Business Analytics

Simple Linear Regression Analysis

Faton Berisha

Chapter 12

Simple Linear Regression Analysis

Simple Linear Regression

Simple Linear Regression Continued

The Simple Linear Regression Model

- The *dependent* (or response) variable is the variable we wish to understand or predict
- The *independent* (or predictor) variable is the variable we will use to understand or predict the dependent variable
- *Regression analysis* is a statistical technique that uses observed data to relate the dependent variable to one or more independent variables

Objective of Regression Analysis

The objective of regression analysis is to build a *regression model* (or predictive equation) that can be used to describe, predict and control the dependent variable on the basis of the independent variable

Form of The Simple Linear Regression Model

$$y = \mu_{y|x} + \varepsilon = \beta_0 + \beta_1 x + \varepsilon$$

$\mu_{y|x} = \beta_0 + \beta_1 x + \varepsilon$ is the mean value of the dependent variable y when the value of the independent variable is x

β_0 is the y-intercept; the mean of y when x is 0

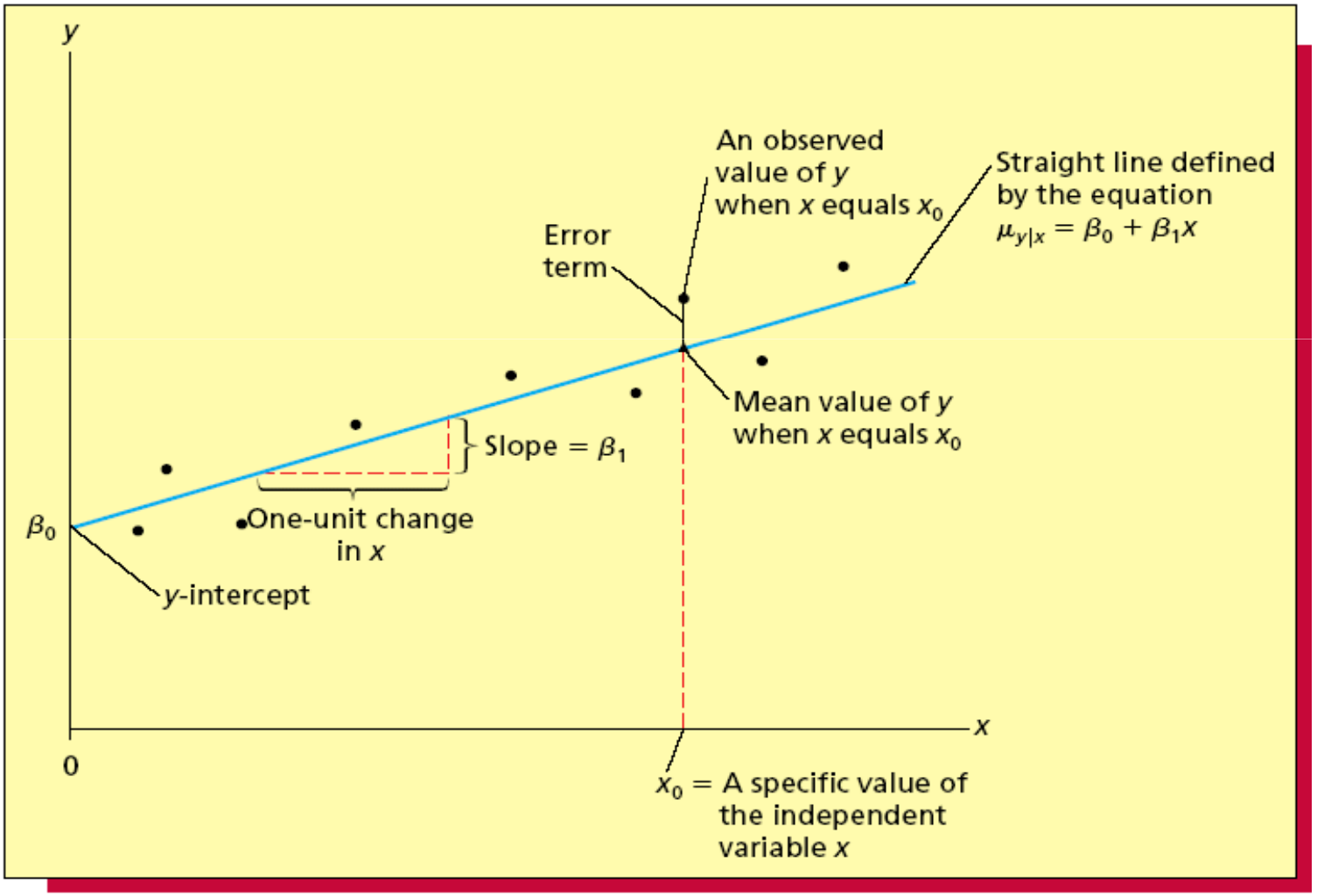
β_1 is the slope; the change in the mean of y per unit change in x

ε is an error term that describes the effect on y of all factors other than x

Regression Terms

- β_0 and β_1 are called regression parameters
- β_0 is the y-intercept and β_1 is the slope
- We do not know the true values of these parameters
- So, we must use sample data to estimate them
- b_0 is the estimate of β_0 and b_1 is the estimate of β_1

