Statistics for Business and Economics

Chapter 11

Simple Regression

11.1

Overview of Linear Models

- An equation can be fit to show the best linear relationship between two variables:

$$\boxed{Y = \beta_0 + \beta_1 X}$$

Where Y is the dependent variable and
X is the independent variable
β_0 is the Y-intercept
β_1 is the slope

Least Squares Regression

- Estimates for coefficients β_0 and β_1 are found using a Least Squares Regression technique
- The least-squares regression line, based on sample data, is

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

- Where b_1 is the slope of the line and b_0 is the y-intercept:

$$b_1 = \frac{\mathrm{Cov}(x, y)}{s_x^2}$$

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

Introduction to Regression Analysis

- Regression analysis is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable
  - Explain the impact of changes in an independent variable on the dependent variable

Dependent variable: the variable we wish to explain
(also called the endogenous variable)

Independent variable: the variable used to explain the dependent variable
(also called the exogenous variable)

11.2

Linear Regression Model

- The relationship between X and Y is described by a linear function
- Changes in Y are assumed to be caused by changes in X
- Linear regression population equation model

$$\boxed{Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i}$$

- Where β_0 and β_1 are the population model coefficients and ε is a random error term.

Simple Linear Regression Model

The population regression model:

Dependent Variable

Population Y intercept

Population Slope Coefficient

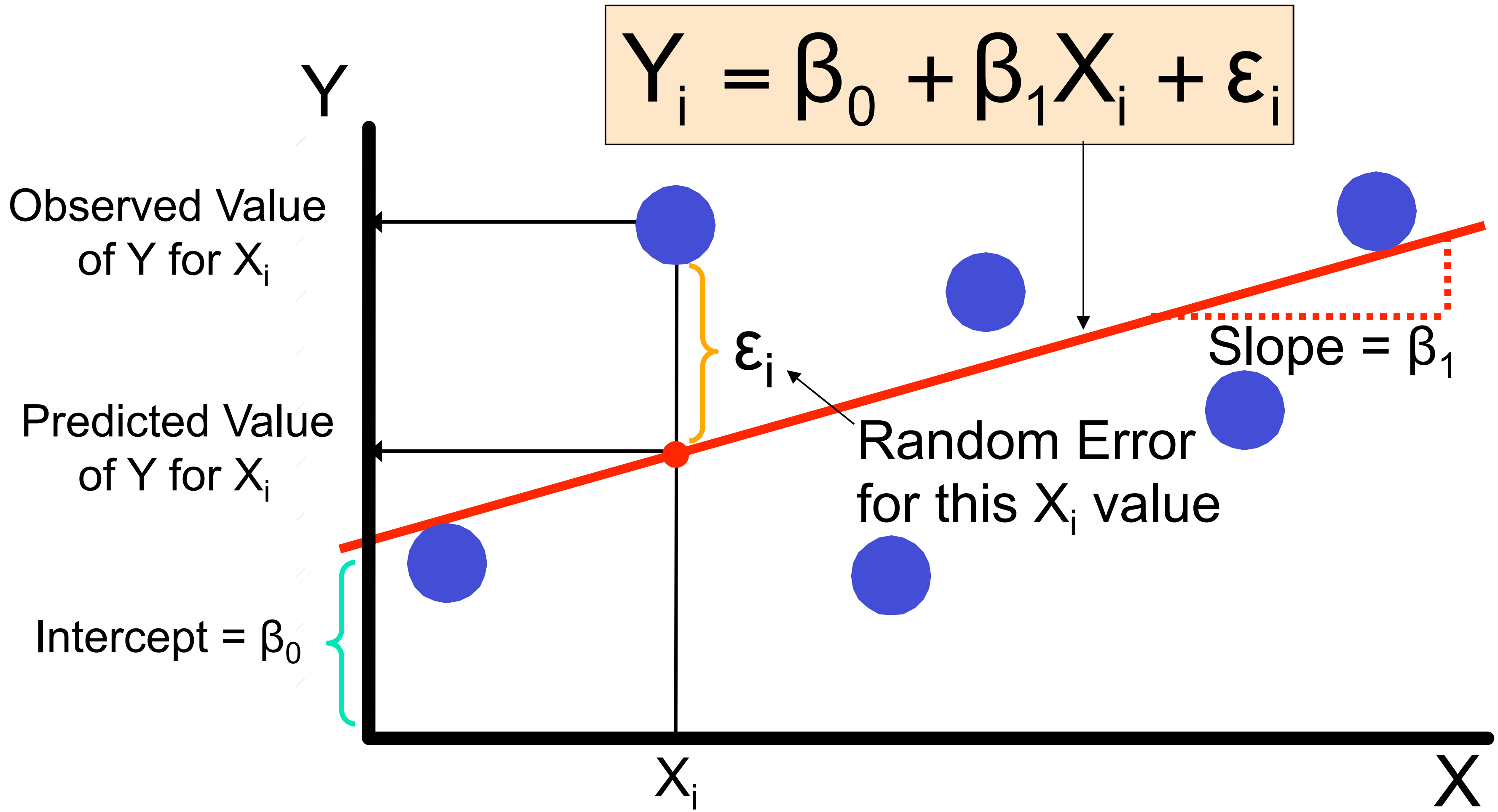
Independent Variable

Random Error term

$$\underbrace{Y_i = \beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

Simple Linear Regression Model

(continued)



Simple Linear Regression Equation

The simple linear regression equation provides an estimate of the population regression line

Estimated (or predicted) y value for observation i

Estimate of the regression intercept

Estimate of the regression slope

Value of x for observation i

