CHAPTER 7

(Part 2)

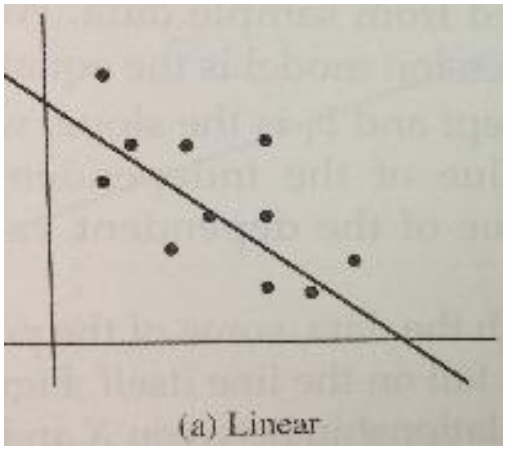
Regression Analysis

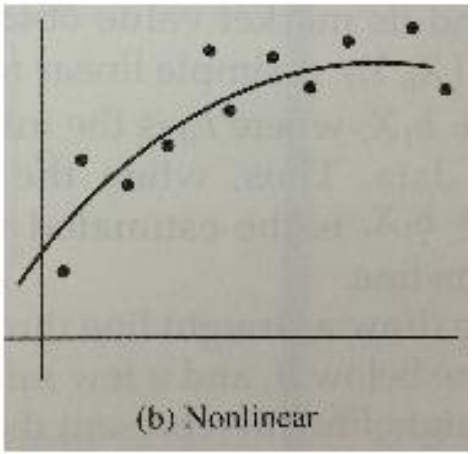
Introduction to Regression Analysis

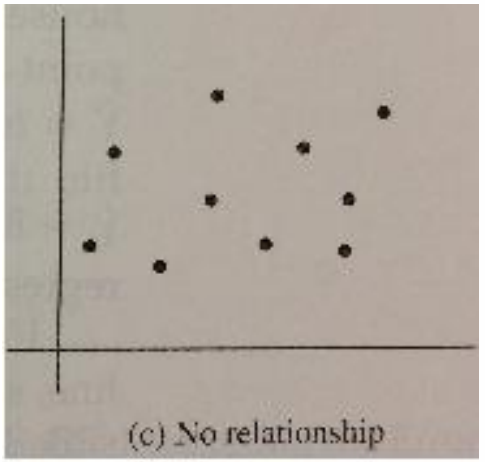
- Regression analysis is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable.
  - Explain the impact of changes in an independent variable on the dependent variable.
- Dependent variable: the variable we wish to explain.
- Independent variable: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

- A regression model that involves a single independent variable is called **simple regression**.
  - **Example:** Imagine that your company wants to understand how past advertising expenditures have related to sales in order to make future decisions about advertising. The dependent variable in this instance is sales and the independent variable is advertising expenditures.
- Usually, more than one independent variable influences the dependent variable.
- A regression model that involves two or more independent variables is called **multiple regression**.
  - **Example:** Sales are influenced by advertising as well as other factors, such as the number of sales representatives and the commission percentage paid to sales representatives

- Regression models can be either linear or nonlinear.
- A linear model assumes the relationships between variables are straight-line relationships, while a nonlinear model assumes the relationships between variables are represented by curved lines.







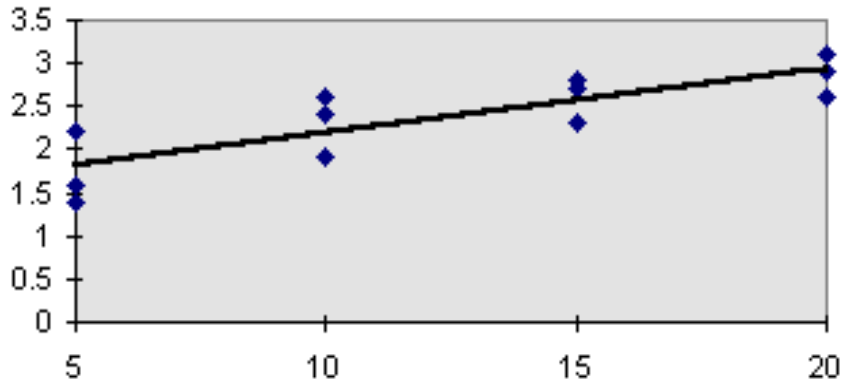
- The most basic type of regression is that of simple linear regression.
- A simple linear regression uses only one independent variable, and it describes the relationship between the independent variable and dependent variable as a straight line.
- This chapter will focus on the basic case of a **simple linear regression**.

