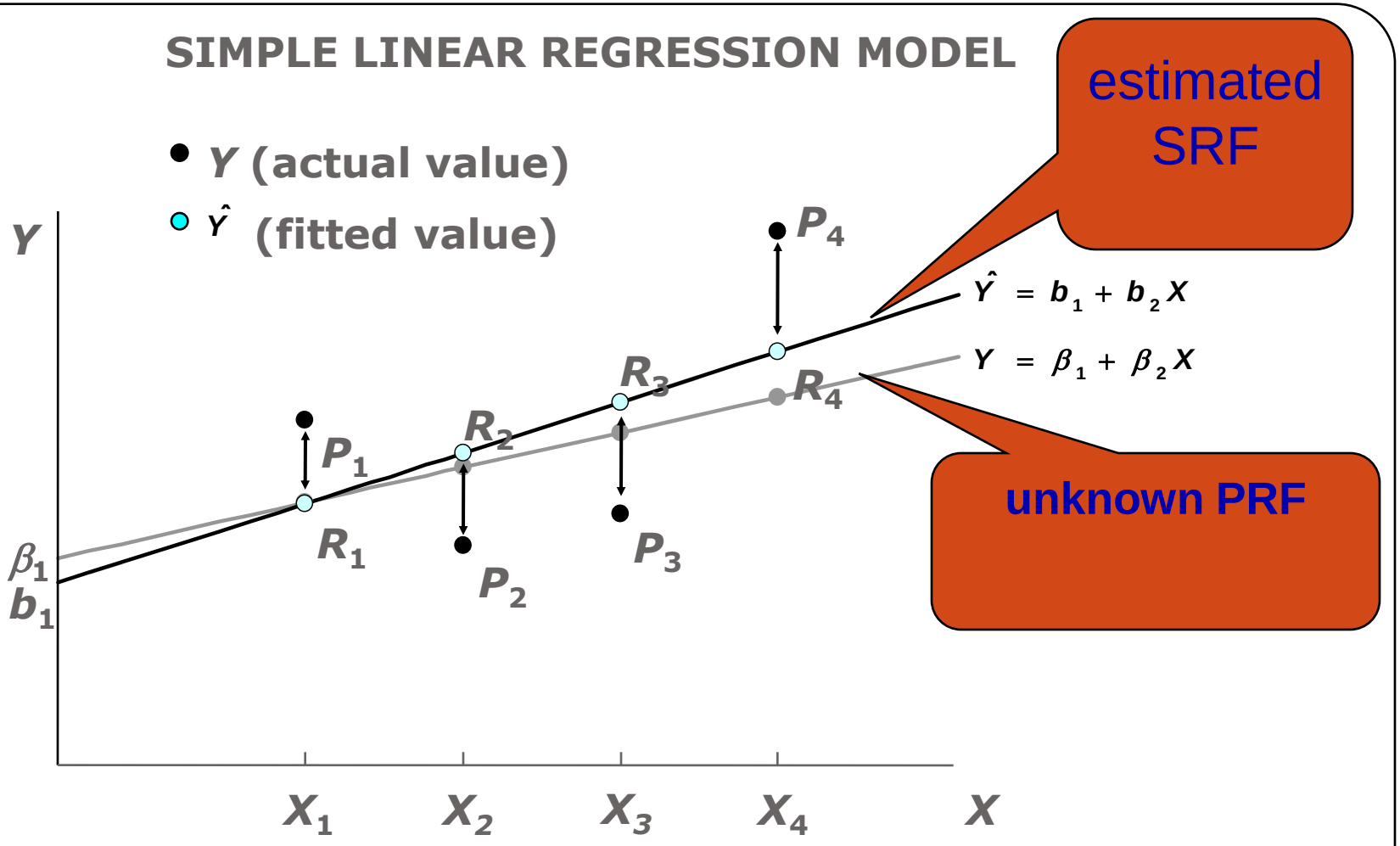
SIMPLE LINEAR REGRESSION MODEL



Note that the values of the residuals are not the same as the values of the disturbance term. The diagram now shows the true unknown relationship as well as the fitted line.

The disturbance term in each observation is responsible for the divergence between the nonrandom component of the true relationship and the actual observation.

SIMPLE LINEAR REGRESSION MODEL



The residuals are the discrepancies between the actual and the fitted values.

If the fit is a good one, the residuals and the values of the disturbance term will be similar, but they must be kept apart conceptually.

- Functional Relationship

$$\Delta y = \beta_2 \Delta x \; if \; \Delta \mu = 0$$

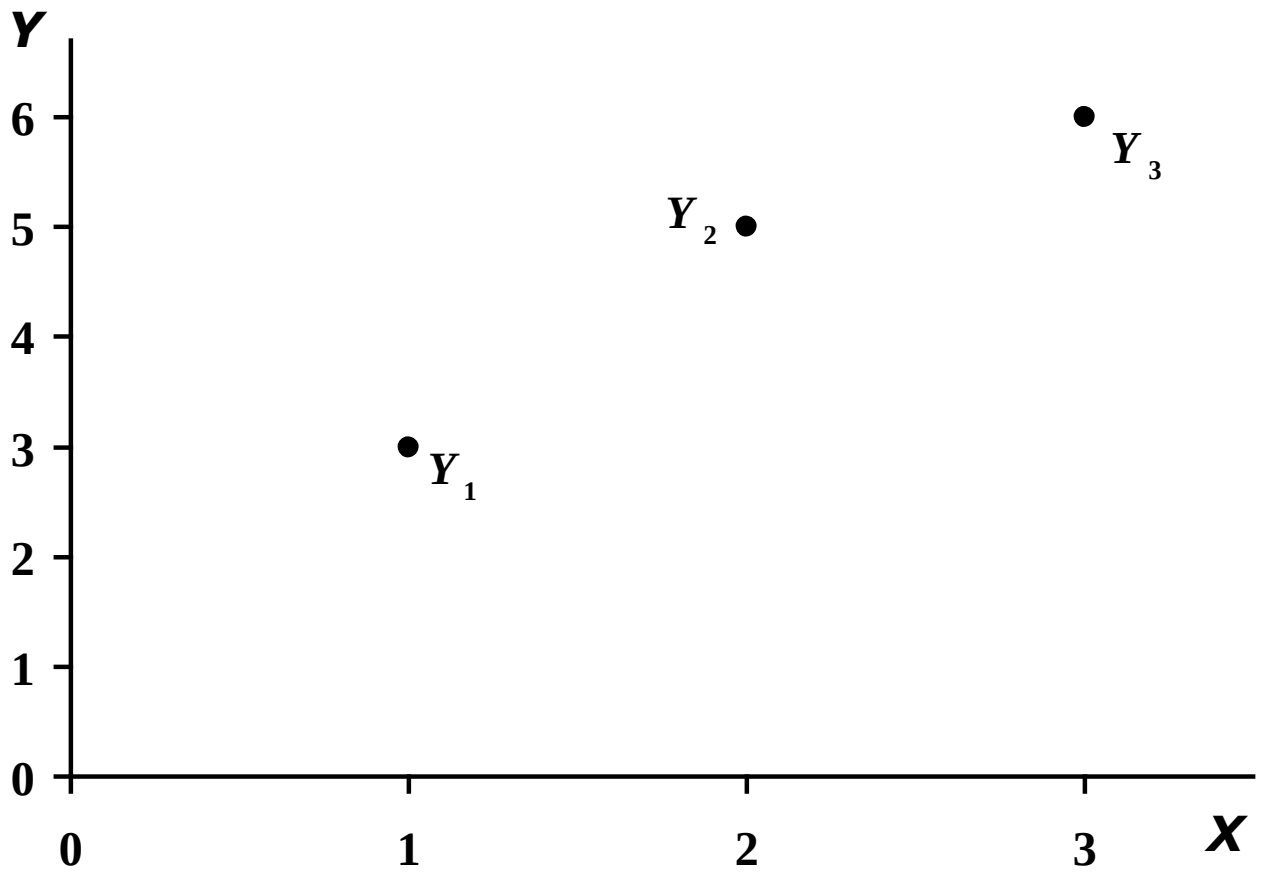
If the other factors in *disturbance term* are held fixed, so that the change in μ is zero, then *independent variable* has a *linear* effect on *dependent variable*

- Zero conditional mean assumption $E(\mu/x) = E(\mu) = 0$
- Population regression function (PRF) $E(y/x) = \beta_1 + \beta_2 x$

A one-unit increase in x changes the expected value of y by the amount β_2

DERIVING LINEAR REGRESSION COEFFICIENTS

$$\text{True model}: \quad Y = \beta_1 + \beta_2 X + u$$



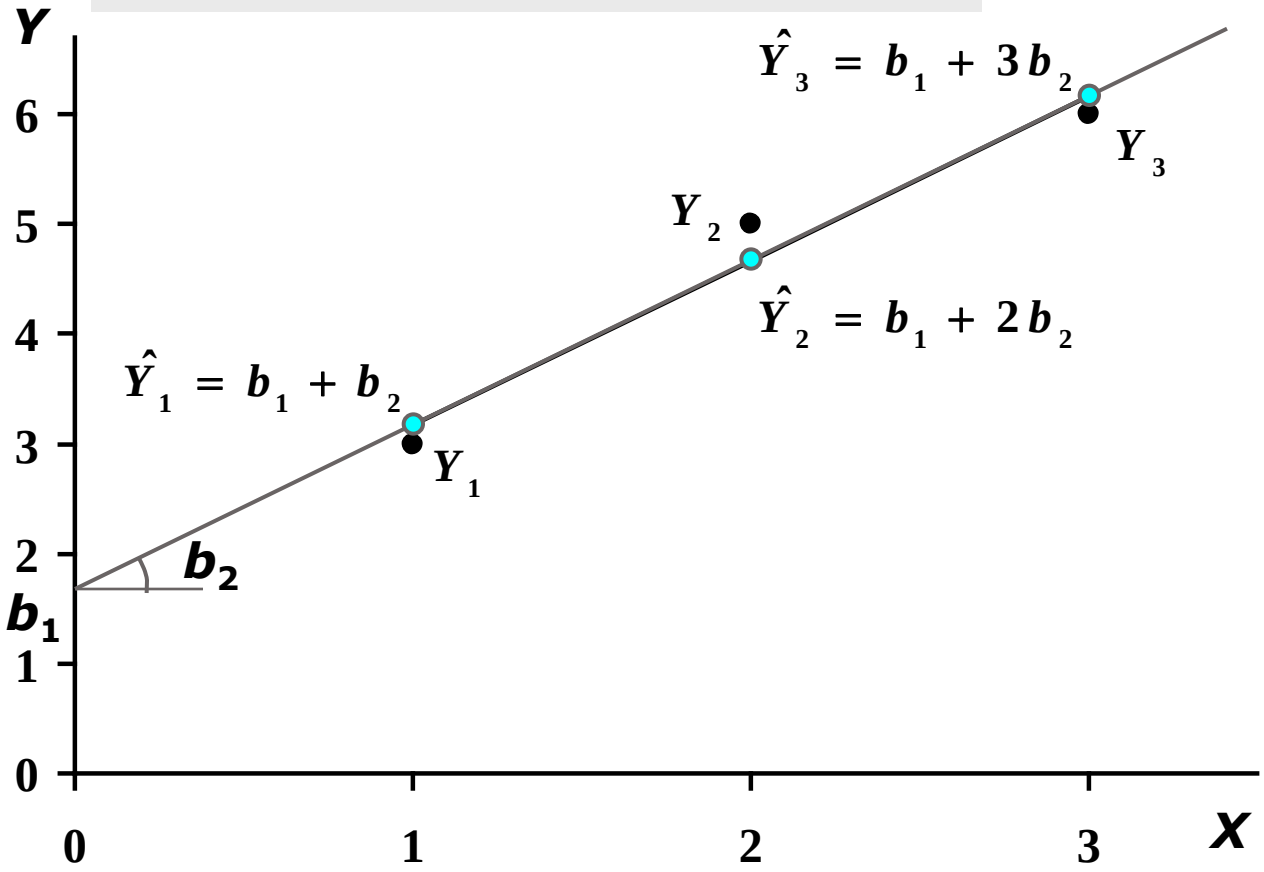
This sequence shows how the regression coefficients for a simple regression model are derived, using the least squares criterion (OLS, for ordinary least squares)

We will start with a numerical example with just three observations: (1,3), (2,5), and (3,6)

DERIVING LINEAR REGRESSION COEFFICIENTS

True model : $Y = \beta_1 + \beta_2 X + u$

Fitted line : $\hat{Y} = b_1 + b_2 X$

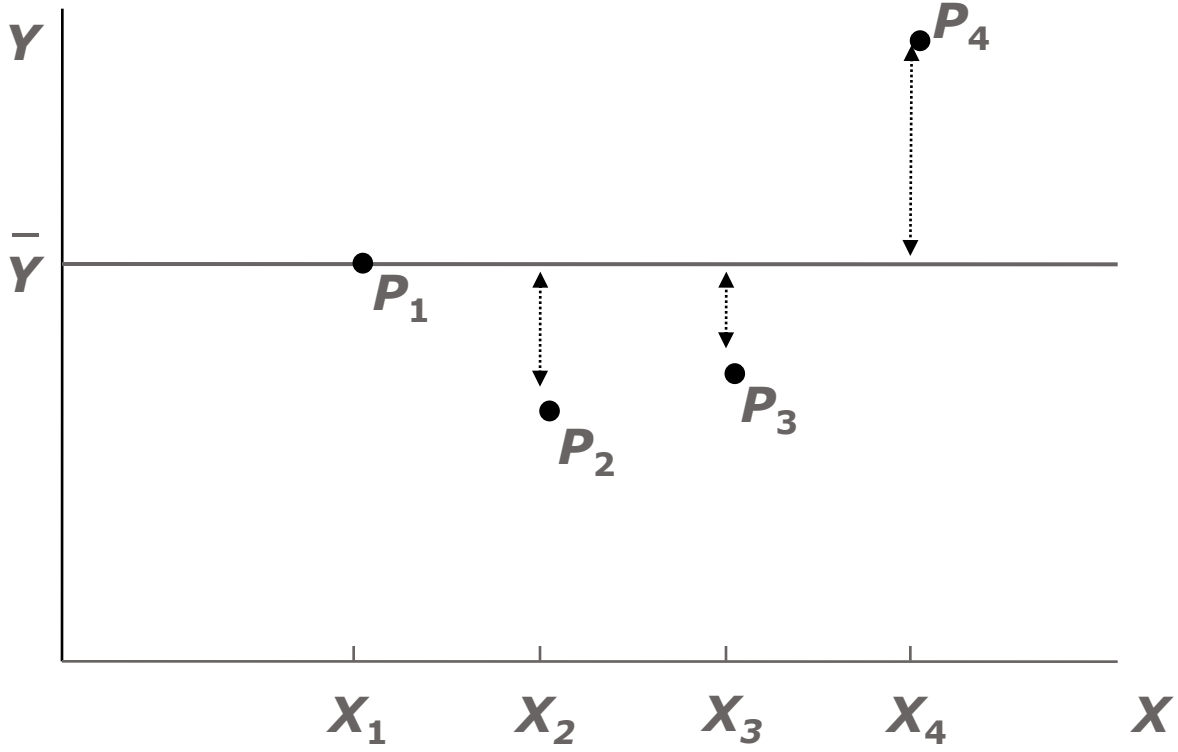


Writing the fitted regression as $\hat{Y} = b_1 + b_2X$, we will determine the values of b_1 and b_2 that minimize *RSS*, the sum of the squares of the residuals.

Least squares criterion:

Minimize *RSS* (residual sum of squares), where

$$RSS = \sum_{i=1}^{n} e_i^2 = e_1^2 + \ldots + e_n^2$$



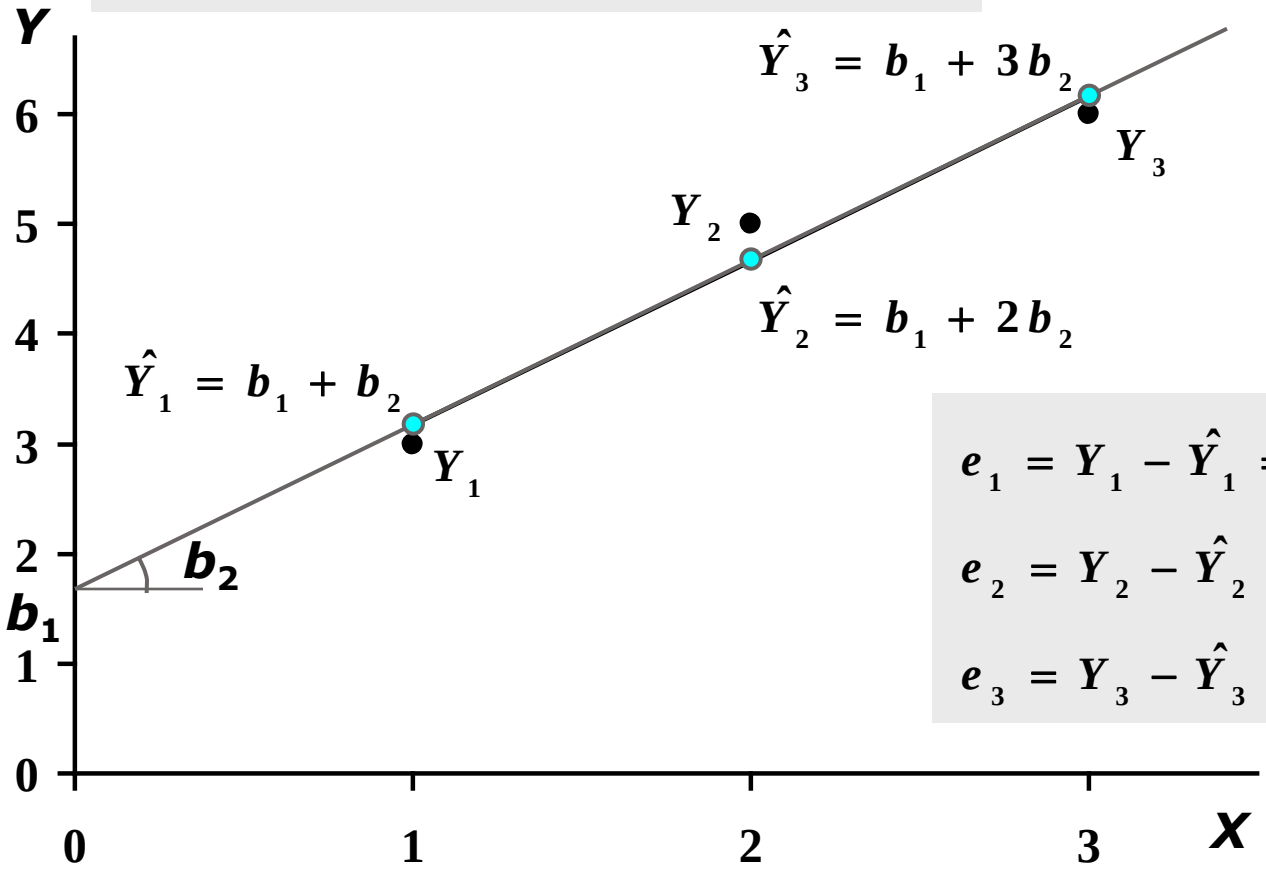
You would get an apparently perfect fit by drawing a horizontal line through the mean value of *Y*. The sum of the residuals would be zero.

You must prevent negative residuals from cancelling positive ones, and one way to do this is to use the squares of the residuals.

DERIVING LINEAR REGRESSION COEFFICIENTS

$$\text{True model} : Y = \beta_1 + \beta_2 X + u$$

$$\text{Fitted line} : \hat{Y} = b_1 + b_2 X$$



$$e_1 = Y_1 - \hat{Y}_1 = 3 - b_1 - b_2$$

$$e_2 = Y_2 - \hat{Y}_2 = 5 - b_1 - 2b_2$$

$$e_3 = Y_3 - \hat{Y}_3 = 6 - b_1 - 3b_2$$

Given our choice of b_1 and b_2, the residuals are as shown.

$$RSS = e_1^2 + e_2^2 + e_3^2 = (3 - b_1 - b_2)^2 + (5 - b_1 - 2b_2)^2 + (6 - b_1 - 3b_2)^2$$

SIMPLE REGRESSION ANALYSIS

$$RSS = e_1^2 + e_2^2 + e_3^2 = (3 - b_1 - b_2)^2 + (5 - b_1 - 2b_2)^2 + (6 - b_1 - 3b_2)^2$$

$$= 9 + b_1^2 + b_2^2 - 6b_1 - 6b_2 + 2b_1b_2$$

$$+ 25 + b_1^2 + 4b_2^2 - 10b_1 - 20b_2 + 4b_1b_2$$

$$+ 36 + b_1^2 + 9b_2^2 - 12b_1 - 36b_2 + 6b_1b_2$$

$$= 70 + 3b_1^2 + 14b_2^2 - 28b_1 - 62b_2 + 12b_1b_2$$

$$\frac{\partial RSS}{\partial b_1} = 0 \quad \Rightarrow \quad 6b_1 + 12b_2 - 28 = 0$$

$$\frac{\partial RSS}{\partial b_2} = 0 \quad \Rightarrow \quad 12b_1 + 28b_2 - 62 = 0$$

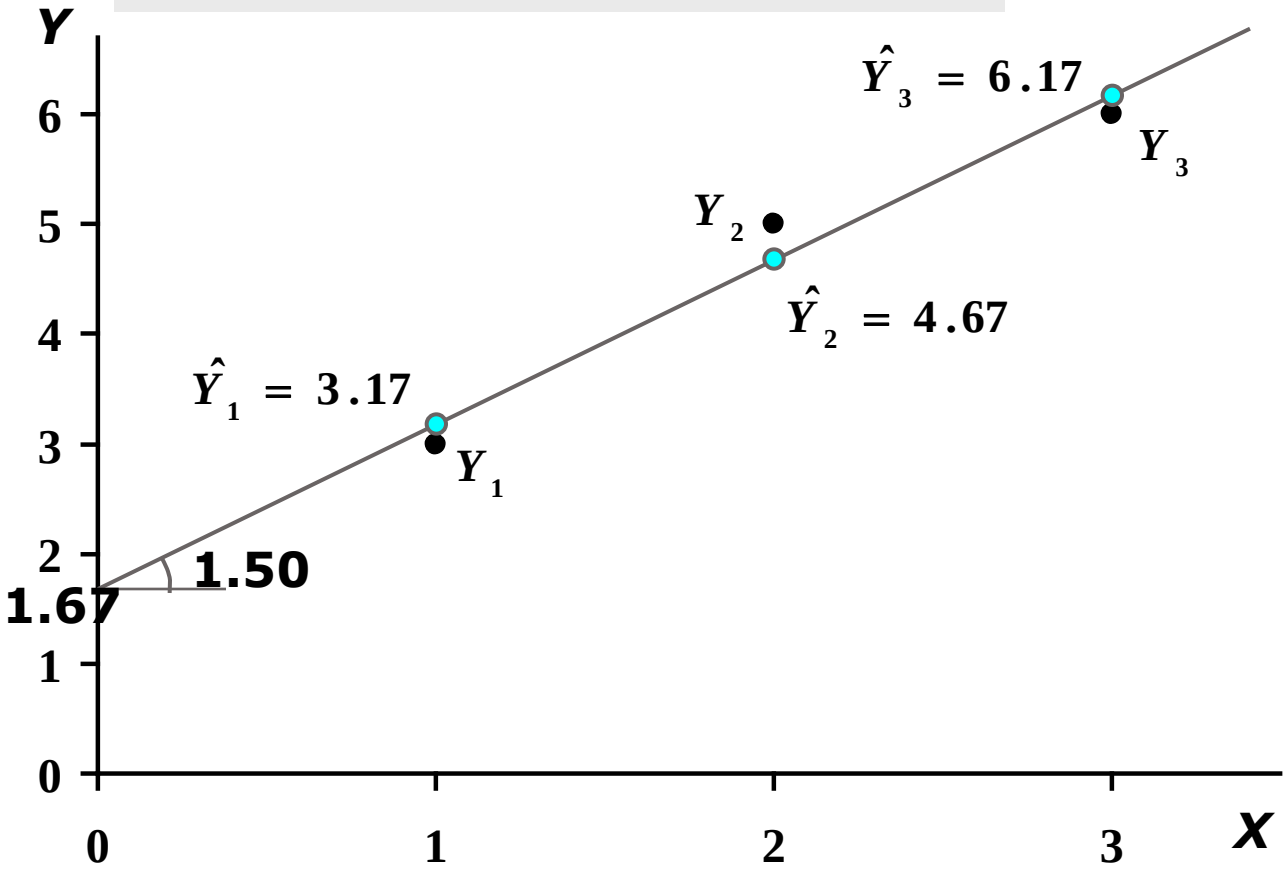
$$\therefore b_1 = 1.67, \quad b_2 = 1.50$$

The first-order conditions give us two equations in two unknowns. Solving them, we find that *RSS* is minimized when b_1 and b_2 are equal to 1.67 and 1.50, respectively.

DERIVING LINEAR REGRESSION COEFFICIENTS

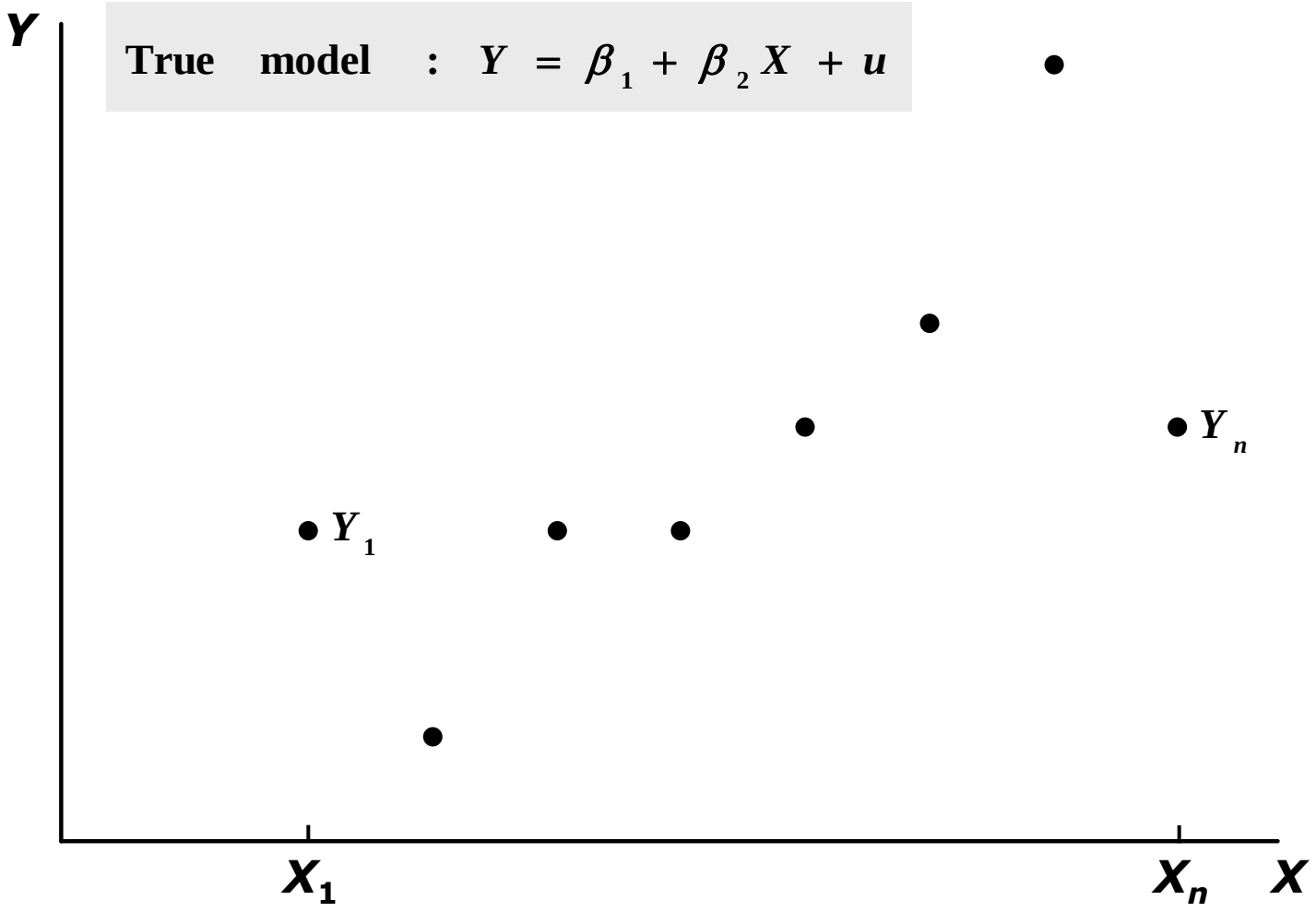
True model : $Y = \beta_1 + \beta_2 X + u$

Fitted line : $\hat{Y} = 1.67 + 1.50\ X$



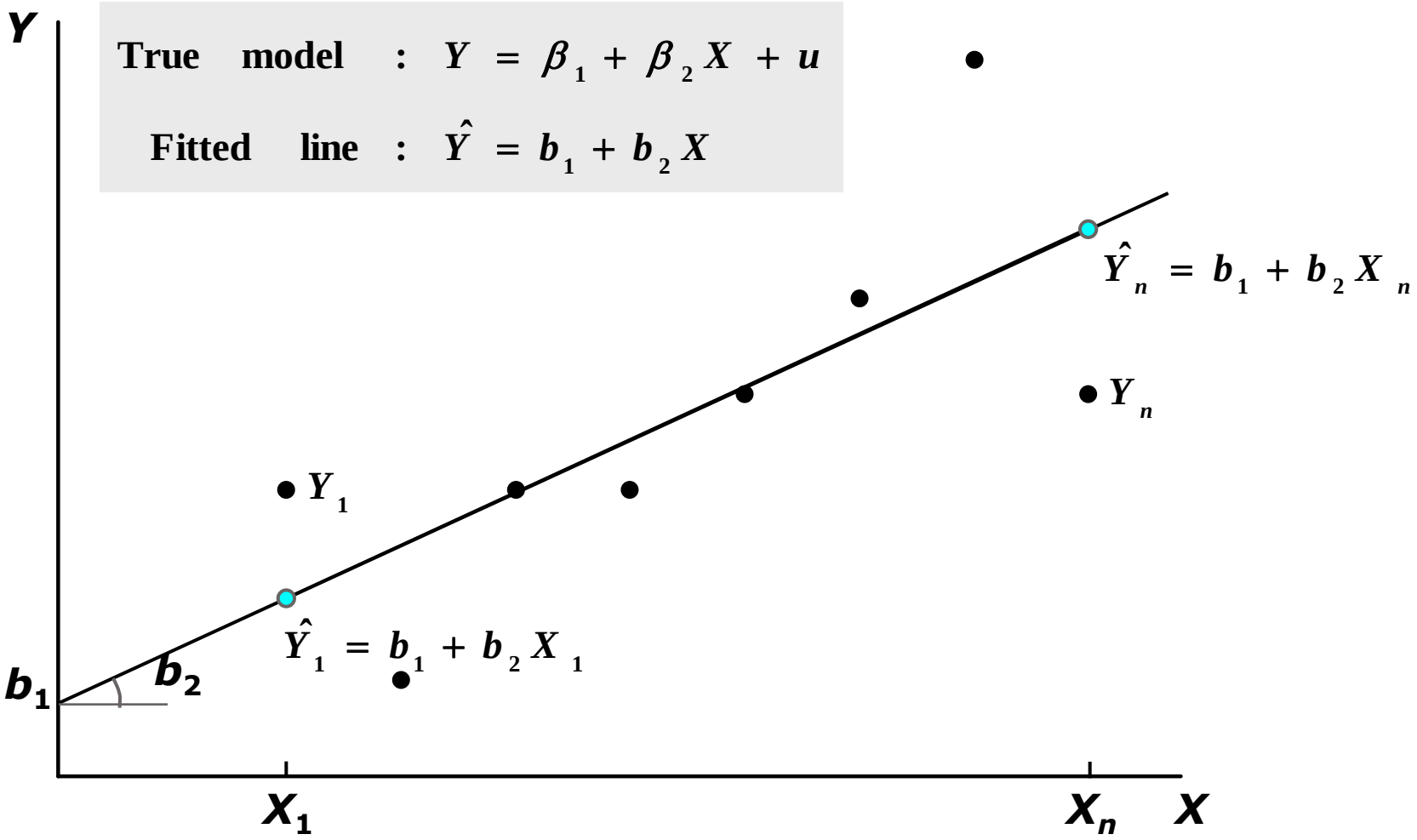
The fitted line and the fitted values of *Y* are as shown.

DERIVING LINEAR REGRESSION COEFFICIENTS



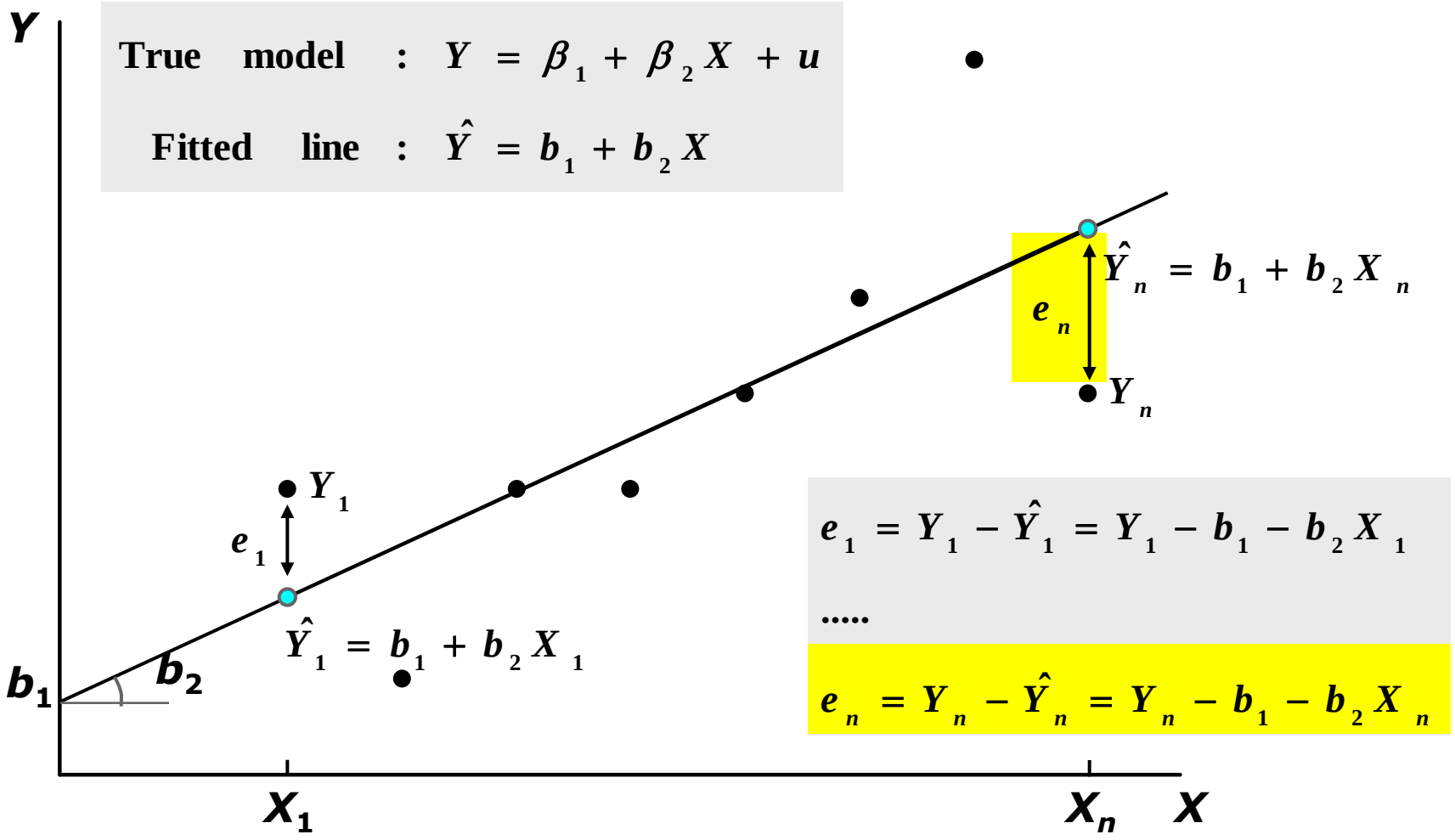
Now we will do the same thing for the general case with *n* observations.

DERIVING LINEAR REGRESSION COEFFICIENTS



Given our choice of b_1 and b_2, we will obtain a fitted line as shown.

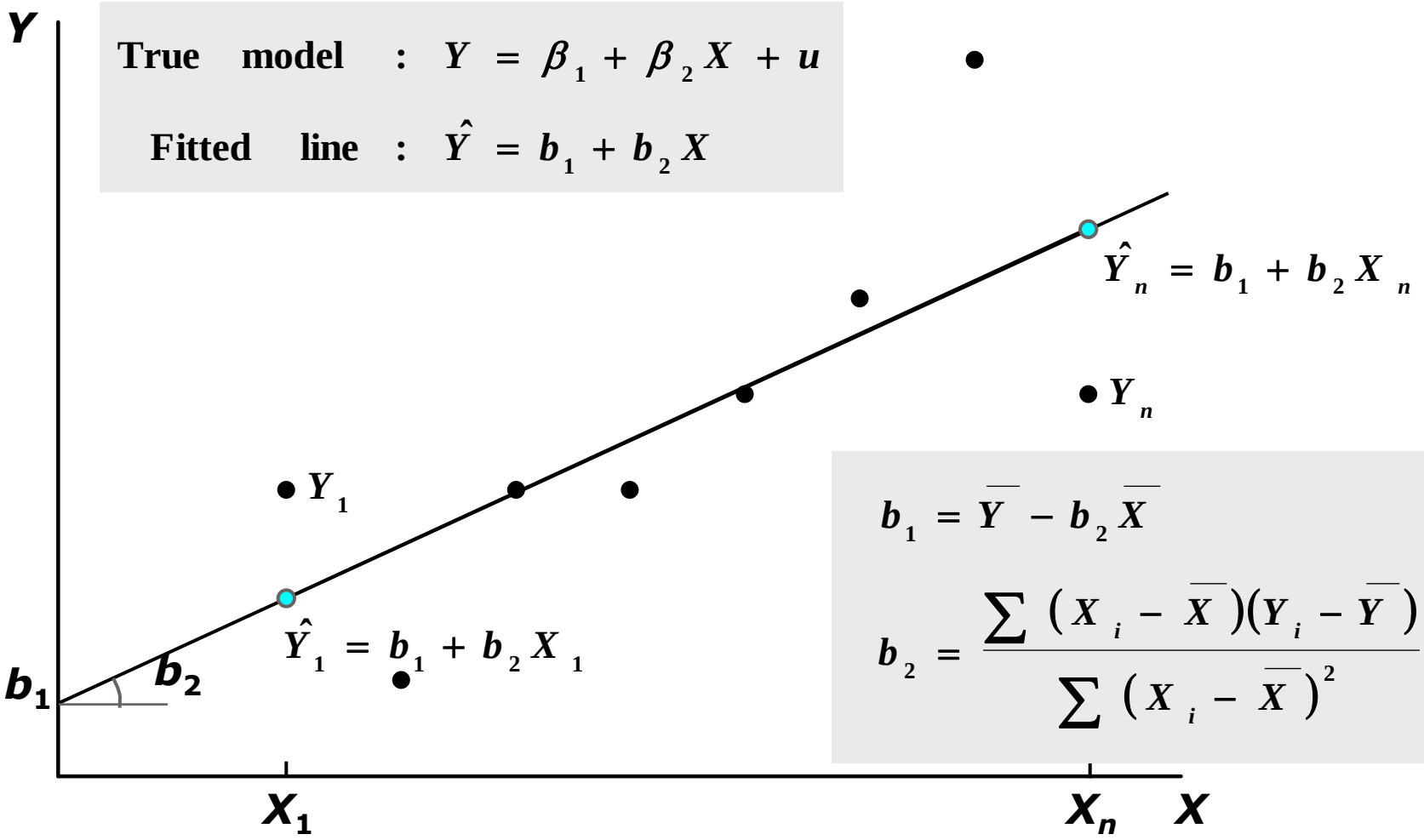
DERIVING LINEAR REGRESSION COEFFICIENTS



The residual for the first observation is defined.