$$\hat{y}_i = b_0 + b_1 x_i$$

The individual random error terms e_i have a mean of zero

$$e_i = (y_i - \hat{y}_i) = y_i - (b_0 + b_1 x_i)$$

11.3

Least Squares Estimators

- b_0 and b_1 are obtained by finding the values of b_0 and b_1 that minimize the sum of the squared differences between y and $\hat{y}$:

$$\begin{aligned} \min\ SSE &= \min \sum e_i^2 \\ &= \min \sum (y_i - \hat{y}_i)^2 \\ &= \min \sum [y_i - (b_0 + b_1 x_i)]^2 \end{aligned}$$

Differential calculus is used to obtain the coefficient estimators b_0 and b_1 that minimize SSE

Least Squares Estimators

(continued)

- The slope coefficient estimator is

$$\boxed{b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\mathrm{Cov}(x, y)}{s_x^2} = r_{xy}\frac{s_y}{s_x}}$$

- And the constant or y-intercept is

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

- The regression line always goes through the mean $\bar{x}, \bar{y}$

Finding the Least Squares Equation

- The coefficients b_0 and b_1 , and other regression results in this chapter, will be found using a computer
  - Hand calculations are tedious
  - Statistical routines are built into Excel
  - Other statistical analysis software can be used

Linear Regression Model Assumptions

- The true relationship form is linear (Y is a linear function of X, plus random error)
- The error terms, ε_i are independent of the x values
- The error terms are random variables with mean 0 and constant variance, σ^2

  (the constant variance property is called homoscedasticity)

$$\boxed{E[\varepsilon_i] = 0 \quad \text{and} \quad E[\varepsilon_i^{\,2}] = \sigma^2 \qquad \text{for } (i = 1, \ldots, n)}$$

- The random error terms, ε_i, are not correlated with one another, so that

$$\boxed{E[\varepsilon_i \varepsilon_j] = 0 \qquad \text{for all } i \neq j}$$

Interpretation of the Slope and the Intercept

- b_0 is the estimated average value of y when the value of x is zero (if $x = 0$ is in the range of observed x values)

- b_1 is the estimated change in the average value of y as a result of a one-unit change in x

Simple Linear Regression Example

- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)