The Simple Linear Regression Model
Illustrated



The Least Squares Point Estimates

Estimation/prediction equation

$$\hat{y} = b_0 + b_1 x$$

Least squares point estimate of the slope β_1

$$b_1 = \frac{SS_{xy}}{SS_{xx}}$$

$$SS_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$$

$$SS_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

The Least Squares Point Estimates
Continued

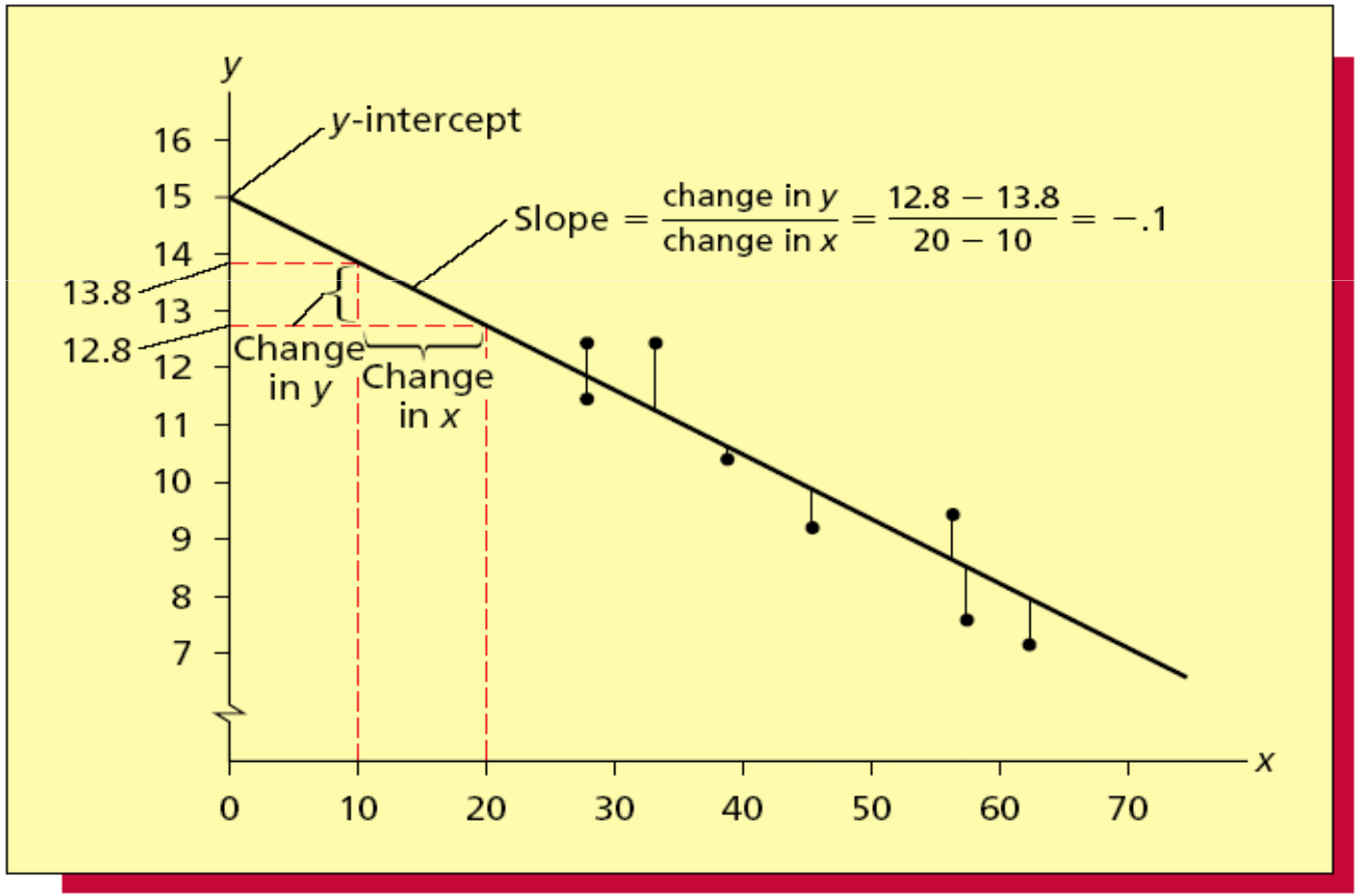
Least squares point estimate of the y-intercept β_0

$$b_0 = \bar{y} - b_1\bar{x} \qquad \bar{y} = \frac{\sum y_i}{n} \qquad \bar{x} = \frac{\sum x_i}{n}$$