Simple Linear Regression Model

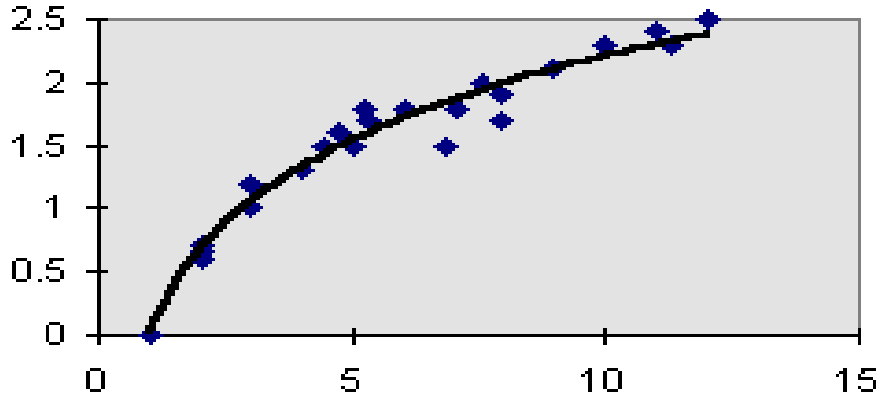
- Only **one** independent variable, x.
- Relationship between x and y is described by a linear function.
- Changes in y are assumed to be caused by changes in x.