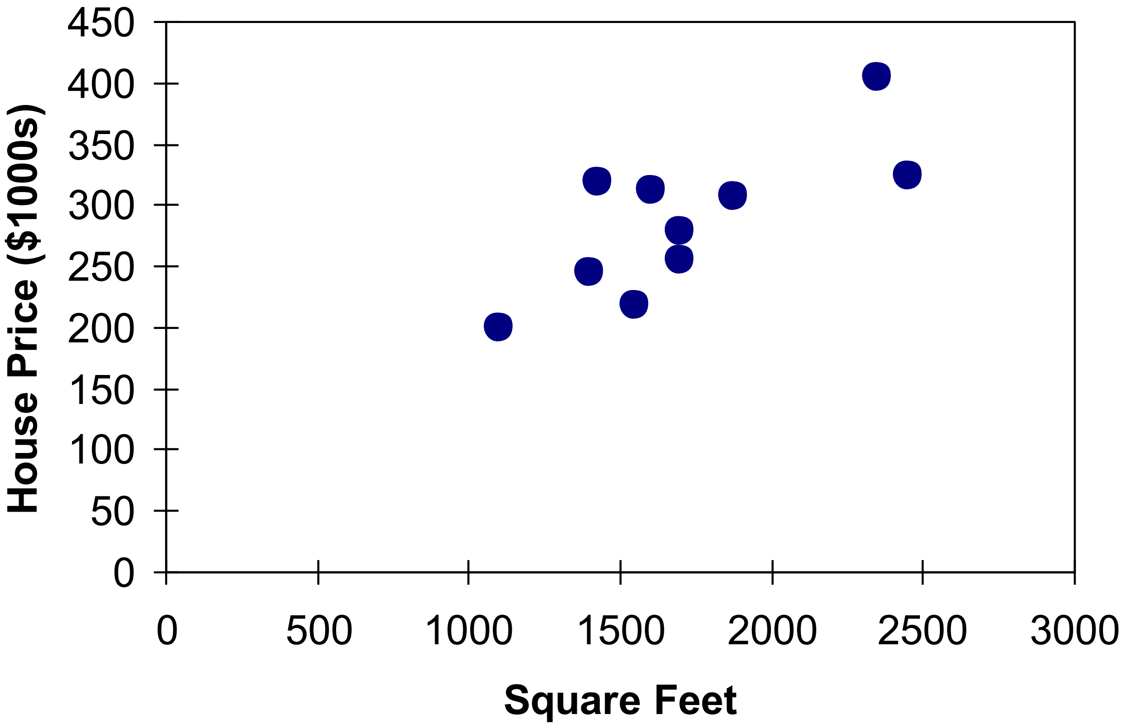
- A random sample of 10 houses is selected
  - Dependent variable (Y) = house price in $1000s
  - Independent variable (X) = square feet

Sample Data for House Price Model

House Price in $1000s (Y)	Square Feet (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