Example 12.3: Fuel Consumption

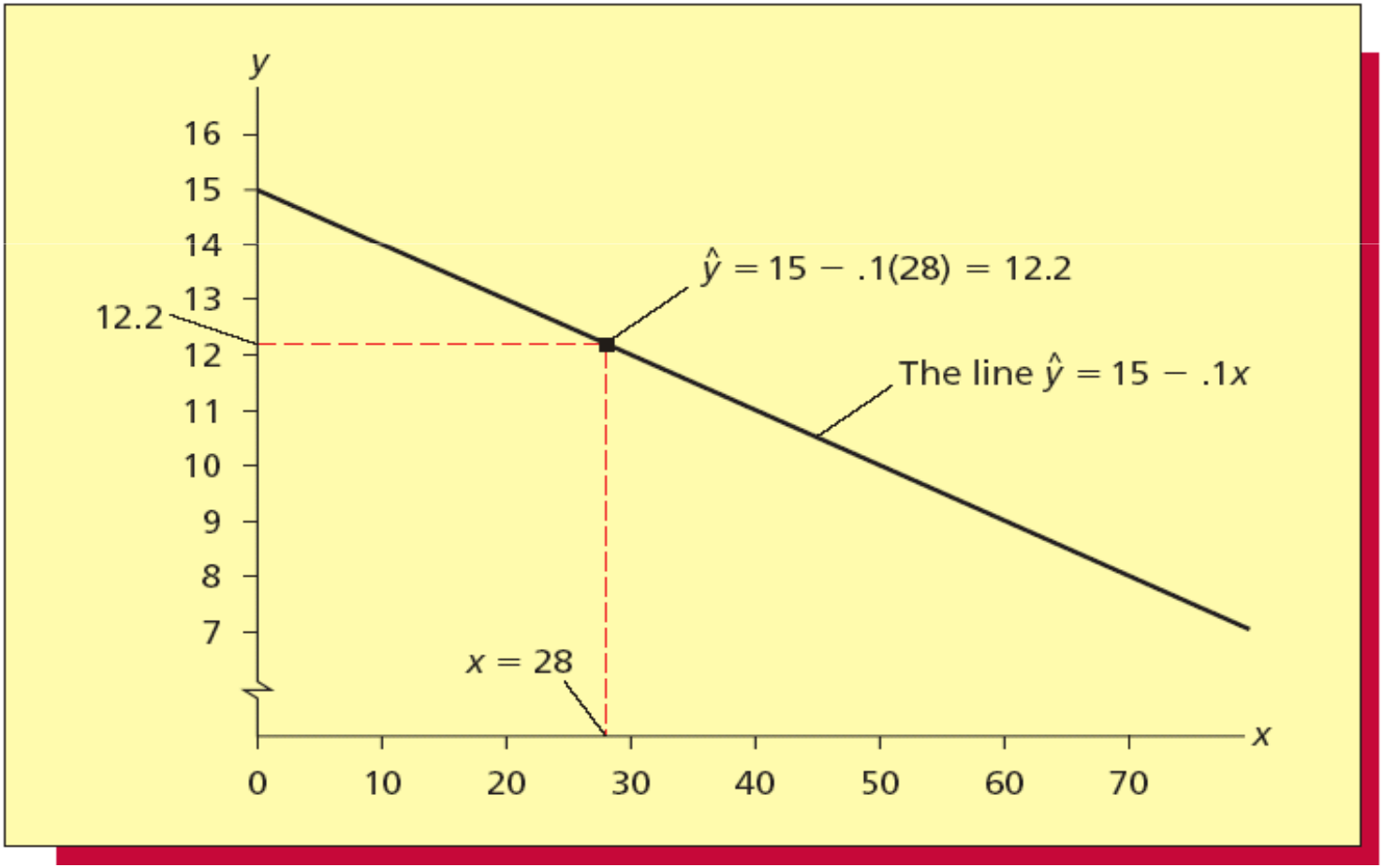
Case #1

Visually fitting a line to the fuel consumption data



Example 12.3: Fuel Consumption
Case #2

Using the visually fitted line to predict when x=28



Example 12.4: Fuel Consumption

Case #1

y	x	x^2	xy
12.4	28.0	784.00	347.20
11.7	28.0	784.00	327.60
12.4	32.5	1056.25	403.00
10.8	39.0	1521.00	421.20
9.4	45.9	2106.81	431.46
9.5	57.8	3340.84	549.10
8.0	58.1	3375.61	464.80
7.5	62.5	3906.25	468.75
81.7	**351.8**	**16874.76**	**3413.11**

Example 12.4: Fuel Consumption
Case #2

- From last slide,
  - $\Sigma y_i = 81.7$
  - $\Sigma x_i = 351.8$
  - $\Sigma x^2_i = 16{,}874.76$
  - $\Sigma x_i y_i = 3{,}413.11$
- Once we have these values, we no longer need the raw data
- Calculation of b_0 and b_1 uses these totals

Example 12.4: Fuel Consumption
Case #3

Slope b_1

$$SS_{xy} = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$$

$$= 3413.11 - \frac{(351.8)(81.7)}{8} = -179.6475$$

$$SS_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

$$= 16874.76 - \frac{(351.8)^2}{8} = 1404.355$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{-179.6475}{1404.355} = -0.1279$$

Example 12.4: Fuel Consumption

Case #4

y-Intercept b_0

$$\bar{y} = \frac{\sum y_i}{n} = \frac{81.7}{8} = 10.2125$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{351.8}{8} = 43.98$$

$$b_0 = \bar{y} - b_1\bar{x}$$

$$= 10.2125 - (-0.1279)(43.98)$$

$$= 15.84$$

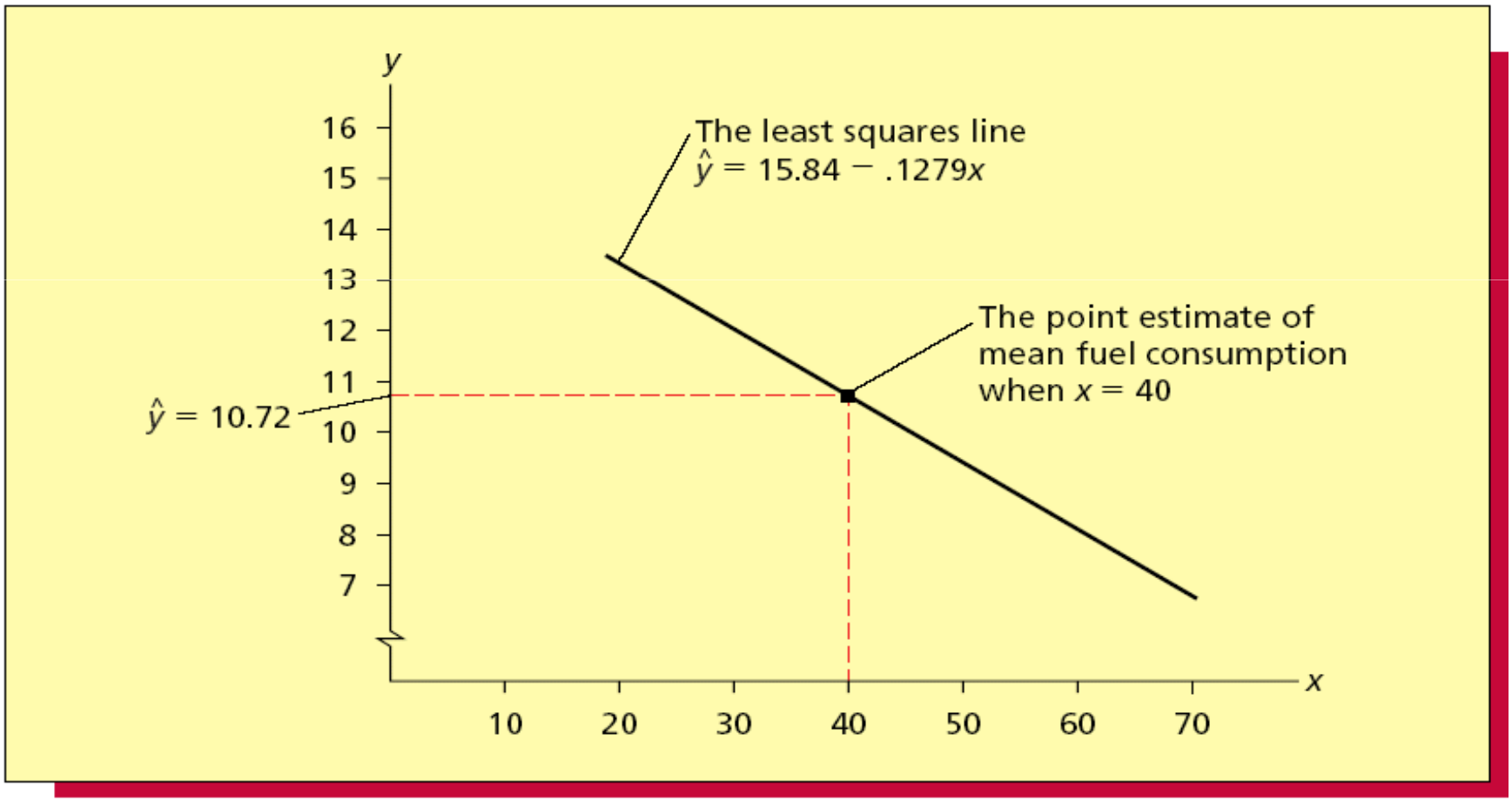
Example 12.4: Fuel Consumption
Case #5

Prediction (x = 40)