Types of Regression Models

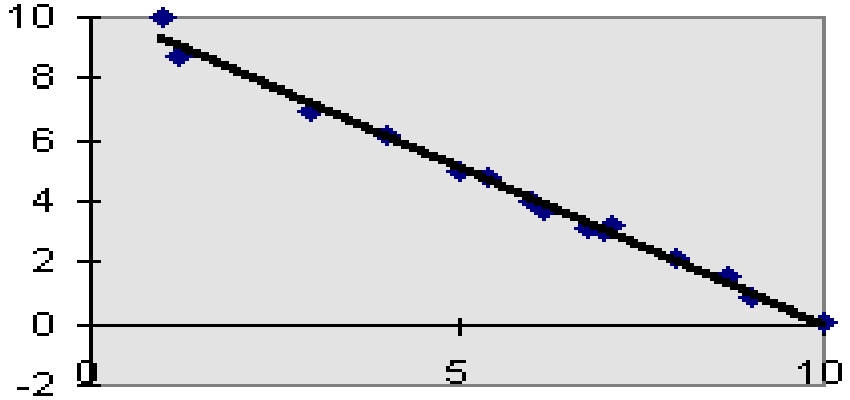
Positive Linear Relationship



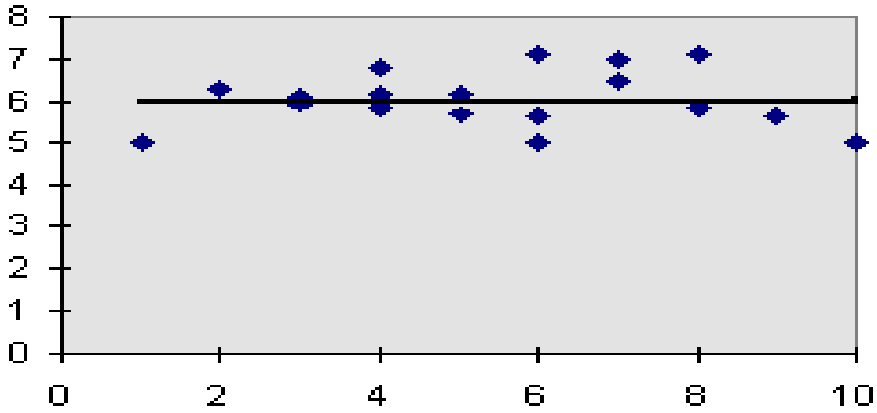
Relationship NOT Linear