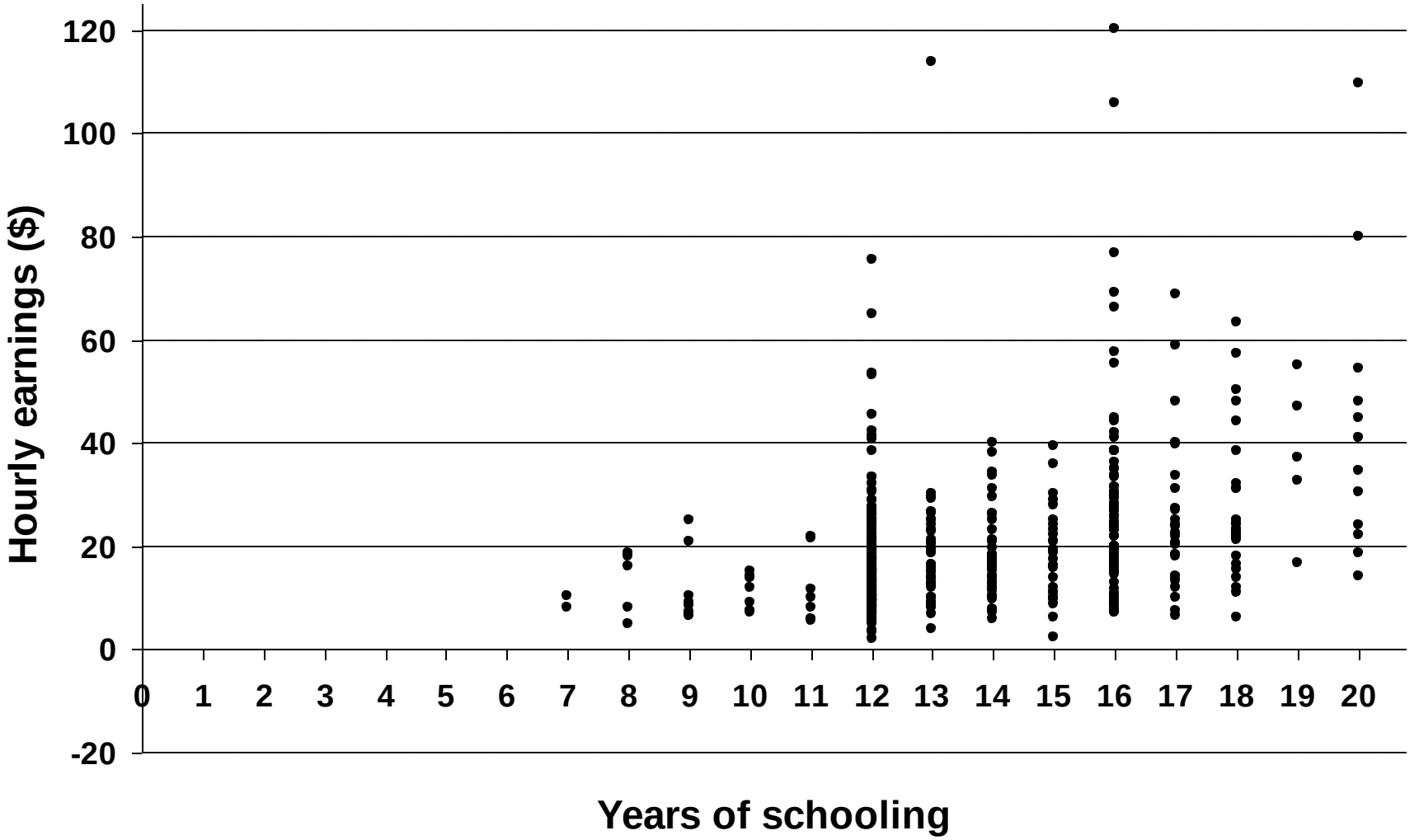
Similarly we define the residuals for the remaining observations. That for the last one is marked.

DERIVING LINEAR REGRESSION COEFFICIENTS



We chose the parameters of the fitted line so as to minimize the sum of the squares of the residuals. As a result, we derived the expressions for b_1 and b_2 using the first order condition

INTERPRETATION OF A REGRESSION EQUATION



The scatter diagram shows hourly earnings in 2002 plotted against years of schooling for a sample of 540 respondents from the National Longitudinal Survey of Youth.

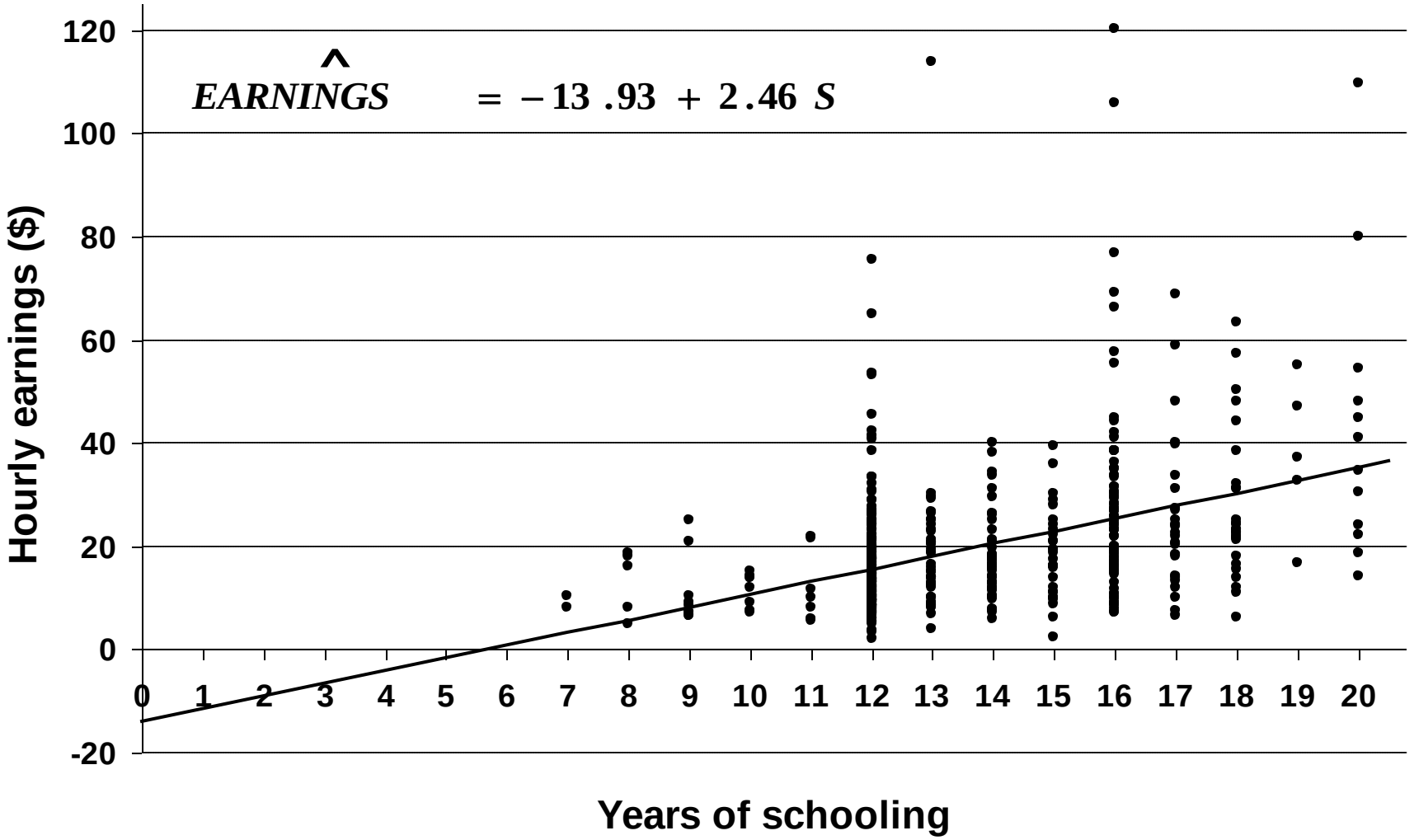
INTERPRETATION OF A REGRESSION EQUATION

```
. reg EARNINGS S

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  1,   538) =  112.15
       Model |  19321.5589     1  19321.5589            Prob > F      =  0.0000
    Residual |  92688.6722   538  172.283777            R-squared     =  0.1725
-------------+------------------------------           Adj R-squared =  0.1710
       Total |  112010.231   539  207.811189            Root MSE      =  13.126

------------------------------------------------------------------------------
    EARNINGS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |   2.455321   .2318512    10.59   0.000     1.999876    2.910765
       _cons |  -13.93347   3.219851    -4.33   0.000    -20.25849   -7.608444
------------------------------------------------------------------------------
```

This is the output from a regression of earnings on years of schooling, using Stata.

In this case there is only one variable, *S*, and its coefficient is 2.46. _cons, in Stata, refers to the constant. The estimate of the intercept is -13.93.

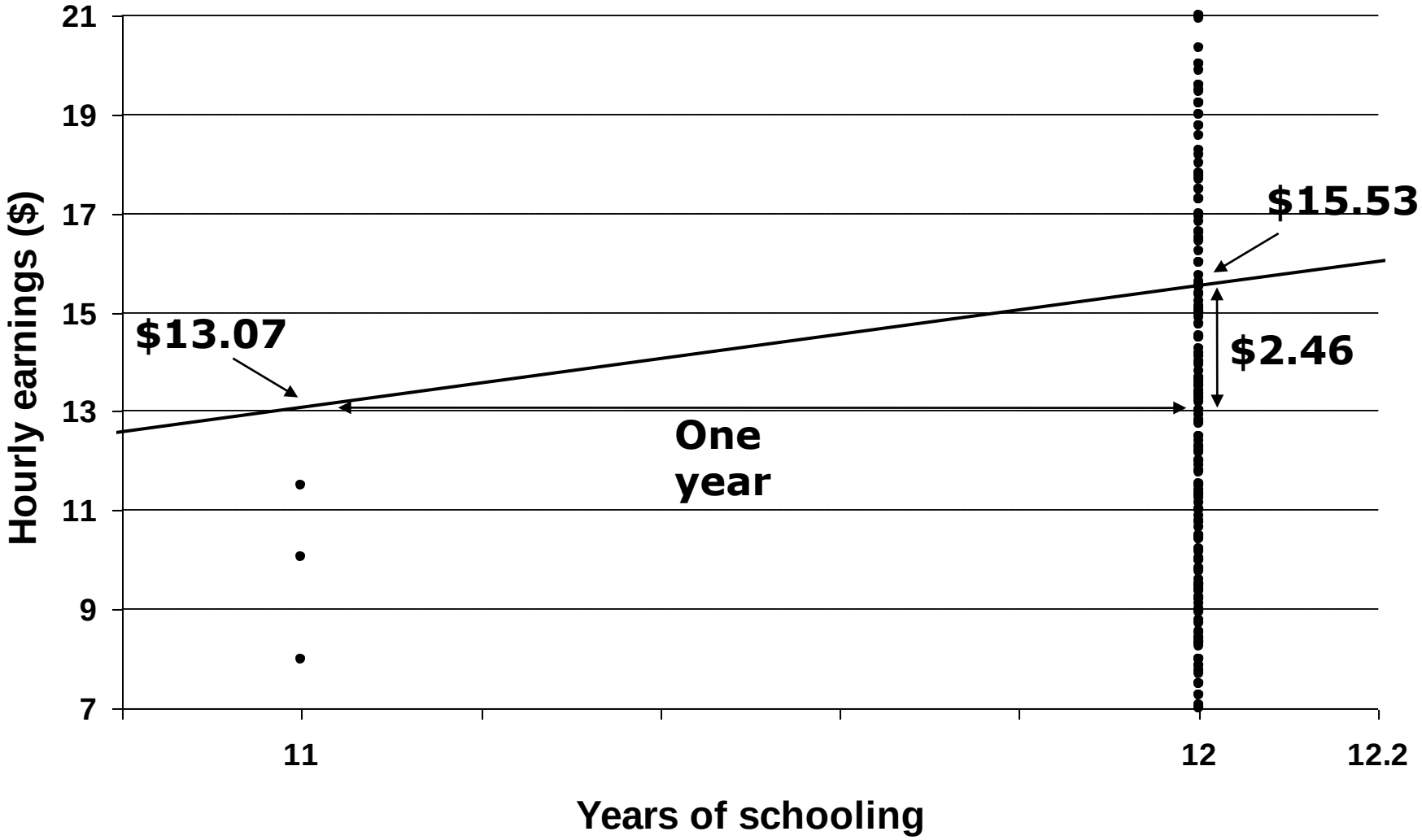
INTERPRETATION OF A REGRESSION EQUATION



Here is the scatter diagram again, with the regression line shown.

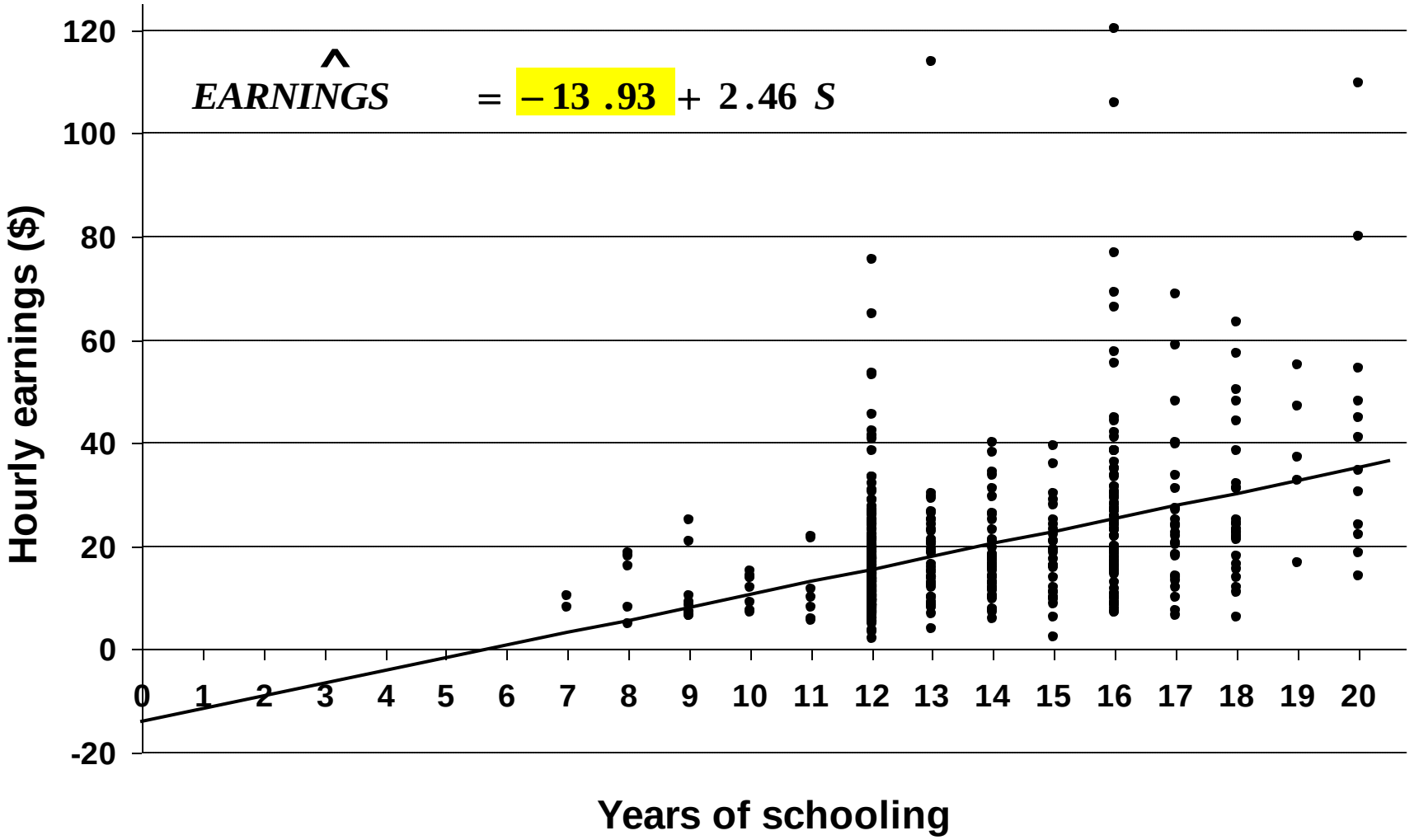
***S* is measured in years (strictly speaking, grades completed), *EARNINGS* in dollars per hour. So the slope coefficient implies that hourly earnings increase by $2.46 for each extra year of schooling.**

INTERPRETATION OF A REGRESSION EQUATION



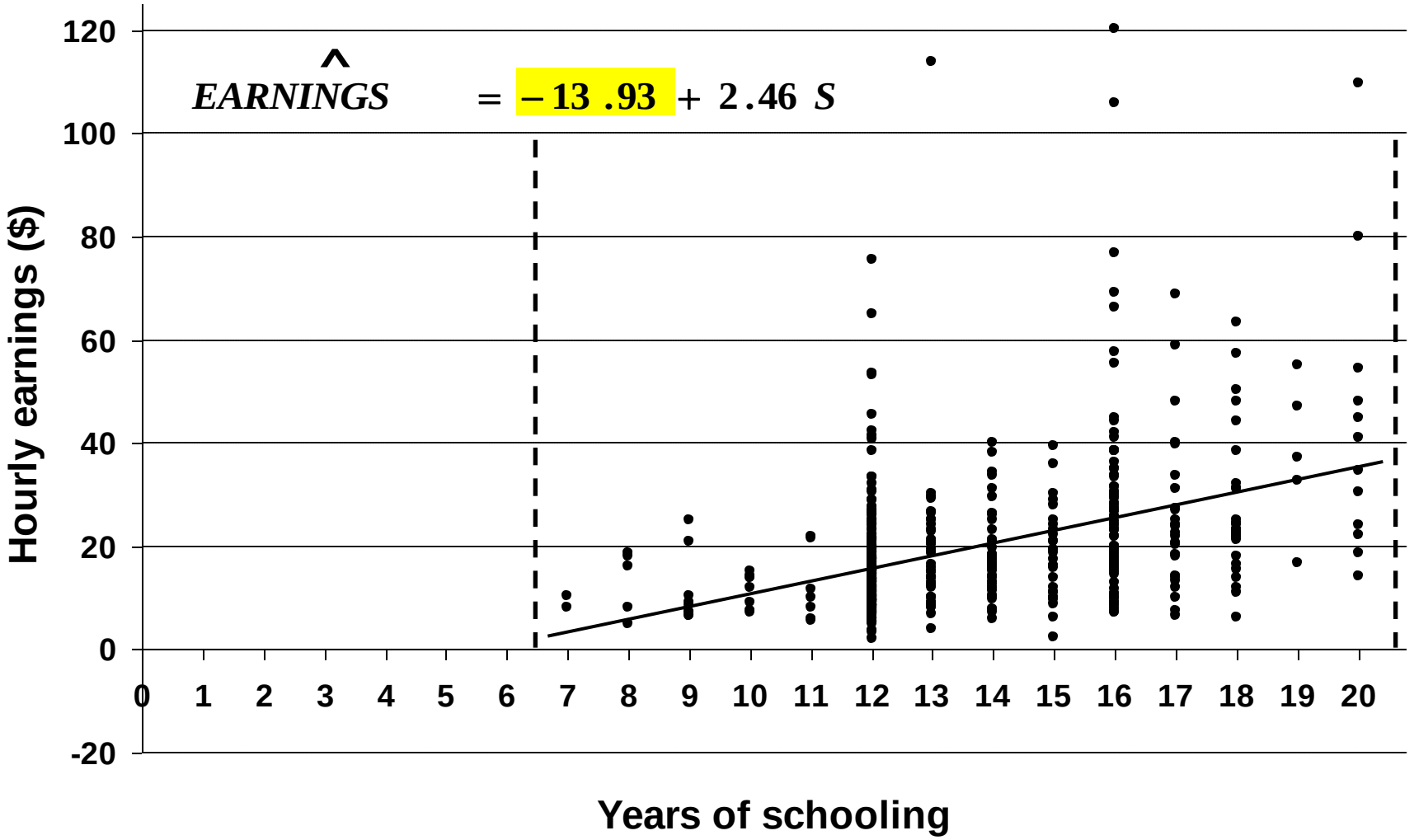
Geometrical representation: The regression line indicates that completing 12th grade instead of 11th grade would increase earnings by $2.46, from $13.07 to $15.53, as a general tendency.

INTERPRETATION OF A REGRESSION EQUATION



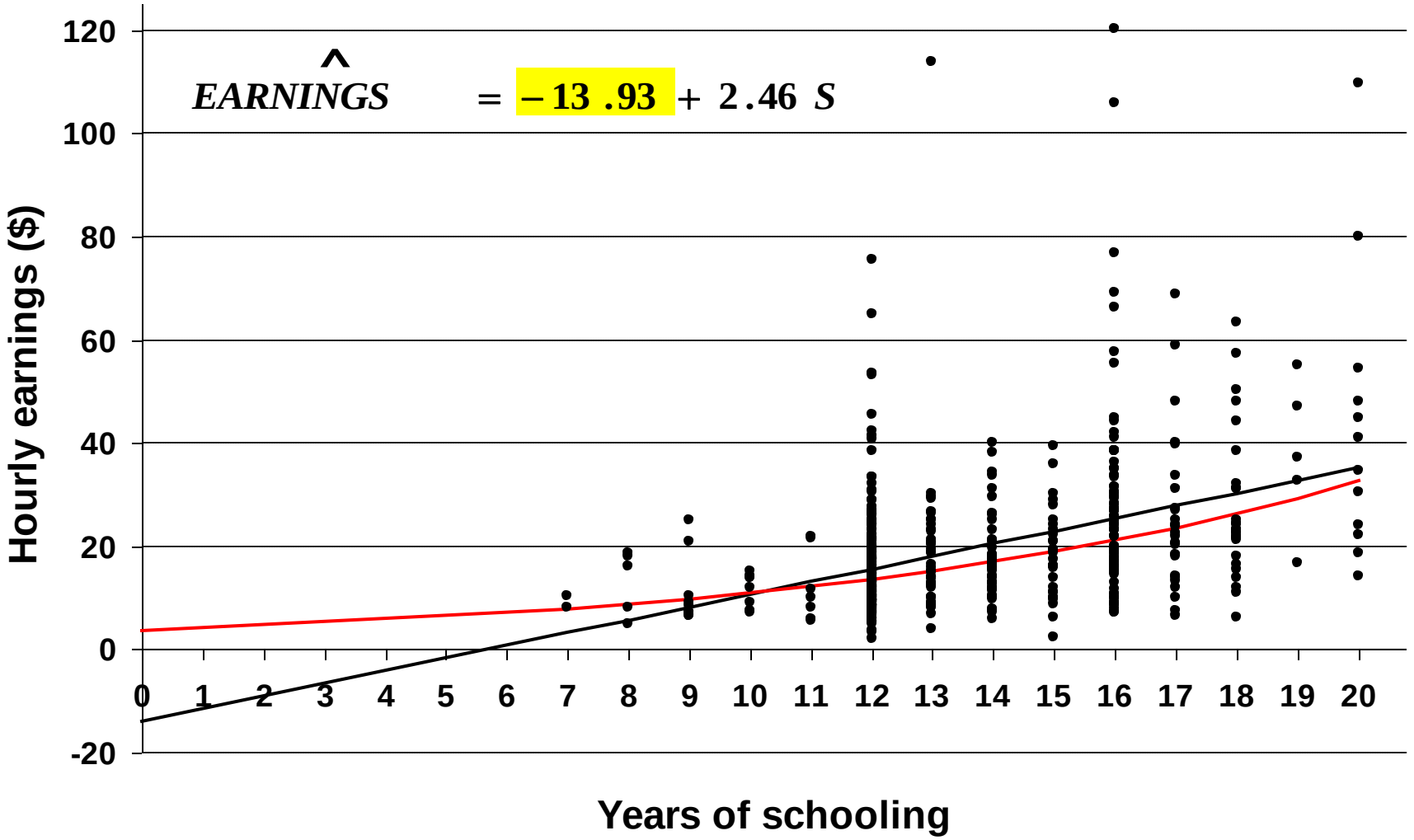
Literally, the constant indicates that an individual with no years of education would have to pay $13.93 per hour to be allowed to work. => it does not make any sense

INTERPRETATION OF A REGRESSION EQUATION



A safe solution to the problem is to limit the interpretation to the range of the sample data, and to refuse to extrapolate on the ground that we have no evidence outside the data range.

INTERPRETATION OF A REGRESSION EQUATION



Another solution is to explore the possibility that the true relationship is nonlinear and that we are approximating it with a linear regression. We will soon extend the regression technique to fit nonlinear models.

BASIC ASSUMPTION OF THE OLS

- zero systematic error: $E(u_i) = 0$
- Homoscedasticity: $var(u_i) = \delta^2$ for all i
- No autocorrelation: $cov(u_i; u_j) = 0$ for all i #j
- X is non-stochastic
- $u \sim N(0, \delta^2)$

GOODNESS OF FIT

Useful results:

$$\bar{e} = 0 \qquad \bar{\hat{Y}} = \bar{Y} \qquad \sum X_i e_i = 0 \qquad \sum \hat{Y}_i e_i = 0$$

$$TSS = ESS + RSS$$

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum e_i^2$$

$$R^2 = \frac{ESS}{TSS} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}$$

The main criterion of goodness of fit, formally described as the coefficient of determination, but usually referred to as R^2, is defined to be the ratio of *ESS* to *TSS*, that is, the proportion of the variance of *Y* explained by the regression equation.

GOODNESS OF FIT

$$R^2 = \frac{ESS}{TSS} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}$$

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}$$

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum e_i^2 \qquad TSS = ESS + RSS$$

The OLS regression coefficients are chosen in such a way as to minimize the sum of the squares of the residuals. Thus it automatically follows that they maximize R^2.

UNBIASEDNESS OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$b_2 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}$$

We will now demonstrate that the ordinary least squares (OLS) estimator of the slope coefficient in a simple regression model is unbiased.

UNBIASEDNESS OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$b_2 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} = \beta_2 + \sum a_i u_i$$

$$a_i = \frac{X_i - \overline{X}}{\sum (X_j - \overline{X})^2}$$

We saw in a previous slideshow that the slope coefficient may be decomposed into the true value and a weighted sum of the values of the disturbance term.

UNBIASEDNESS OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$b_2 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} = \beta_2 + \sum a_i u_i$$

$$a_i = \frac{X_i - \overline{X}}{\sum (X_j - \overline{X})^2}$$

$$E(b_2) = E(\beta_2) + E\left(\sum a_i u_i\right)$$

Hence the expected value of b_2 is equal to the expected value of β_2 and the expected value of the weighted sum of the values of the disturbance term.

UNBIASEDNESS OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$b_2 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} = \beta_2 + \sum a_i u_i$$

$$a_i = \frac{X_i - \overline{X}}{\sum (X_j - \overline{X})^2}$$

$$E(b_2) = E(\beta_2) + E\left(\sum a_i u_i\right)$$

$$= \beta_2 + \sum E(a_i u_i)$$

$$E\left(\sum a_i u_i\right) = E(a_1 u_1 + \ldots + a_n u_n) = E(a_1 u_1) + \ldots + E(a_n u_n) = \sum E(a_i u_i)$$

β_2 is fixed so it is unaffected by taking expectations. The first expectation rule states that the expectation of a sum of several quantities is equal to the sum of their expectations.

UNBIASEDNESS OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$b_2 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} = \beta_2 + \sum a_i u_i$$

$$a_i = \frac{X_i - \overline{X}}{\sum (X_j - \overline{X})^2}$$

$$E(b_2) = E(\beta_2) + E\left(\sum a_i u_i\right)$$

$$= \beta_2 + \sum E(a_i u_i) = \beta_2 + \sum a_i E(u_i)$$

Now for each i, $E(a_i u_i) = a_i E(u_i)$

UNBIASEDNESS OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$b_2 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \beta_2 + \sum a_i u_i$$

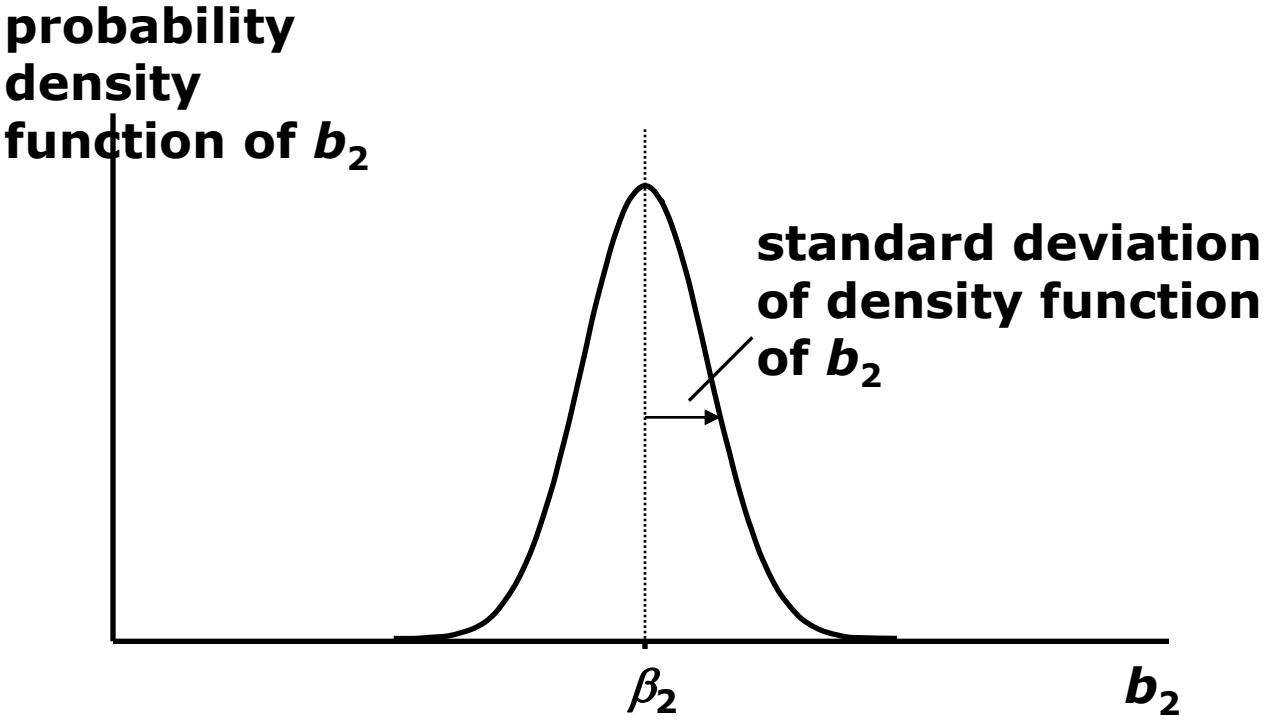
$$a_i = \frac{X_i - \bar{X}}{\sum (X_j - \bar{X})^2}$$

$$\begin{aligned} E(b_2) &= E(\beta_2) + E\left(\sum a_i u_i\right) \\ &= \beta_2 + \sum E(a_i u_i) = \beta_2 + \sum a_i E(u_i) \\ &= \beta_2 \end{aligned}$$

Under zero conditional mean, $E(u_i) = 0$ for all i, and so the estimator is unbiased

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$



In this sequence we will see that we can also obtain estimates of the standard deviations of the distributions. These will give some idea of their likely reliability and will provide a basis for tests of hypotheses.

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$\sigma_{b_1}^2 = \sigma_u^2 \left\{ \frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2} \right\}$$

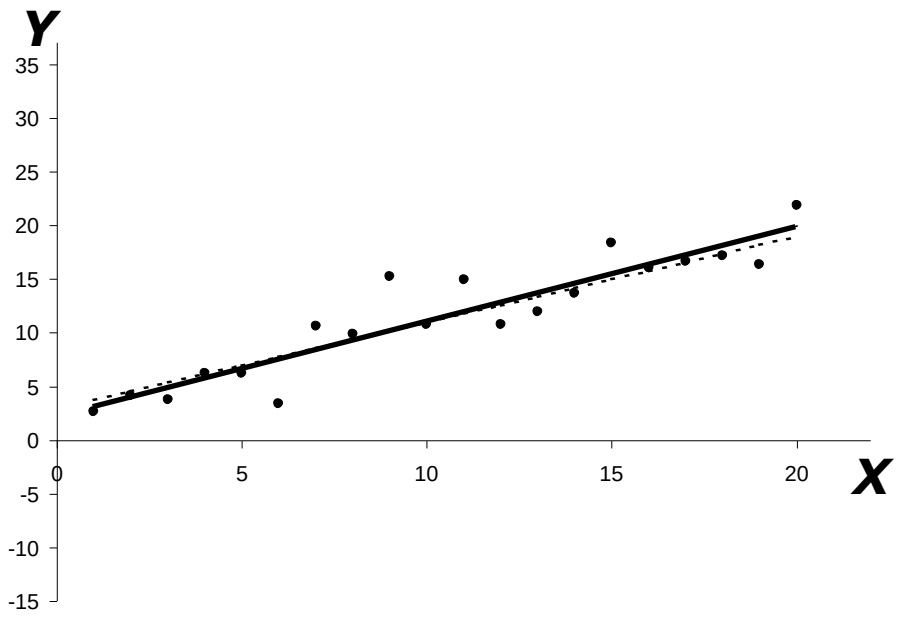
$$\sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum (X_i - \bar{X})^2}$$

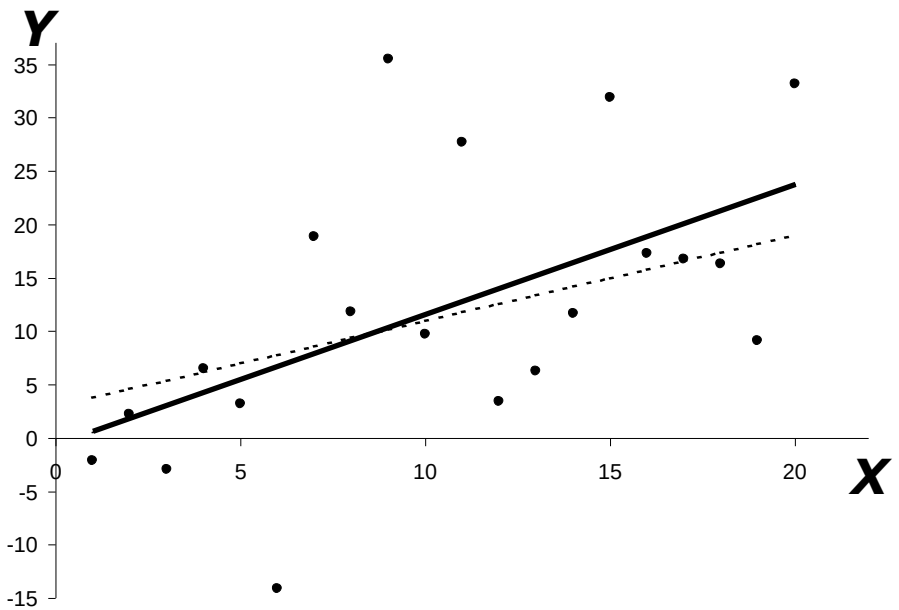
Expressions (which will not be derived) for the variances of their distributions are shown above.

We will focus on the implications of the expression for the variance of b_2. Looking at the numerator, we see that the variance of b_2 is proportional to σ_u^2. This is as we would expect. The more noise there is in the model, the less precise will be our estimates.

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$





$$Y = 3.0 + 0.8X$$

This is illustrated by the diagrams above. The nonstochastic component of the relationship, $Y = 3.0 + 0.8X$, represented by the dotted line, is the same in both diagrams.

However, in the right-hand diagram the random numbers have been multiplied by a factor of 5. As a consequence, the regression line, the solid line, is a much poorer approximation to the nonstochastic relationship.

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

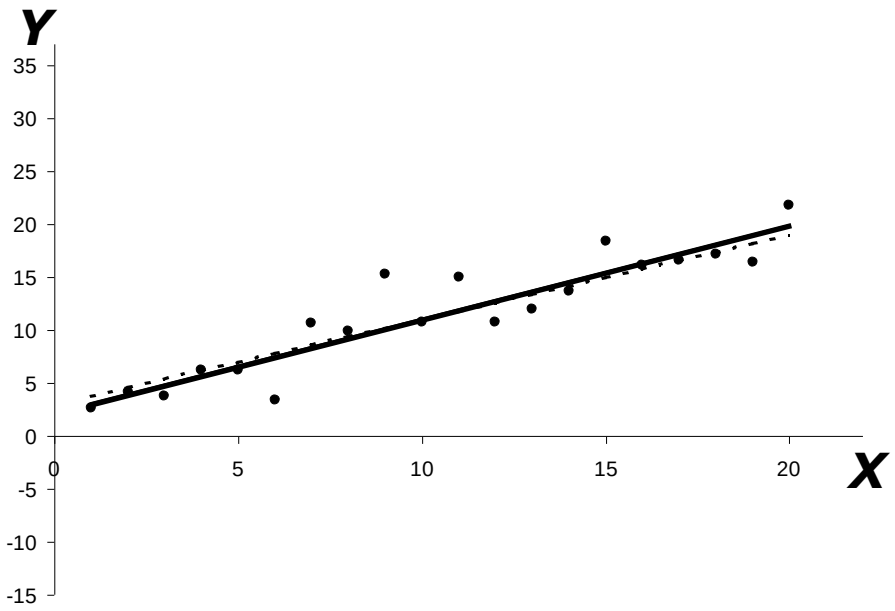
$$\sigma_{b_1}^2 = \sigma_u^2 \left\{ \frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2} \right\}$$

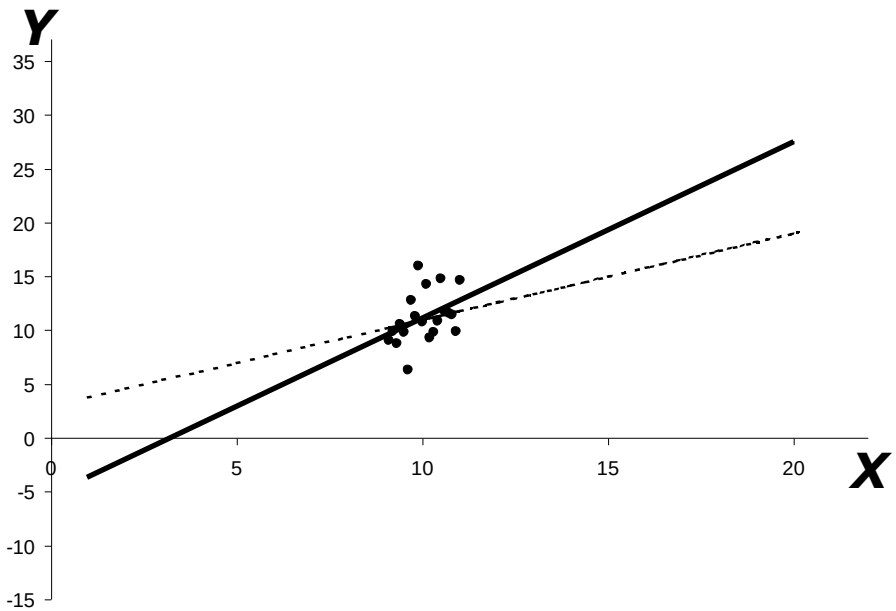
$$\sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum (X_i - \bar{X})^2}$$

Looking at the denominator, the larger is the sum of the squared deviations of X, the smaller is the variance of b_2.

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$



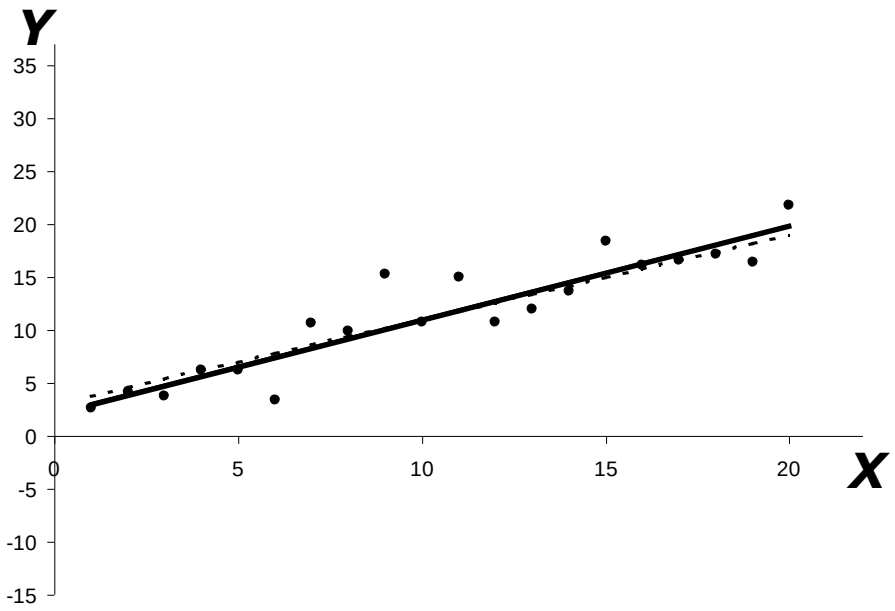


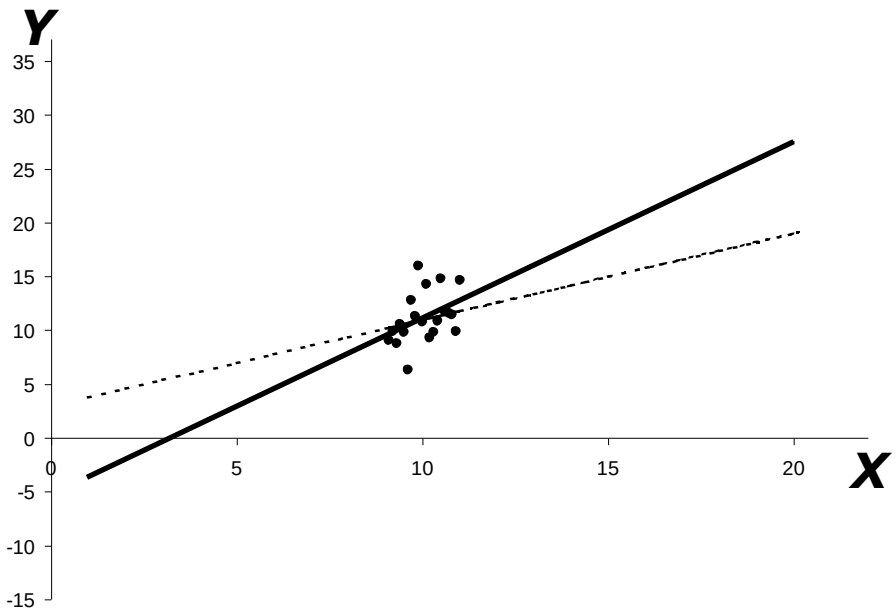
$$Y = 3.0 + 0.8X$$

In the diagrams above, the nonstochastic component of the relationship is the same and the same random numbers have been used for the 20 values of the disturbance term.