$$\hat{y} = b_0 + b_1 x = 15.84 - 0.1279(40)$$
$$= 10.72 \text{ MMcf of Gas}$$

Example 12.4: Fuel Consumption

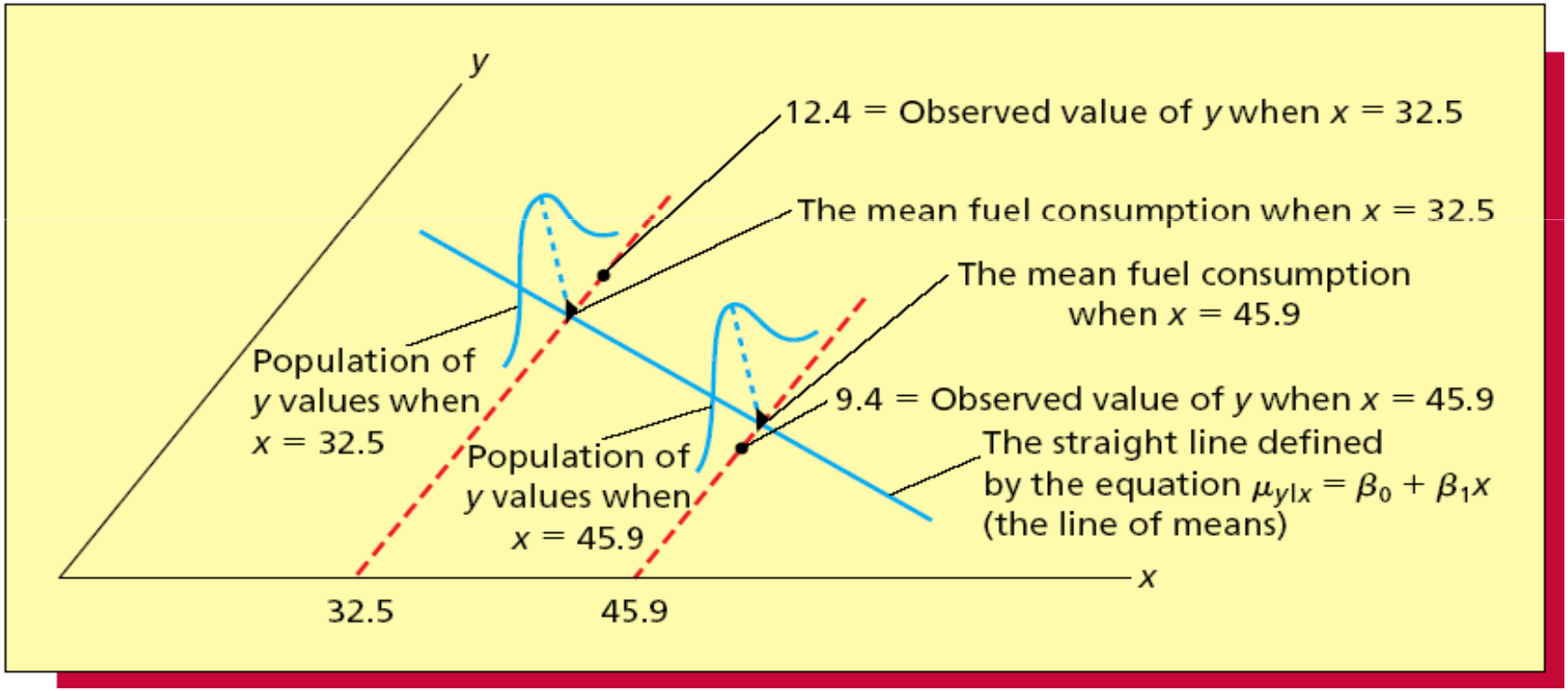
Case #6



Model Assumptions

1. **Mean of Zero**
   At any given value of x, the population of potential error term values has a mean equal to zero
2. **Constant Variance Assumption**
   At any given value of x, the population of potential error term values has a variance that does not depend on the value of x
3. **Normality Assumption**
   At any given value of x, the population of potential error term values has a normal distribution
4. **Independence Assumption**
   Any one value of the error term ε is statistically independent of any other value of ε

Model Assumptions Illustrated



Sum of Squared Errors

$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

Mean Square Error

- This is the point estimate of the residual variance σ^2
- SSE is from last slide

$$s^2 = MSE = \frac{SSE}{n-2}$$

Standard Error

- This is the point estimate of the residual standard deviation σ
- MSE is from last slide

$$s = \sqrt{MSE} = \sqrt{\frac{SSE}{n\text{-}2}}$$

Significance Test and Estimation for Slope

- A regression model is not likely to be useful unless there is a significant relationship between x and y
- To test significance, we use the null hypothesis:

  $H_0: \beta_1 = 0$

- Versus the alternative hypothesis:

  $H_a: \beta_1 \neq 0$

Significance Test and Estimation for Slope #2

Alternative	Reject H_0 If	*p*-Value
$H_a: \beta_1 > 0$	$t > t_\alpha$	Area under t distribution right of t
$H_a: \beta_1 < 0$	$t < -t_\alpha$	Area under t distribution left of t
$H_a: \beta_1 \neq 0$	$\lvert t \rvert > t_{\alpha/2}$*	Twice area under t distribution right of $\lvert t \rvert$

* That is $t > t_{\alpha/2}$ or $t < -t_{\alpha/2}$

Significance Test and Estimation for Slope #3

Test Statistics

$$t = \frac{b_1}{s_{b_1}} \quad \text{where} \quad s_{b_1} = \frac{s}{\sqrt{SS_{xx}}}$$