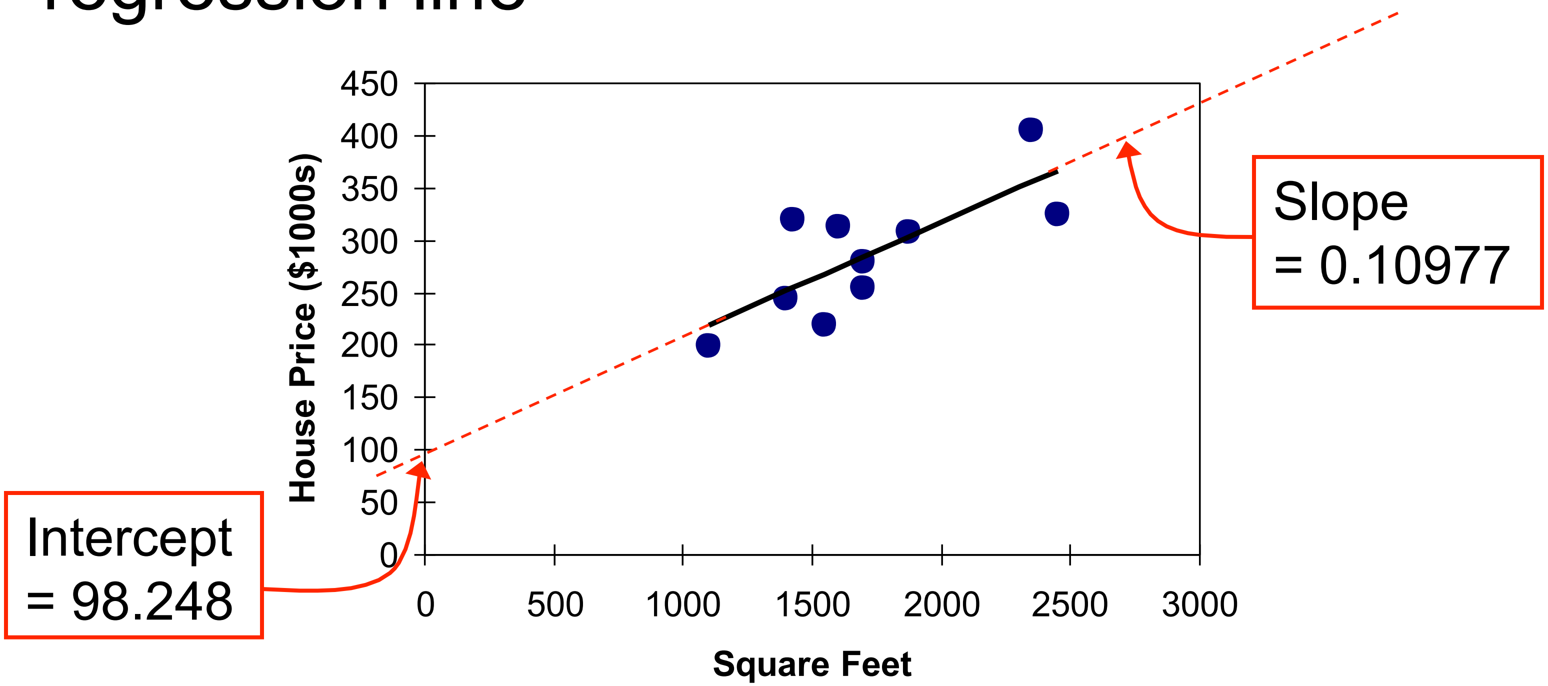
Graphical Presentation

- House price model: scatter plot



Graphical Presentation

- House price model: scatter plot and regression line



$$\widehat{\text{house price}} = 98.24833 + 0.10977\ (\text{square feet})$$

Interpretation of the Intercept, b_0

$$\widehat{\text{house price}} = \boxed{98.24833} + 0.10977\,(\text{square feet})$$

- b_0 is the estimated average value of Y when the value of X is zero (if X = 0 is in the range of observed X values)

Interpretation of the Slope Coefficient, b_1

$$\widehat{\text{house price}} = 98.24833 + \boxed{0.10977}\,(\text{square feet})$$

- b_1 measures the estimated change in the average value of Y as a result of a one-unit change in X
  - Here, $\boxed{b_1 = .10977}$ tells us that the average value of a house increases by .10977($1000) = $109.77, on average, for each additional one square foot of size

11.4 Measures of Variation

- Total variation is made up of two parts:

$$SST = SSR + SSE$$

Total Sum of Squares	Regression Sum of Squares	Error Sum of Squares
$SST = \sum (y_i - \bar{y})^2$	$SSR = \sum (\hat{y}_i - \bar{y})^2$	$SSE = \sum (y_i - \hat{y}_i)^2$

where:

$\bar{y}$ = Average value of the dependent variable

y_i = Observed values of the dependent variable

$\hat{y}_i$ = Predicted value of y for the given x_i value

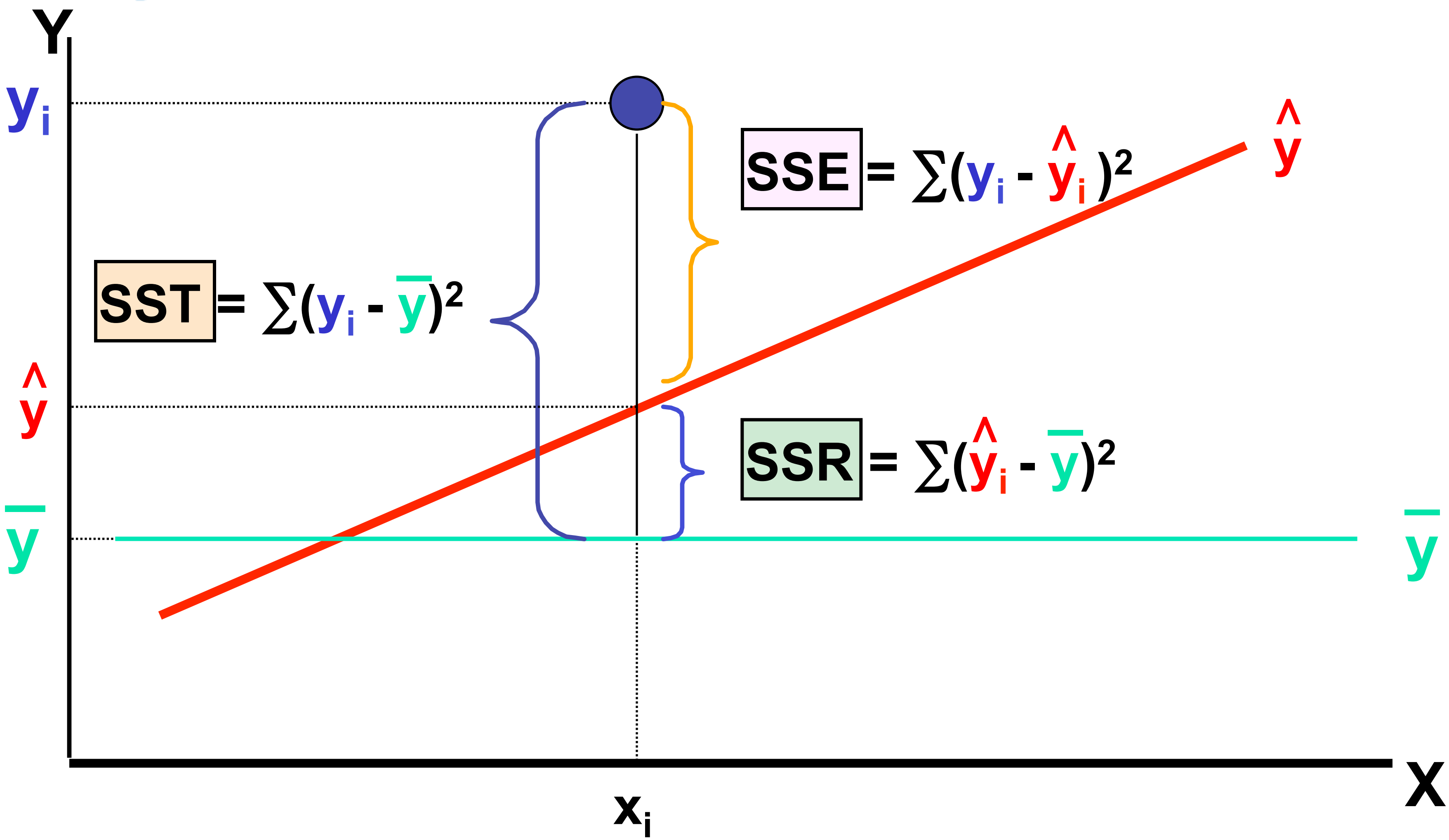
Measures of Variation

(continued)

- SST = total sum of squares
  - Measures the variation of the y_i values around their mean, $\bar{y}$
- SSR = regression sum of squares
  - Explained variation attributable to the linear relationship between x and y
- SSE = error sum of squares
  - Variation attributable to factors other than the linear relationship between x and y

Measures of Variation

(continued)



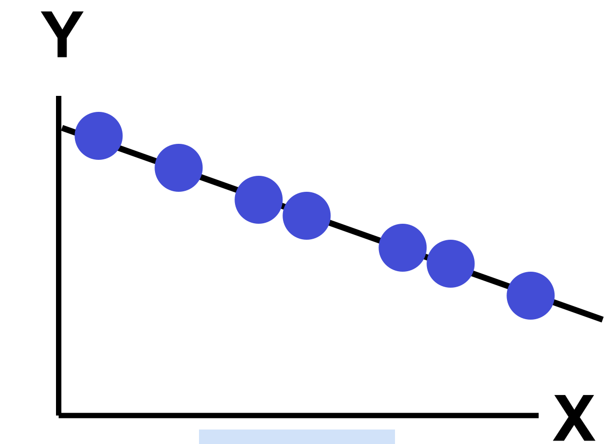
Coefficient of Determination, R^2

- The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable
- The coefficient of determination is also called R-squared and is denoted as R^2