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$



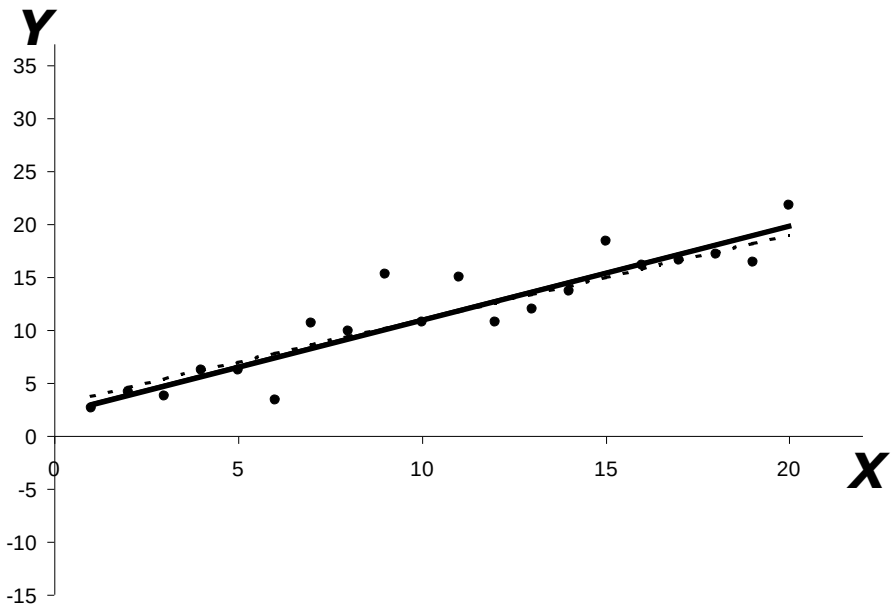


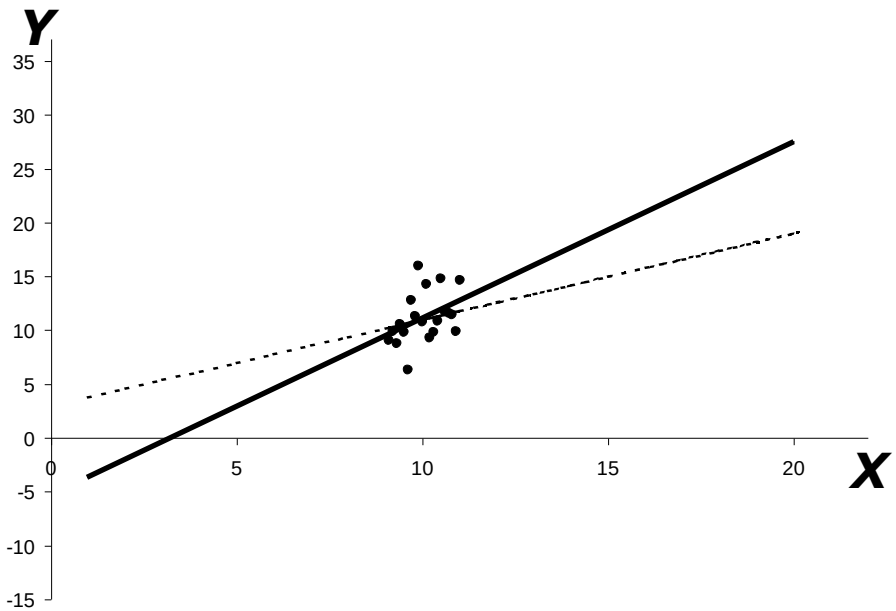
$$Y = 3.0 + 0.8X$$

However, MSD(*X*) is much smaller in the right-hand diagram because the values of *X* are much closer together.

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$





$$Y = 3.0 + 0.8X$$

Hence in that diagram the position of the regression line is more sensitive to the values of the disturbance term, and as a consequence the regression line is likely to be relatively inaccurate.

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$\sigma_{b_1}^2 = \sigma_u^2 \left\{ \frac{1}{n} + \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2} \right\} \qquad \sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum (X_i - \overline{X})^2} = \frac{\sigma_u^2}{n \text{ MSD } (X)}$$

We cannot calculate the variances exactly because we do not know the variance of the disturbance term. However, we can derive an estimator of σ_u^2 from the residuals.

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

$$\sigma_{b_1}^2 = \sigma_u^2 \left\{ \frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2} \right\} \qquad \sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum (X_i - \bar{X})^2} = \frac{\sigma_u^2}{n \text{ MSD } (X)}$$

$$\text{MSD } (e) = \frac{1}{n} \sum (e_i - \bar{e})^2 = \frac{1}{n} \sum e_i^2$$

Clearly the scatter of the residuals around the regression line will reflect the unseen scatter of *u* about the line $Y_i = \beta_1 + b_2 X_i$, although in general the residual and the value of the disturbance term in any given observation are not equal to one another.

One measure of the scatter of the residuals is their mean square error, MSD(*e*), defined as shown.

PRECISION OF THE REGRESSION COEFFICIENTS

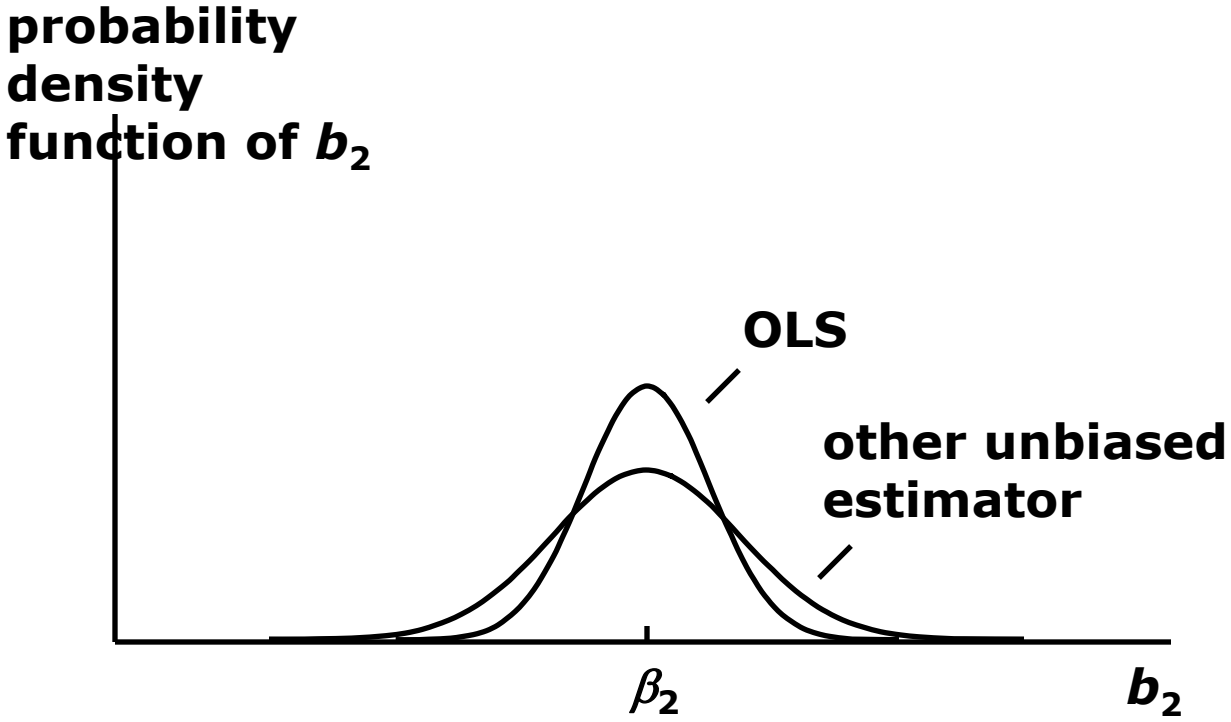
```
. reg EARNINGS S

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  1,   538) =  112.15
       Model |  19321.5589     1  19321.5589            Prob > F      =  0.0000
    Residual |  92688.6722   538  172.283777           R-squared     =  0.1725
-------------+------------------------------           Adj R-squared =  0.1710
       Total |  112010.231   539  207.811189           Root MSE      =  13.126

------------------------------------------------------------------------------
    EARNINGS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |    2.455321   .2318512    10.59   0.000     1.999876    2.910765
       _cons |   -13.93347   3.219851    -4.33   0.000    -20.25849   -7.608444
------------------------------------------------------------------------------
```

The standard errors of the coefficients always appear as part of the output of a regression. The standard errors appear in a column to the right of the coefficients.

PRECISION OF THE REGRESSION COEFFICIENTS

Simple regression model: $Y = \beta_1 + \beta_2 X + u$

Efficiency



The Gauss–Markov theorem states that, provided that the regression model assumptions are valid, the OLS estimators are BLUE: Linear, Unbiased, Minimum variance in the class of all unbiased estimators

Summing up

- Simple Linear Regression model:
  - Verify dependent, independent variables, parameters, and the error terms
  - Interpret estimated parameters b_1 & b_2 as they show the relationship between X and Y.
  - OLS provides BUE estimators for the parameters under 5 Gauss-Makov ass.
- What next:

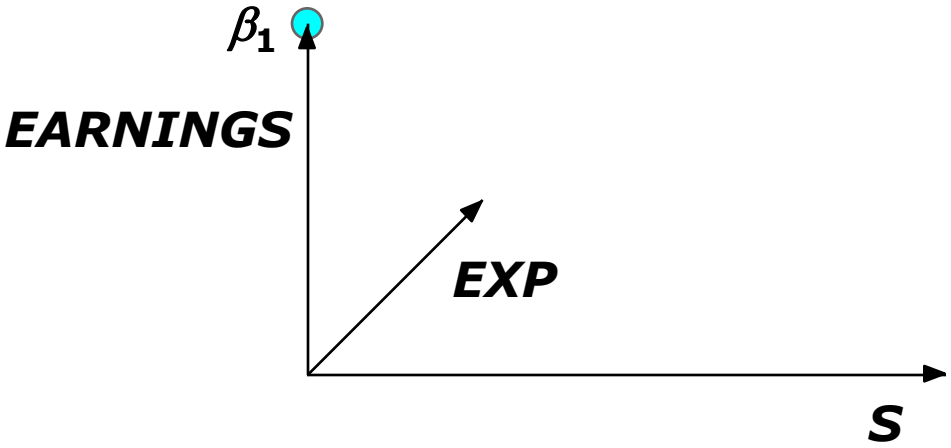
  Estimation of multiple regression model

MULTIPLE REGRESSION ANALYSIS: ESTIMATION

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES: EXAMPLE

$$EARNINGS = \beta_1 + \beta_2 S + \beta_3 EXP + u$$

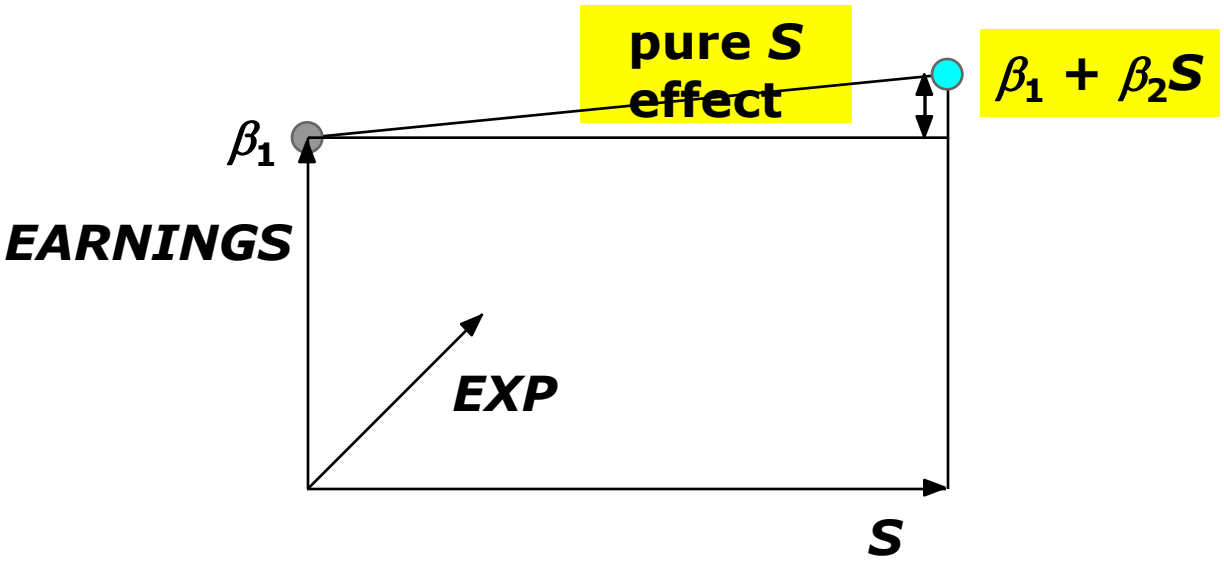
The model has three dimensions, one each for *EARNINGS, S*, and *EXP*. The starting point for investigating the determination of *EARNINGS* is the intercept, β_1.



Literally the intercept gives *EARNINGS* for those respondents who have no schooling and no work experience. However, there were no respondents with less than 6 years of schooling. Hence a literal interpretation of β_1 would be unwise.

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES: EXAMPLE

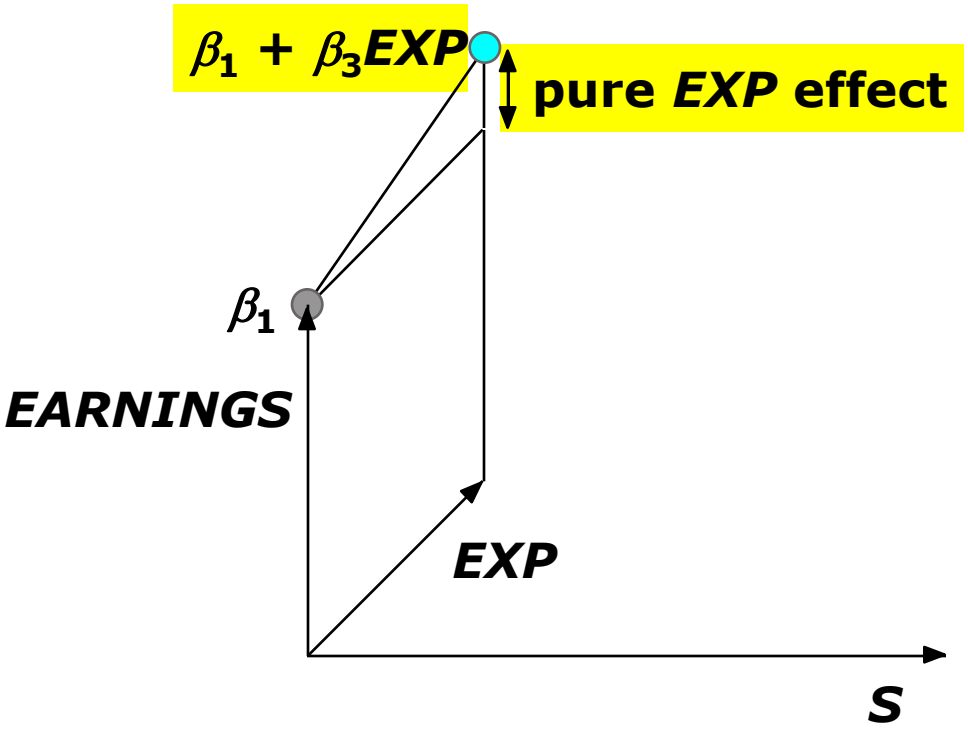
$$EARNINGS = \beta_1 + \beta_2 S + \beta_3 EXP + u$$



The next term on the right side of the equation gives the effect of variations in *S*. A one year increase in *S* causes *EARNINGS* to increase by β_2 dollars, holding *EXP* constant.

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES: EXAMPLE

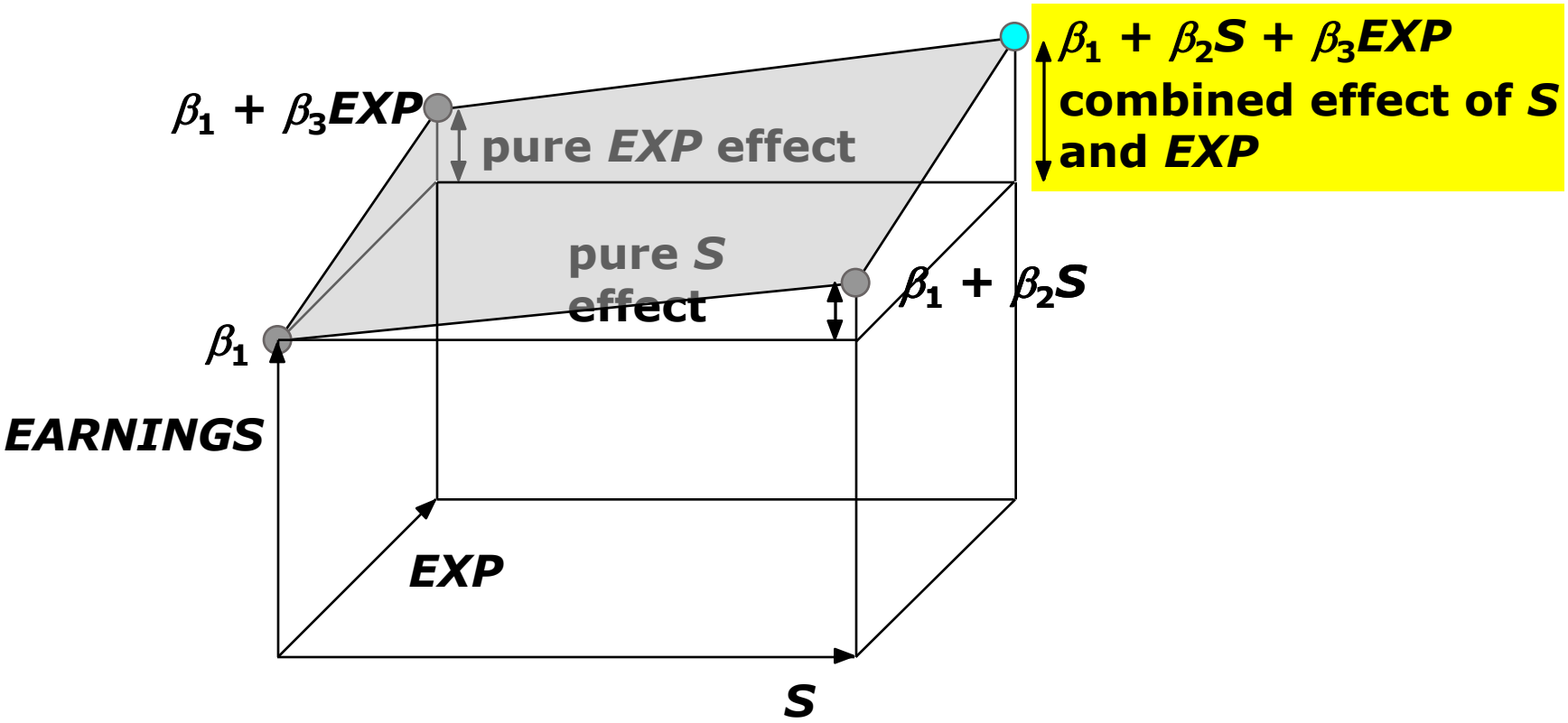
$$EARNINGS = \beta_1 + \beta_2 S + \beta_3 EXP + u$$



Similarly, the third term gives the effect of variations in *EXP*. A one year increase in *EXP* causes earnings to increase by β_3 dollars, holding *S* constant.

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES: EXAMPLE

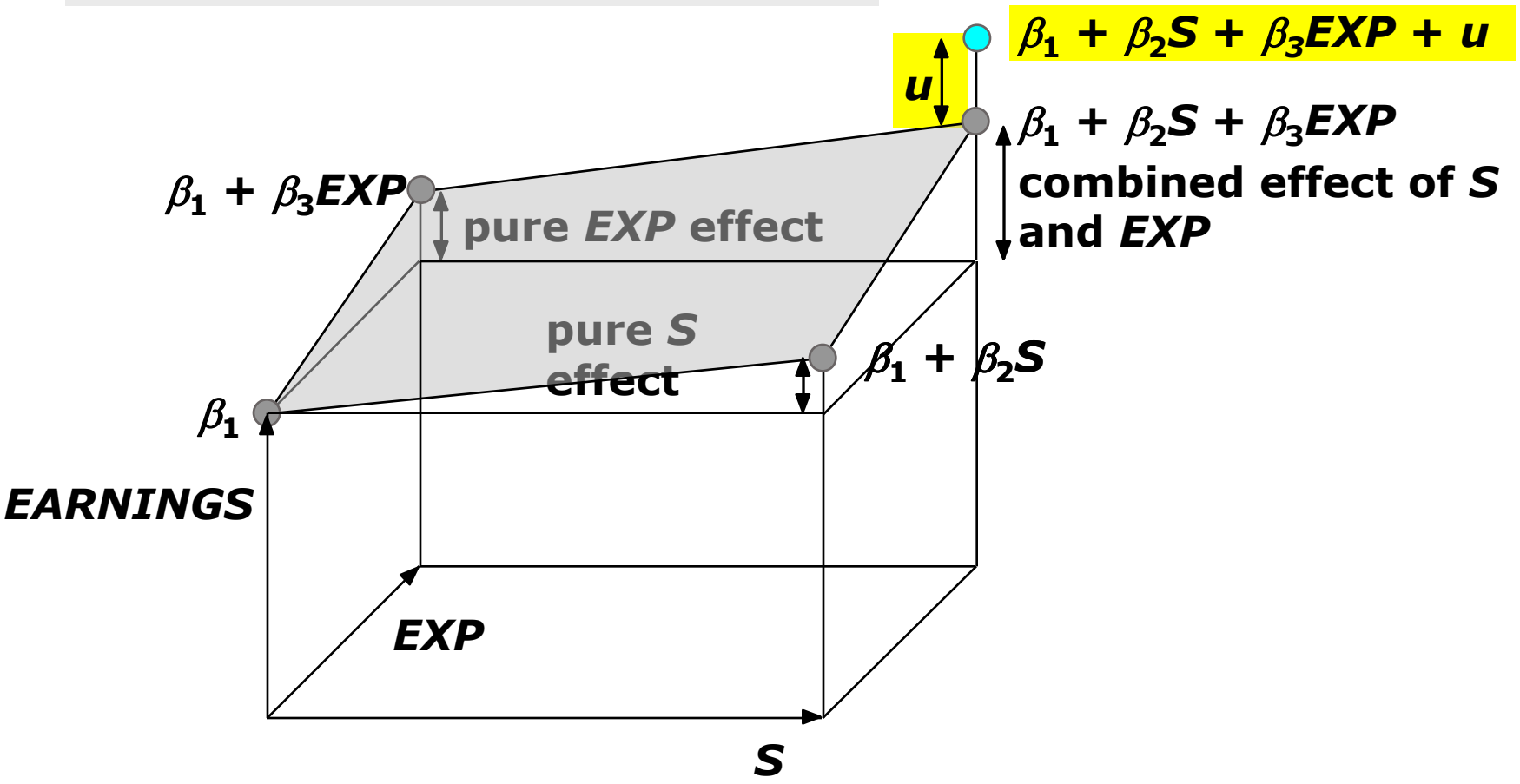
$$EARNINGS = \beta_1 + \beta_2 S + \beta_3 EXP + u$$



Different combinations of *S* and *EXP* give rise to values of *EARNINGS* which lie on the plane shown in the diagram, defined by the equation $EARNINGS = \beta_1 + \beta_2 S + \beta_3 EXP$.

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES: EXAMPLE

$$EARNINGS = \beta_1 + \beta_2 S + \beta_3 EXP + u$$



The final element of the model is the disturbance term, *u*. This causes the actual values of *EARNINGS* to deviate from the plane. In this observation, *u* happens to have a positive value.

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES: EXAMPLE

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

$$\hat{Y}_i = b_1 + b_2 X_{2i} + b_3 X_{3i}$$

$$e_i = Y_i - \hat{Y}_i = Y_i - b_1 - b_2 X_{2i} - b_3 X_{3i}$$

$$RSS = \sum e_i^2 = \sum (Y_i - b_1 - b_2 X_{2i} - b_3 X_{3i})^2$$

The regression coefficients are derived using the same least squares principle used in simple regression analysis. The fitted value of *Y* in observation *i* depends on our choice of b_1, b_2, and b_3.

The residual e_i in observation *i* is the difference between the actual and fitted values of *Y*.

We define *RSS*, the sum of the squares of the residuals, and choose b_1, b_2, and b_3 so as to minimize it, using first order condition.

MULTIPLE REGRESSION WITH TWO EXPLANATORY VARIABLES: EXAMPLE

```
. reg EARNINGS S EXP

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  2,   537) =   67.54
       Model |  22513.6473     2  11256.8237            Prob > F      =  0.0000
    Residual |  89496.5838   537  166.660305            R-squared     =  0.2010
-------------+------------------------------           Adj R-squared =  0.1980
       Total |  112010.231   539  207.811189            Root MSE      =   12.91

------------------------------------------------------------------------------
    EARNINGS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |    2.678125   .2336497    11.46   0.000     2.219146    3.137105
         EXP |    .5624326   .1285136     4.38   0.000     .3099816    .8148837
       _cons |   -26.48501    4.27251    -6.20   0.000    -34.87789   -18.09213
------------------------------------------------------------------------------
```

$$\widehat{EARNINGS} = -26.49 + 2.68\,S + 0.56\,EXP$$

It indicates that earnings increase by $2.68 for every extra year of schooling and by $0.56 for every extra year of work experience.

PROPERTIES OF THE MULTIPLE REGRESSION COEFFICIENTS

A.1: ***The model is linear in parameters and correctly specified.***

$$Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k + u$$

A.2: ***There does not exist an exact linear relationship among the regressors in the sample.***

A.3 ***The disturbance term has zero expectation***

A.4 ***The disturbance term is homoskedastic***

A.5 ***The values of the disturbance term have independent distributions***

A.6 ***The disturbance term has a normal distribution***

Provided that the regression model assumptions are valid, the OLS estimators in the multiple regression model are unbiased and efficient, as in the simple regression model.

MULTICOLLINEARITY

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad X_3 = \lambda + \mu X_2$$

What would happen if you tried to run a regression when there is an exact linear relationship among the explanatory variables? The coefficient is not defined

$$EARNINGS = \beta_1 + \beta_2 S + \beta_3 EXP + \beta_4 EXPSQ + u$$

For example, when relating earnings to schooling and work experience, it if often reasonable to suppose that the effect of work experience is subject to diminishing returns. β_4 should be negative

MULTICOLLINEARITY

```
. reg EARNINGS S EXP EXPSQ

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  3,   536) =   45.57
       Model |  22762.4472     3  7587.48241            Prob > F      =  0.0000
    Residual |  89247.7839   536  166.507059            R-squared     =  0.2032
-------------+------------------------------           Adj R-squared =  0.1988
       Total |  112010.231   539  207.811189            Root MSE      =  12.904

------------------------------------------------------------------------------
    EARNINGS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |    2.754372   .2417286    11.39   0.000     2.279521    3.229224
         EXP |   -.2353907    .665197    -0.35   0.724    -1.542103    1.071322
       EXPSQ |    .0267843   .0219115     1.22   0.222    -.0162586    .0698272
       _cons |   -22.21964   5.514827    -4.03   0.000    -33.05297   -11.38632
------------------------------------------------------------------------------
```

$$EARNINGS = \beta_1 + \beta_2 S + \beta_3 EXP + \beta_4 EXPSQ + u$$

The schooling component of the regression results is not much affected by the inclusion of the *EXPSQ* term. Another year of schooling increases earnings by 2.75 usd

By contrast, the inclusion of the new term has had a dramatic effect on the coefficient of *EXP*. Now it is negative, which makes little sense, and insignificant.

MULTICOLLINEARITY

```
. reg EARNINGS S EXP EXPSQ

------------------------------------------------------------------------------
    EARNINGS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |    2.754372   .2417286    11.39   0.000     2.279521    3.229224
         EXP |   -.2353907    .665197    -0.35   0.724    -1.542103    1.071322
       EXPSQ |    .0267843   .0219115     1.22   0.222    -.0162586    .0698272
       _cons |   -22.21964   5.514827    -4.03   0.000    -33.05297   -11.38632
------------------------------------------------------------------------------

. reg EARNINGS S EXP

------------------------------------------------------------------------------
    EARNINGS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |    2.678125   .2336497    11.46   0.000     2.219146    3.137105
         EXP |    .5624326   .1285136     4.38   0.000     .3099816    .8148837
       _cons |   -26.48501    4.27251    -6.20   0.000    -34.87789   -18.09213
------------------------------------------------------------------------------
```

The high correlation causes the standard error of *EXP* to be larger than it would have been if *EXP* and *EXPSQ* had been less highly correlated, warning us that the point estimate is unreliable.

When high correlations among the explanatory variables lead to erratic point estimates of the coefficients, large standard errors and unsatisfactorily low *t* statistics, the regression is said to be suffering from multicollinearity.

MULTICOLLINEARITY

```
. reg EARNINGS S EXP

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  2,   537) =   67.54
       Model |  22513.6473     2  11256.8237            Prob > F      =  0.0000
    Residual |  89496.5838   537  166.660305            R-squared     =  0.2010
-------------+------------------------------           Adj R-squared =  0.1980
       Total |  112010.231   539  207.811189            Root MSE      =   12.91

------------------------------------------------------------------------------
    EARNINGS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |    2.678125   .2336497    11.46   0.000     2.219146    3.137105
         EXP |    .5624326   .1285136     4.38   0.000     .3099816    .8148837
       _cons |  -26.48501    4.27251    -6.20   0.000    -34.87789   -18.09213
------------------------------------------------------------------------------
```

$$EARNINGS = \beta_1 + \beta_2 S + \beta_3 EXP + \beta_4 EXPSQ + u$$

In the specification without *EXPSQ* it is 2.68, not much different.

But experience is positive and highly significant.

MULTIPLE REGRESSION ANALYSIS: INFERENCE

TESTING A HYPOTHESIS RELATING TO A REGRESSION COEFFICIENT

Model: $Y = \beta_1 + \beta_2 X + u$

Null hypothesis: $H_0 : \beta_2 = \beta_2^0$

Alternative hypothesis: $H_1 : \beta_2 \neq \beta_2^0$

We will suppose that we have the standard simple regression model and that we wish to test the hypothesis H_0 that the slope coefficient is equal to some value β_2^0. We test it against the alternative hypothesis H_1, which is simply that β_2 is not equal to β_2^0

TESTING A HYPOTHESIS RELATING TO A REGRESSION COEFFICIENT

Model: $Y = \beta_1 + \beta_2 X + u$

Null hypothesis: $H_0 : \beta_2 = \beta_2^0$

Alternative hypothesis: $H_1 : \beta_2 \neq \beta_2^0$

Example model: $p = \beta_1 + \beta_2 w + u$

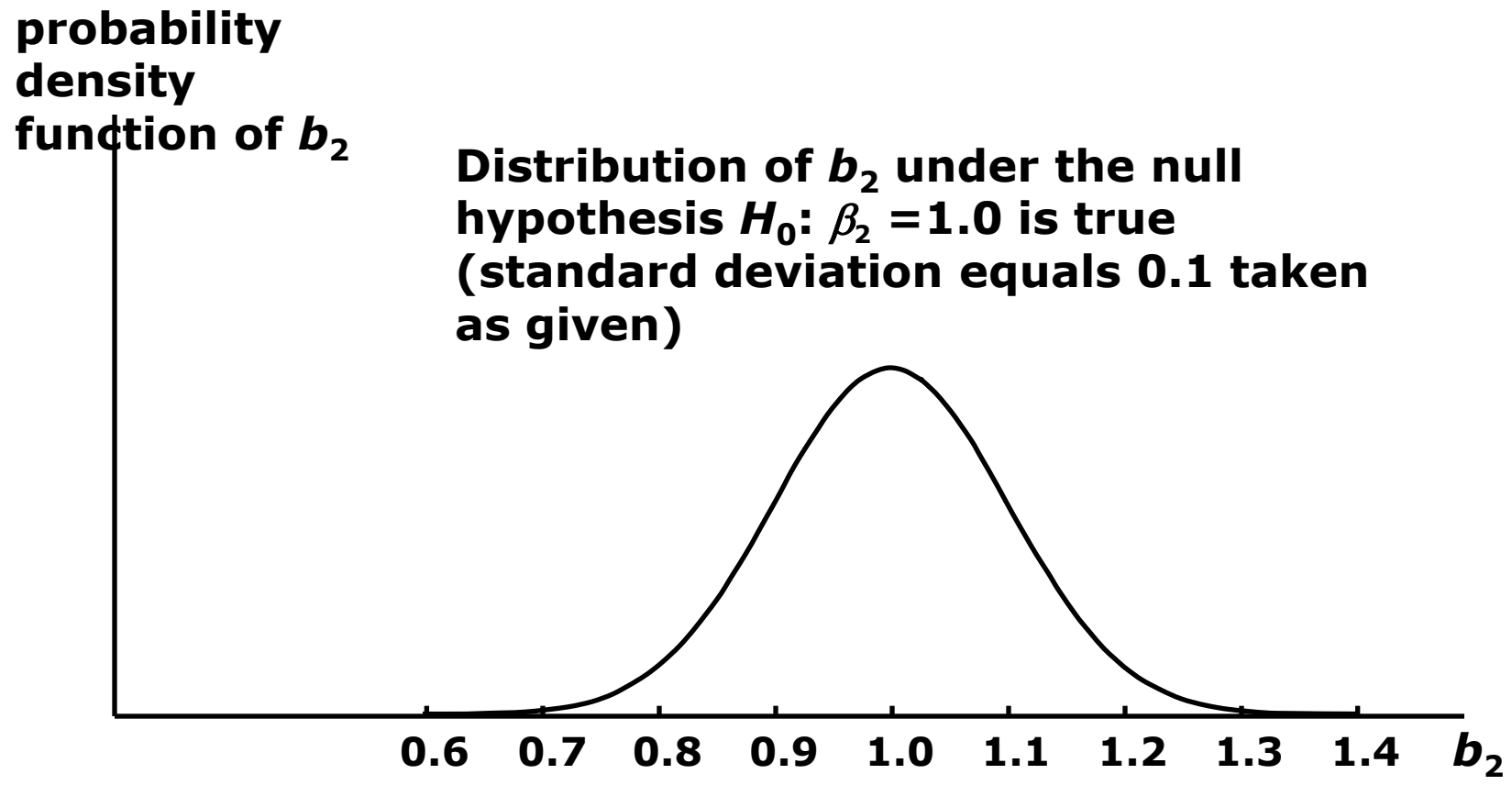
Null hypothesis: $H_0 : \beta_2 = 1.0$

Alternative hypothesis: $H_1 : \beta_2 \neq 1.0$

As an illustration, we will consider a model relating price inflation to wage inflation. *p* is the rate of growth of prices and *w* is the rate of growth of wages.

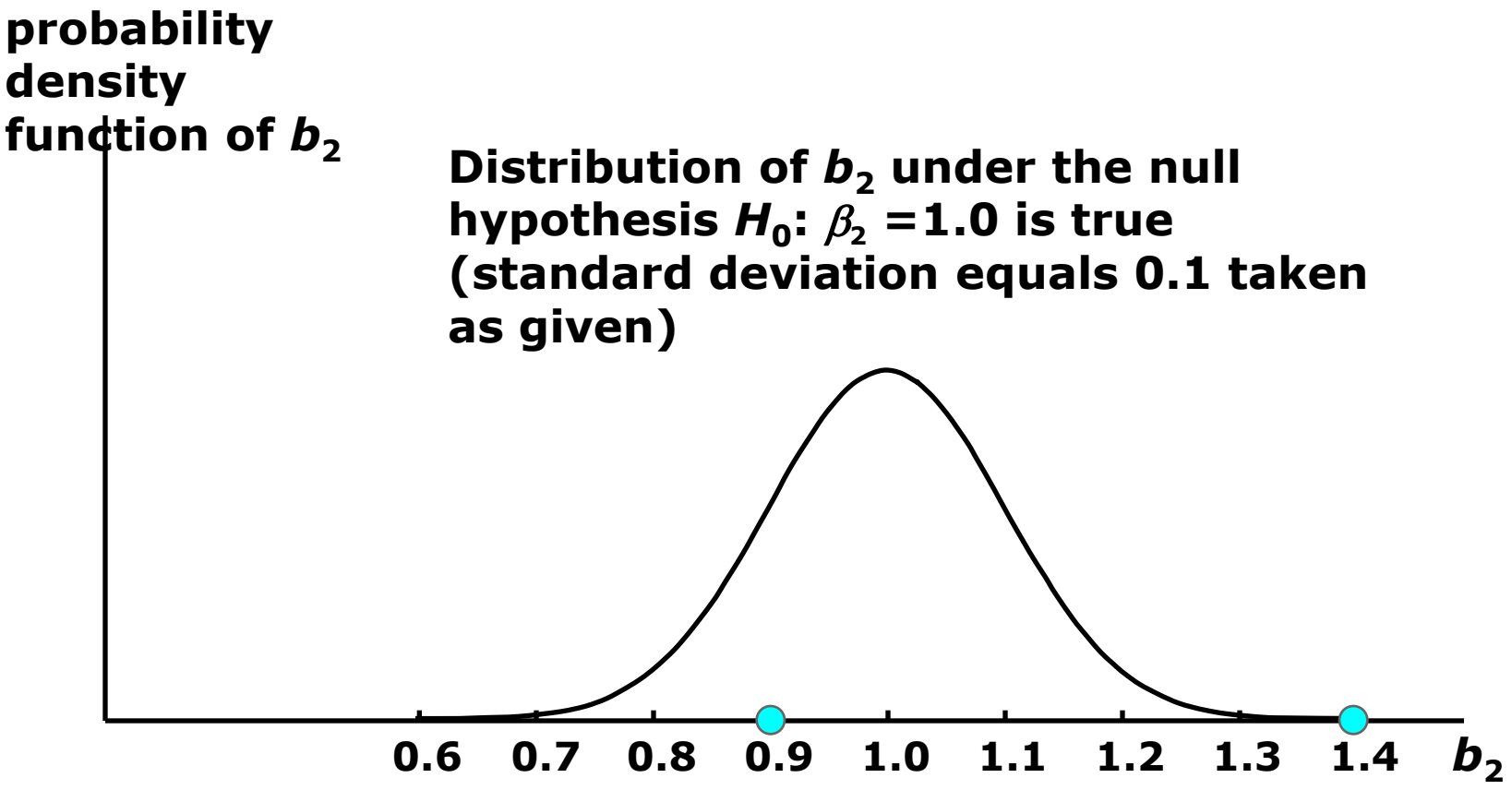
We will test the hypothesis that the rate of price inflation is equal to the rate of wage inflation. The null hypothesis is therefore H_0: $\beta_2 = 1.0$.

TESTING A HYPOTHESIS RELATING TO A REGRESSION COEFFICIENT



We will assume that we know the standard deviation and that it is equal to 0.1. This is a very unrealistic assumption. In practice you have to estimate it.