100(1-α)% Confidence Interval for β_1

$$[b_1 \pm t_{\alpha/2} s_{b_1}]$$

t_α, $t_{\alpha/2}$ and p-values are based on n–2 degrees of freedom

Confidence and Prediction Intervals

- The point on the regression line corresponding to a particular value of x_0 of the independent variable x is

$$\hat{y} = b_0 + b_1 x_0$$

- It is unlikely that this value will equal the mean value of y when x equals x_0
- Therefore, we need to place bounds on how far the predicted value might be from the actual value
- We can do this by calculating a confidence interval mean for the value of y and a prediction interval for an individual value of y

Distance Value

- Both the confidence interval for the mean value of y and the prediction interval for an individual value of y employ a quantity called the *distance value*
- The distance value for a particular value x_0 of x is

$$\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}$$

- The distance value is a measure of the distance between the value x_0 of x and $\bar{x}$
- Notice that the further x_0 is from $\bar{x}$, the larger the distance value

A Confidence Interval for a Mean
Value of y

- Assume that the regression assumption holds
- The formula for a 100(1-α) confidence interval for the mean value of y is as follows:

$$[\hat{y} \pm t_{\alpha/2} s \sqrt{\text{Distance value}}]$$

- This is based on n-2 degrees of freedom

A Prediction Interval for an Individual
Value of y

- Assume that the regression assumption holds
- The formula for a 100(1-α) prediction interval for an individual value of y is as follows:

$$[\hat{y} \pm t_{\alpha/2} s\sqrt{1 + \text{Distance value}}\,]$$

- This is based on n-2 degrees of freedom

Which to Use?

- The prediction interval is useful if it is important to predict an individual value of the dependent variable
- A confidence interval is useful if it is important to estimate the mean value
- The prediction interval will always be wider than the confidence interval

The Simple Coefficient of Determination and Correlation

- How useful is a particular regression model?
- One measure of usefulness is the simple coefficient of determination
- It is represented by the symbol r^2

Calculating The Simple Coefficient of Determination

1. Total variation is given by the formula
$$\sum (y_i - \bar{y})^2$$
2. Explained variation is given by the formula
$$\sum (\hat{y}_i - \bar{y})^2$$
3. Unexplained variation is given by the formula
$$\sum (y_i - \hat{y}_i)^2$$
4. Total variation is the sum of explained and unexplained variation
5. r^2 is the ratio of explained variation to total variation

The Simple Correlation Coefficient

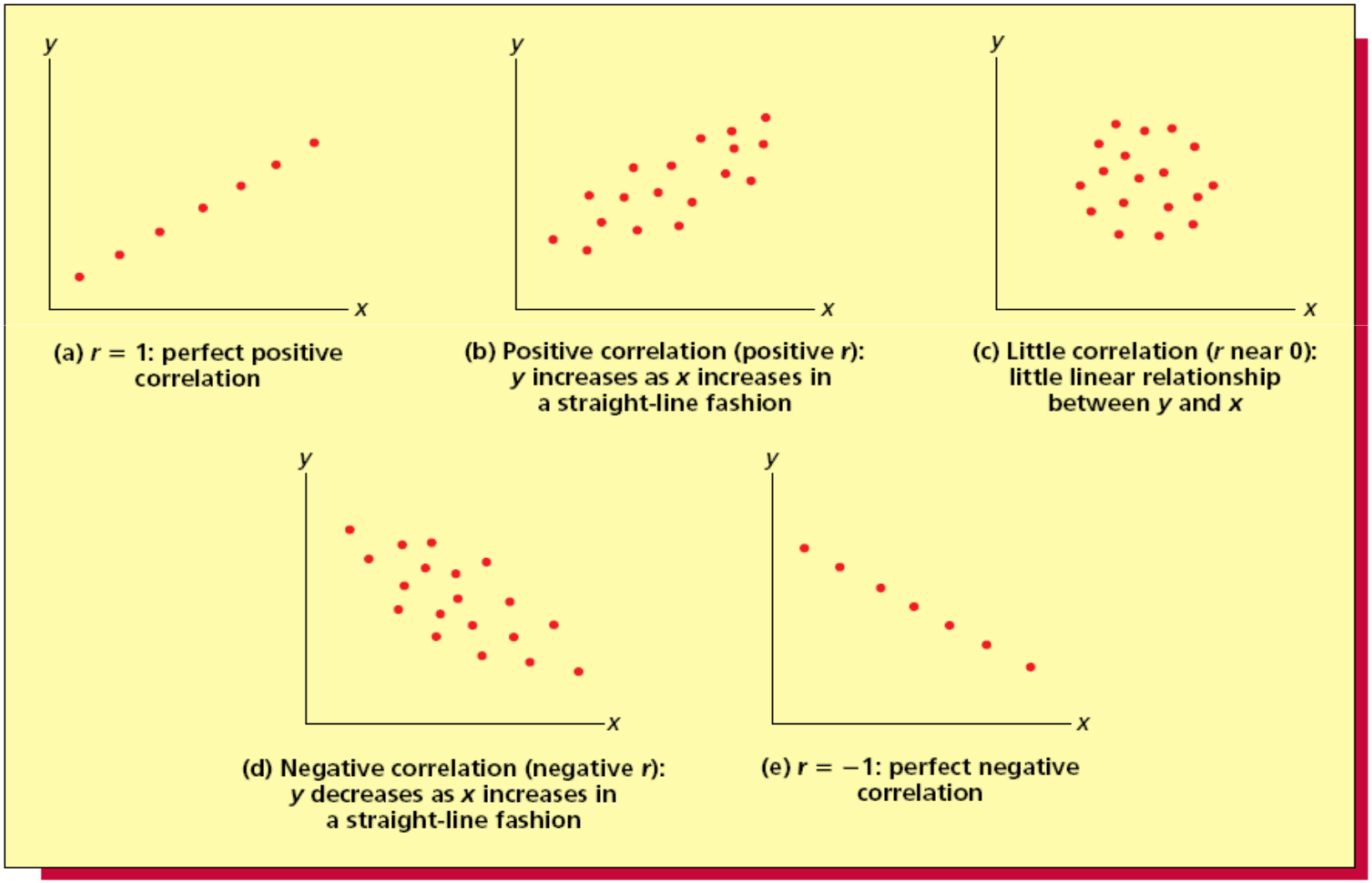
The simple correlation coefficient measures the strength of the linear relationship between y and x and is denoted by r