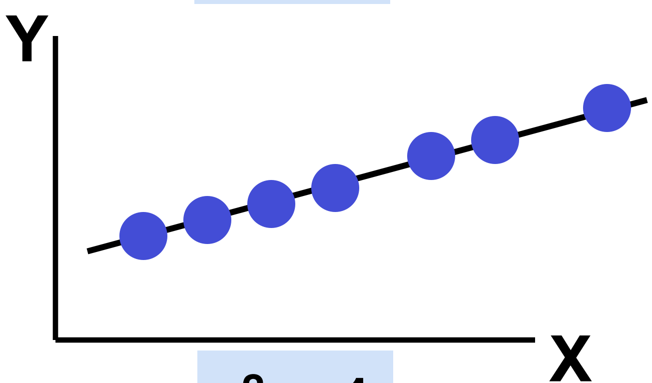
$$\boxed{R^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}}$$

note: $\boxed{0 \le R^2 \le 1}$

Examples of Approximate r^2 Values



$r^2 = 1$



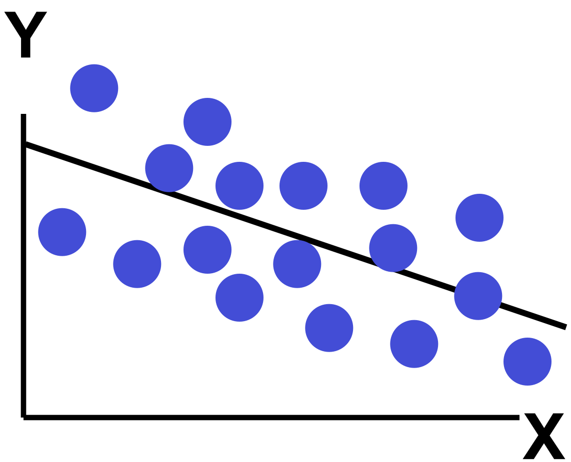
$r^2 = 1$

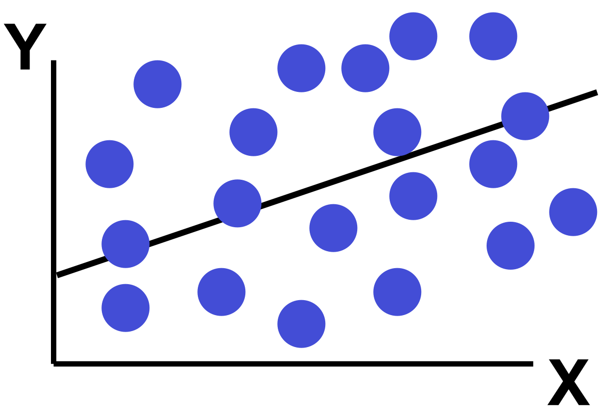
$r^2 = 1$

Perfect linear relationship between X and Y:

100% of the variation in Y is explained by variation in X

Examples of Approximate r^2 Values



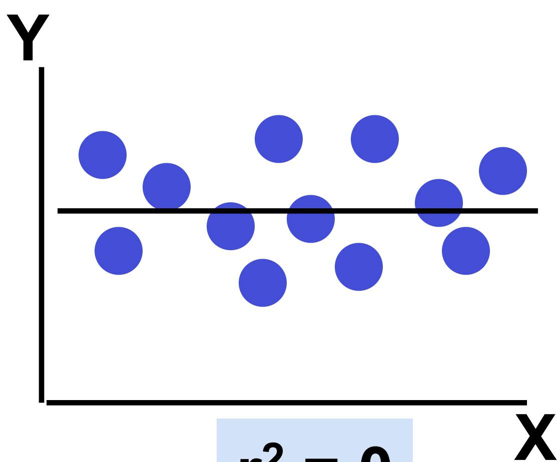


$0 < r^2 < 1$

Weaker linear relationships between X and Y:

Some but not all of the variation in Y is explained by variation in X

Examples of Approximate r^2 Values



$r^2 = 0$

$r^2 = 0$

No linear relationship between X and Y:

The value of Y does not depend on X. (None of the variation in Y is explained by variation in X)

Correlation and R_2

- The coefficient of determination, R^2, for a simple regression is equal to the simple correlation squared

$$\boxed{R^2 = r_{xy}^2}$$

Estimation of Model Error Variance

- An estimator for the variance of the population model error is

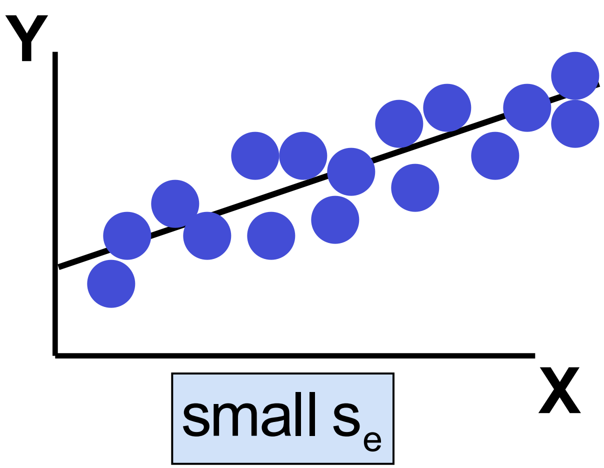
$$\hat{\sigma}^2 = s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SSE}{n-2}$$