TESTING A HYPOTHESIS RELATING TO A REGRESSION COEFFICIENT



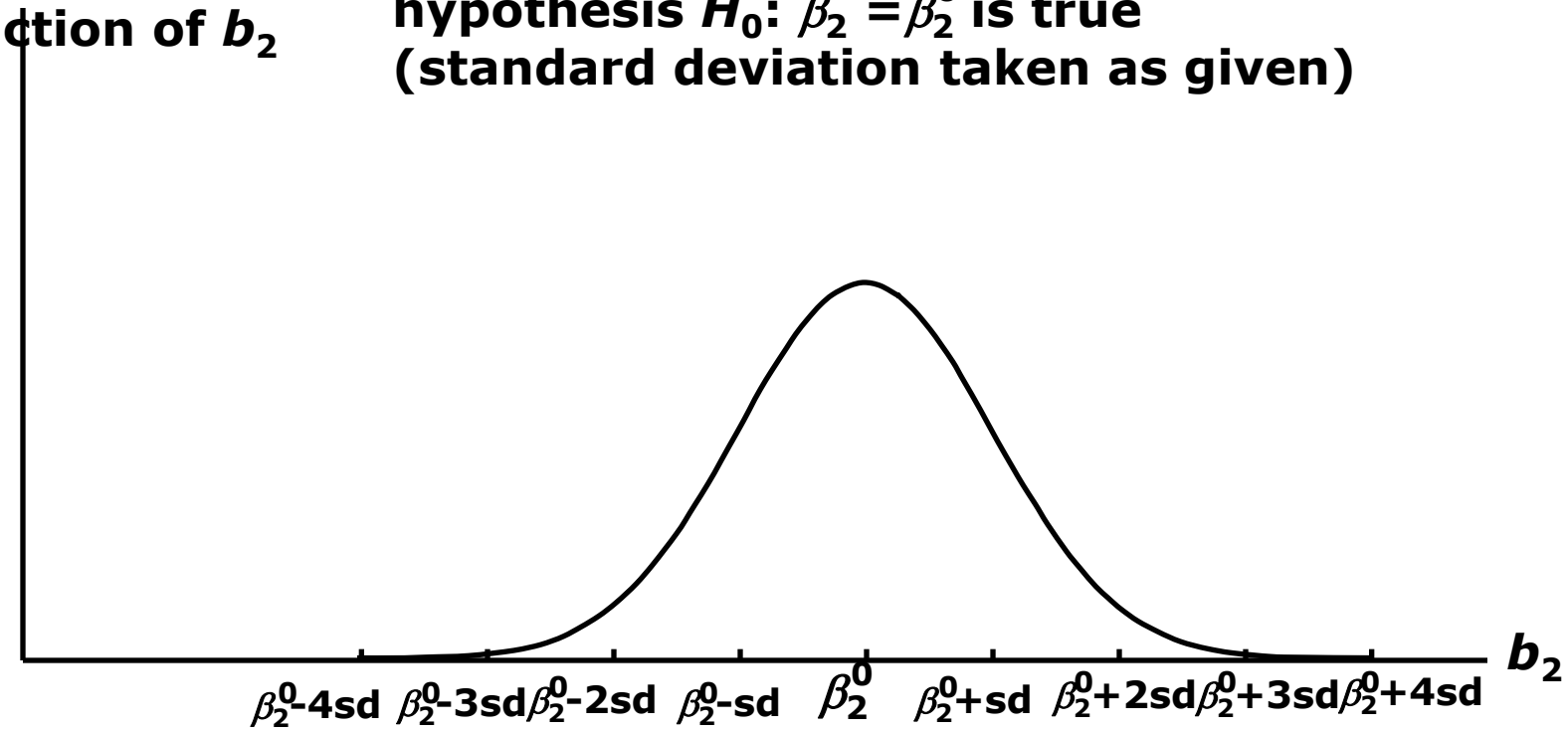
Suppose that we have a sample of data for the price inflation/wage inflation model and the estimate of the slope coefficient, b_2, is 0.9. Would this be evidence against the null hypothesis β_2 = 1.0?

And what if b_2 =1.4?

TESTING A HYPOTHESIS RELATING TO A REGRESSION COEFFICIENT

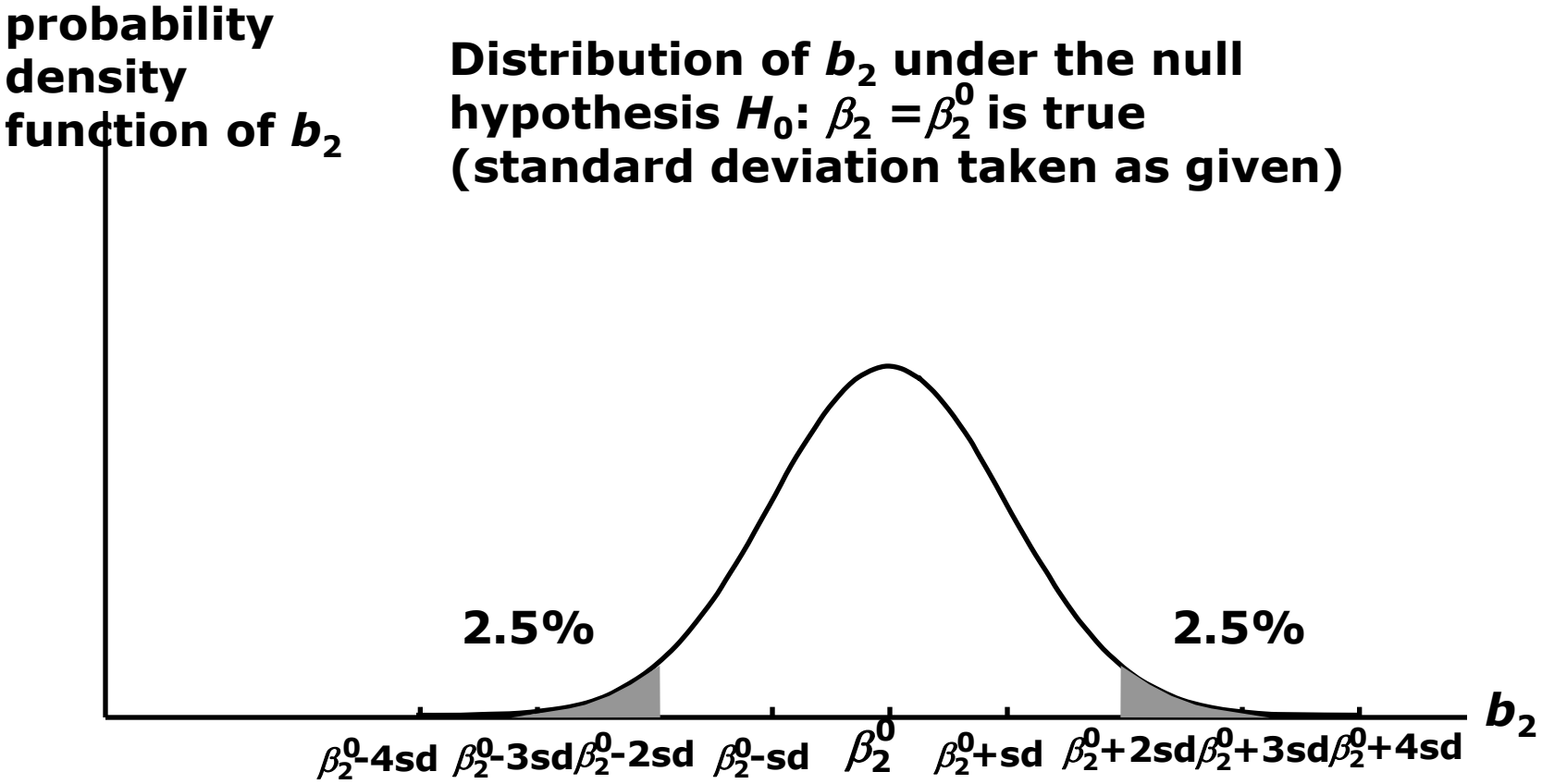
probability density function of b_2

Distribution of b_2 under the null hypothesis H_0: $\beta_2 = \beta_2^0$ is true (standard deviation taken as given)



The usual procedure for making decisions is to reject the null hypothesis if it implies that the probability of getting such an extreme estimate is less than some (small) probability *p*.

TESTING A HYPOTHESIS RELATING TO A REGRESSION COEFFICIENT



For example, we might choose to reject the null hypothesis if it implies that the probability of getting such an extreme estimate is less than 0.05 (5%).

According to this decision rule, we would reject the null hypothesis if the estimate fell in the upper or lower 2.5% tails.

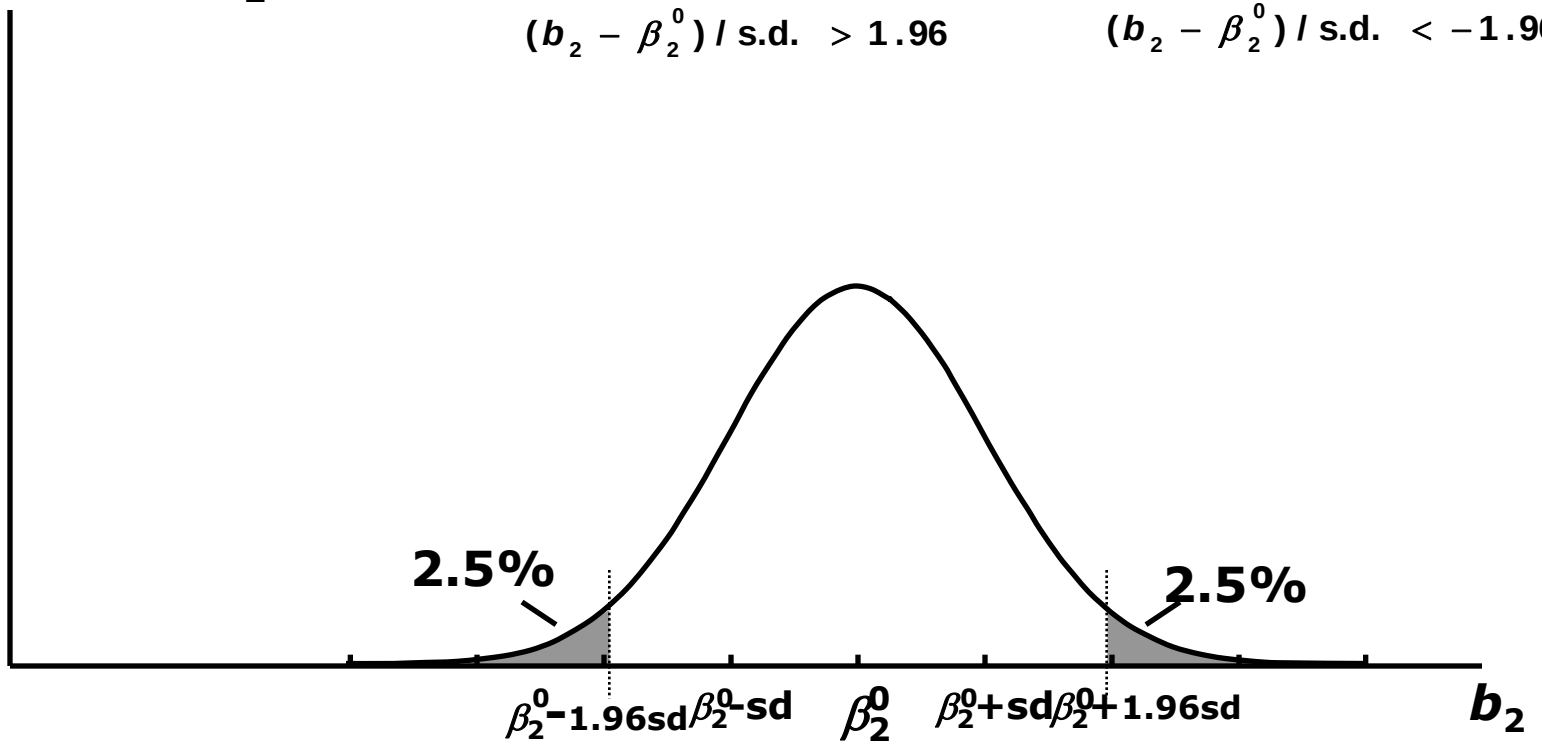
TESTING A HYPOTHESIS RELATING TO A REGRESSION COEFFICIENT

Decision rule (5% significance level):

reject $H_0 : \beta_2 = \beta_2^0$

(1) if $b_2 > \beta_2^0 + 1.96$ s.d. $\quad (b_2 - \beta_2^0) / \text{s.d.} > 1.96$

(2) if $b_2 < \beta_2^0 - 1.96$ s.d. $\quad (b_2 - \beta_2^0) / \text{s.d.} < -1.96$



The 2.5% tails of a normal distribution always begin 1.96 standard deviations from its mean. Thus we would reject H_0 if the estimate were 1.96 standard deviations (or more) above or below the hypothetical mean.

Or if the difference, expressed in terms of standard deviations, were more than 1.96 in absolute terms (positive or negative).

TESTING A HYPOTHESIS RELATING TO A REGRESSION COEFFICIENT

Decision rule (5% significance level):

reject $H_0 : \beta_2 = \beta_2^0$

(1) if $b_2 > \beta_2^0 + 1.96$ s.d. **(2) if** $b_2 < \beta_2^0 - 1.96$ s.d.

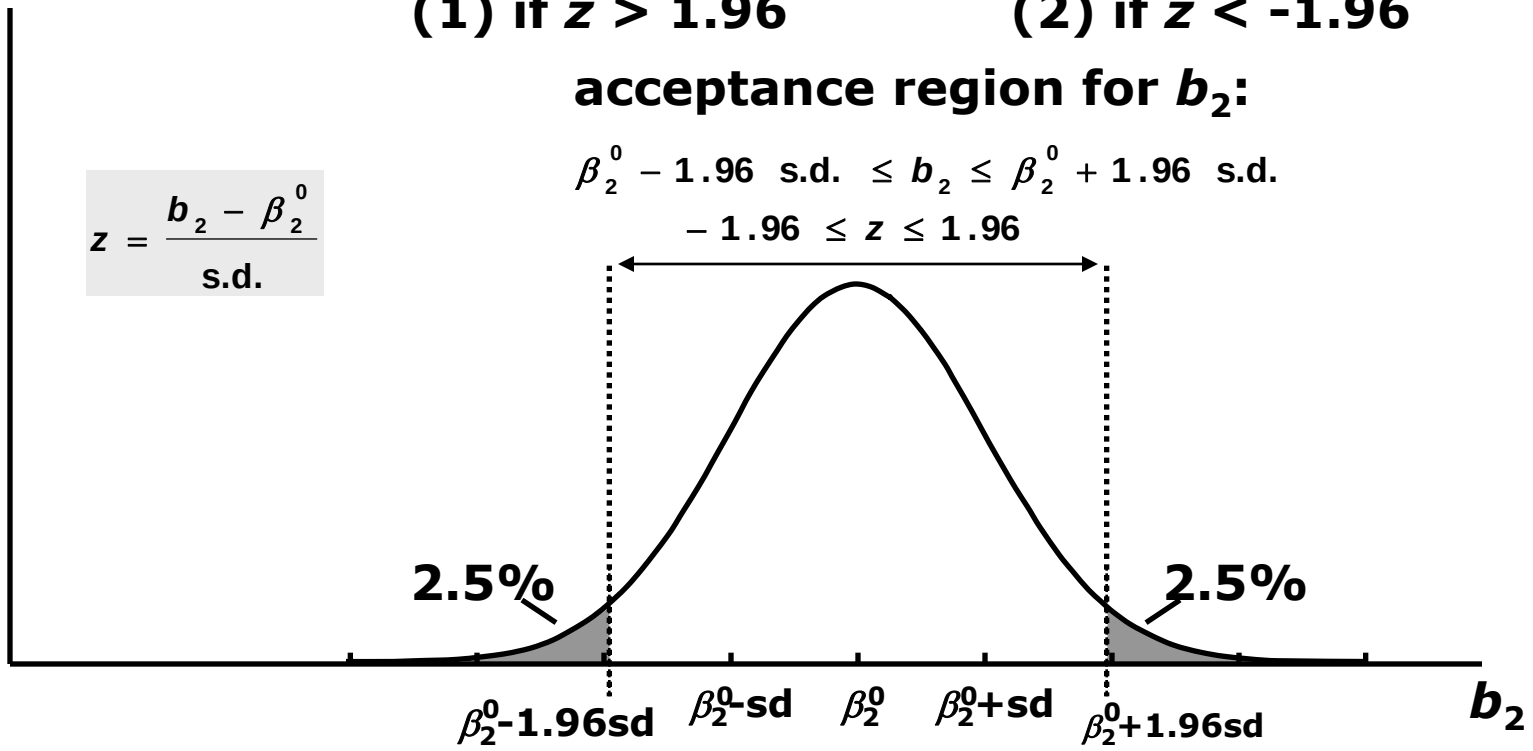
(1) if $z > 1.96$ **(2) if** $z < -1.96$

acceptance region for b_2:

$$\beta_2^0 - 1.96 \text{ s.d.} \leq b_2 \leq \beta_2^0 + 1.96 \text{ s.d.}$$

$$-1.96 \leq z \leq 1.96$$

$$z = \frac{b_2 - \beta_2^0}{\text{s.d.}}$$



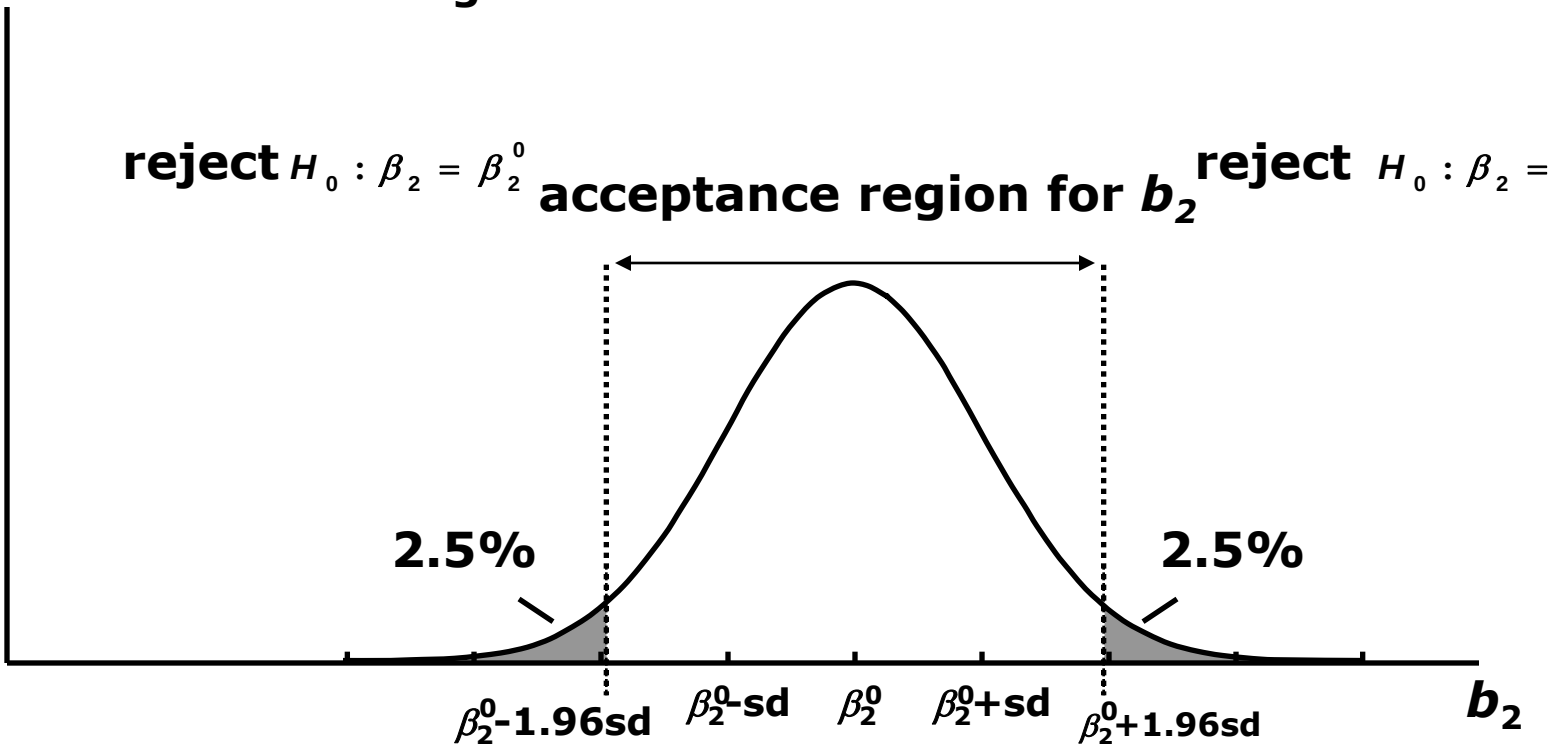
The range of values of b_2 that do not lead to the rejection of the null hypothesis is known as the acceptance region.

The limiting values of z for the acceptance region are 1.96 and -1.96 (for a 5% significance test).

TESTING A HYPOTHESIS RELATING TO A REGRESSION COEFFICIENT

probability density function of b_2

Type I error: rejection of H_0 when it is in fact true.
Probability of Type I error: in this case, 5%
Significance level of the test is 5%.

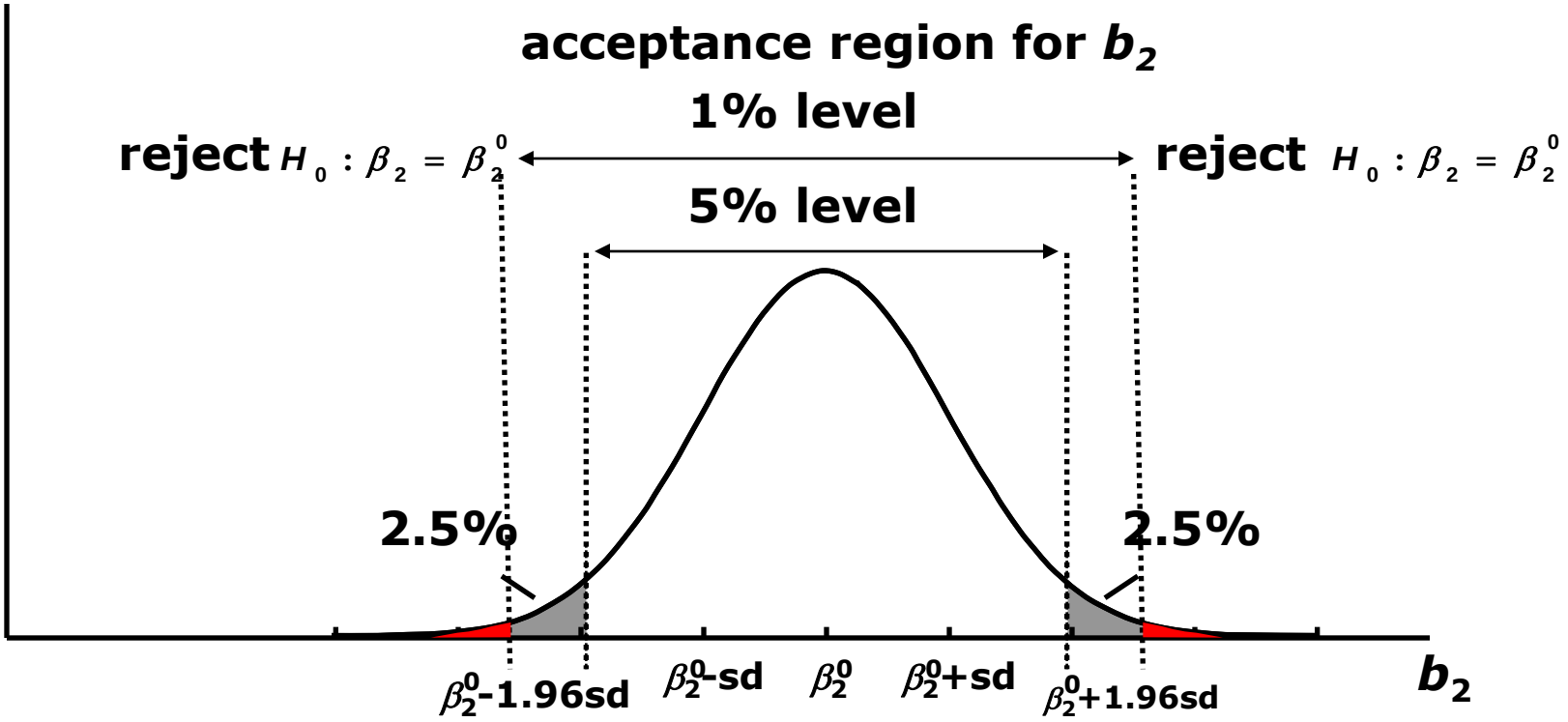


Rejection of the null hypothesis when it is in fact true is described as a Type I error.

With the present test, if the null hypothesis is true, a Type I error will occur 5% of the time because 5% of the time we will get estimates in the upper or lower 2.5% tails. The significance level of a test is defined to be the probability of making a Type I error if the null hypothesis is true.

TESTING A HYPOTHESIS RELATING TO A REGRESSION COEFFICIENT

probability density function of b_2



We could change the decision rule to "reject the null hypothesis if it implies that the probability of getting the sample estimate is less than 0.01 (1%)".

The rejection region now becomes the upper and lower 0.5% tails

t TEST OF A HYPOTHESIS RELATING TO A REGRESSION COEFFICIENT

s.d. of b_2 known

discrepancy between hypothetical value and sample estimate, in terms of s.d.:

$$z = \frac{b_2 - \beta_2^0}{\text{s.d.}}$$

5% significance test:

reject H_0: $\beta_2 = \beta_2^0$ if

$z > 1.96$ or $z < -1.96$

s.d. of b_2 not known

discrepancy between hypothetical value and sample estimate, in terms of s.e.:

$$t = \frac{b_2 - \beta_2^0}{\text{s.e.}}$$

5% significance test:

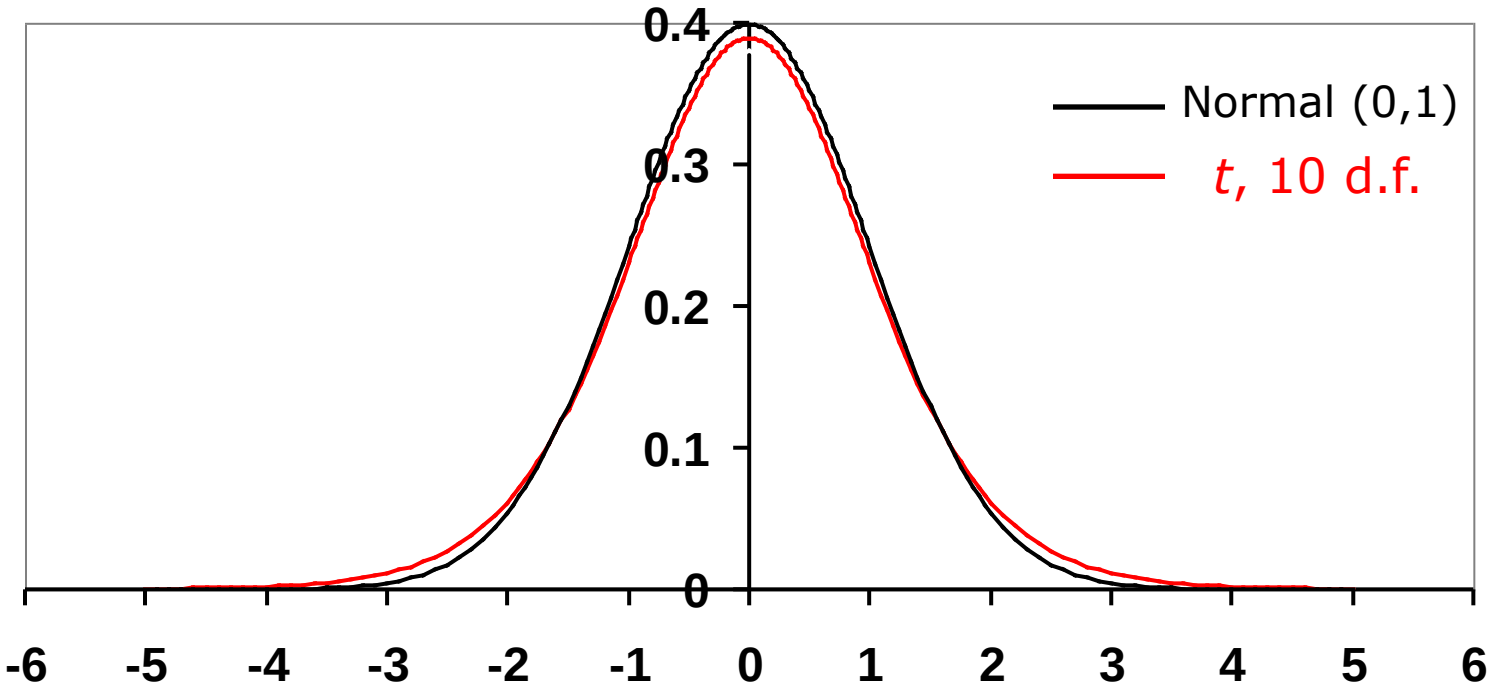
reject H_0: $\beta_2 = \beta_2^0$ if

$t > t_{crit}$ or $t < -t_{crit}$

We replace the standard deviation in its denominator with the standard error, the test statistic has a *t* distribution instead of a normal distribution.

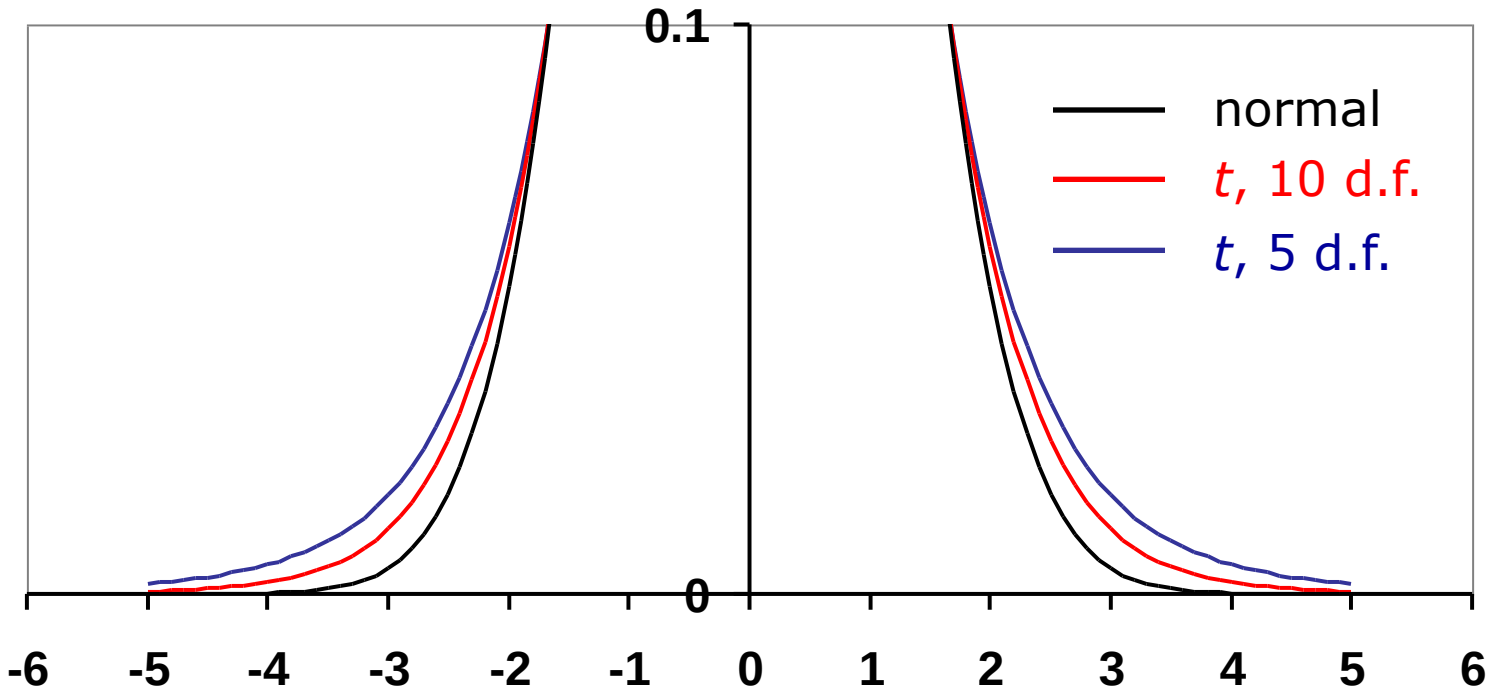
We look up the critical value of *t* and if the *t* statistic is greater than it, positive or negative, we reject the null hypothesis. If it is not, we do not.

t TEST OF A HYPOTHESIS RELATING TO A REGRESSION COEFFICIENT



A graph of a *t* distribution with 10 degrees of freedom. When the number of degrees of freedom is large, the *t* distribution looks very much like a normal distribution

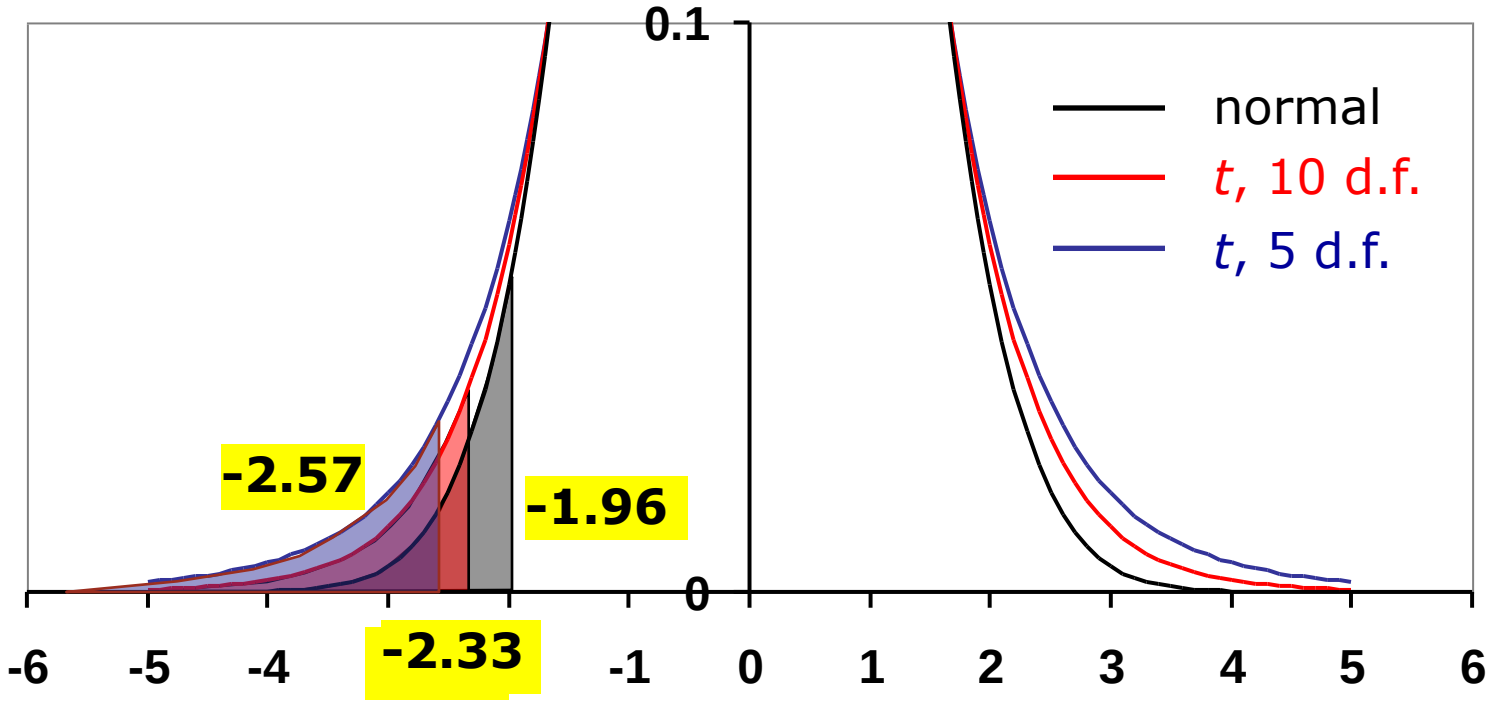
t TEST OF A HYPOTHESIS RELATING TO A REGRESSION COEFFICIENT



***t* distribution has longer tails than the normal distribution, the difference being the greater, the smaller the number of degrees of freedom**

This means that the rejection regions have to start more standard deviations away from zero for a *t* distribution than for a normal distribution.

t TEST OF A HYPOTHESIS RELATING TO A REGRESSION COEFFICIENT



The 2.5% tail of a *t* distribution with 10 degrees of freedom starts 2.33 standard deviations from its mean.

That for a *t* distribution with 5 degrees of freedom starts 2.57 standard deviations from its mean.

t TEST OF A HYPOTHESIS RELATING TO A REGRESSION COEFFICIENT

t Distribution: Critical values of *t*

Degrees of freedom	Two-sided test / One-sided test	10% / 5%	5% / 2.5%	2% / 1%	1% / 0.5%	0.2% / 0.1%	0.1% / 0.05%
1		6.314	12.706	31.821	63.657	318.31	636.62
2		2.920	4.303	6.965	9.925	22.327	31.598
3		2.353	3.182	4.541	5.841	10.214	12.924
4		2.132	2.776	3.747	4.604	7.173	8.610
5		2.015	2.571	3.365	4.032	5.893	6.869
...		...	...	...	...	...	...
...		...	...	...	...	...	...
18		1.734	2.101	2.552	2.878	3.610	3.922
19		1.729	2.093	2.539	2.861	3.579	3.883
20		1.725	2.086	2.528	2.845	3.552	3.850
...		...	...	...	...	...	...
...		...	...	...	...	...	...
600		1.647	1.964	2.333	2.584	3.104	3.307
∞		1.645	1.960	2.326	2.576	3.090	3.291

For this reason we need to refer to a table of critical values of *t* when performing significance tests on the coefficients of a regression equation.

Number of degrees of freedom in a regression

= number of observations − number of parameters estimated.

t TEST OF A HYPOTHESIS RELATING TO A REGRESSION COEFFICIENT

Example:

$$p = \beta_1 + \beta_2 w + u$$

$$H_0 : \beta_2 = 1; \quad H_0 : \beta_2 \neq 1$$

$$\hat{p} = 1.21 + 0.82$$
$$(0.05)\ (0.10)$$

$$t = \frac{b_2 - \beta_2^0}{\text{s.e.}\ (b_2)} = \frac{0.82 - 1.00}{0.10} = -1.80 .$$

$$n = 20; \quad \text{degrees of freedom} = 18$$

$$t_{\text{crit}, 5\%} = 2.101$$

The critical value of *t* with 18 degrees of freedom is 2.101 at the 5% level. The absolute value of the *t* statistic is less than this, so we do not reject the null hypothesis.

t TEST OF A HYPOTHESIS RELATING TO A REGRESSION COEFFICIENT

```
. reg EARNINGS S

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  1,   538) =  112.15
       Model |  19321.5589     1  19321.5589            Prob > F      =  0.0000
    Residual |  92688.6722   538  172.283777           R-squared     =  0.1725
-------------+------------------------------           Adj R-squared =  0.1710
       Total |  112010.231   539  207.811189           Root MSE      =  13.126

------------------------------------------------------------------------------
    EARNINGS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |    2.455321   .2318512    10.59   0.000     1.999876    2.910765
       _cons |   -13.93347   3.219851    -4.33   0.000    -20.25849   -7.608444
------------------------------------------------------------------------------
```

You can see that the *t* statistic for the coefficient of *S* is enormous. We would reject the null hypothesis that schooling does not affect earnings at the 0.1% significance level without even looking at the table of critical values of *t*.

The next column in the output gives what are known as the *p* values for each coefficient. This is the probability of obtaining the corresponding *t* statistic as a matter of chance, if the null hypothesis H_0: $\beta = 0$ is true.