Negative Linear Relationship

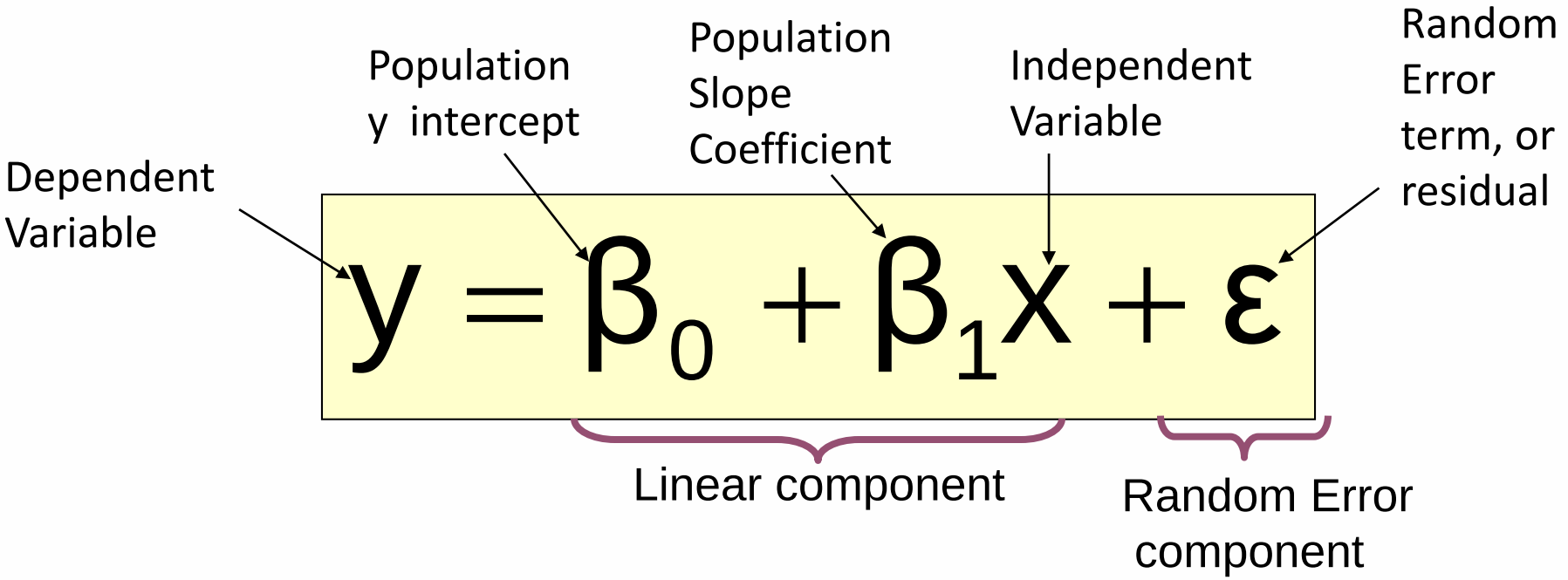


No Relationship



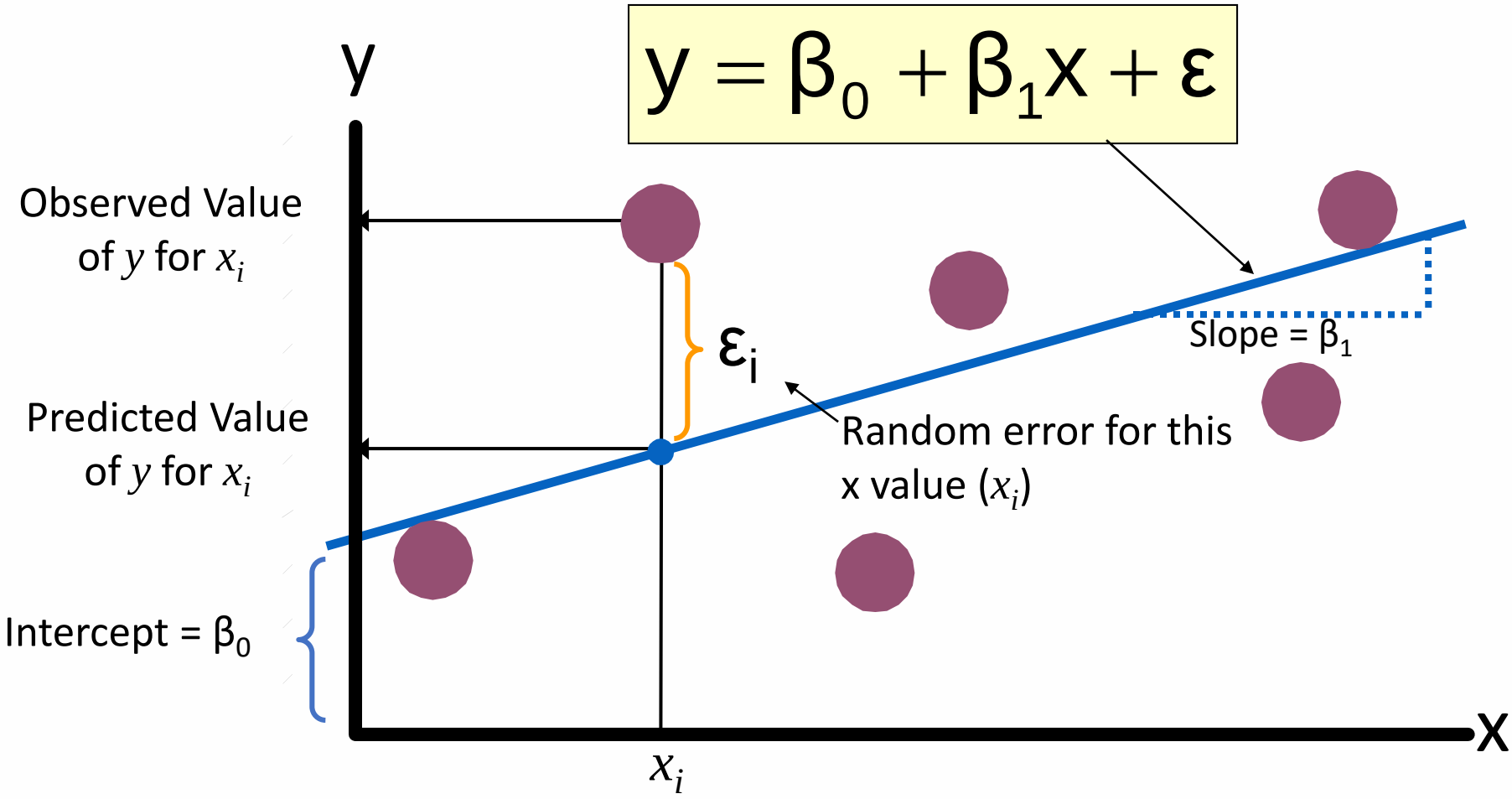
Population Linear Regression

The population regression model:



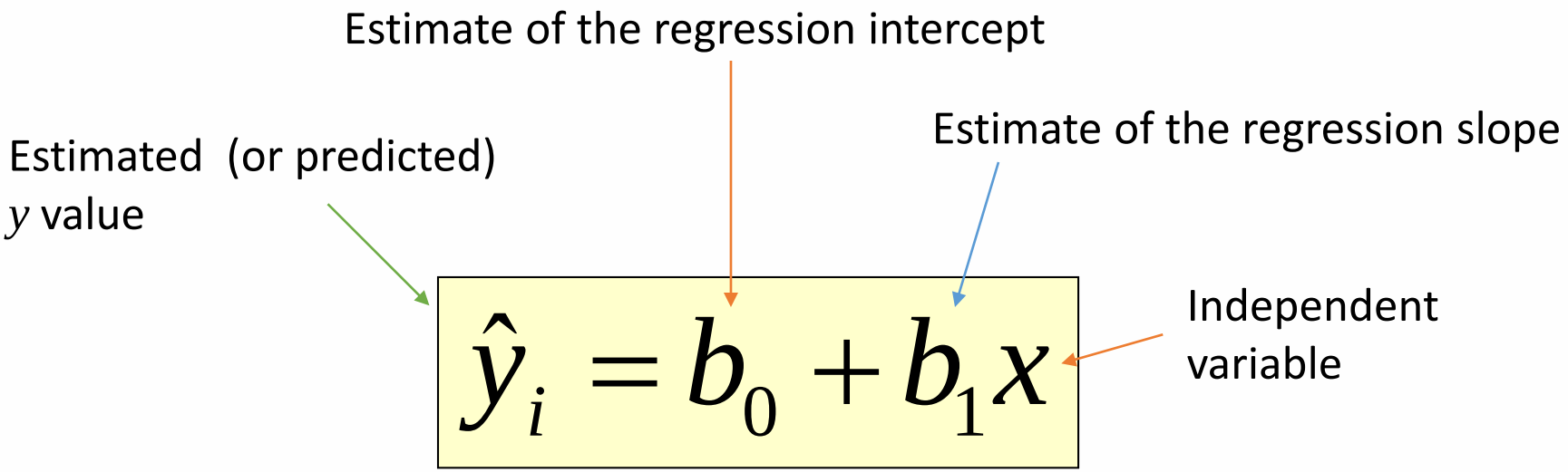
Linear Regression Assumptions

- Error values (ε) are statistically independent.
- Error values are normally distributed for any given value of x.
- The probability distribution of the errors is normal.
- The probability distribution of the errors has constant variance.
- The underlying relationship between the x variable and the y variable is linear.


$y = \beta_0 + \beta_1 x + \varepsilon$
y
Observed Value of y for x_i
Predicted Value of y for x_i
Intercept = β_0
ε_i
Slope = β_1
Random error for this x value (x_i)
x
x_i

Estimated Regression Model

The sample regression line provides an estimate of the population regression line



The individual random error terms e_i have a mean of zero.

Least Squares Criterion

- b_0 and b_1 are obtained by finding the values of b_0 and b_1 that minimize the sum of the squared residuals:

$$\sum e^2 = \sum (y - \hat{y})^2$$
$$= \sum (y - (b_0 + b_1 x))^2$$

Note:
residual is the difference between the observed value and the **mean** value that the model predicts for that observation.

The Least Squares Equation

- The formulas for b_1 and b_0 are:

$$b_1 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

algebraic equivalent:

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

and

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

Interpretation of the Slope and the Intercept

- b_0 is the estimated average value of y when the value of x is zero.

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

Finding the Least Squares Equation

- The coefficients b_0 and b_1 will usually be found using computer software such as R, Excel or SPSS.

- Other regression measures will also be computed as part of computer-based regression analysis.

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

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

Example - Solution

y	x	xy	x^2
245	1400	343000	1960000
312	1600	499200	2560000
279	1700	474300	2890000
308	1875	577500	3515625
199	1100	218900	1210000
219	1550	339450	2402500
405	2350	951750	5522500
324	2450	793800	6002500
319	1425	454575	2030625
255	1700	433500	2890000
$\Sigma y=2865$	$\Sigma x= 17150$	$\Sigma xy= 5085975$	$\Sigma x^2=30983750$

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

$$b_1 = \frac{5085975 - \frac{(17150)(2865)}{10}}{30983750 - \frac{(17150)^2}{10}}$$

$$= \frac{172500}{1571500} = 0.109767737$$

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

$$b_0 = 286.5 - 0.109767737(1715)$$

$$= 98.24832962$$

Output from software R:

```
x <- c(1400, 1600, 1700, 1875,
1100, 1550, 2350, 2450, 1425,
1700)

y <- c(245, 312, 279, 308, 199,
219, 405, 324, 319, 255)

model <- lm(y~x)
model
```

```
> model

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
    98.2483       0.1098
```

$$\hat{y} = 98.2483 + 0.1098x$$

Output from software SPSS:

Coefficients[a]