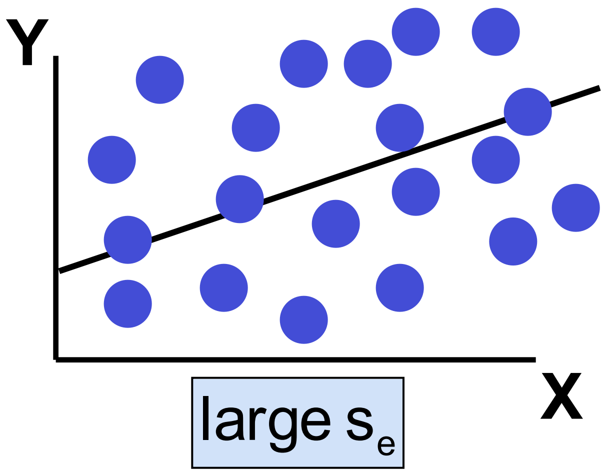
- Division by n – 2 instead of n – 1 is because the simple regression model uses two estimated parameters, b_0 and b_1, instead of one

$s_e = \sqrt{s_e^2}$ is called the standard error of the estimate

Comparing Standard Errors

s_e is a measure of the variation of observed y values from the regression line





The magnitude of s_e should always be judged relative to the size of the y values in the sample data

i.e., s_e = \$41.33K is moderately small relative to house prices in the \$200 - \$300K range

Inferences About the Regression Model

- The variance of the regression slope coefficient (b_1) is estimated by

$$\boxed{s_{b_1}^2 = \frac{s_e^2}{\sum (x_i - \bar{x})^2} = \frac{s_e^2}{(n-1)s_x^2}}$$

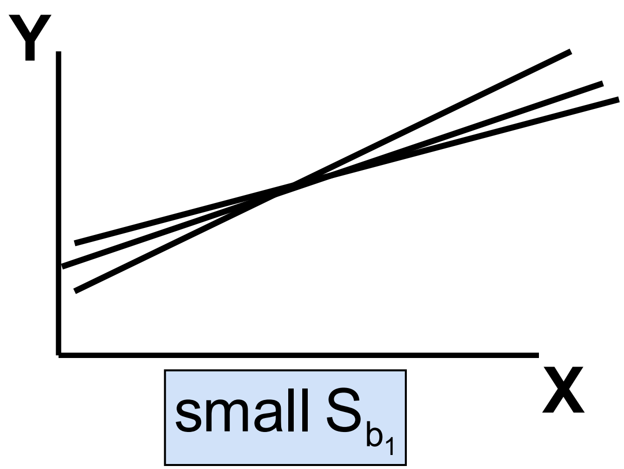
where:

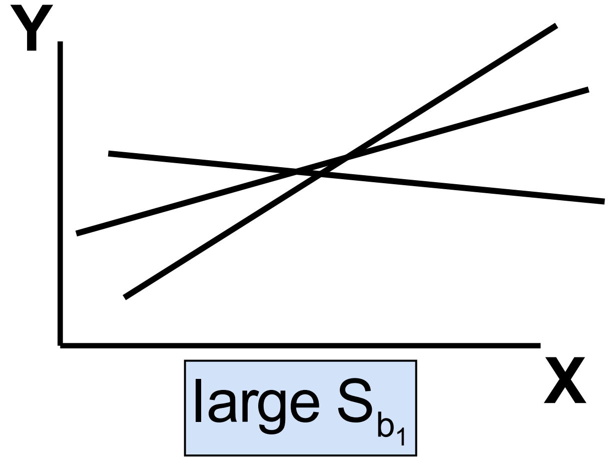
s_{b_1} = Estimate of the standard error of the least squares slope