$r = +\sqrt{r^2}$ if b_1 is positive, and

$r = -\sqrt{r^2}$ if b_1 is negative

Where b_1 is the slope of the least squares line

Different Values of the Correlation Coefficient



Testing the Significance of the Population Correlation Coefficient

- The simple correlation coefficient (r) measures the linear relationship between the observed values of x and the observed values of y from the sample
- The population correlation coefficient (ρ) measures the linear relationship between all possible combinations of observed values of x and y
- r is an estimate of ρ

Testing ρ

- We can test to see if the correlation is significant using the hypotheses

  H_0: ρ = 0
  H_a: ρ ≠ 0

- The statistic is

$$t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}$$

- This test will give the same results as the test for significance on the slope coefficient β_1

An F Test for Model

- For simple regression, this is another way to test the null hypothesis

  $H_0: \beta_1 = 0$

- That will not be the case for multiple regression
- The F test tests the significance of the overall regression relationship between x and y

Mechanics of the F Test

To test H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$ at the α level of significance

Test statistics based on F

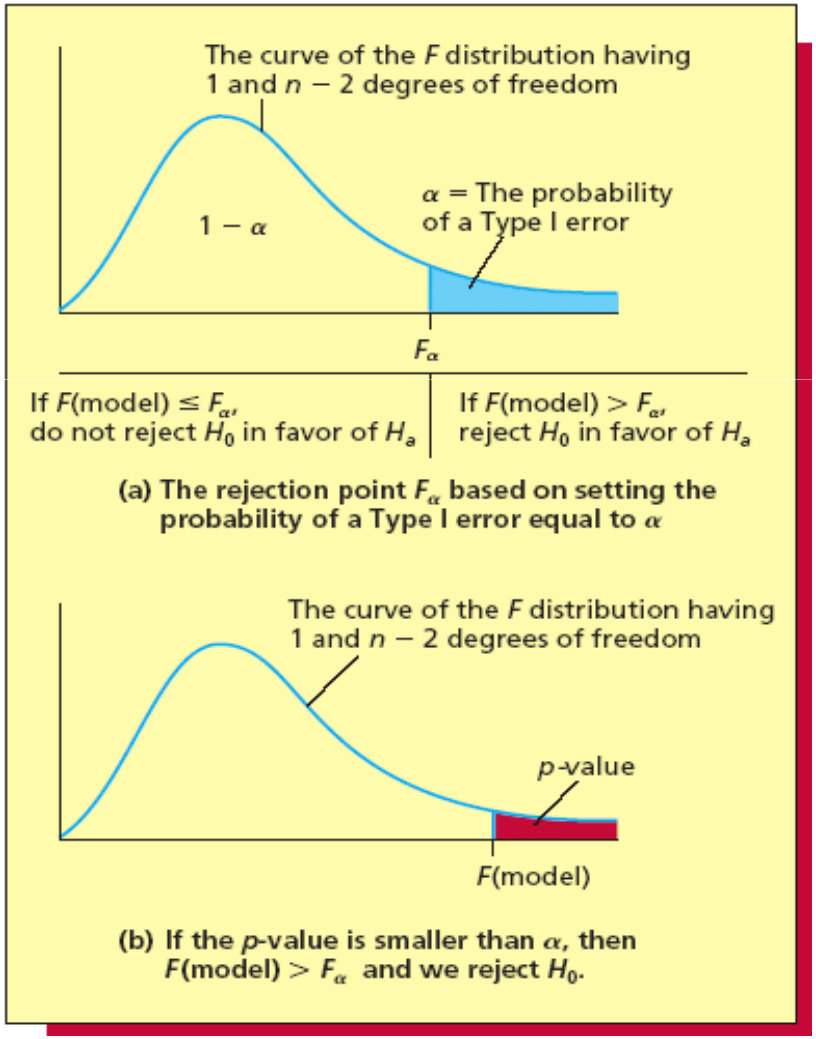
$$F = \frac{\text{Explained variation}}{(\text{Unexplained variation})/(n-2)}$$

Reject H_0 if

F(model) > F_α or

p-value < α

F_α is based on 1 numerator and n-2 denominator degrees of freedom



Residual Analysis #1

- Checks of regression assumptions are performed by analyzing the regression residuals
- Residuals (*e*) are defined as the difference between the observed value of y and the predicted value of y

$$e = y - \hat{y}$$

- Note that e is the point estimate of ε
- If the regression assumptions are valid, the population of potential error terms will be normally distributed with a mean of zero and a variance σ^2
- Furthermore, the different error terms will be statistically independent

Residual Analysis #2

- The residuals should look like they have been randomly and independently selected from normally distributed populations having mean zero and variance σ^2
- With any real data, assumptions will not hold exactly
- Mild departures do not affect our ability to make statistical inferences
- In checking assumptions, we are looking for pronounced departures from the assumptions
- So, only require residuals to approximately fit the description above