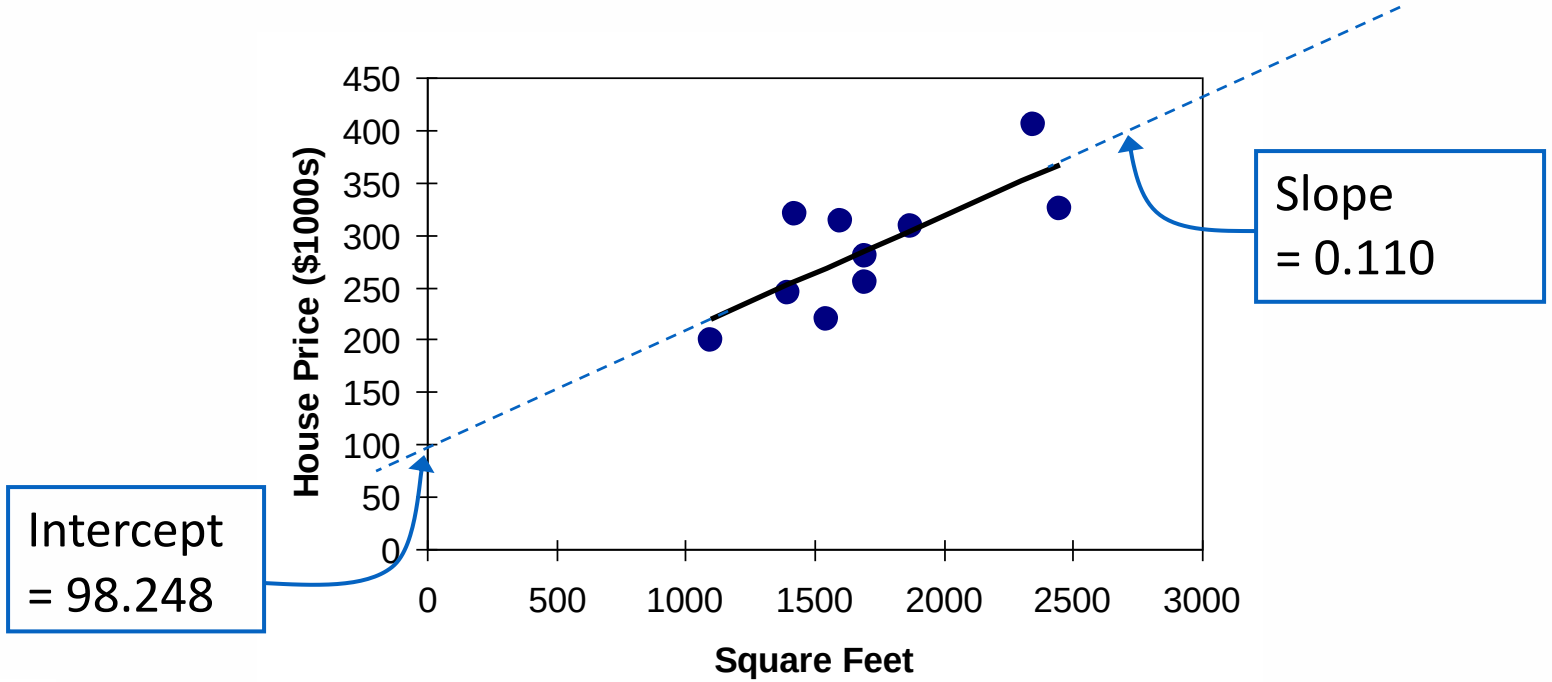
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	98.248	58.033		1.693	.129
	SQR_FT	.110	.033	.762	3.329	.010

a. Dependent Variable: H_PRICE

The regression equation is: $\hat{y} = 98.248 + 0.110\,x$

Graphical Presentation

- House price model: scatter plot and regression line



$$\hat{y} = 98.248 + 0.110\ x$$

$$\hat{y} = 98.248 + 0.110\,x$$

- 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)
- Here, no houses had zero square feet, so b_0 = 98.248 just indicates that, for houses within the range of sizes observed, $98,248.33 is the portion of the house price not explained by square feet.

Interpretation of the Slope Coefficient, b_1

$$\hat{y} = 98.248 + 0.110x$$

- b_1 measures the estimated change in the average value of y as a result of a one-unit change in x.
- Here, b_1 = 0.110 tells us that the average value of a house increases by 0.110($1000) = $110, on average, for each additional one square feet of size.

Least Squares Regression Properties

- The sum of the residuals from the least squares regression line is 0 ($\sum (y - \hat{y}) = 0$)
- The sum of the squared residuals is a minimum (minimized $\sum (y - \hat{y})^2$)
- The simple regression line always passes through the mean of the y variable and the mean of the x variable.
- The least squares coefficients are unbiased estimates of β_0 and β_1