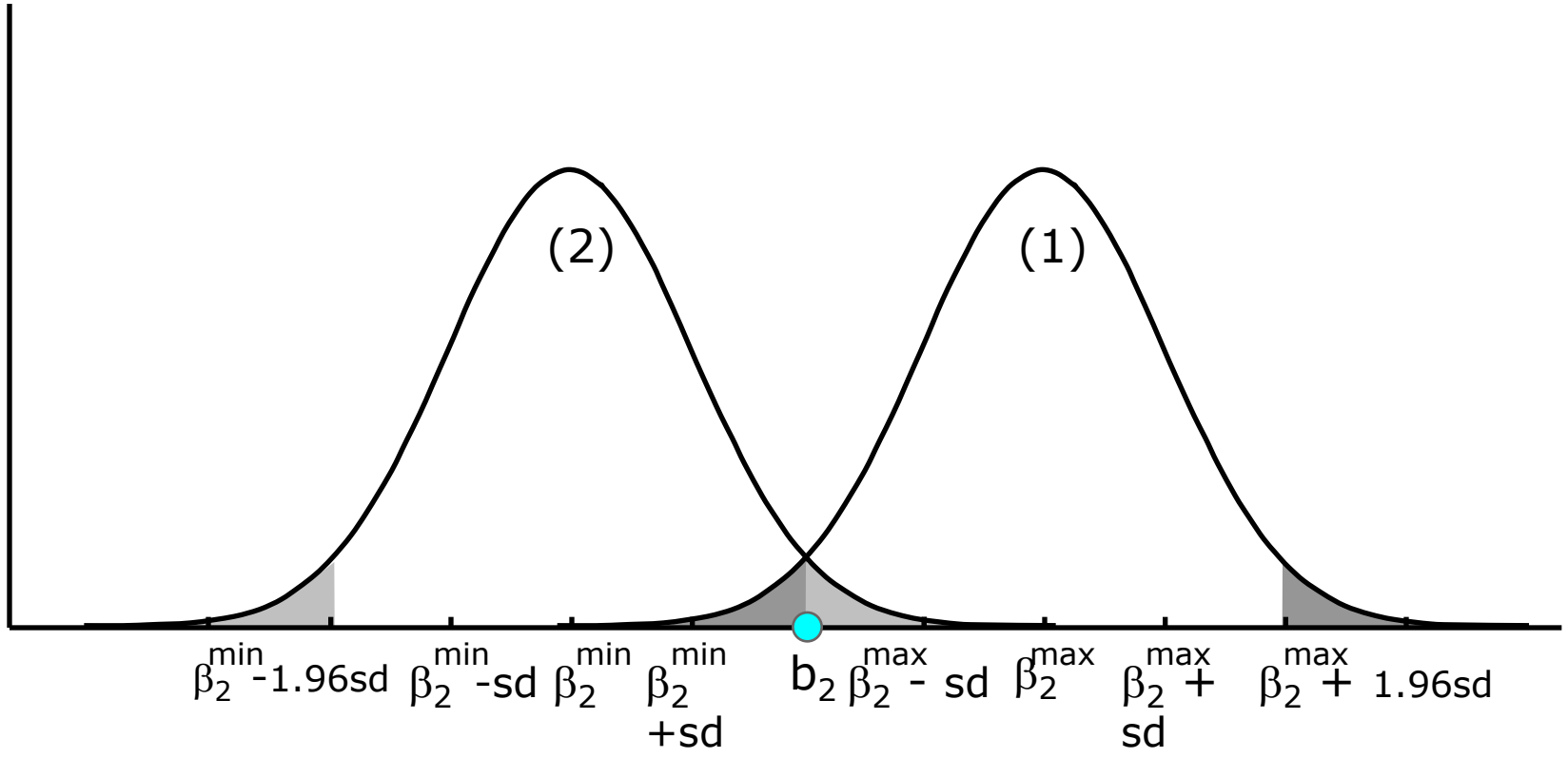
In the present case *p* = 0. This means that we can reject the null hypothesis H_0: $\beta_2 = 0$ at the 0.1% level.

CONFIDENCE INTERVALS

probability density function of b_2

(1) conditional on $\beta_2 = \beta_2^{max}$ being true

(2) conditional on $\beta_2 = \beta_2^{min}$ being true



The diagram shows the limiting values of the hypothetical values of β_2, together with their associated probability distributions for b_2.

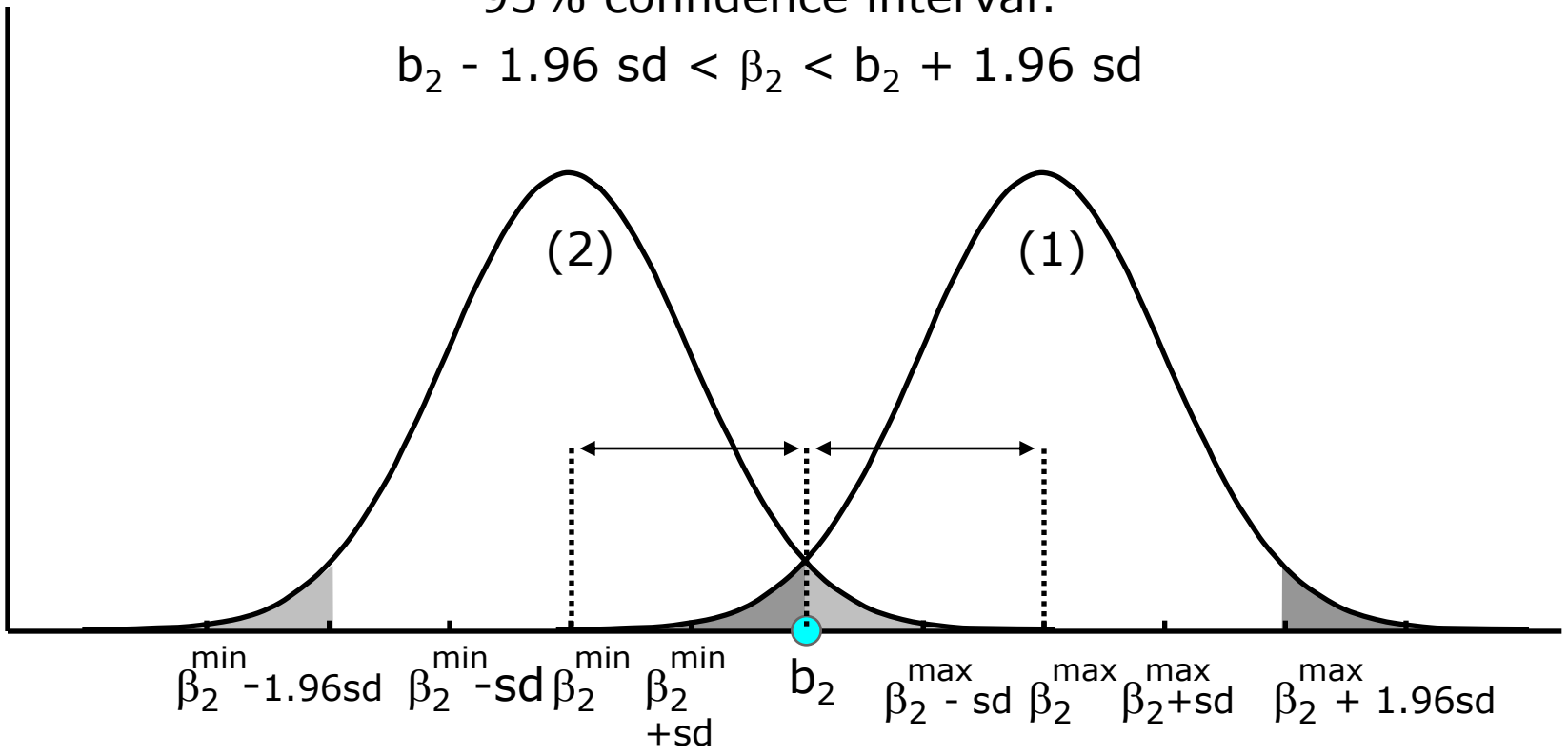
CONFIDENCE INTERVALS

reject any $\beta_2 > \beta_2^{max} = b_2 + 1.96\ sd$

reject any $\beta_2 < \beta_2^{min} = b_2 - 1.96\ sd$

95% confidence interval:

$b_2 - 1.96\ sd < \beta_2 < b_2 + 1.96\ sd$



Any hypothesis lying in the interval from β_2^{min} to β_2^{max} would be compatible with the sample estimate (not be rejected by it). We call this interval the 95% confidence interval.

CONFIDENCE INTERVALS

Standard deviation known

95% confidence interval

$$b_2 - 1.96\ sd \leqslant \beta_2 \leqslant b_2 + 1.96\ sd$$

99% confidence interval

$$b_2 - 2.58\ sd \leqslant \beta_2 \leqslant b_2 + 2.58\ sd$$

Standard deviation estimated by standard error

95% confidence interval

$$b_2 - t_{crit\ (5\%)}\ se \leqslant \beta_2 \leqslant b_2 + t_{crit\ (5\%)}\ se$$

99% confidence interval

$$b_2 - t_{crit\ (1\%)}\ se \leqslant \beta_2 \leqslant b_2 + t_{crit\ (1\%)}\ se$$

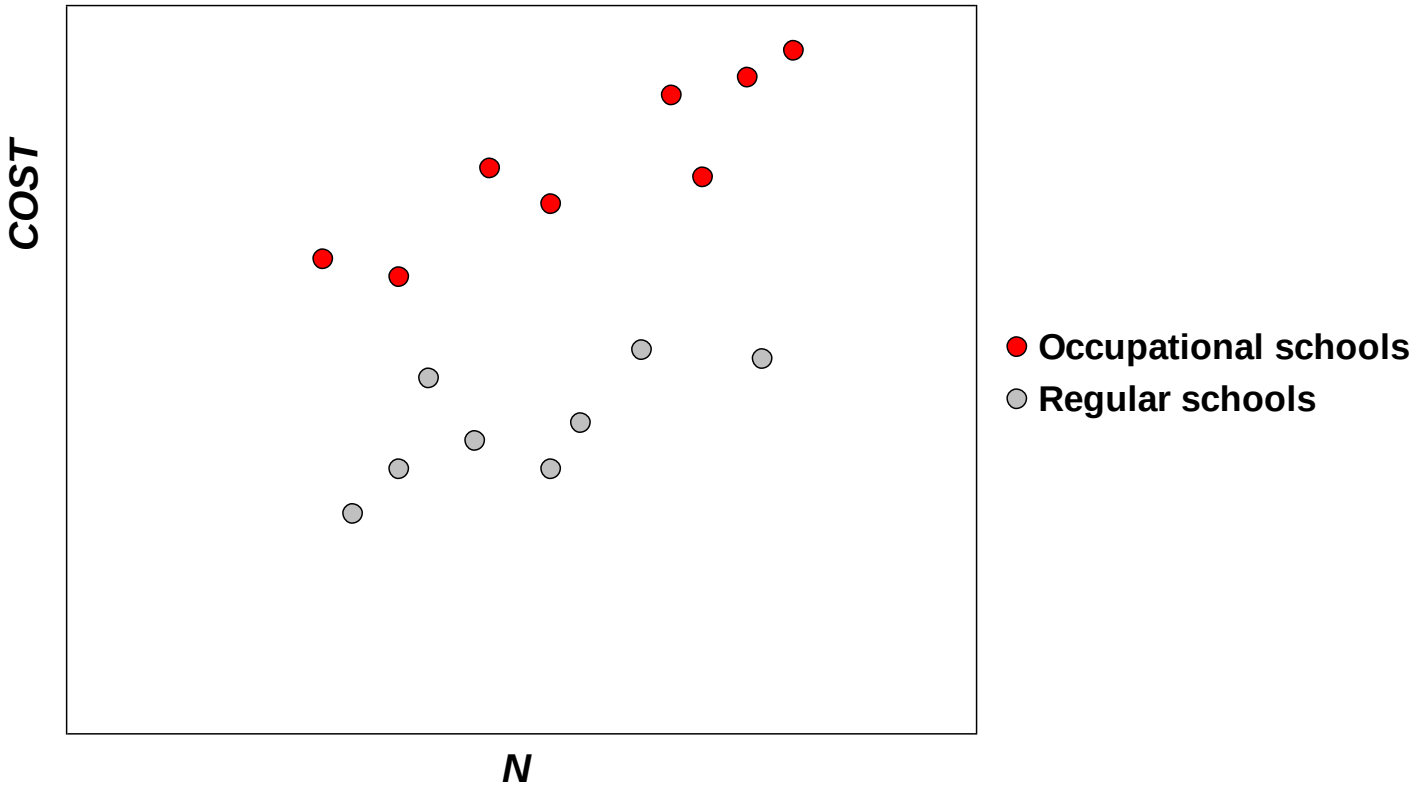
β_2^{min} and β_2^{max} will now be 2.58 standard deviations to the left and to the right of b_2 for 99% confidence interval

In practice, the t distribution has to be used instead of the normal distribution when locating β_2^{min} and β_2^{max}.

This implies that the standard error should be multiplied by the critical value of t, given the significance level and number of degrees of freedom, when determining the limits of the interval.

MULTIPLE REGRESSION WITH QUALITATIVE INFORMATION

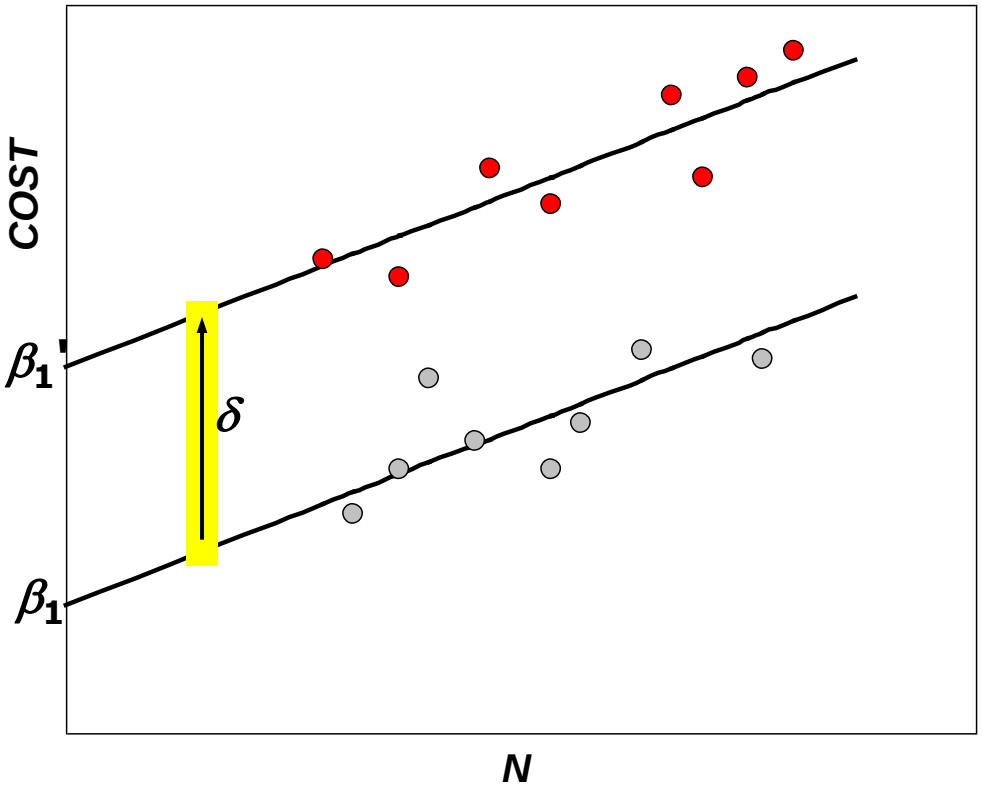
DUMMY VARIABLE CLASSIFICATION WITH TWO CATEGORIES



Suppose that you have data on the annual recurrent expenditure, *COST*, and the number of students enrolled, *N*, for a sample of secondary schools, of which there are two types: regular and occupational.

One way of dealing with the difference in the costs would be to run separate regressions for the two types of school. This is unadvisable.

DUMMY VARIABLE CLASSIFICATION WITH TWO CATEGORIES



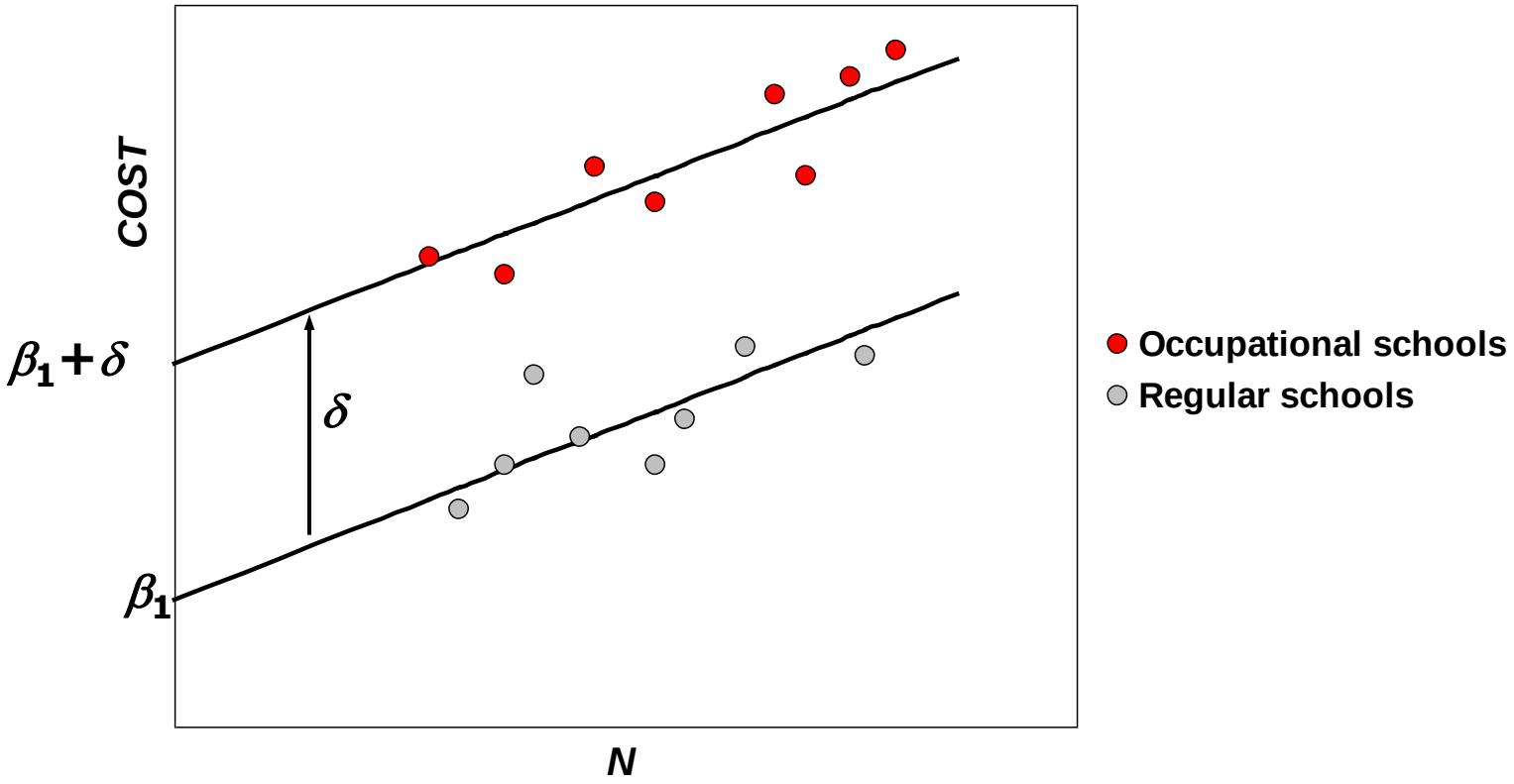
Regular school $COST = \beta_1 + \beta_2 N + u$

Occupational school $COST = \beta_1' + \beta_2 N + u$

Another way of handling the difference would be to hypothesize that the cost function for occupational schools has an intercept β_1' that is greater than that for regular schools, the marginal cost is the same.

Let us define δ to be the difference in the intercepts: $\delta = \beta_1' - \beta_1$

DUMMY VARIABLE CLASSIFICATION WITH TWO CATEGORIES



Combined equation $COST = \beta_1 + \delta OCC + \beta_2 N + u$

$OCC = 0$ **Regular school** $COST = \beta_1 + \beta_2 N + u$

$OCC = 1$ **Occupational school** $COST = \beta_1 + \delta + \beta_2 N + u$

Then $\beta_1' = \beta_1 + \delta$. We can now combine the two cost functions by defining a dummy variable *OCC* that has value 0 for regular schools and 1 for occupational schools.

DUMMY VARIABLE CLASSIFICATION WITH TWO CATEGORIES

```
. reg COST N OCC

  Source |       SS       df       MS              Number of obs =      74
---------+------------------------------           F(  2,    71) =   56.86
   Model |  9.0582e+11     2  4.5291e+11           Prob > F      =  0.0000
Residual |  5.6553e+11    71  7.9652e+09           R-squared     =  0.6156
---------+------------------------------           Adj R-squared =  0.6048
   Total |  1.4713e+12    73  2.0155e+10           Root MSE      =   89248

------------------------------------------------------------------------------
    COST |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
       N |   331.4493   39.75844      8.337   0.000       252.1732    410.7254
     OCC |   133259.1   20827.59      6.398   0.000       91730.06    174788.1
   _cons |  -33612.55   23573.47     -1.426   0.158      -80616.71    13391.61
------------------------------------------------------------------------------
```

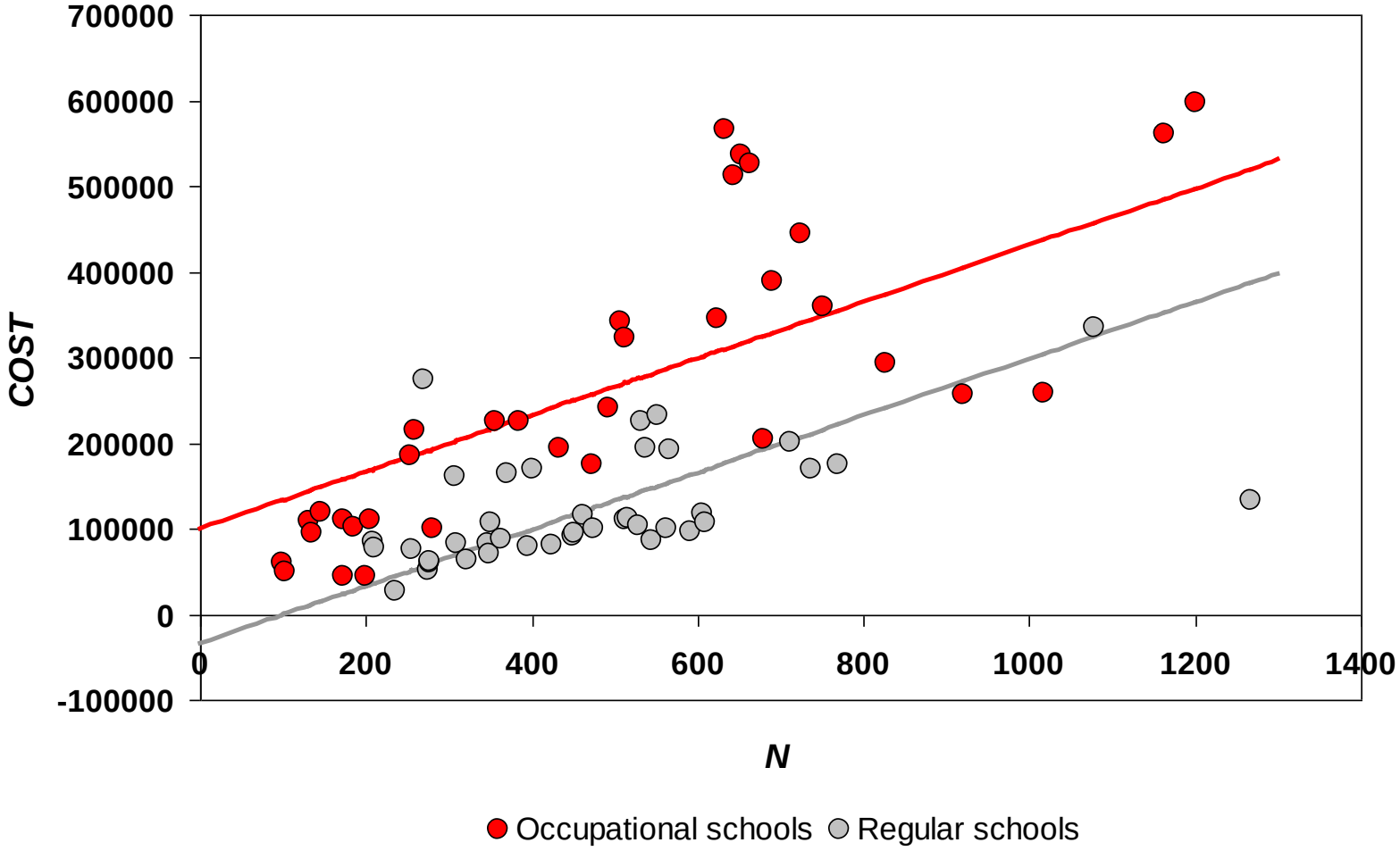
We now run the regression of *COST* on *N* and *OCC*, treating *OCC* just like any other explanatory variable, despite its artificial nature. The Stata output is shown.

$$COST = -34{,}000 + 133{,}000OCC + 331N$$

$$\widehat{COST}_{regular} = -34{,}000 + 331N$$

$$\widehat{COST}_{occupational} = -34{,}000 + 133{,}000 + 331N = 99{,}000 + 331N$$

DUMMY VARIABLE CLASSIFICATION WITH TWO CATEGORIES



The scatter diagram shows the data and the two cost functions derived from the regression results.

DUMMY CLASSIFICATION WITH MORE THAN TWO CATEGORIES

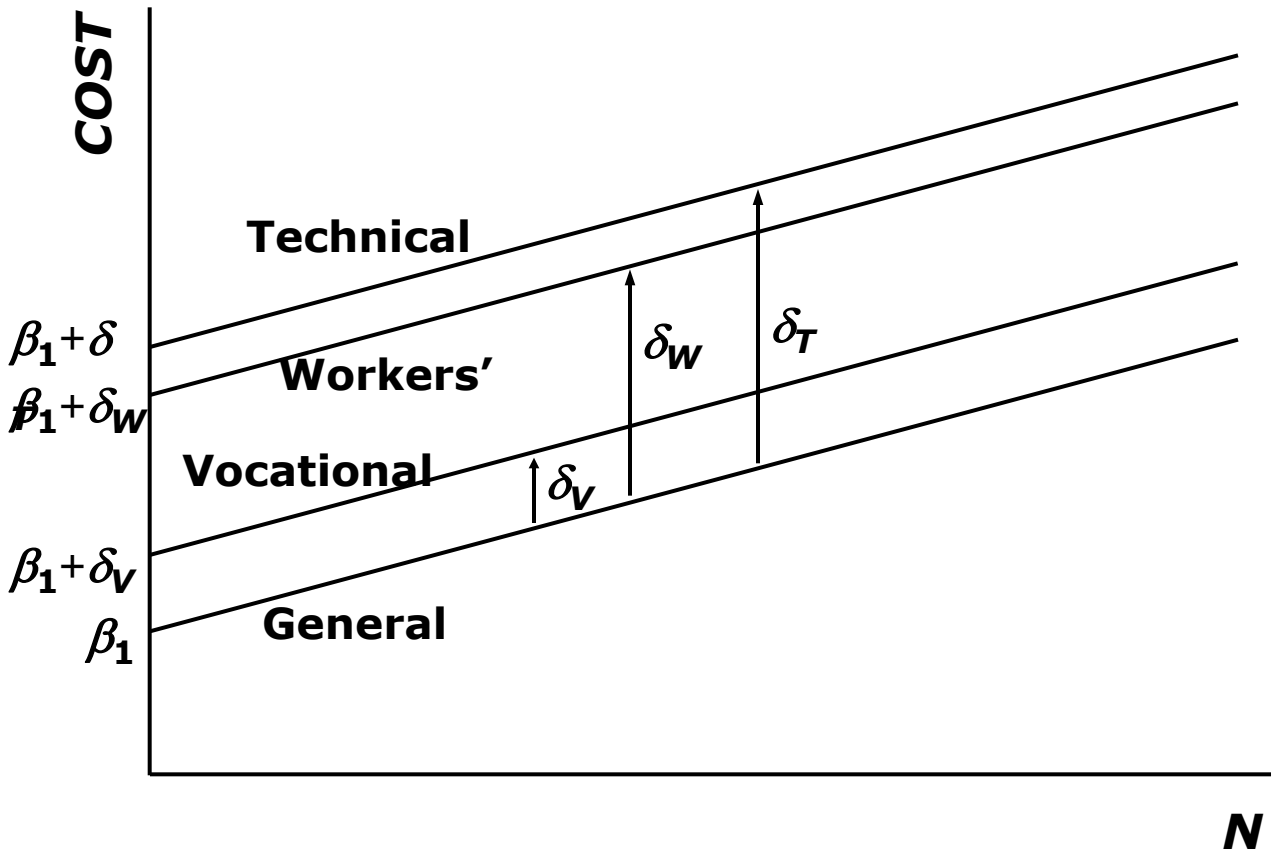
$$COST = \beta_1 + \delta_T TECH + \delta_W WORKER + \delta_V VOC + \beta_2 N + u$$

General School (*TECH* = *WORKER* = *VOC* = 0)	$COST = \beta_1 + \beta_2 N + u$
Technical School (*TECH* = 1; *WORKER* = *VOC* = 0)	$COST = (\beta_1 + \delta_T) + \beta_2 N + u$
Skilled Workers' School (*WORKER* = 1; *TECH* = *VOC* = 0)	$COST = (\beta_1 + \delta_W) + \beta_2 N + u$
Vocational School (*VOC* = 1; *TECH* = *WORKER* = 0)	$COST = (\beta_1 + \delta_V) + \beta_2 N + u$

The reference category: General School (do not include a dummy variable for the reference category)

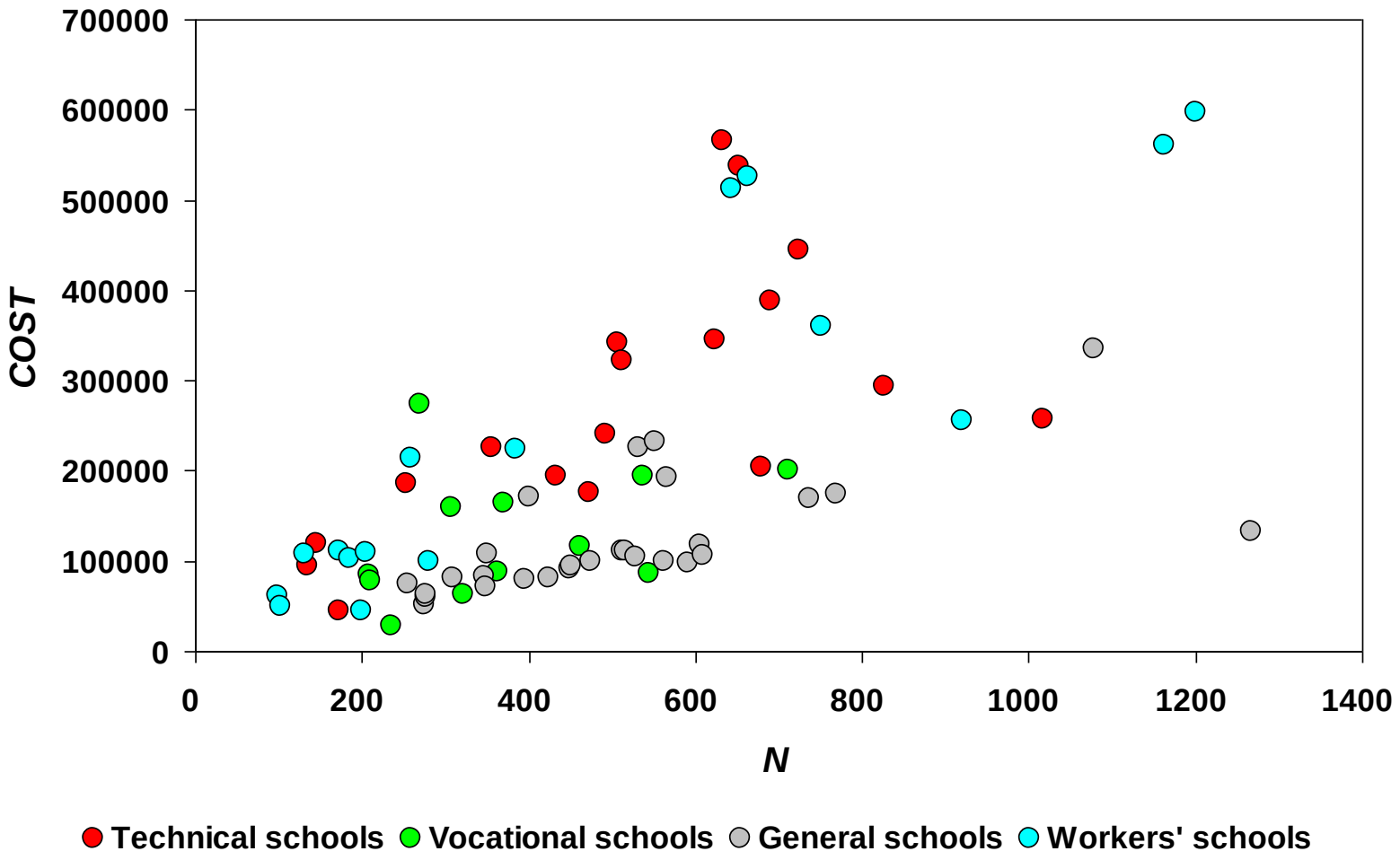
The regression model simplifies in a similar manner in the case of observations relating to skilled workers' schools and vocational schools.

DUMMY CLASSIFICATION WITH MORE THAN TWO CATEGORIES



The diagram illustrates the model graphically. The δ coefficients are the extra overhead costs of running technical, skilled workers', and vocational schools, relative to the overhead cost of general schools.

DUMMY CLASSIFICATION WITH MORE THAN TWO CATEGORIES



The scatter diagram shows the data for the entire sample, differentiating by type of school.

DUMMY CLASSIFICATION WITH MORE THAN TWO CATEGORIES

```
. reg COST N TECH WORKER VOC
```

Source	SS	df	MS		
Model	9.2996e+11	4	2.3249e+11	Prob > F	= 0.0000
Residual	5.4138e+11	69	7.8461e+09	R-squared	= 0.6320
Total	1.4713e+12	73	2.0155e+10	Adj R-squared	= 0.6107
				Root MSE	= 88578

Number of obs = 74
F(4, 69) = 29.63

COST	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
N	342.6335	40.2195	8.519	0.000	262.3978	422.8692
TECH	154110.9	26760.41	5.759	0.000	100725.3	207496.4
WORKER	143362.4	27852.8	5.147	0.000	87797.57	198927.2
VOC	53228.64	31061.65	1.714	0.091	-8737.646	115194.9
_cons	-54893.09	26673.08	-2.058	0.043	-108104.4	-1681.748

The coefficient of *N* indicates that the marginal cost per student per year is 343 yuan.

The coefficients of *TECH*, *WORKER,* and *VOC* are 154,000, 143,000, and 53,000, respectively, and should be interpreted as the additional annual overhead costs, relative to those of general schools.

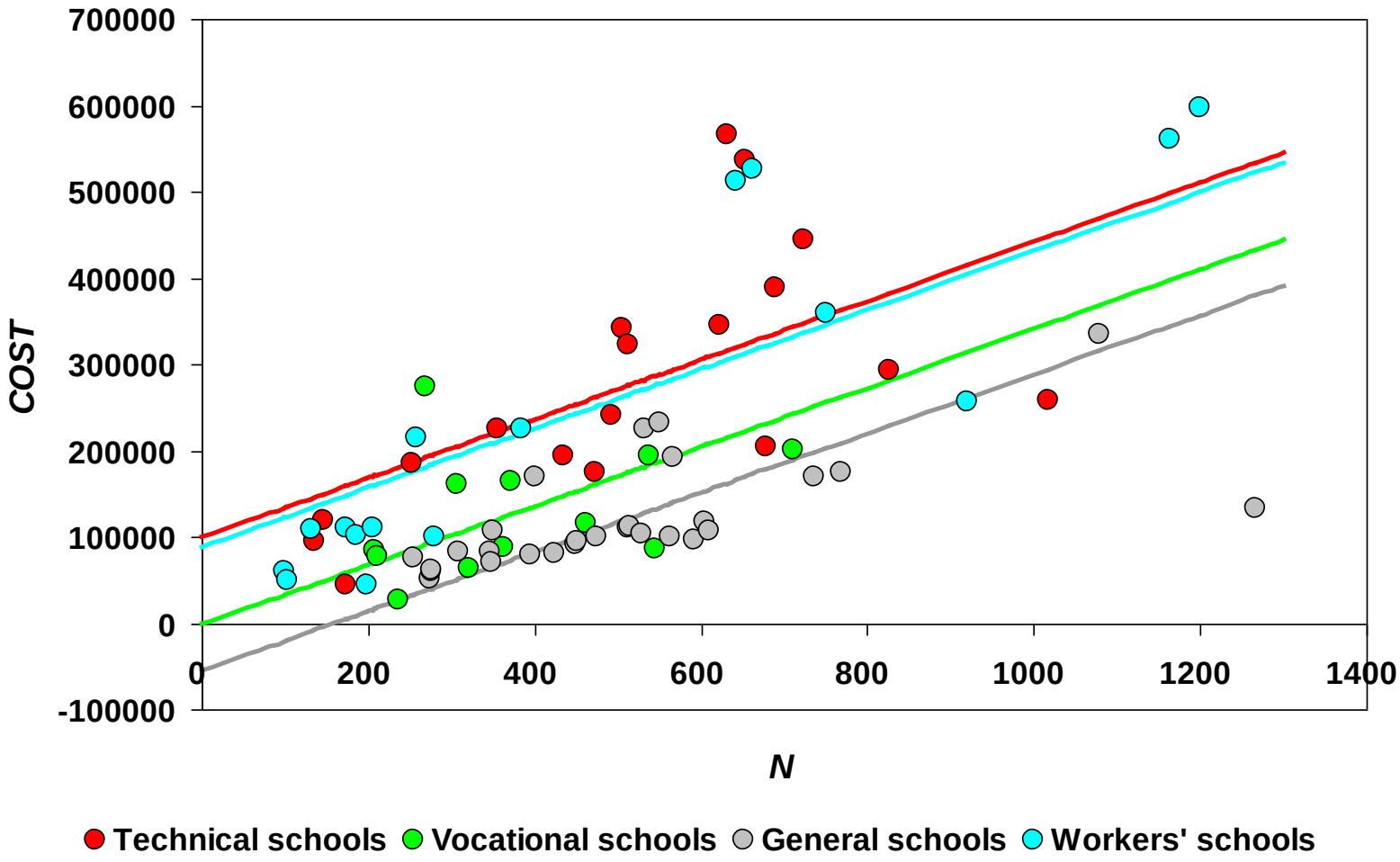
DUMMY CLASSIFICATION WITH MORE THAN TWO CATEGORIES

$\hat{COST} = -55{,}000 + 154{,}000TECH + 143{,}000WORKER + 53{,}000VOC + 343N$

General School (*TECH* = *WORKER* = *VOC* = 0)	$\hat{COST} = -55{,}000 + 343N$
Technical School (*TECH* = 1; *WORKER* = *VOC* = 0)	$\hat{COST} = -55{,}000 + 154{,}000 + 343N$ $= 99{,}000 + 343N$
Skilled Workers' School (*WORKER* = 1; *TECH* = *VOC* = 0)	$\hat{COST} = -55{,}000 + 143{,}000 + 343N$ $= 88{,}000 + 343N$
Vocational School (*VOC* = 1; *TECH* = *WORKER* = 0)	$\hat{COST} = -55{,}000 + 53{,}000 + 343N$ $= -2{,}000 + 343N$

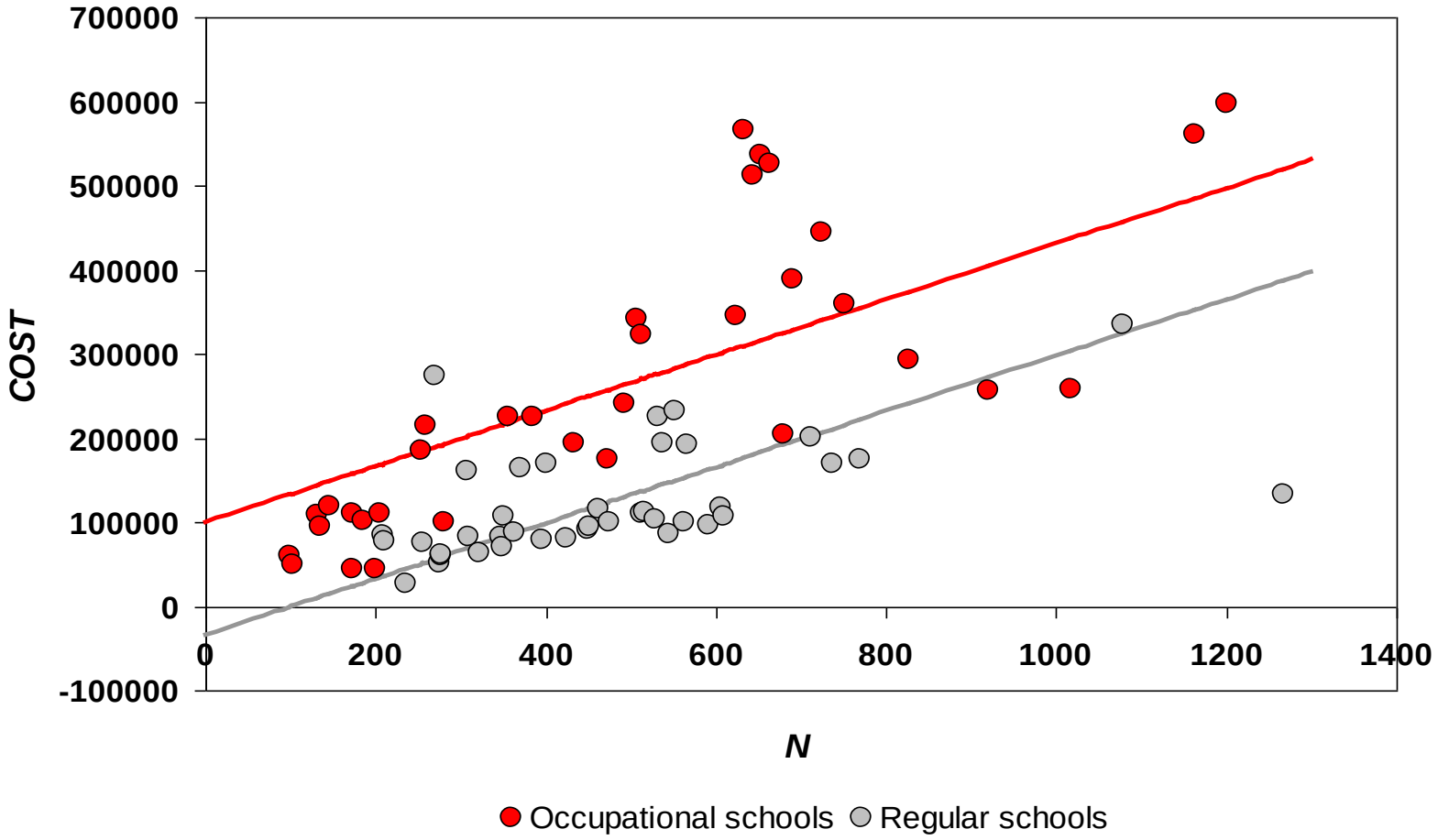
Note that in each case the annual marginal cost per student is estimated at 343 yuan. The model specification assumes that this figure does not differ according to type of school.