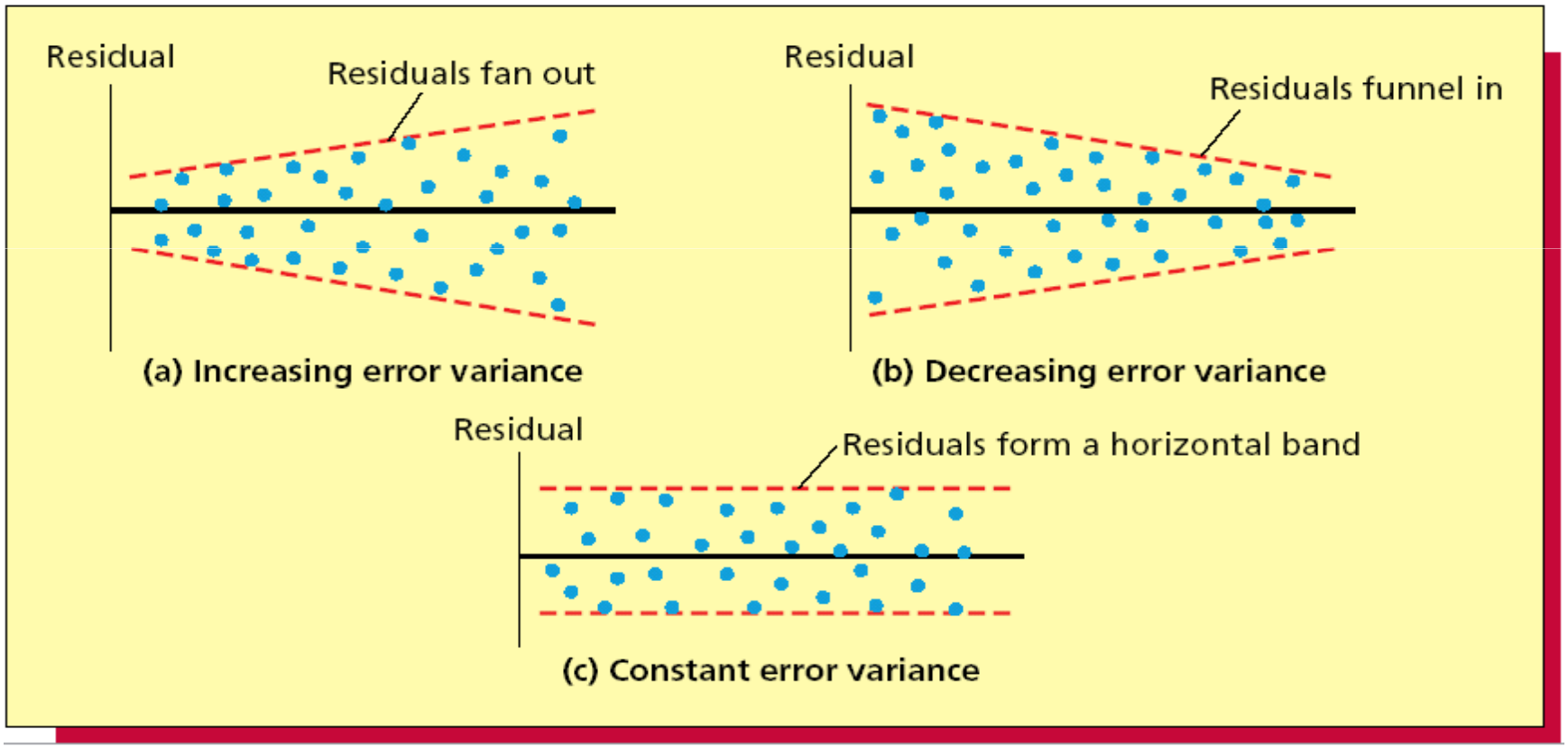
Residual Plots

1. Residuals versus independent variable
2. Residuals versus predicted y's
3. Residuals in time order (if the response is a time series)
4. Histogram of residuals
5. Normal plot of the residuals

Constant Variance Assumptions

- To check the validity of the constant variance assumption, we examine plots of the residuals against
  - The x values
  - The predicted y values
  - Time (when data is time series)
- A pattern that fans out says the variance is increasing rather than staying constant
- A pattern that funnels in says the variance is decreasing rather than staying constant
- A pattern that is evenly spread within a band says the assumption has been met

Constant Variance Visually



Assumption of Correct Functional Form

- If the relationship between x and y is something other than a linear one, the residual plot will often suggest a form more appropriate for the model
- For example, if there is a curved relationship between x and y, a plot of residuals will often show a curved relationship

Normality Assumption

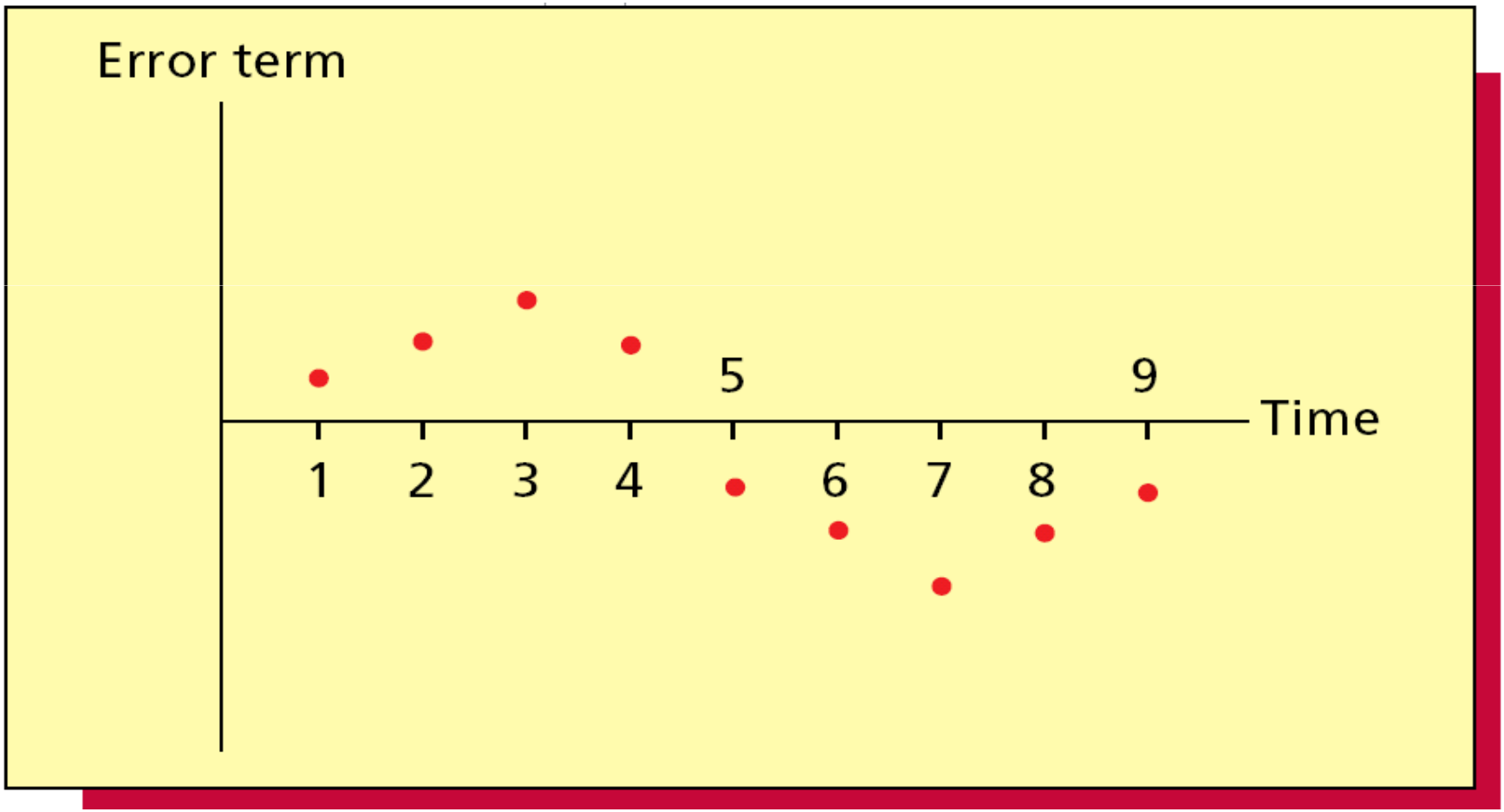
- If the normality assumption holds, a histogram or stem-and-leaf display of residuals should look bell-shaped and symmetric
- Another way to check is a normal plot of residuals
  1. Order residuals from smallest to largest
  2. Plot $e_{(i)}$ on vertical axis against $z_{(i)}$
     - $Z_{(i)}$ is the point on the horizontal axis under the z curve so that the area under this curve to the left is (3i-1)/(3n+1)
- If the normality assumption holds, the plot should have a straight-line appearance