Explained and Unexplained Variation

- Total variation is made up of two parts:

$$SST = SSE + SSR$$

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

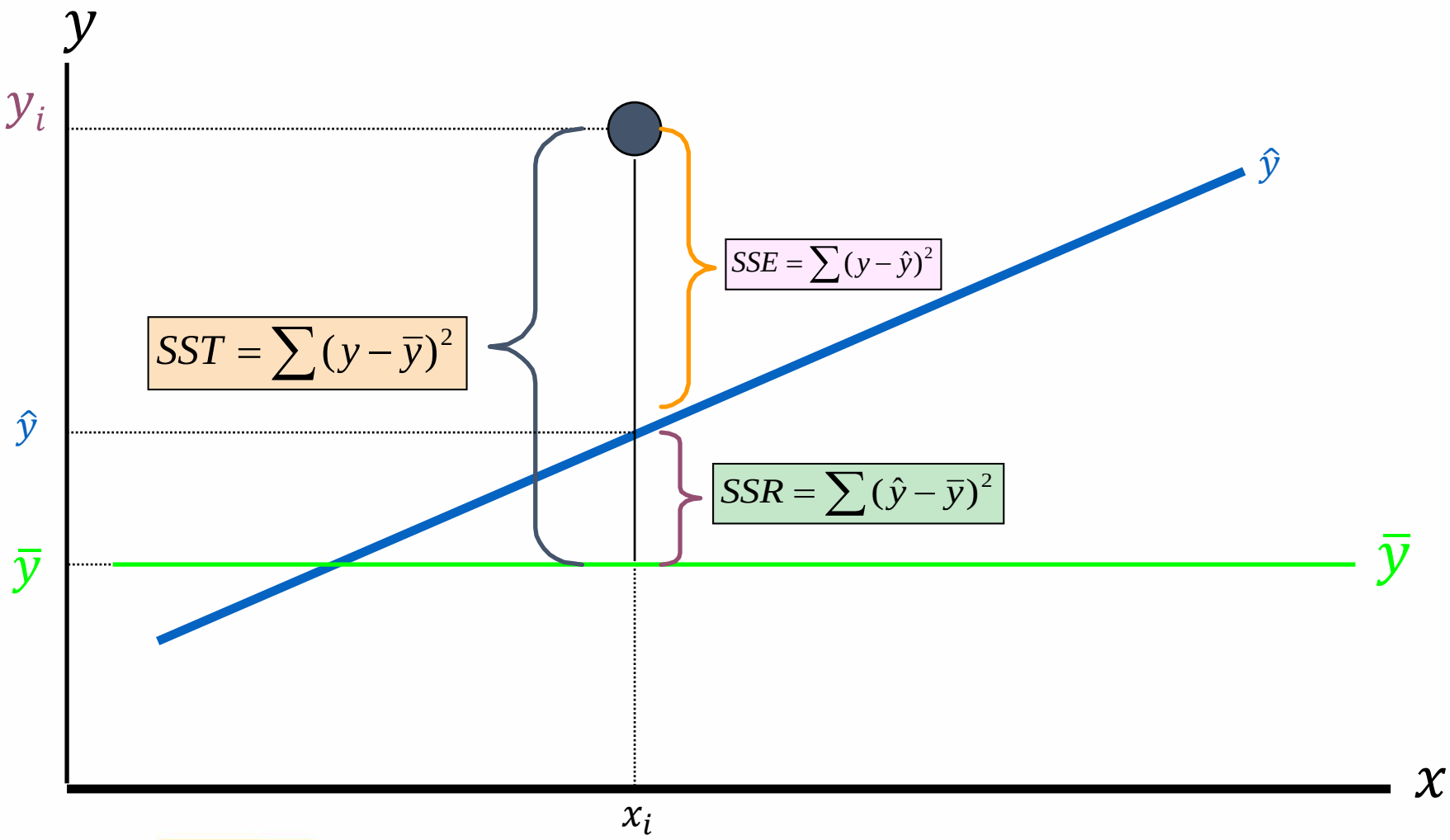
where:

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

y = Observed values of the dependent variable

$\hat{y}$ = Estimated value of y for the given x value

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


y
y_i
$\hat{y}$
$\bar{y}$
$SST = \sum (y - \bar{y})^2$
$SSE = \sum (y - \hat{y})^2$
$SSR = \sum (\hat{y} - \bar{y})^2$
$\hat{y}$
$\bar{y}$
x
x_i

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

$$R^2 = \frac{SSR}{SST}$$

where

$$0 \leq R^2 \leq 1$$

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

Note: In the single independent variable case, the coefficient of determination is