DUMMY CLASSIFICATION WITH MORE THAN TWO CATEGORIES



The four cost functions are illustrated graphically.

SLOPE DUMMY VARIABLES



Previously, we have the assumption that the marginal cost per student is the same for occupational and regular schools. Hence the cost functions are parallel. This is unrealistic.

In practice, the cost function for the occupational schools should be steeper, and that for the regular schools should be flatter.

SLOPE DUMMY VARIABLES

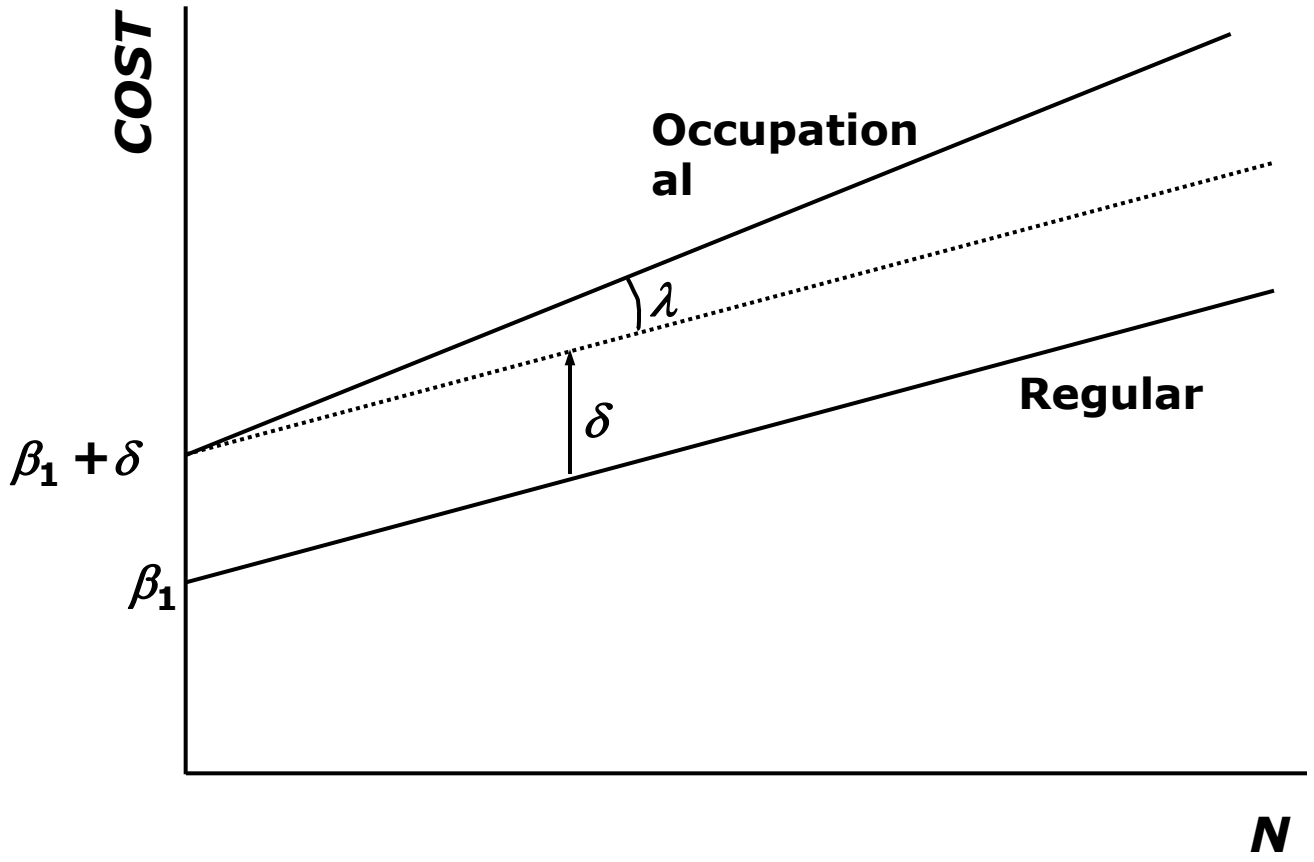
$$COST = \beta_1 + \delta OCC + \beta_2 N + \lambda NOCC + u$$

Regular school (*OCC* = *NOCC* = 0)	$COST = \beta_1 + \beta_2 N + u$
Occupational school (*OCC* = 1; *NOCC* = *N*)	$COST = (\beta_1 + \delta) + (\beta_2 + \lambda) N + u$

We will relax the assumption of the same marginal cost by introducing what is known as a slope dummy variable. This is *NOCC*, defined as the product of *N* and *OCC*.

The model now allows the marginal cost per student to be an amount λ greater than that in regular schools, as well as allowing the overhead costs to be different.

SLOPE DUMMY VARIABLES



The diagram illustrates the model graphically.

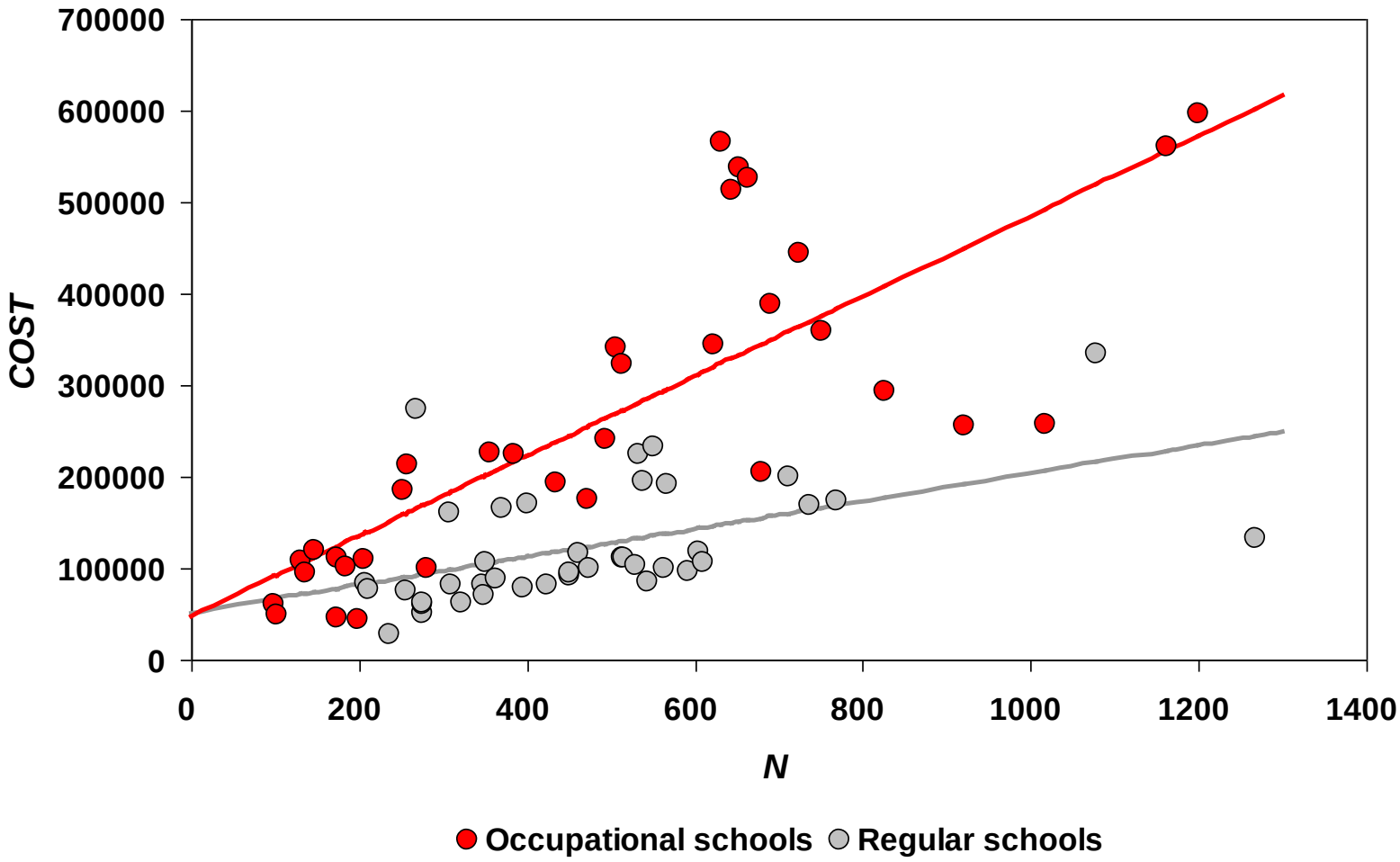
SLOPE DUMMY VARIABLES

$$\hat{COST} = 51{,}000 - 4{,}000\,OCC + 152N + 284\,NOCC$$

Regular school (*OCC* = *NOCC* = 0)	$\hat{COST} = 51{,}000 + 152N$
Occupational school (*OCC* = 1; *NOCC* = *N*)	$\hat{COST} = 51{,}000 - 4{,}000 + 152N + 284N$ $= 47{,}000 + 436N$

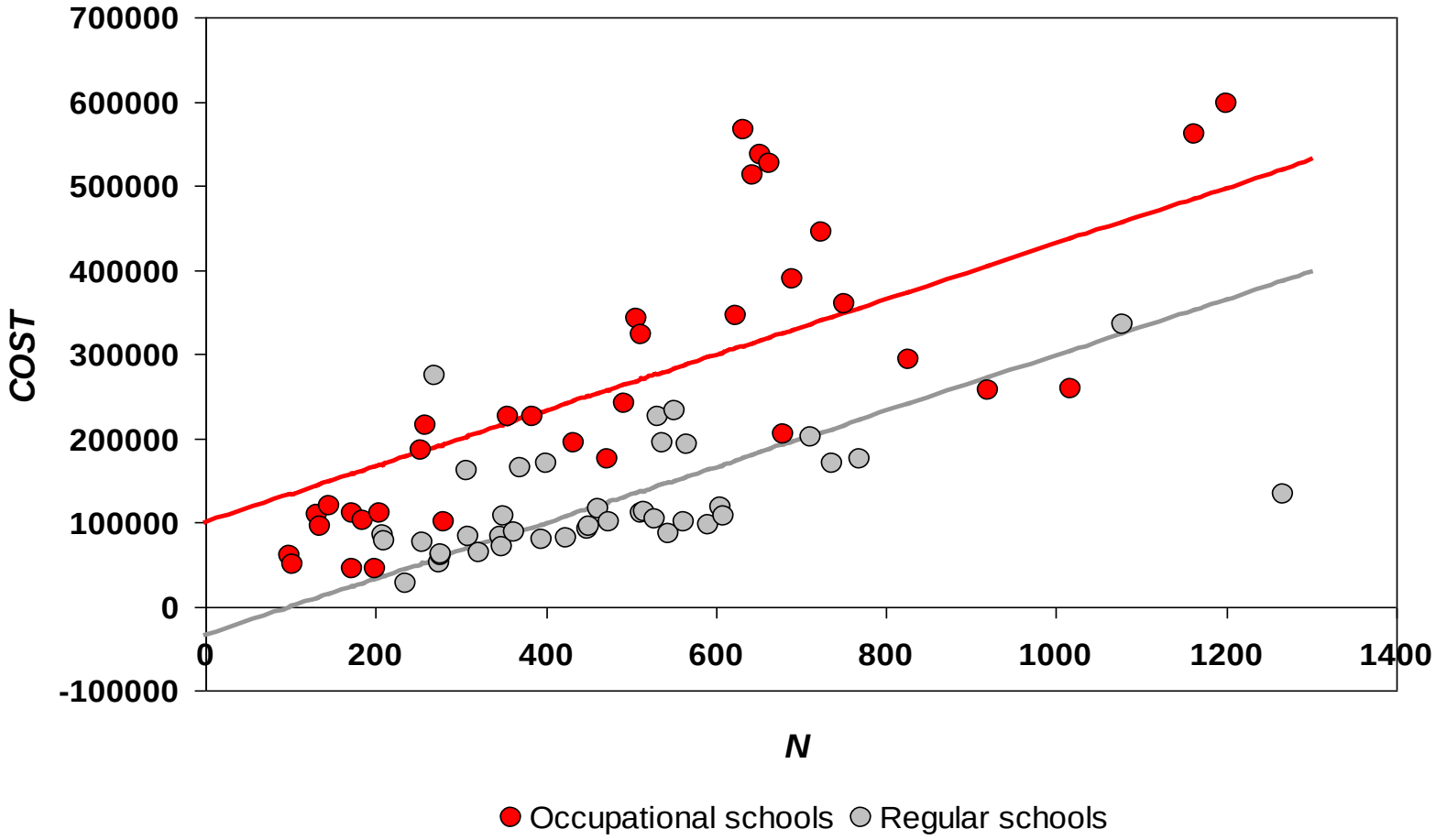
Here is the regression in equation form.

SLOPE DUMMY VARIABLES



You can see that the cost functions fit the data much better than before and that the real difference is in the marginal cost, not the overhead cost.

SLOPE DUMMY VARIABLES



The assumption of the same marginal cost led to an estimate of the marginal cost that was a compromise between the marginal costs of occupational and regular schools.

The cost function for regular schools was too steep and as a consequence the intercept was underestimated, actually becoming negative and indicating that something must be wrong with the specification of the model.

SPECIFICATION AND DATA PROBLEMS

VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

Consequences of Variable Misspecification

		True Model	
		$Y = \beta_1 + \beta_2 X_2 + u$	$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$
Fitted Model	$\hat{Y} = b_1 + b_2 X_2$		
	$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$		

To keep the analysis simple, we will assume that there are only two possibilities. Either Y depends only on X_2, or it depends on both X_2 and X_3.

VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

Consequences of Variable Misspecification

		True Model	
		$Y = \beta_1 + \beta_2 X_2 + u$	$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$
Fitted Model	$\hat{Y} = b_1 + b_2 X_2$	Correct specification, no problems	
	$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$		Correct specification, no problems

If Y depends only on X_2, and we fit a simple regression model, we will not encounter any problems, assuming of course that the regression model assumptions are valid.

Likewise we will not encounter any problems if Y depends on both X_2 and X_3 and we fit the multiple regression.

VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

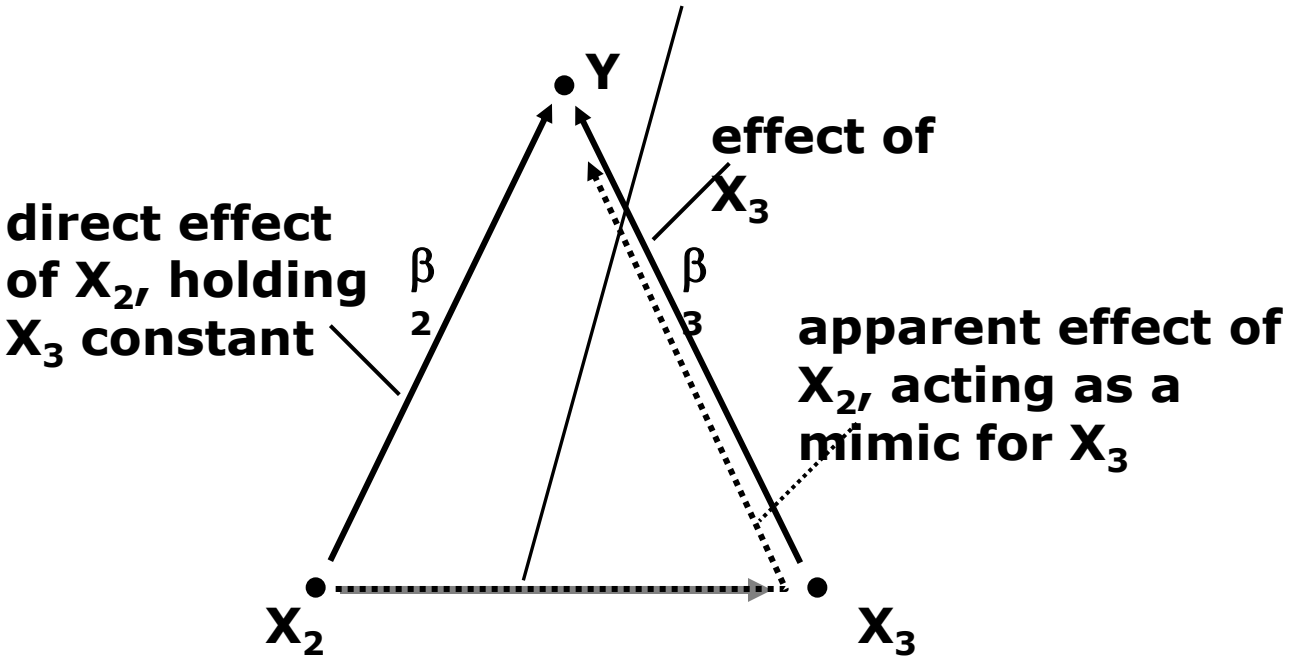
Consequences of Variable Misspecification			
		True Model	
		$Y = \beta_1 + \beta_2 X_2 + u$	$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$
Fitted Model	$\hat{Y} = b_1 + b_2 X_2$	Correct specification, no problems	Coefficients are biased (in general). Standard errors are invalid.
	$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$		Correct specification, no problems

In this sequence we will examine the consequences of fitting a simple regression when the true model is multiple. The omission of a relevant explanatory variable causes the regression coefficients to be biased and the standard errors to be invalid.

VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad \hat{Y} = b_1 + b_2 X_2$$

$$E(b_2) = \beta_2 + \beta_3 \frac{\sum (X_{2i} - \overline{X}_2)(X_{3i} - \overline{X}_3)}{\sum (X_{2i} - \overline{X}_2)^2}$$



The strength of the proxy effect depends on two factors: the strength of the effect of X_3 on Y, which is given by β_3, and the ability of X_2 to mimic X_3.

The ability of X_2 to mimic X_3 is determined by the slope coefficient obtained when X_3 is regressed on X_2, the term highlighted in yellow.

VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

```
. reg LGEARN S EXP
------------------------------------------------------------------------------
      LGEARN |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |    .1235911   .0090989    13.58   0.000     .1057173     .141465
         EXP |    .0350826   .0050046     7.01   0.000     .0252515    .0449137
       _cons |    .5093196   .1663823     3.06   0.002     .1824796    .8361596

. reg LGEARN S
------------------------------------------------------------------------------
      LGEARN |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |    .1096934   .0092691    11.83   0.000     .0914853    .1279014
       _cons |   1.292241   .1287252    10.04   0.000     1.039376    1.545107

. reg LGEARN EXP
------------------------------------------------------------------------------
      LGEARN |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         EXP |    .0202708   .0056564     3.58   0.000     .0091595     .031382
       _cons |    2.44941   .0988233    24.79   0.000     2.255284    2.643537
```

As can be seen, the coefficients of S and EXP are indeed lower in the simple regressions.

VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

```
. reg LGEARN S EXP
      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  2,   537) =  100.86
       Model |  50.9842581     2  25.492129              Prob > F      =  0.0000
    Residual |  135.723385   537  .252743734             R-squared     =  0.2731
-------------+------------------------------           Adj R-squared =  0.2704
       Total |  186.707643   539   .34639637            Root MSE      =  .50274

. reg LGEARN S
      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  1,   538) =  140.05
       Model |  38.5643833     1  38.5643833              Prob > F      =  0.0000
    Residual |   148.14326   538  .275359219             R-squared     =  0.2065
-------------+------------------------------           Adj R-squared =  0.2051
       Total |  186.707643   539   .34639637            Root MSE      =  .52475

. reg LGEARN EXP
      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  1,   538) =   12.84
       Model |  4.35309315     1  4.35309315              Prob > F      =  0.0004
    Residual |   182.35455   538  .338948978             R-squared     =  0.0233
-------------+------------------------------           Adj R-squared =  0.0215
       Total |  186.707643   539   .34639637            Root MSE      =  .58219
```

A comparison of R^2 for the three regressions shows that the sum of R^2 in the simple regressions is actually less than R^2 in the multiple regression.

VARIABLE MISSPECIFICATION II: INCLUSION OF AN IRRELEVANT VARIABLE

Consequences of Variable Misspecification			
		True Model	
		$Y = \beta_1 + \beta_2 X_2 + u$	$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$
Fitted Model	$\hat{Y} = b_1 + b_2 X_2$	**Correct specification, no problems**	**Coefficients are biased (in general). Standard errors are invalid.**
	$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$	**Coefficients are unbiased (in general), but inefficient. Standard errors are valid (in general)**	**Correct specification, no problems**

Including irrelevant variables: The effects are different from those of omitted variable misspecification. In this case the coefficients in general remain unbiased, but they are inefficient. The standard errors remain valid, but are needlessly large.

VARIABLE MISSPECIFICATION II: INCLUSION OF AN IRRELEVANT VARIABLE

$$Y = \beta_1 + \beta_2 X_2 + u$$

$$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$$

$$Y = \beta_1 + \beta_2 X_2 + 0\, X_3 + u$$

$$\sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum \left(X_{2i} - \overline{X}_2 \right)^2} \times \frac{1}{1 - r_{X_2, X_3}^2}$$

Rewrite the true model adding X_3 as an explanatory variable, with a coefficient of 0. Now the true model and the fitted model coincide. Hence b_2 will be an unbiased estimator of β_2 and b_3 will be an unbiased estimator of 0.

However, the variance of b_2 will be larger than it would have been if the correct simple regression had been run because it includes the factor 1 / (1 – r^2), where r is the correlation between X_2 and X_3.

The standard errors remain valid, but they will tend to be larger than those obtained in a simple regression, reflecting the loss of efficiency.

VARIABLE MISSPECIFICATION II: INCLUSION OF AN IRRELEVANT VARIABLE

```
. reg LGFDHO LGEXP LGSIZE
```

LGFDHO	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
LGEXP	.2866813	.0226824	12.639	0.000	.2421622	.3312003
LGSIZE	.4854698	.0255476	19.003	0.000	.4353272	.5356124
_cons	4.720269	.2209996	21.359	0.000	4.286511	5.154027

```
. reg LGFDHO LGEXP LGSIZE LGHOUS
```

LGFDHO	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
LGEXP	.2673552	.0370782	7.211	0.000	.1945813	.340129
LGSIZE	.4868228	.0256383	18.988	0.000	.4365021	.5371434
LGHOUS	.0229611	.0348408	0.659	0.510	-.0454214	.0913436
_cons	4.708772	.2217592	21.234	0.000	4.273522	5.144022

The inclusion does not cause the coefficients of those variables to be biased.

But it does increase their standard errors, particularly that of *LGEXP*, as you would expect, reflecting the loss of efficiency.

PROXY VARIABLES

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + u$$

$$X_2 = \lambda + \mu Z$$

$$Y = \beta_1 + \beta_2(\lambda + \mu Z) + \beta_3 X_3 + \ldots + \beta_k X_k + u$$

$$= (\beta_1 + \beta_2 \lambda) + \beta_2 \mu Z + \beta_3 X_3 + \ldots + \beta_k X_k + u$$

Suppose that a variable *Y* is hypothesized to depend on a set of explanatory variables X_2, ..., X_k as shown above, and suppose that for some reason there are no data on X_2. A regression of *Y* on X_3, ..., X_k would yield biased estimates of the coefficients and invalid standard errors and tests (omitted var. bias).

These problems can be reduced or eliminated by using a proxy variable in the place of X_2. A proxy variable is one that is hypothesized to be linearly related to the missing variable. Here *Z* could act as a proxy for X_2.

We thus obtain a model with all variables observable. If the proxy relationship is an exact one, and we fit this relationship, most of the regression results will be rescued.

TESTING A LINEAR RESTRICTION

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + \beta_4 SF + u$$

$$\beta_4 = \beta_3$$

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 (SM + SF) + u$$

$$= \beta_1 + \beta_2 ASVABC + \beta_3 SP + u$$

In the last sequence it was argued that educational attainment might be related to cognitive ability and family background, with mother's and father's educational attainment proxying for the latter.

It was suggested that the impact of parental education might be the same for both parents, that is, that β_3 and β_4 might be equal.

We now have a total parental education variable, *SP*, instead of separate variables for mother's and father's education, and the multicollinearity caused by the correlation between the latter has been eliminated.

TESTING A LINEAR RESTRICTION

```
. reg S ASVABC SM SF
```

S	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ASVABC	.1257087	.0098533	12.76	0.000	.1063528	.1450646
SM	.0492424	.0390901	1.26	0.208	-.027546	.1260309
SF	.1076825	.0309522	3.48	0.001	.04688	.1684851
_cons	5.370631	.4882155	11.00	0.000	4.41158	6.329681

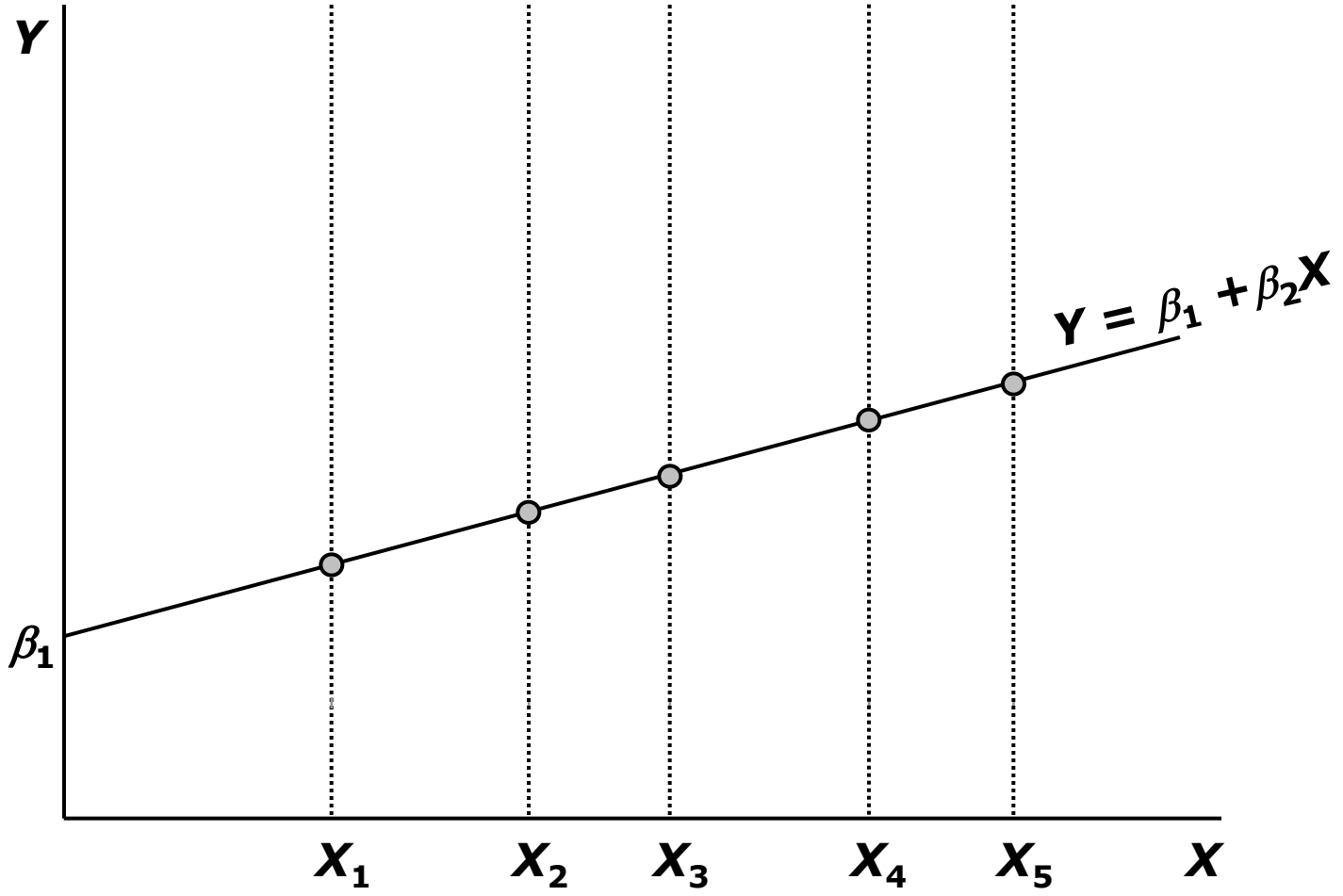
```
. reg S ASVABC SP
```

S	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ASVABC	.1253106	.0098434	12.73	0.000	.1059743	.1446469
SP	.0828368	.0164247	5.04	0.000	.0505722	.1151014
_cons	5.29617	.4817972	10.99	0.000	4.349731	6.242608

A comparison of the regressions reveals that the standard error of the coefficient of *SP* is much smaller than those of *SM* and *SF*, and consequently its *t* statistic is higher. Its coefficient is a compromise between those of *SM* and *SF*, as might be expected.

Make sure the restriction is valid

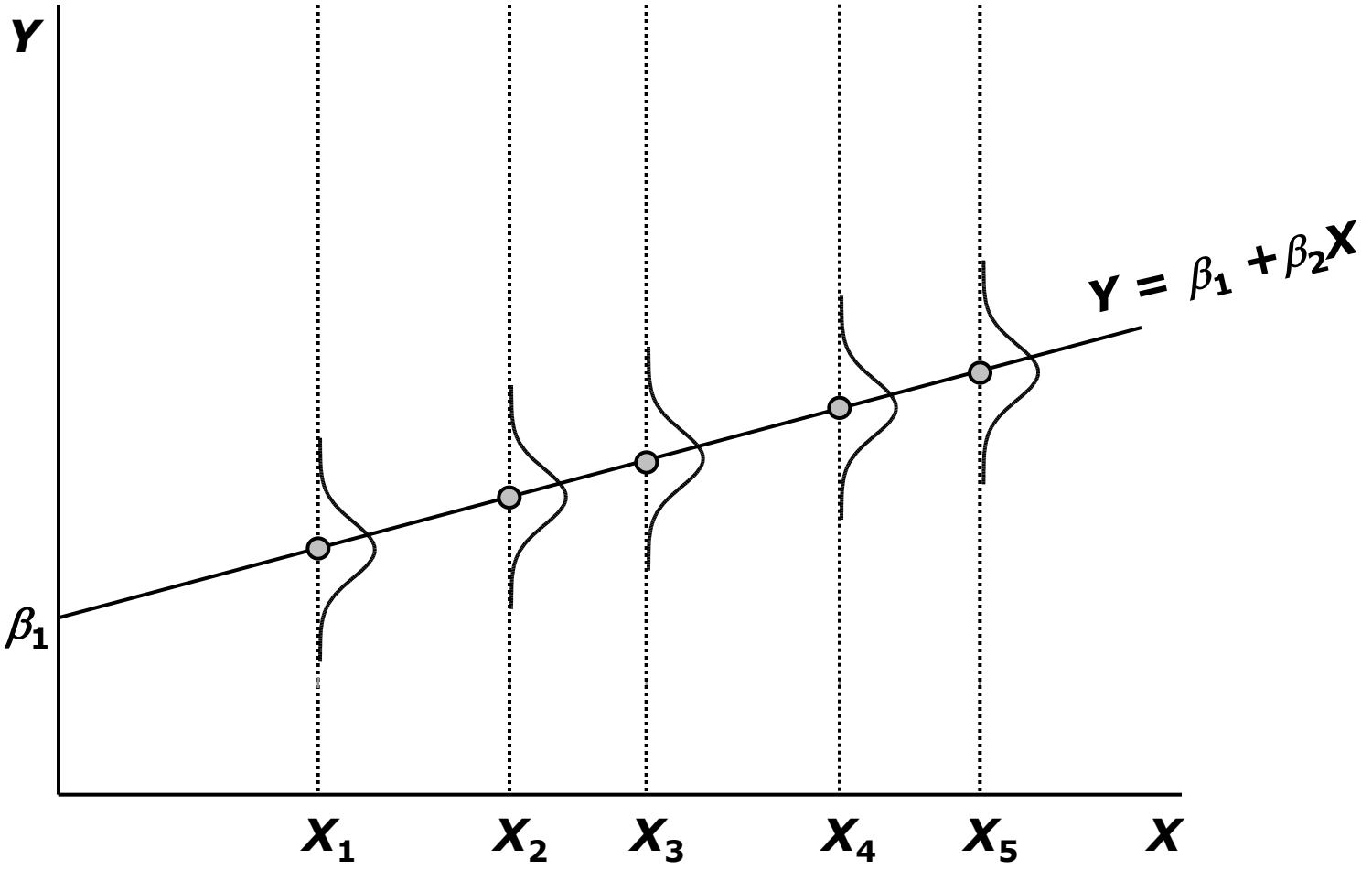
HETEROSKEDASTICITY



Heteroskedasticity relates to the distribution of the disturbance term in a regression model.

We will discuss it in the context of the regression model $Y = \beta_1 + \beta_2 X + u$. If there were no disturbance term in the model, the observations would lie on the line as shown.

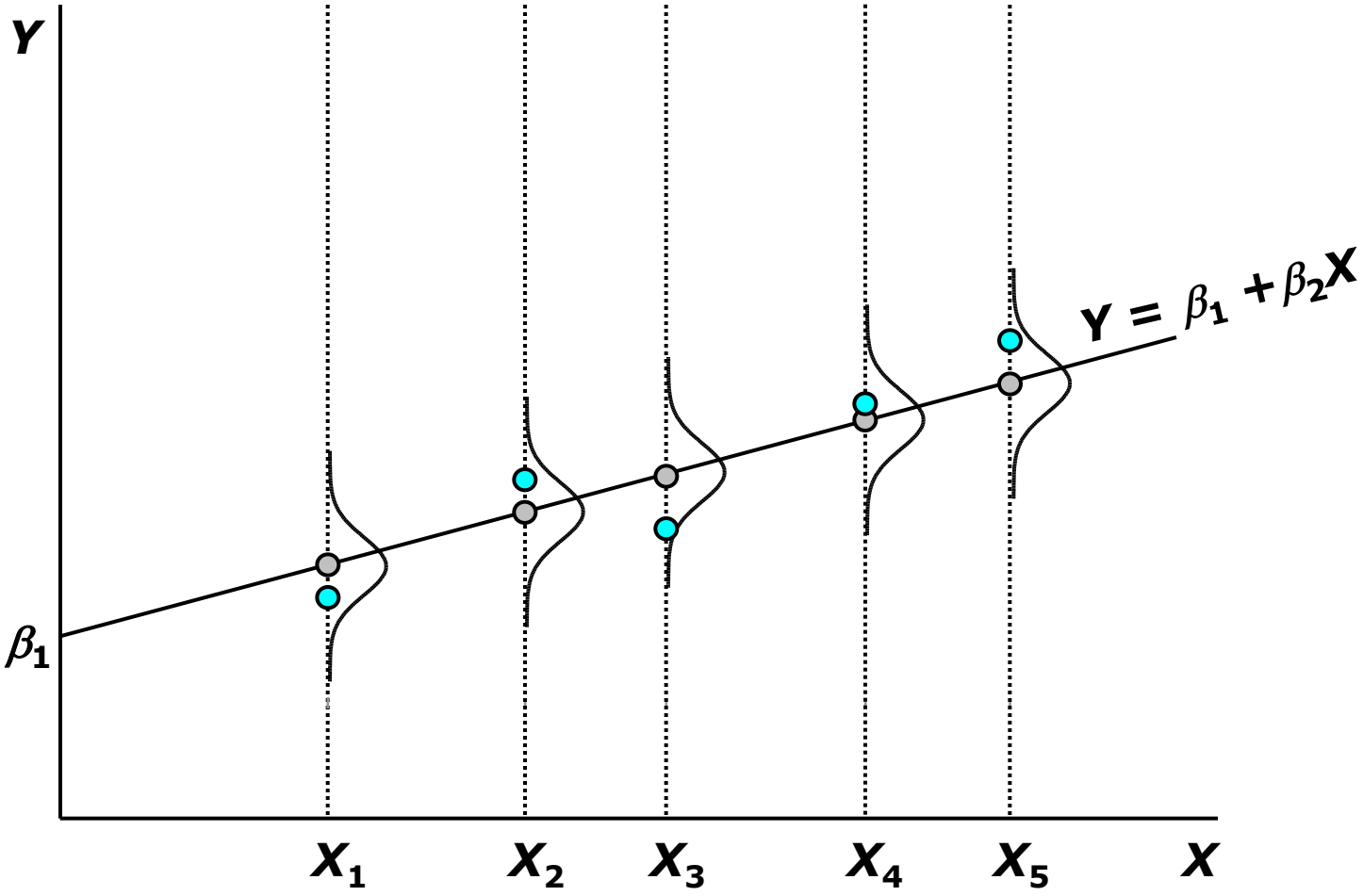
HETEROSKEDASTICITY



Now we take account of the effect of the disturbance term. It will displace each observation in the vertical dimension, since it modifies the value of *Y* without affecting *X*.

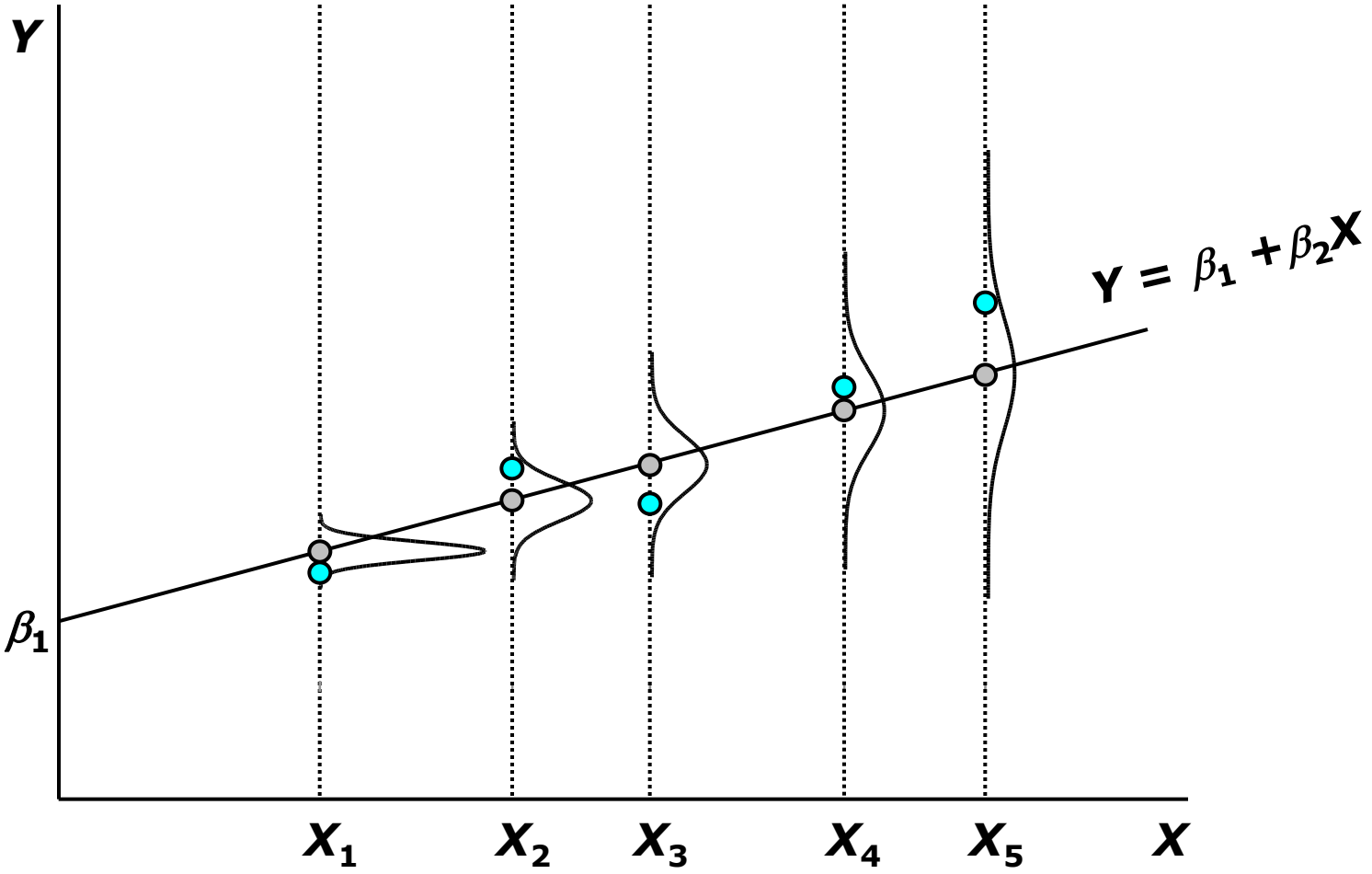
Assumptions: the expected value of *u* in each observation is 0; the distribution in each observation is normal; the variance of the distribution of the disturbance term is the same for each observation (homoskedasticity).