Independence Assumption

- Independence assumption is most likely to be violated when the data are time-series data
  - If the data is not time series, then it can be reordered without affecting the data
  - Changing the order would change the interdependence of the data
- For time-series data, the time-ordered error terms can be autocorrelated
  - Positive autocorrelation is when a positive error term in time period i tends to be followed by another positive value in i+k
  - Negative autocorrelation is when a positive error term in time period i tends to be followed by a negative value in i+k
- Either one will cause a cyclical error term over time

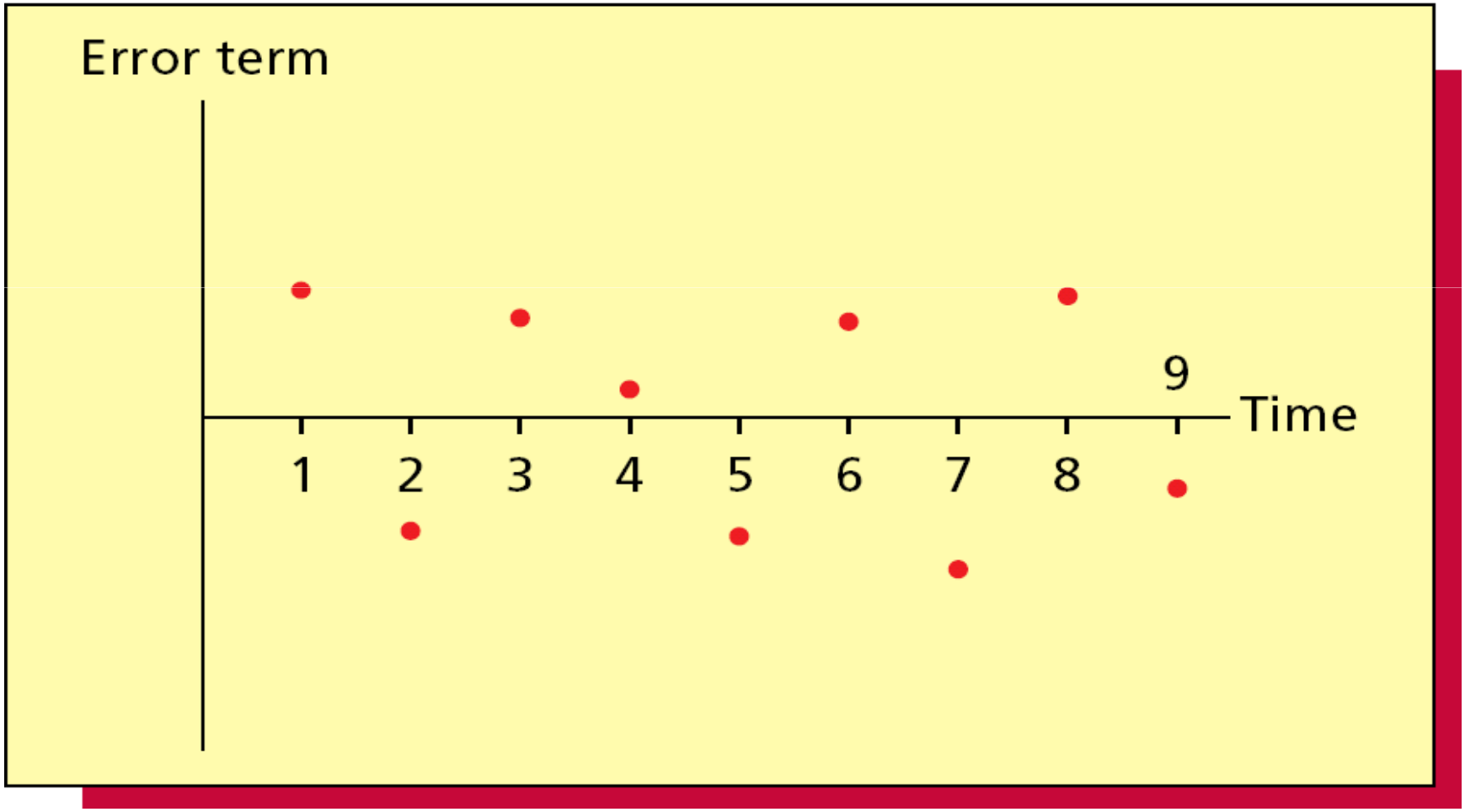
Independence Assumption Visually

Positive Autocorrelation



Independence Assumption Visually

Negative Autocorrelation



Durbin-Watson Test

- One type of autocorrelation is called *first-order autocorrelation*
- This is when the error term in time period t (ε_t) is related to the error term in time period t-1 (ε_{t-1})
- The Durbin-Watson statistic checks for first-order autocorrelation

Some Shortcut Formulas

$$\text{Total variation} = SSTO = SS_{yy}$$

$$\text{Explained variation} = SSR = \frac{SS_{xy}^2}{SS_{xx}}$$

$$\text{Unexplained variation} = SSE = SS_{yy} - \frac{SS_{xy}^2}{SS_{xx}}$$

where

$$SS_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$$

$$SS_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

$$SS_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}$$