$s_e = \sqrt{\dfrac{SSE}{n-2}}$ = Standard error of the estimate

Comparing Standard Errors of the Slope

S_{b_1} is a measure of the variation in the slope of regression lines from different possible samples





Inference about the Slope: t Test

- t test for a population slope
  - Is there a linear relationship between X and Y?
- Null and alternative hypotheses

H_0:	$\beta_1 = 0$	(no linear relationship)
H_1:	$\beta_1 \neq 0$	(linear relationship does exist)

- Test statistic

$$\boxed{t = \frac{b_1 - \beta_1}{s_{b_1}}}$$

$$d.f. = n - 2$$

where:

b_1 = regression slope coefficient

β_1 = hypothesized slope

s_{b1} = standard error of the slope

Inference about the Slope: t Test

(continued)

House Price in $1000s (y)	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Estimated Regression Equation:

$$\widehat{\text{house price}} = 98.25 + 0.1098\,(\text{sq.ft.})$$

The slope of this model is 0.1098

Does square footage of the house affect its sales price?

Inferences about the Slope: t Test Example

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

b_1 $\qquad$ s_{b_1}

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{0.10977 - 0}{0.03297} = 3.32938$$

Inferences about the Slope: t Test Example

(continued)

Test Statistic: **t = 3.329**

H_0: $\beta_1 = 0$

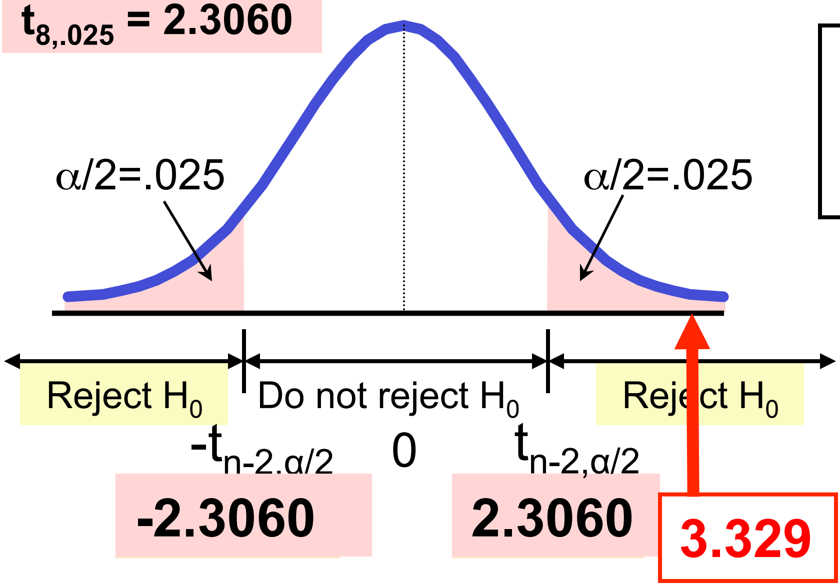
H_1: $\beta_1 \neq 0$

d.f. = 10-2 = 8

$\mathbf{t_{8,.025} = 2.3060}$

b_1 s_{b_1} t

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**



Decision:

Reject H_0

Conclusion:

There is sufficient evidence that square footage affects house price

Inferences about the Slope: t Test Example

(continued)

P-value = **0.01039**

H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$

P-value

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**

This is a two-tail test, so the p-value is

$P(t > 3.329)+P(t < -3.329) = 0.01039$

(for 8 d.f.)

Decision: P-value < α so Reject H_0

Conclusion:

There is sufficient evidence that square footage affects house price

Confidence Interval Estimate for the Slope

Confidence Interval Estimate of the Slope:

$$b_1 - t_{n-2,\alpha/2} s_{b_1} < \beta_1 < b_1 + t_{n-2,\alpha/2} s_{b_1}$$

d.f. = n - 2

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**	**-35.57720**	**232.07386**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**	**0.03374**	**0.18580**

At 95% level of confidence, the confidence interval for the slope is (0.0337, 0.1858)

Confidence Interval Estimate for the Slope

(continued)

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**	**-35.57720**	**232.07386**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**	**0.03374**	**0.18580**

Since the units of the house price variable is $1000s, we are 95% confident that the average impact on sales price is between $33.70 and $185.80 per square foot of house size

This 95% confidence interval does not include 0.

Conclusion: There is a significant relationship between house price and square feet at the .05 level of significance