HETEROSKEDASTICITY



Once the sample has been drawn, some observations will lie closer to the line than others, but we have no way of anticipating in advance which ones these will be.

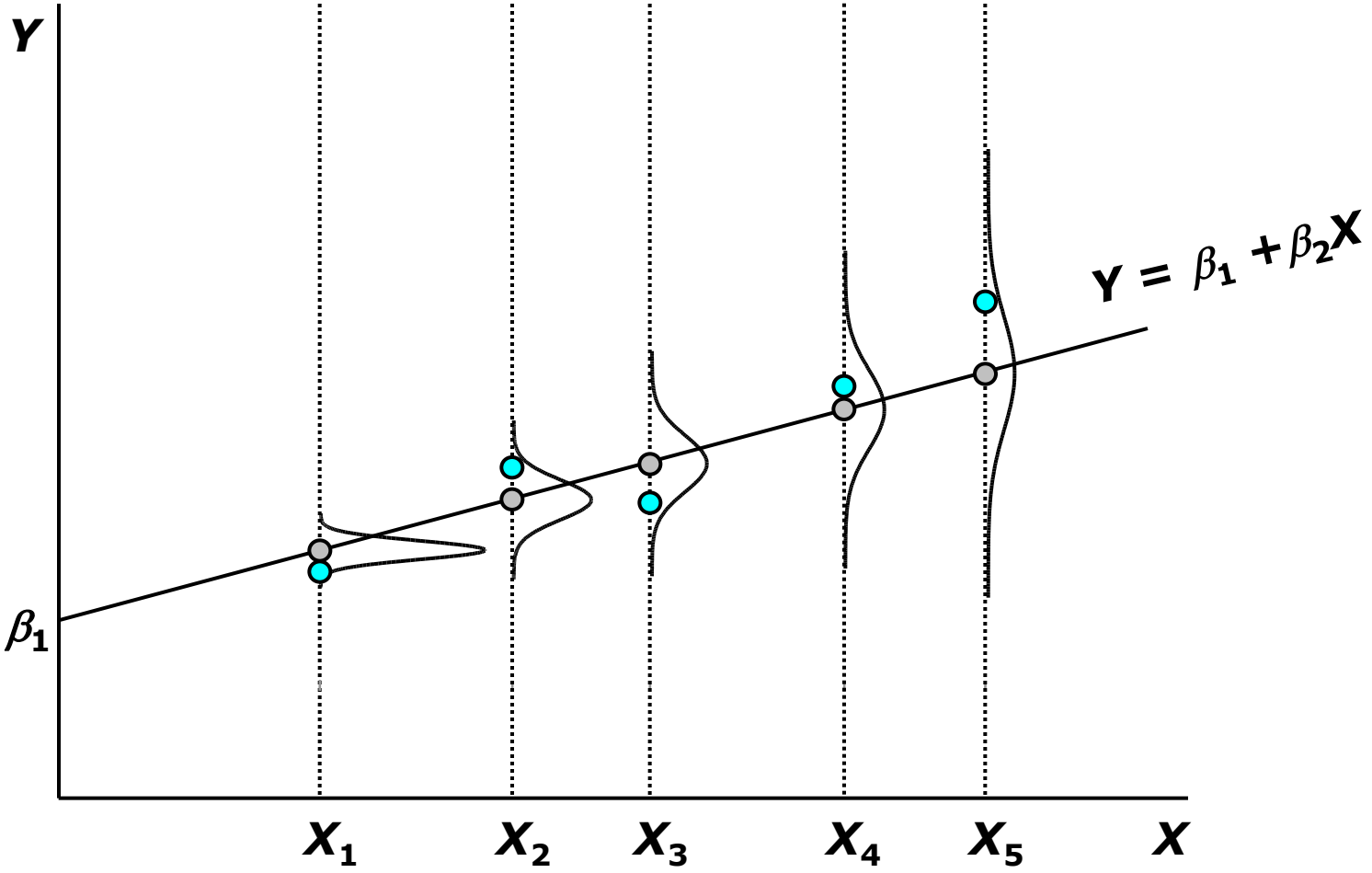
HETEROSKEDASTICITY



The distribution of *u* associated with each observation still has expected value 0 and is normal. However Assumption Homoskedasticity is violated and the variance is no longer constant.

Obviously, observations where *u* has low variance will tend to be better guides to the underlying relationship than those having a relatively high variance. When the distribution is not the same for each observation, the disturbance term is said to be subject to heteroscedasticity.

HETEROSKEDASTICITY



There are two major consequences of heteroscedasticity. One is that the standard errors of the regression coefficients are estimated wrongly and the *t* tests (and *F* test) are invalid.

The other is that OLS is an inefficient estimation technique. An alternative technique which gives relatively high weight to the relatively low-variance observations should tend to yield more accurate estimates.

HETEROSKEDASTICITY-CONSISTENT STANDARD ERRORS

$$b_2^{OLS} = \beta_2 + \sum_{i=1}^{n} a_i u_i$$

$$\text{where } a_i = \frac{(X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

Heteroscedasticity causes OLS standard errors to be biased is finite samples. However it can be demonstrated that they are nevertheless consistent, provided that their variances are distributed independently of the regressors.

HETEROSKEDASTICITY-CONSISTENT STANDARD ERRORS

$$b_2^{OLS} = \beta_2 + \sum_{i=1}^{n} a_i u_i$$

where $$a_i = \frac{(X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$\sigma^2_{b_2^{OLS}} = \sum_{i=1}^{n} a_i^2 E\left(u_i^2\right) = \sum_{i=1}^{n} a_i^2 \sigma^2_{u_i}$$

$$s^2_{b_2^{OLS}} = \sum_{i=1}^{n} a_i^2 e_i^2$$

White (1980) demonstrates that a consistent estimator of $\sigma^2_{b_2^{OLS}}$ is obtained if the squared residual in observation *i* is used as an estimator of $\sigma^2_{u_i}$. Taking the square root, one obtains a heteroscedasticity-consistent standard error.

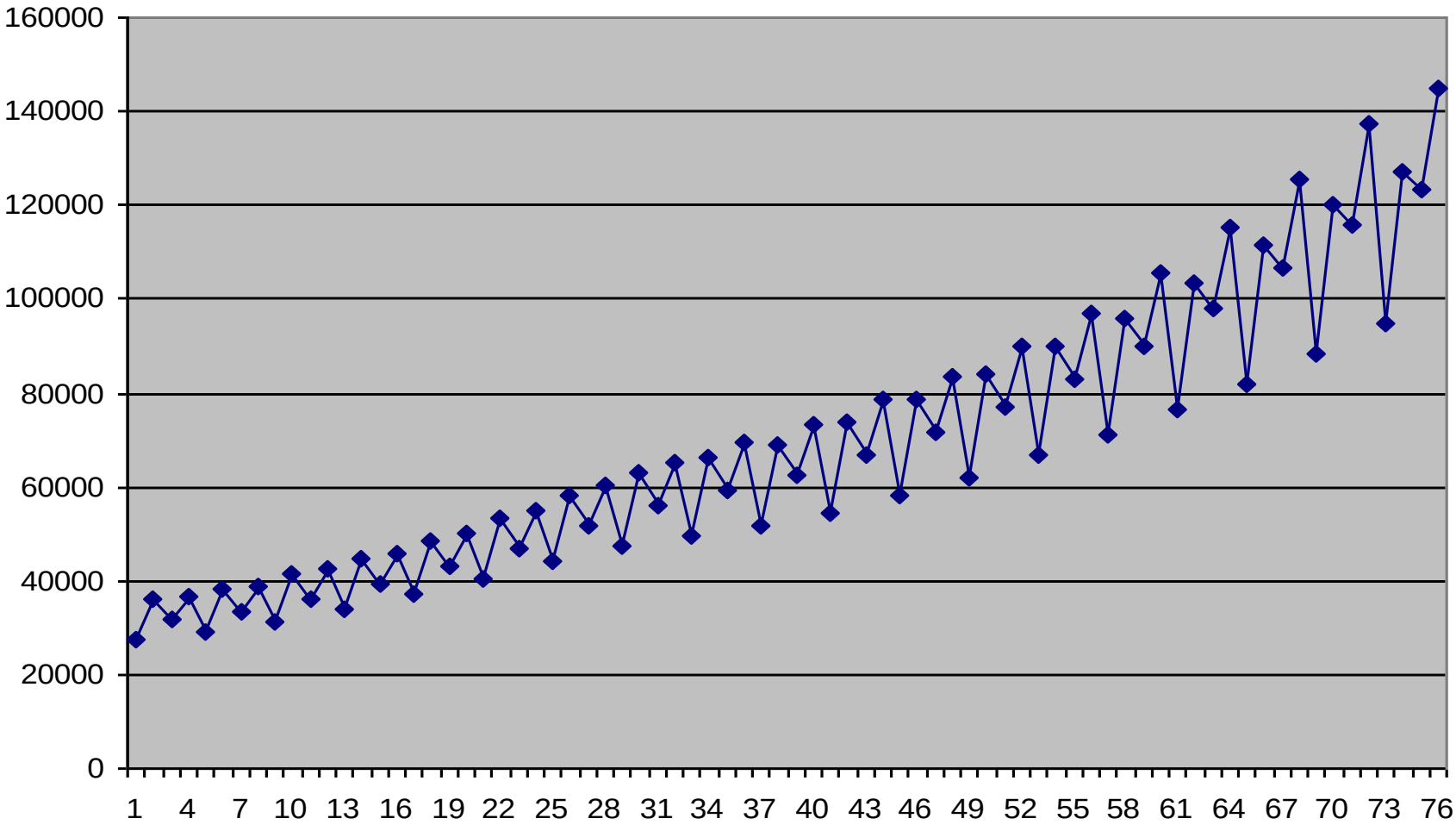
HETEROSKEDASTICITY-CONSISTENT STANDARD ERRORS

```
. reg manu gdp
      Source |       SS       df       MS              Number of obs =      28
-------------+------------------------------           F(  1,    26) =  210.73
       Model |  1.1600e+11     1  1.1600e+11            Prob > F      =  0.0000
    Residual |  1.4312e+10    26   550462775            R-squared     =  0.8902
-------------+------------------------------           Adj R-squared =  0.8859
       Total |  1.3031e+11    27  4.8264e+09            Root MSE      =   23462
------------------------------------------------------------------------------
        manu |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gdp |    .193693   .0133428    14.52   0.000     .1662665    .2211195
       _cons |   603.9453   5699.677     0.11   0.916    -11111.91     12319.8

. reg manu gdp, robust
Regression with robust standard errors                Number of obs =      28
                                                      F(  1,    26) =  116.39
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.8902
                                                      Root MSE      =   23462
------------------------------------------------------------------------------
             |               Robust
        manu |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gdp |    .193693   .0179542    10.79   0.000     .1567877    .2305983
       _cons |   603.9453   3542.388     0.17   0.866    -6677.538    7885.429
```

The point estimates of the coefficients are exactly the same. However the standard error of the coefficient of *GDP* rises from 0.13 to 0.18, indicating that it is underestimated in the original OLS regression.

TIME SERIES-FORECASTING

QUARTERLY GDP



COMPONENTS OF A TIME SERIES

- Time series: *An ordered sequence of values of a variable at equally spaced time intervals*
- Such as: vn index, inflation, gdp growth rate, etc.
- Components:
  - Trend
  - Seasonality
  - Cycle
  - Irregular
- The 4 components may make up a TS in two ways:
  - additive model: $X_t = T_t + S_t + C_t + I_t$
  - multiplicative model: $X_t = T_t * S_t * C_t * I_t$

ASSUMPTIONS FOR TIME SERIES MODEL

C.1 ***The model is linear in parameters and correctly specified.***

$$Y = b_1 + b_2X_2 + \ldots + b_kX_k + u$$

C.2 ***The time series for the regressors are weakly persistent***

C.3 ***There does not exist an exact linear relationship among the regressors***

C.4 ***The disturbance term has zero expectation***

C.5 ***The disturbance term is homoscedastic***

ASSUMPTIONS FOR TIME SERIES MODEL

C.6 ***The values of the disturbance term have independent distributions***

u_t **is distributed independently of** $u_{t'}$ **for** $t' \neq t$

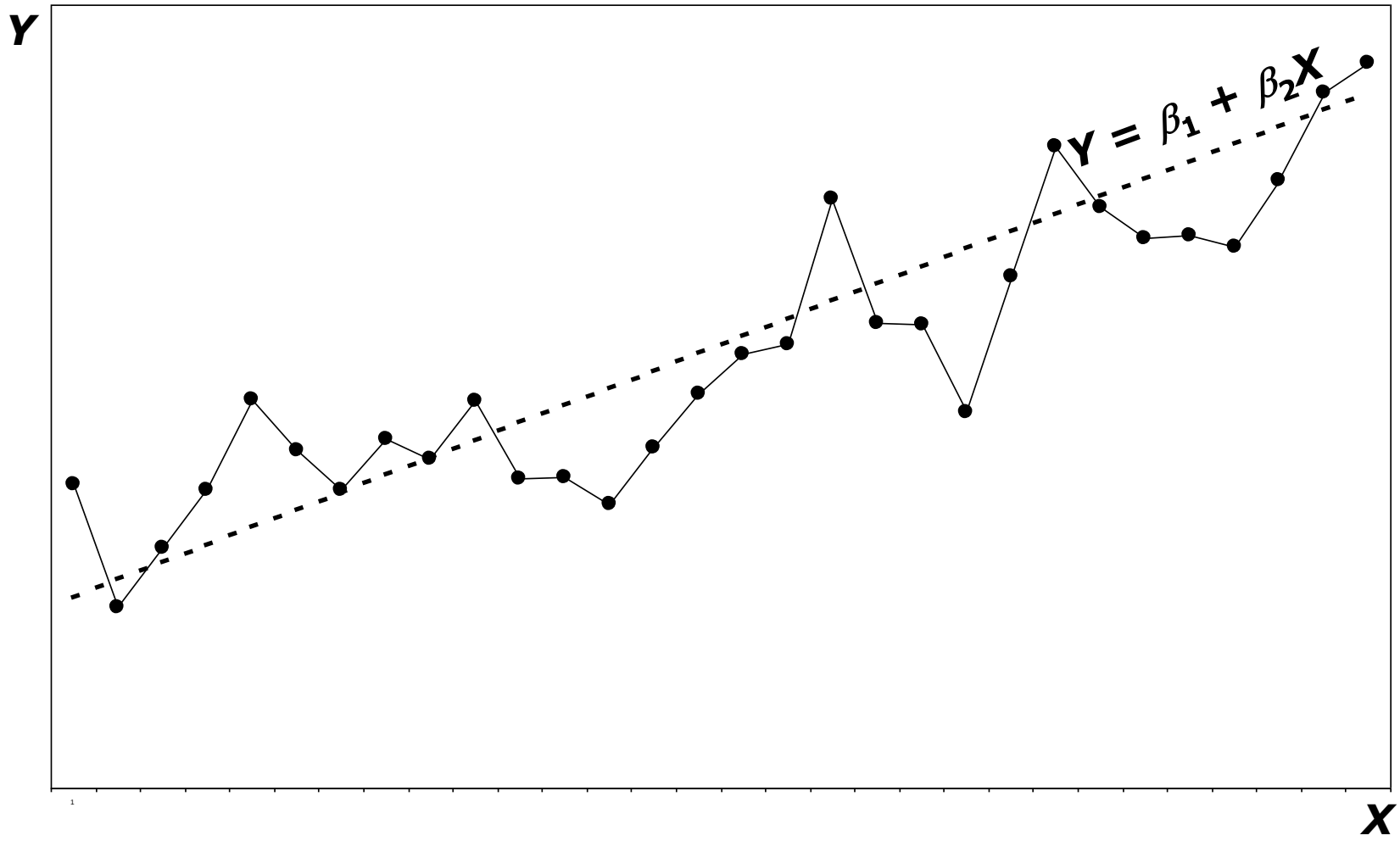
C.7 ***The disturbance term is distributed independently of the regressors***

u_t **is distributed independently of** $X_{jt'}$ **for all** t' **(including** t**) and** j

C.8 ***The disturbance term has a normal distribution***

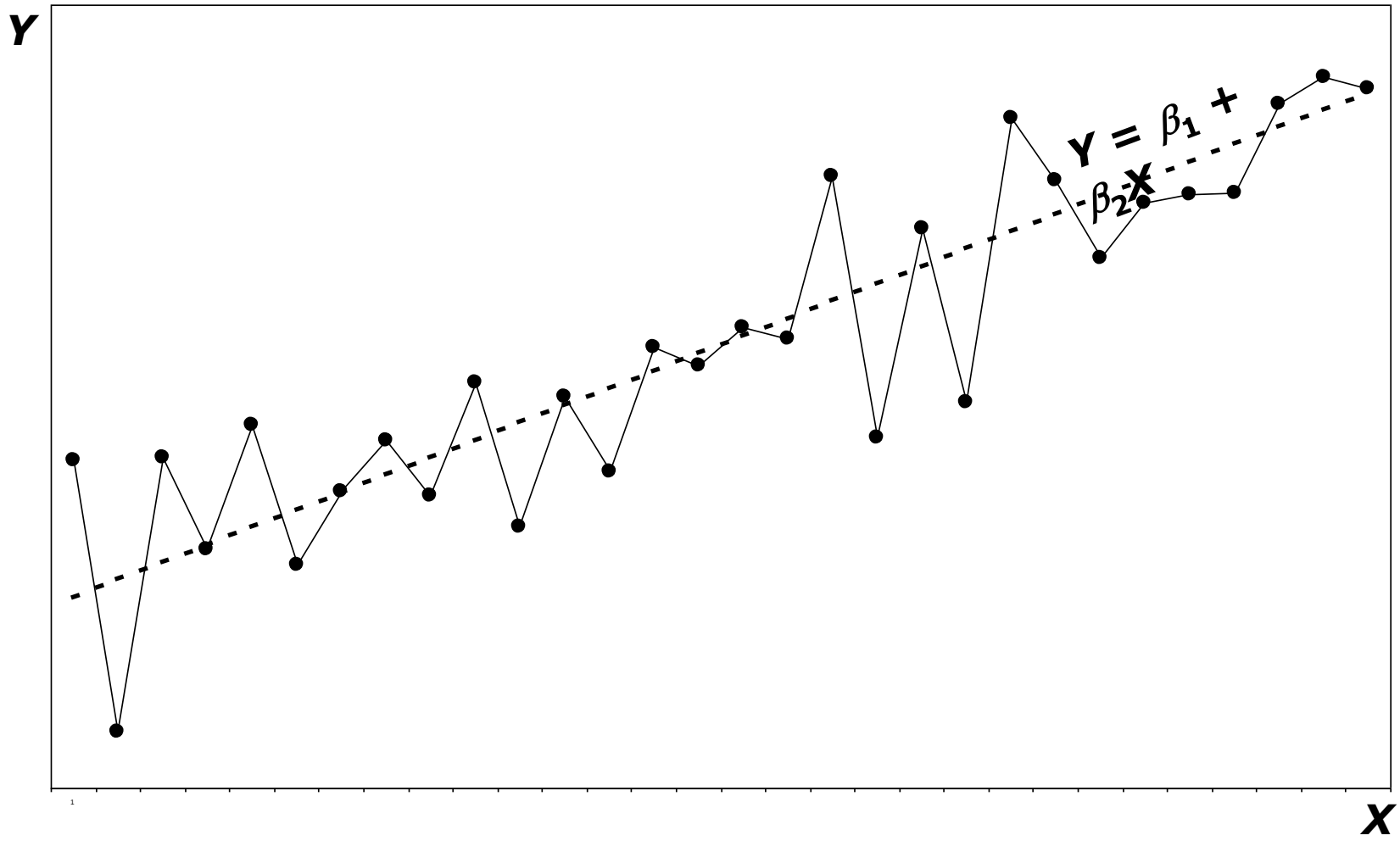
Assumption C.6 is rarely an issue with cross-sectional data. When observations are generated randomly, there is no reason to suppose that there should be any connection between the value of the disturbance term in one observation and its value in any other.

AUTOCORRELATION



In the graph above, it is clear that disturbance terms are not generated independently of each other. Positive values tend to be followed by positive ones, and negative values by negative ones. Successive values tend to have the same sign. This is described as positive autocorrelation.

AUTOCORRELATION



In this graph, positive values tend to be followed by negative ones, and negative values by positive ones. This is an example of negative autocorrelation.

AUTOCORRELATION

$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

First-order autoregressive autocorrelation: AR(1)

$$u_t = \rho u_{t-1} + \varepsilon_t$$

Fifth-order autoregressive autocorrelation: AR(5)

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \rho_3 u_{t-3} + \rho_4 u_{t-4} + \rho_5 u_{t-5} + \varepsilon_t$$

A particularly common type of autocorrelation is first-order autoregressive autocorrelation, usually denoted AR(1) autocorrelation.

It is autoregressive, because u_t depends on lagged values of itself, and first-order, because it depends only on its previous value. u_t also depends on ε_t, an injection of fresh randomness at time *t*, often described as the innovation at time *t*.

Fifth-order autocorrelation AR(5): it depends on lagged values of u_t up to the fifth lag

AUTOCORRELATION

$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

First-order autoregressive autocorrelation: AR(1)

$$u_t = \rho u_{t-1} + \varepsilon_t$$

Fifth-order autoregressive autocorrelation: AR(5)

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \rho_3 u_{t-3} + \rho_4 u_{t-4} + \rho_5 u_{t-5} + \varepsilon_t$$

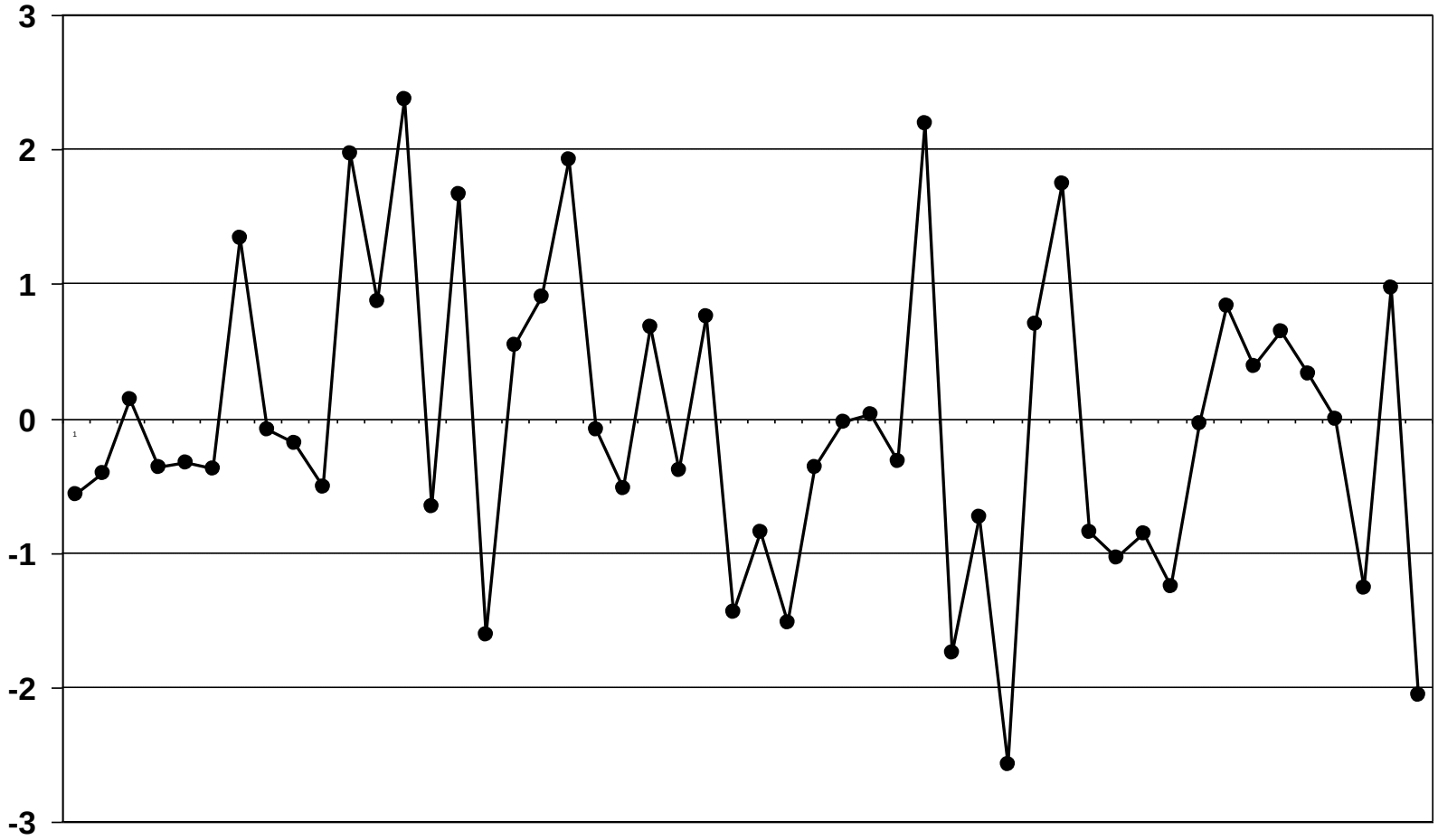
Third-order moving average autocorrelation: MA(3)

$$u_t = \lambda_0 \varepsilon_t + \lambda_1 \varepsilon_{t-1} + \lambda_2 \varepsilon_{t-2} + \lambda_3 \varepsilon_{t-3}$$

Moving average autocorrelation: the disturbance term is a linear combination of the current innovation and a finite number of previous ones.

MA(3): it depends on the three previous innovations as well as the current one.

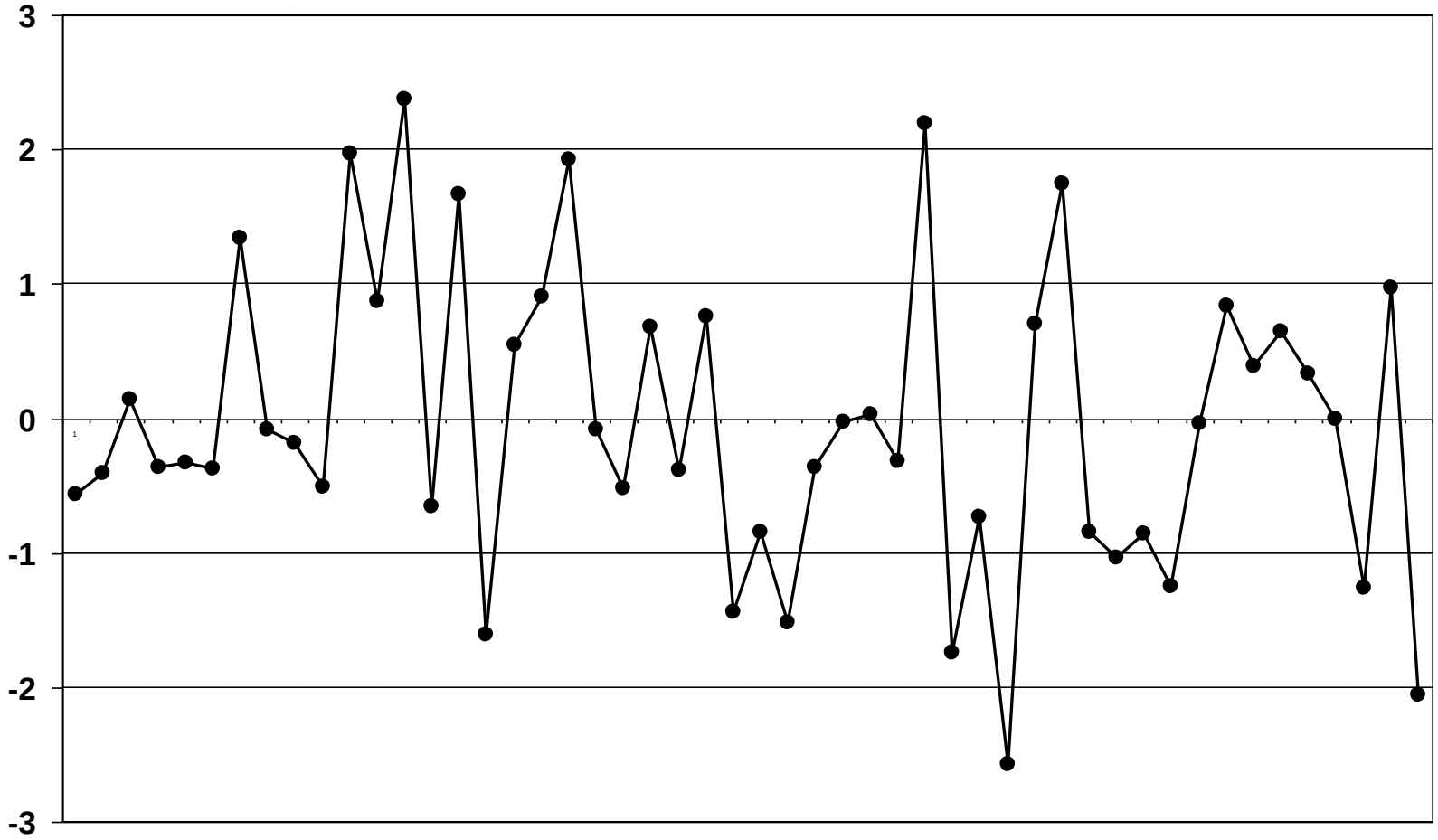
AUTOCORRELATION



$$u_t = \rho u_{t-1} + \varepsilon_t$$

The rest of this sequence gives examples of the patterns that are generated when the disturbance term is subject to AR(1) autocorrelation. The object is to provide some bench-mark images to help you assess plots of residuals in time series regressions.

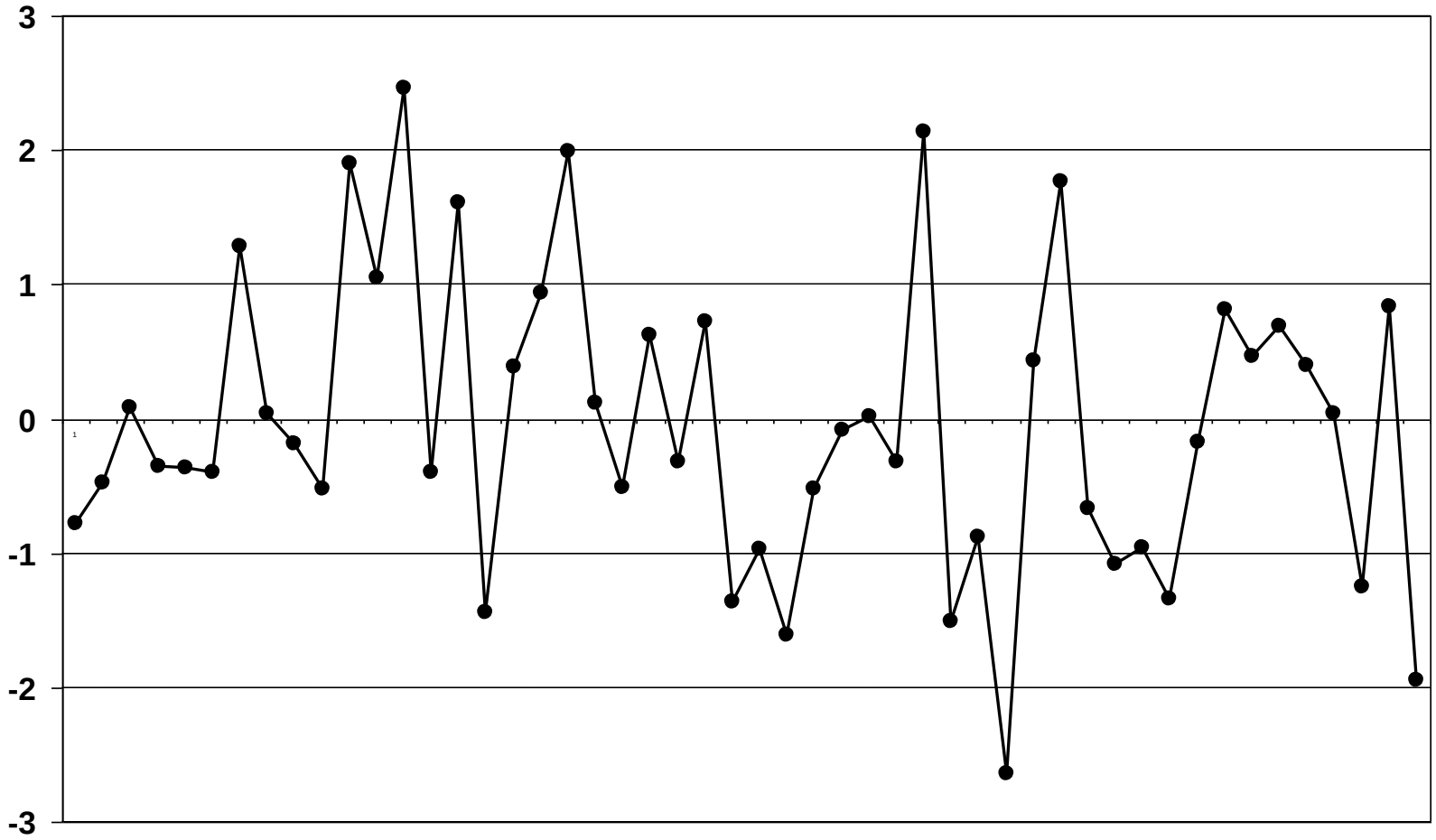
AUTOCORRELATION



$$u_t = 0.0\, u_{t-1} + \varepsilon_t$$

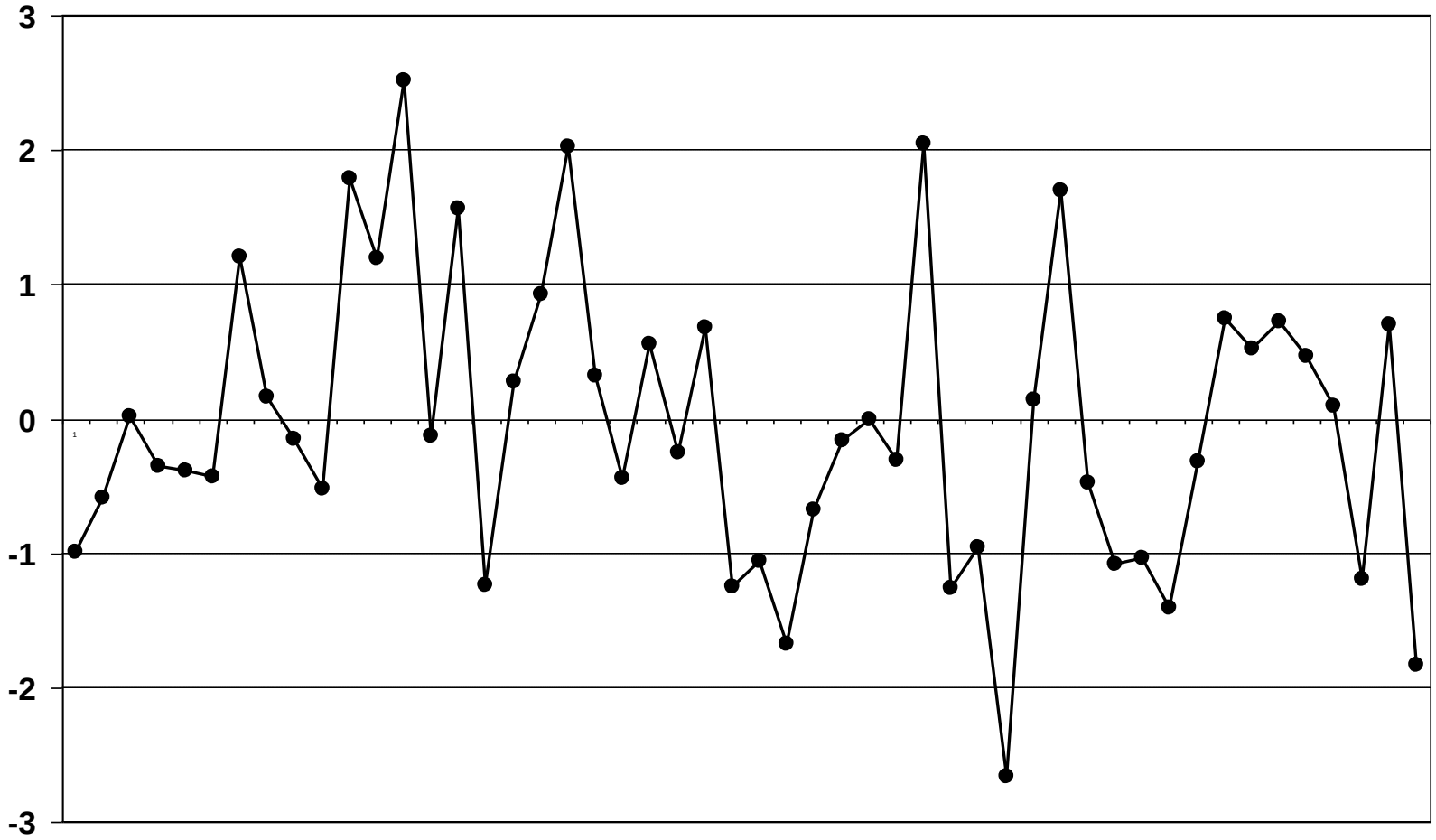
We have started with ρ equal to 0, so there is no autocorrelation. We will increase ρ progressively in steps of 0.1.

AUTOCORRELATION



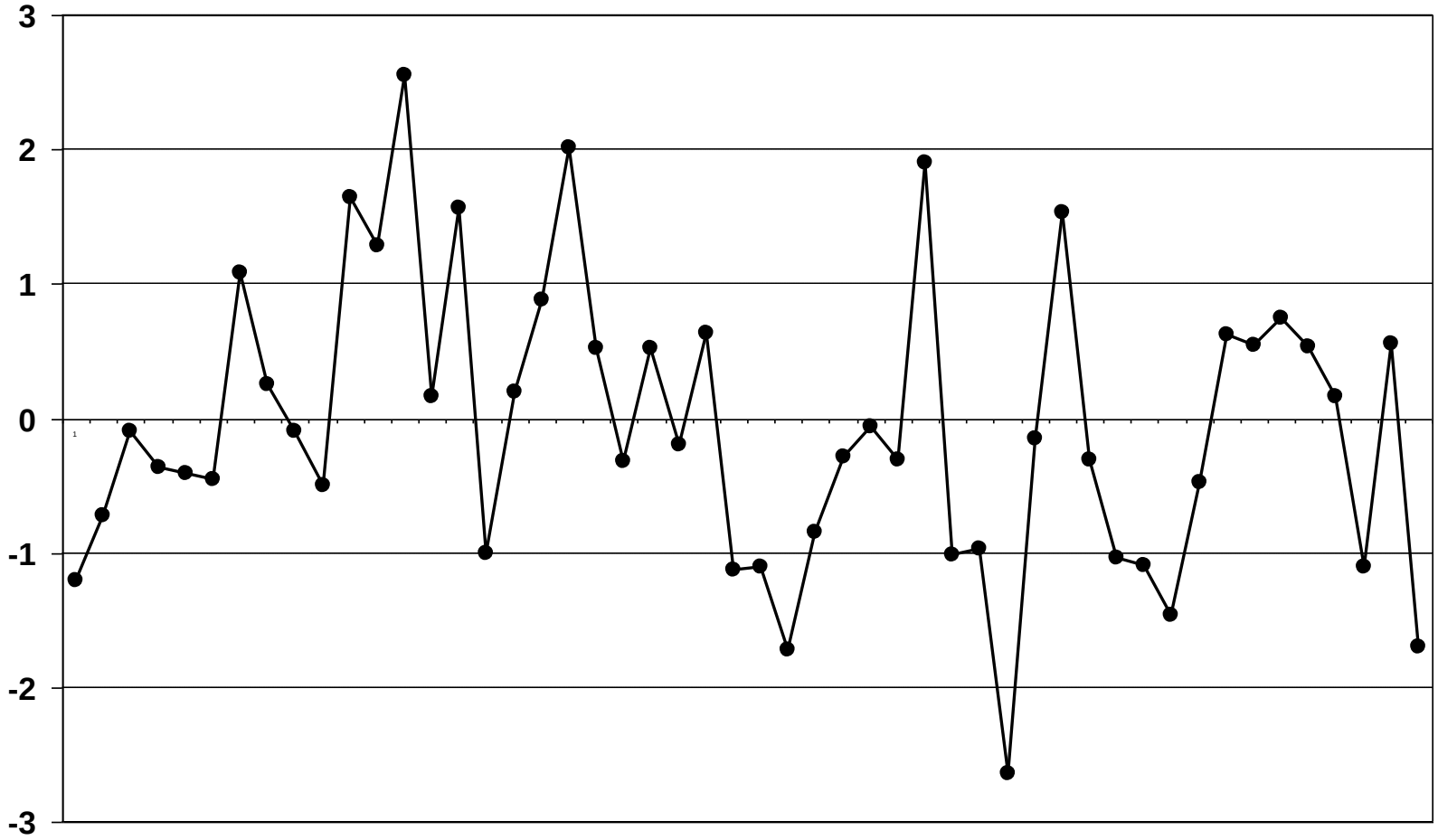
$$u_t = 0.1 u_{t-1} + \varepsilon_t$$

AUTOCORRELATION



$$u_t = 0.2\,u_{t-1} + \varepsilon_t$$

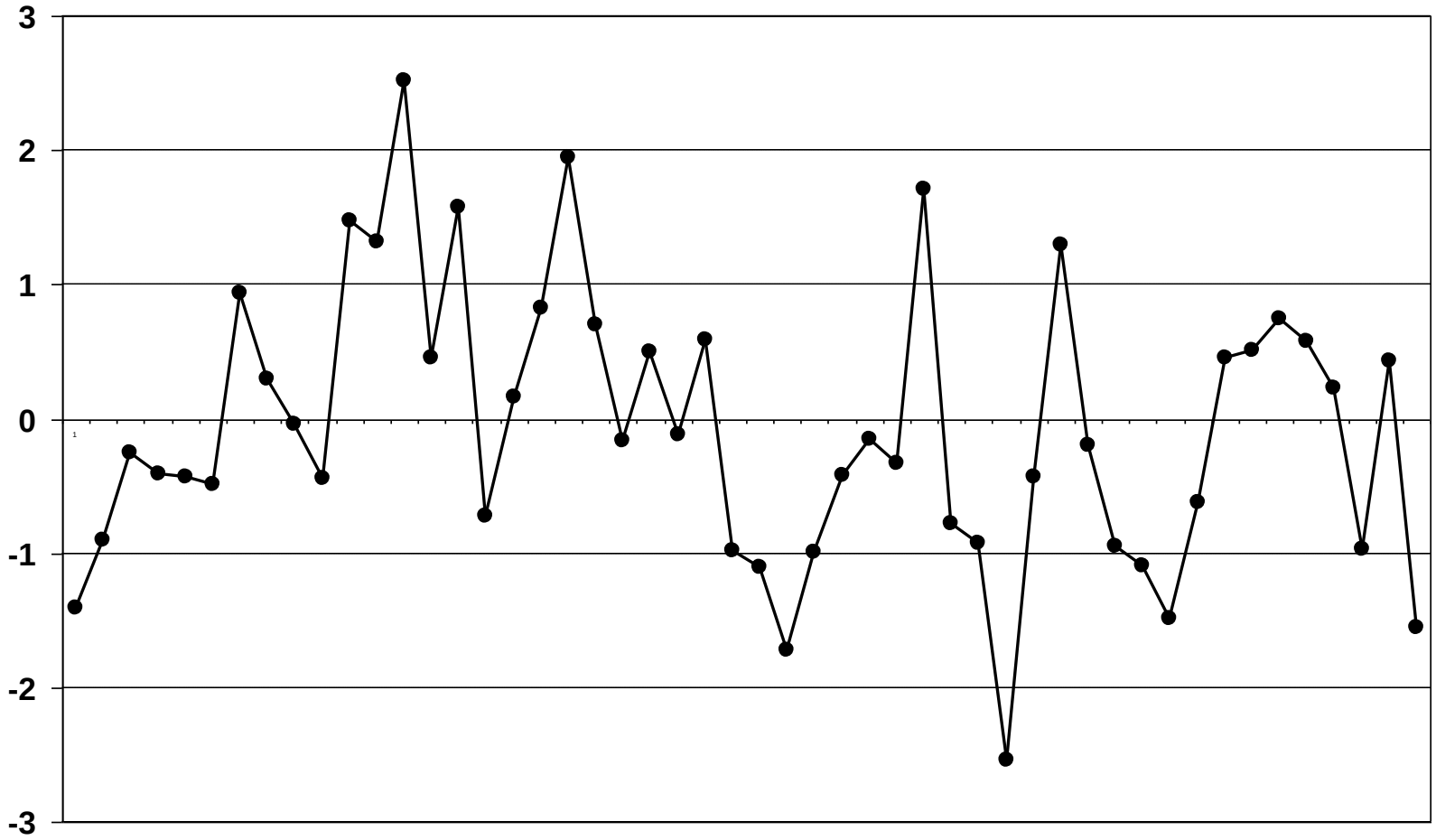
AUTOCORRELATION



$$u_t = 0.3\,u_{t-1} + \varepsilon_t$$

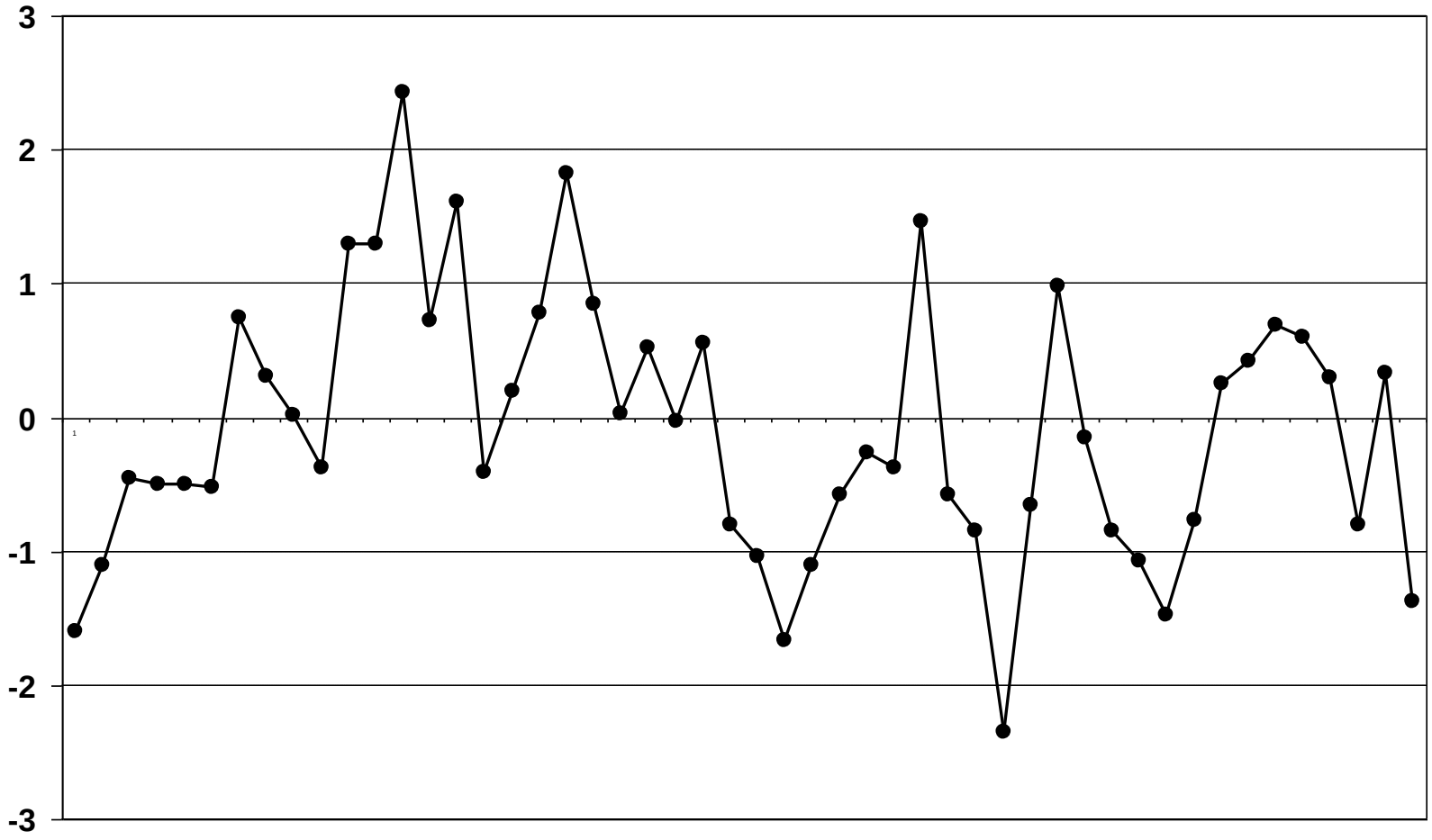
With ρ equal to 0.3, a pattern of positive autocorrelation is beginning to be apparent.

AUTOCORRELATION



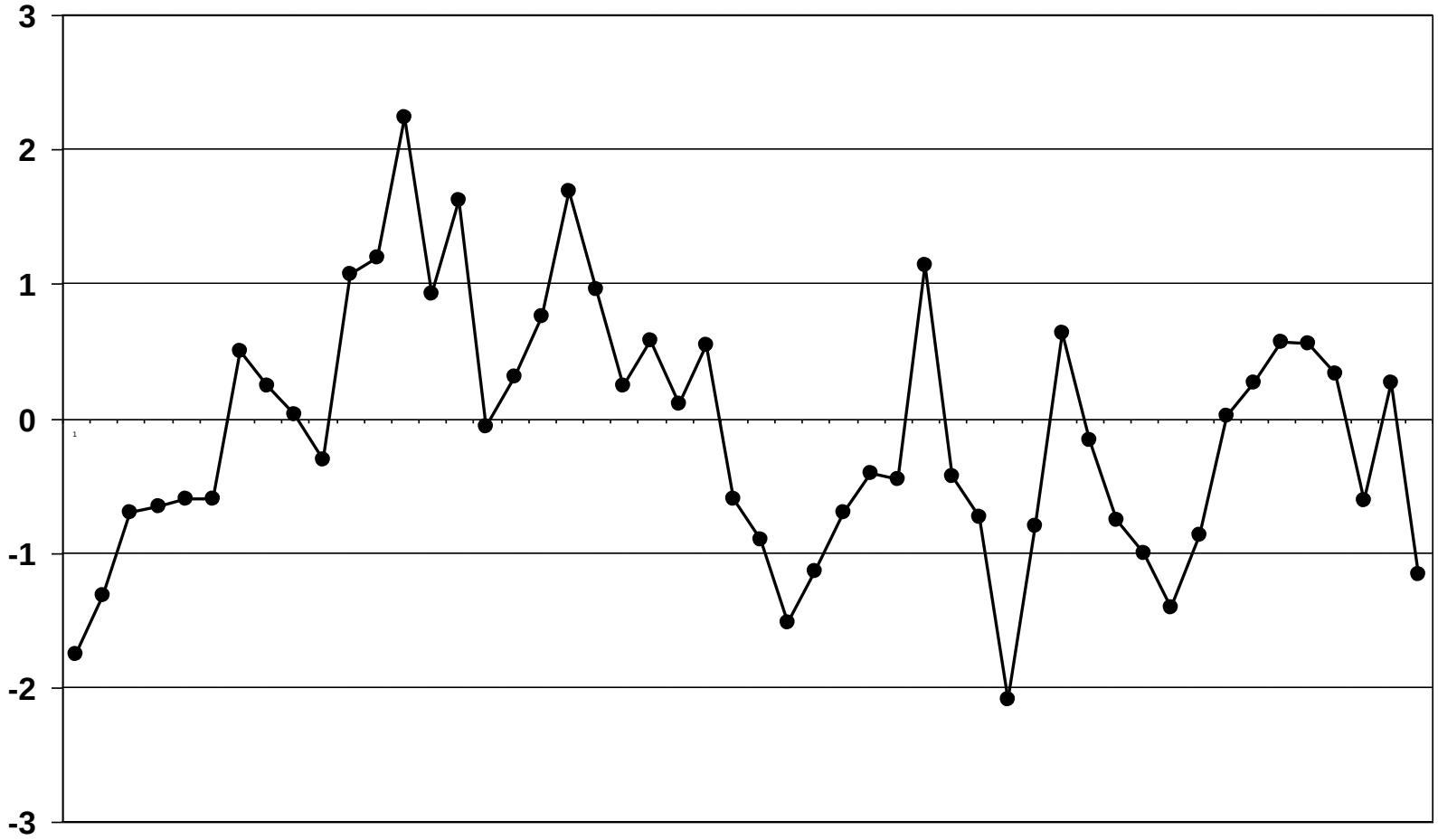
$$u_t = 0.4\,u_{t-1} + \varepsilon_t$$

AUTOCORRELATION



$$u_t = 0.5\,u_{t-1} + \varepsilon_t$$

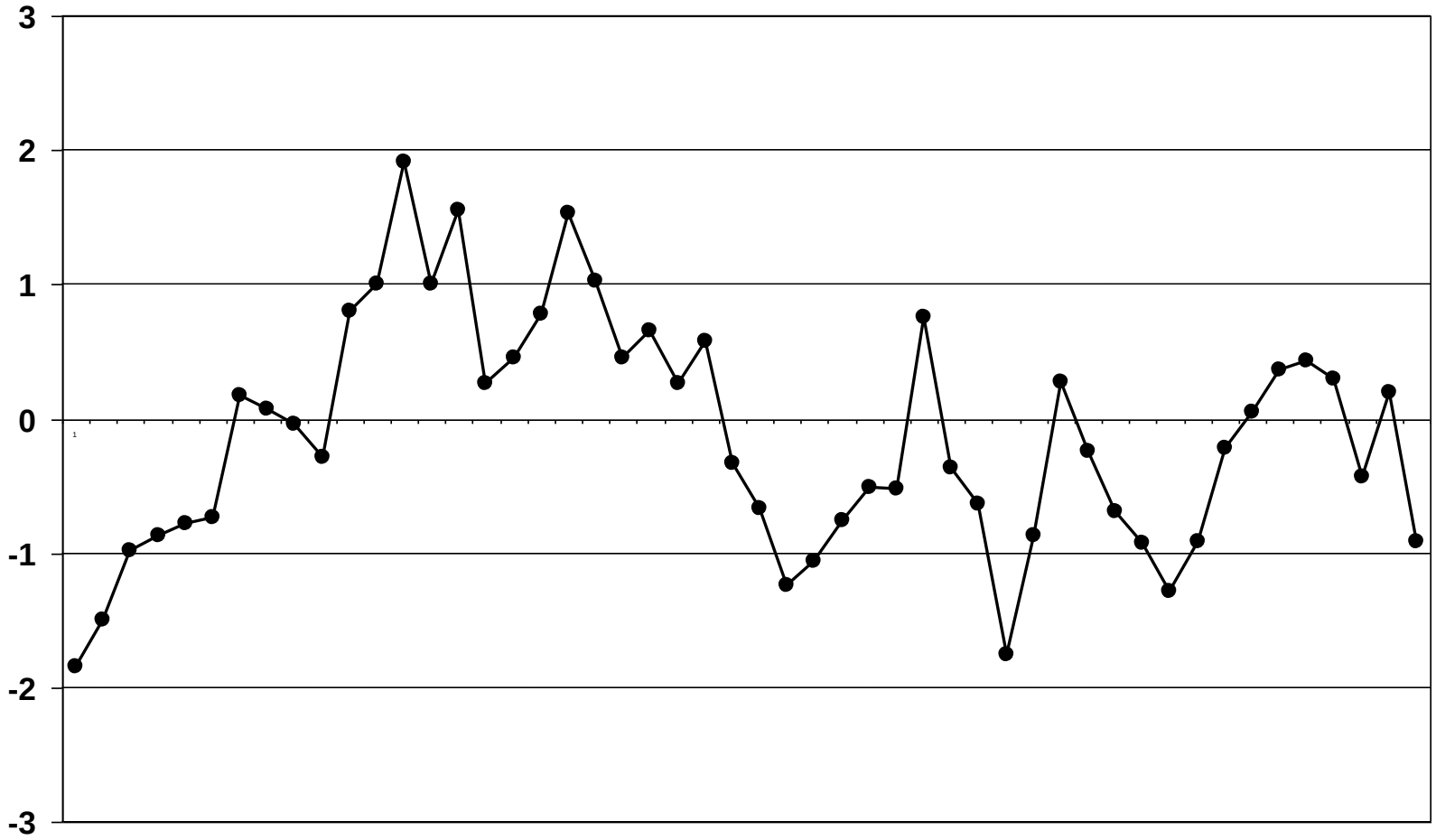
AUTOCORRELATION



$$u_t = 0.6\,u_{t-1} + \varepsilon_t$$

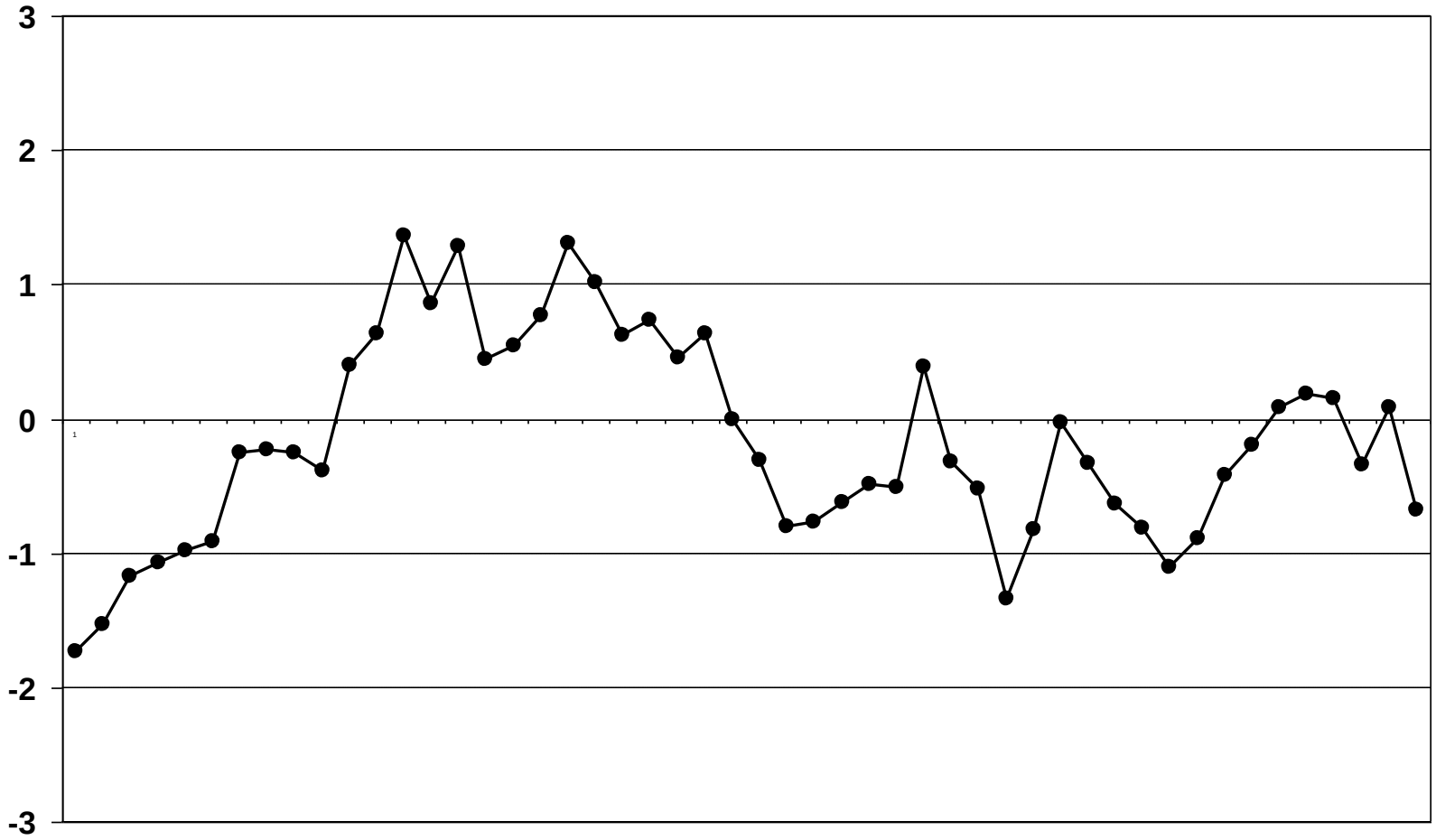
With ρ equal to 0.6, it is obvious that *u* is subject to positive autocorrelation. Positive values tend to be followed by positive ones and negative values by negative ones.

AUTOCORRELATION



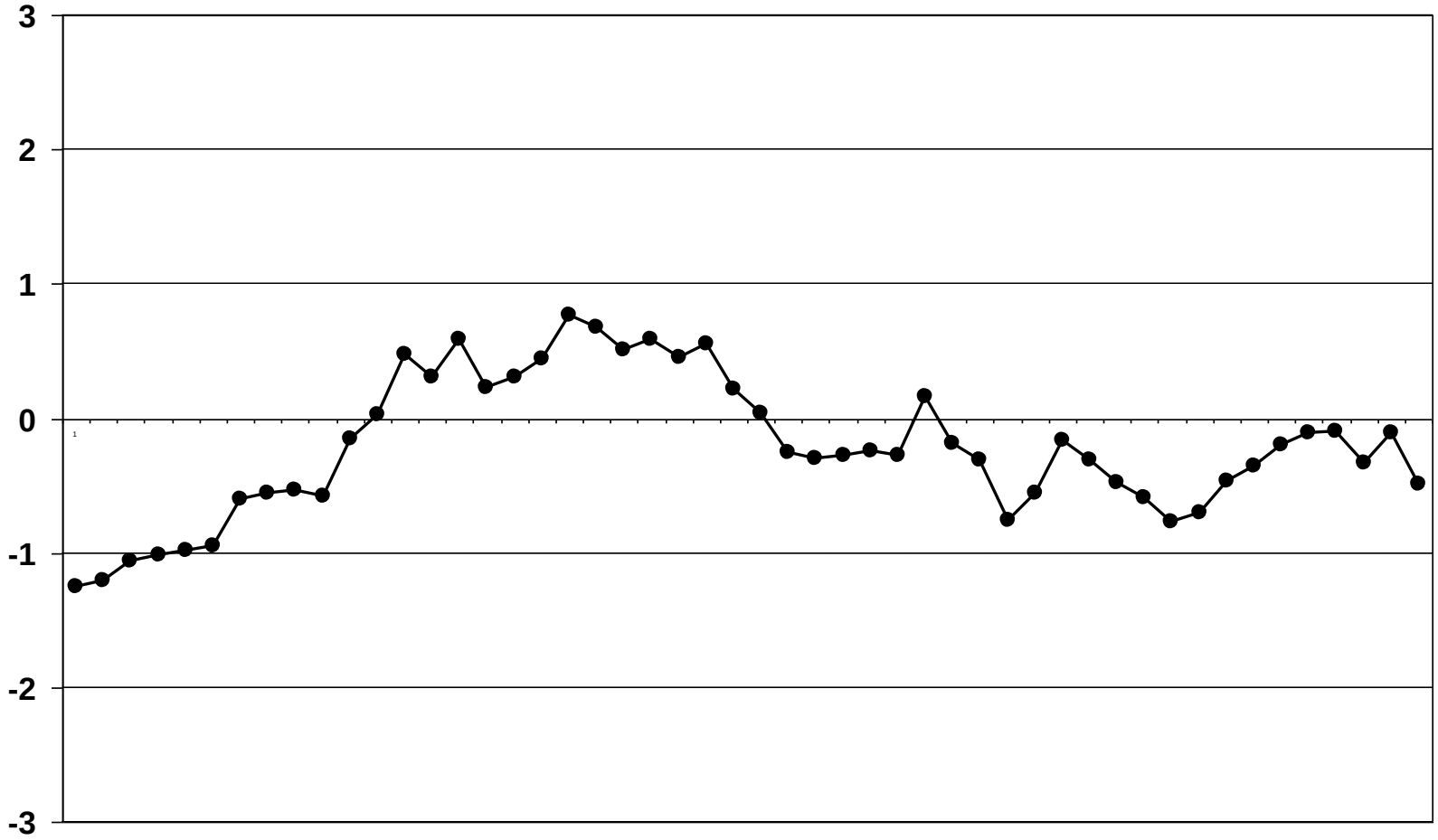
$$u_t = 0.7\,u_{t-1} + \varepsilon_t$$

AUTOCORRELATION



$$u_t = 0.8\,u_{t-1} + \varepsilon_t$$

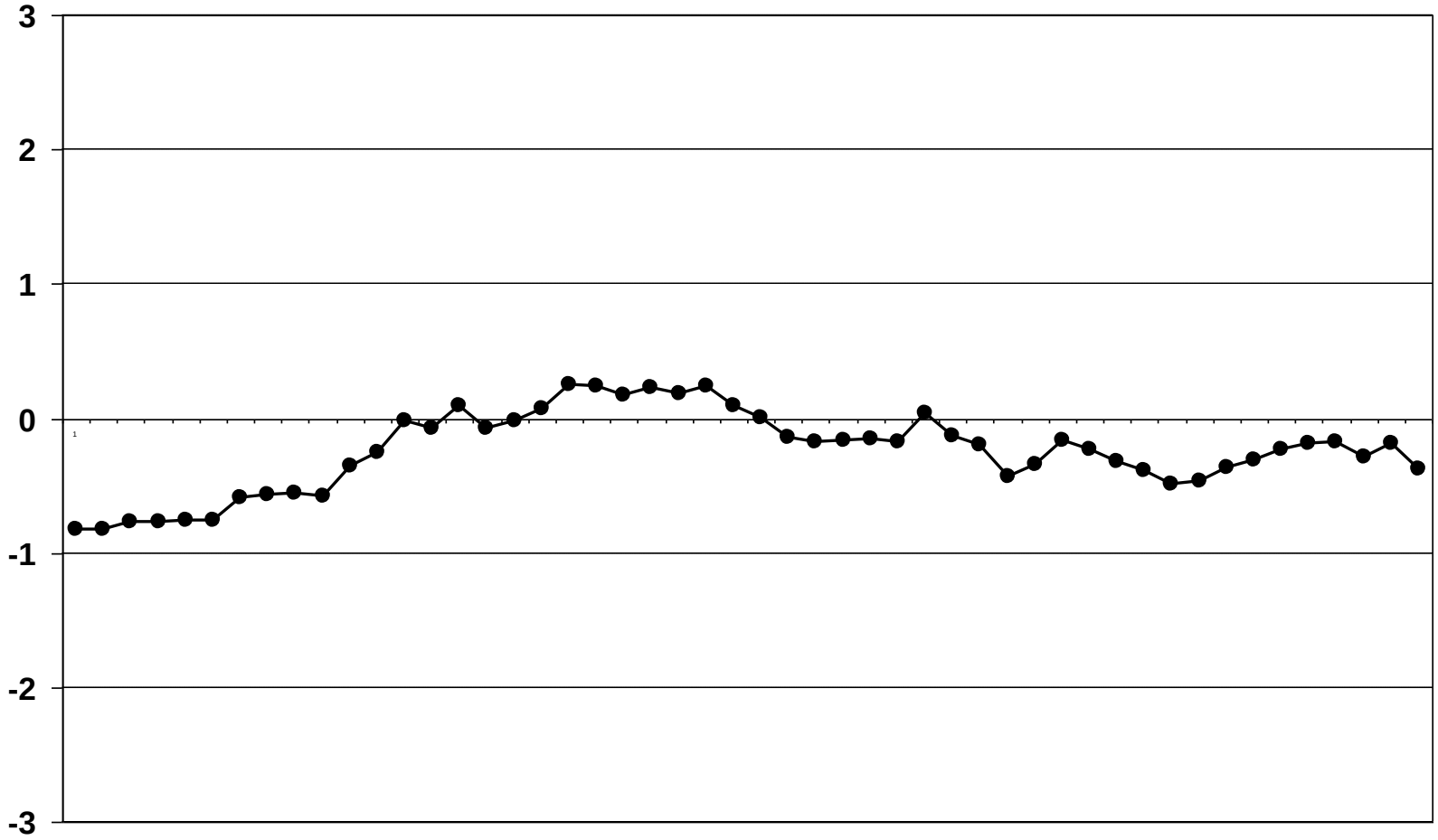
AUTOCORRELATION



$$u_t = 0.9\,u_{t-1} + \varepsilon_t$$

With ρ equal to 0.9, the sequences of values with the same sign have become long and the tendency to return to 0 has become weak.

AUTOCORRELATION



$$u_t = 0.95\, u_{t-1} + \varepsilon_t$$

The process is now approaching what is known as a random walk, where ρ is equal to 1 and the process becomes nonstationary. The terms random walk and nonstationarity will be defined in the next chapter. For the time being we will assume $|\rho| < 1$.

STATIONARY PROCESSES

X_t is stationary if $E(X_t)$, $\sigma^2_{X_t}$, and the population covariance of X_t and X_{t+s} are independent of t

$$X_t = \beta_2 X_{t-1} + \varepsilon_t \qquad -1 < \beta_2 < 1$$

$$E(X_t) = \beta_2^t X_0 \rightarrow 0$$

$$\sigma^2_{X_t} = \frac{1-\beta_2^{2t}}{1-\beta_2^2}\sigma^2_\varepsilon \rightarrow \frac{1}{1-\beta_2^2}\sigma^2_\varepsilon$$

$$\text{population covariance of } X_t \text{ and } X_{t+s} = \frac{\beta_2^s}{1-\beta_2^2}\sigma^2_\varepsilon$$

A time series X_t is said to be stationary if its expected value and population variance are independent of time and if the population covariance between its values at time t and time $t + s$ depends on s but not on t.

An example of a stationary time series is an AR(1) process $X_t = \beta_2 X_{t-1} + \varepsilon_t$, provided that $-1 < \beta_2 < 1$, where ε_t is a random variable with 0 mean and constant variance and not subject to autocorrelation.

NONSTATIONARY PROCESSES

Random walk

$$X_t = X_{t-1} + \varepsilon_t$$

$$X_t = X_0 + \varepsilon_1 + \ldots + \varepsilon_{t-1} + \varepsilon_t$$

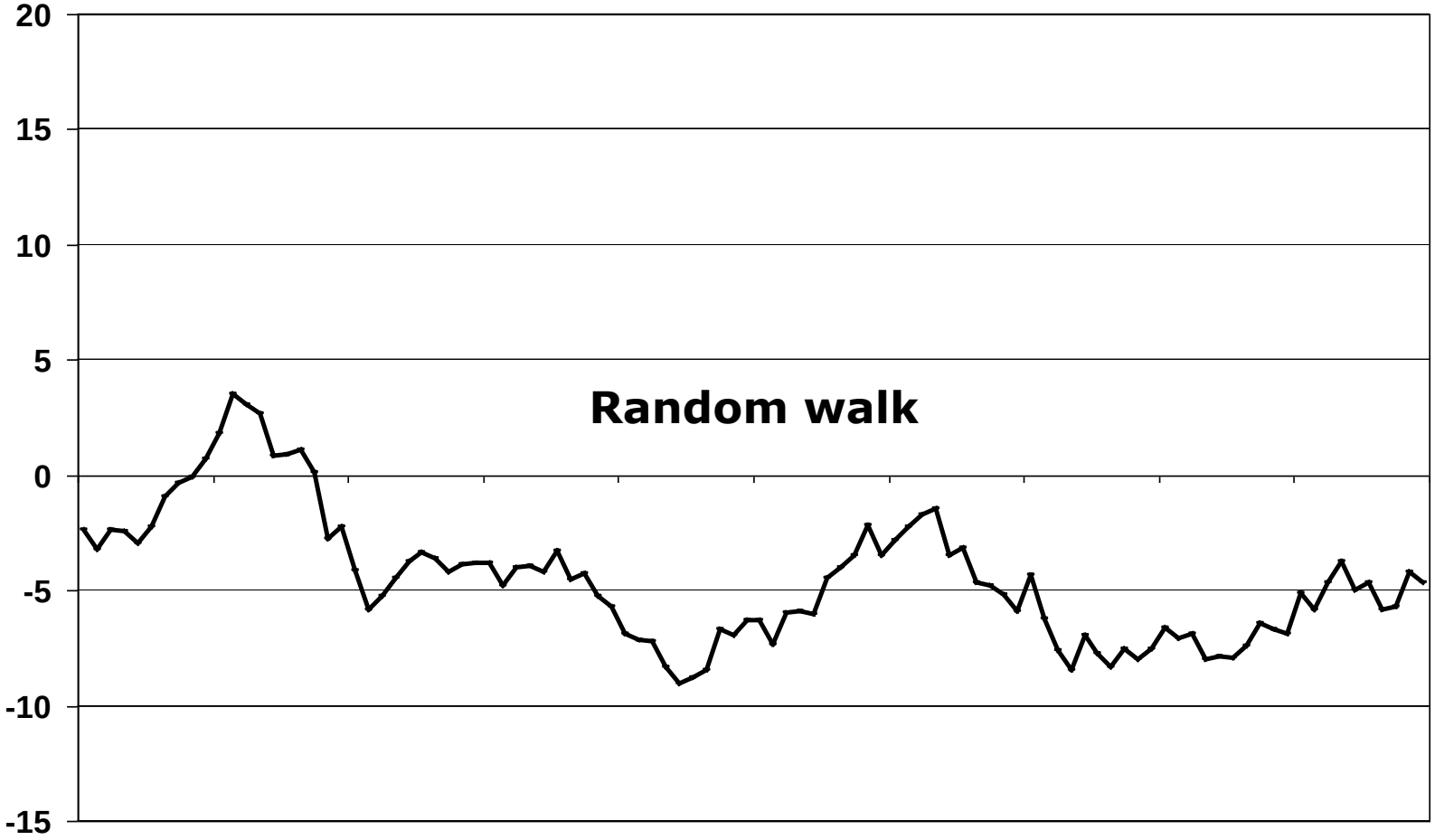
$$E(X_t) = X_0 + E(\varepsilon_1) + \ldots + E(\varepsilon_n) = X_0$$

$$\begin{aligned}
\sigma^2_{X_t} &= \text{population variance of } (X_0 + \varepsilon_1 + \ldots + \varepsilon_{t-1} + \varepsilon_t) \\
&= \text{population variance of } (\varepsilon_1 + \ldots + \varepsilon_{t-1} + \varepsilon_t) \\
&= \sigma^2_\varepsilon + \ldots + \sigma^2_\varepsilon + \sigma^2_\varepsilon \\
&= t\sigma^2_\varepsilon
\end{aligned}$$

The condition $-1 < \beta_2 < 1$ was crucial for stationarity. If $\beta_2 = 1$, the series becomes a nonstationary process known as a random walk, $\varepsilon \sim (0, \sigma_\varepsilon)$

***E*(X_t) is independent of *t* and the first condition for stationarity remains satisfied. However, the condition that the variance of X_t be independent of time is not satisfied.**

NONSTATIONARY PROCESSES



The chart shows a typical random walk. If it were a stationary process, there would be a tendency for the series to return to 0 periodically. Here there is no such tendency.

REGRESSION ANALYSIS WITH PANEL DATA

INTRODUCTION

- **A panel data set, or longitudinal data set, is one where there are repeated observations on the same units.**
- **The units may be individuals, households, enterprises, countries, or any set of entities that remain stable through time.**
- **A *balanced* panel is one where every unit is surveyed in every time period. An *unbalanced* panel is one where some units have not been surveyed (missing observations).**

INTRODUCTION (cont.)

- **Panel data sets have several advantages over cross-section data sets:**
  - **overcome a problem of bias caused by unobserved heterogeneity.**
  - **investigate dynamics without relying on retrospective questions that may yield data subject to measurement error.**
  - **often very large. If there are *n* units and *T* time periods, the potential number of observations is *nT*.**
  - **often well designed and have high response rates.**

REGRESSION ANALYSIS WITH PANEL DATA: INTRODUCTION

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \sum_{p=1}^{s} \gamma_p Z_{pi} + \delta t + \varepsilon_{it}$$

$$\alpha_i = \sum_{p=1}^{s} \gamma_p Z_{pi}$$

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \alpha_i + \delta t + \varepsilon_{it}$$

Index *i* refers to the unit of observation, *t* refers to the time period, and *j* and *p* differentiate observed with unobserved explanatory variables.

X_j : observed variables; Z_p : unobserved variables; ε_{it} : disturbance term.

The X_j variables are usually the variables of interest, while the Z_p variables are responsible for unobserved heterogeneity and constitute a nuisance component of the model.

α_i, known as the unobserved effect, representing the joint impact of the Z_p variables on Y_i

FIXED EFFECTS REGRESSIONS: WITHIN-GROUPS METHOD

Fixed effects estimation (within-groups method)

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \alpha_i + \delta t + \varepsilon_{it}$$

$$\overline{Y_i} = \beta_1 + \sum_{j=2}^{k} \beta_j \overline{X_{ji}} + \alpha_i + \delta \bar{t} + \overline{\varepsilon_i}$$

$$Y_{it} - \overline{Y_i} = \sum_{j=2}^{k} \beta_j (X_{jit} - \overline{X_{ji}}) + \delta (t - \bar{t}) + \varepsilon_{it} - \overline{\varepsilon_i}$$

(1) the mean values of the variables in the observations on a given individual are calculated by averaging the observations for that individual. The unobserved effect α_i is unaffected because it is the same for all observations for that individual.

(2) If the second equation is subtracted from the first, the unobserved effect disappears.

Disadvantage: First, the intercept β_1 and any *X* variable that remains constant for each individual will drop out of the model.

FIXED EFFECTS REGRESSIONS: FIRST-DIFFERENCES METHOD

Fixed effects estimation (first-differences method)

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \alpha_i + \delta t + \varepsilon_{it}$$

$$Y_{it-1} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit-1} + \alpha_i + \delta (t-1) + \varepsilon_{it-1}$$

$$Y_{it} - Y_{it-1} = \sum_{j=2}^{k} \beta_j (X_{jit} - X_{jit-1}) + \delta + \varepsilon_{it} - \varepsilon_{it-1}$$

$$\Delta Y_{it} = \sum_{j=2}^{k} \beta_j \Delta X_{jit} + \delta + \varepsilon_{it} - \varepsilon_{it-1}$$

The unobserved effect is eliminated by subtracting the observation for the previous time period from the observation for the current time period, for all time periods.

Subtracting the second equation from the first, one obtains the third, rewritten as the fourth, and again the unobserved heterogeneity has disappeared.

RANDOM EFFECTS REGRESSIONS

Random effects estimation

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \sum_{p=1}^{s} \gamma_p Z_{pi} + \delta t + \varepsilon_{it}$$

$$Y_{it} = \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \alpha_i + \delta t + \varepsilon_{it}$$

$$= \beta_1 + \sum_{j=2}^{k} \beta_j X_{jit} + \delta t + u_{it} \qquad u_{it} = \alpha_i + \varepsilon_{it}$$

When the observed variables of interest are constant for each individual, a fixed effects regression is not an effective tool because such variables cannot be included. In this case each of the unobserved Z_p variables is treated as being drawn randomly from a given distribution.

Application of Econometrics in Finance

APPLICATION IN FINANCE CAPM

- Example : Suppose you consider investing 100 mill in 3 years either on A or on the G. bond rate, which is given at the rate of 10% per year

Expected return on A	After 3 years
10% =>	133 mill?
11.5%=>	138 mill?
12% =>	140mill
15%=>	150 mill?

Which one would you choose?

APPLICATION IN FINANCE – CAPM

- => Uncertainty (or risk) about the outcome should be compensated
- => Q: how to measure the risk of an asset, a portfolio?
- What about the variance?
- Consider an example:

	Air conditioner	Blanket
Mean	15	15
Variance	4	4

APPLICATION IN FINANCE
CAPM

- Options to choose

	A	A	(A+B)/2
Mean	14	14	14
Variance	4	4	?

- => some part of risk can be reduced by diversifying

APPLICATION IN FINANCE – CAPM

- => Can it be reduced to zero? if you still wanna to make a return rate at 14%?
- Total risk consists of (market risk & firm's specific risk)
- Market risk: you can't reduce
- => beta: measure of a stock's volatility in relation to the market's / "to base line"
- => beta(A) =1.5 means: when market return (+/-) 1% then we expect the A's return (+/-) 1.5%
- => large beta: the stock swings more than the market.
- => small beta: the stock is more stable than the market.
- How to measure the beta?=> CAPM (go to word.file)