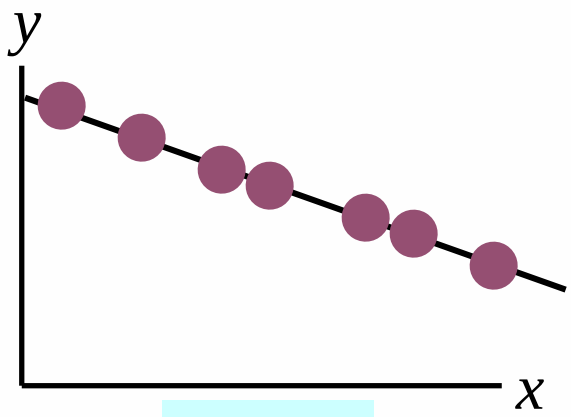
$$R^2 = r^2$$

where:

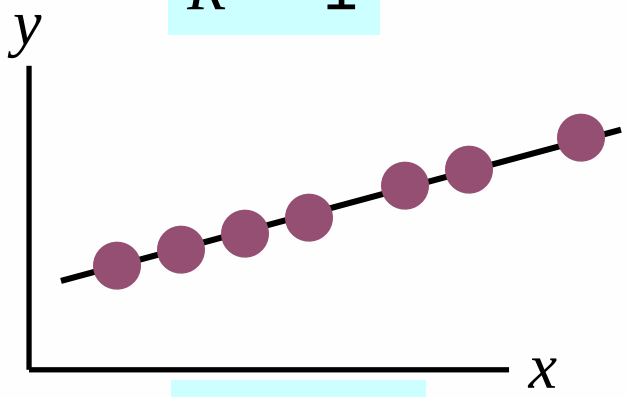
R^2 = Coefficient of determination

r = Simple correlation coefficient

Examples of Approximate R^2 Values



$R^2 = 1$



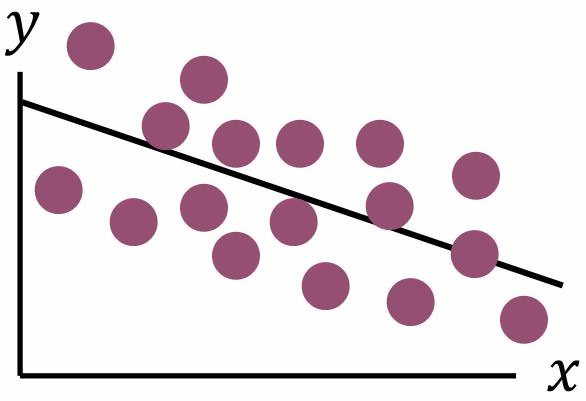
$R^2 = +1$

$R^2 = 1$

Perfect linear relationship between x and y:

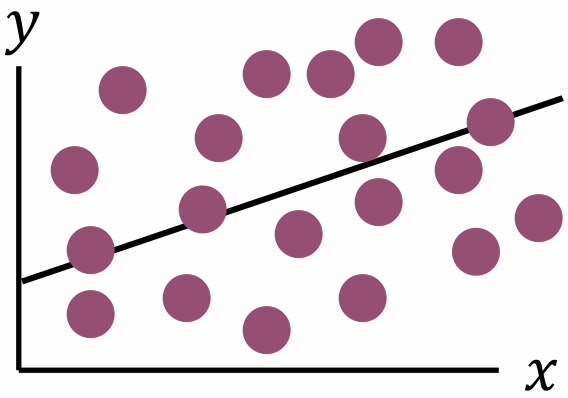
100% of the variation in y is explained by variation in x

Examples of Approximate R^2 Values



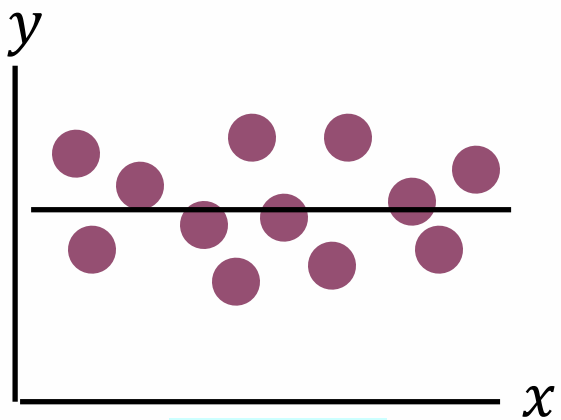
$0 < R^2 < 1$

Weaker linear relationship 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)

Example – R Programming

58.1% of the variation in house prices is explained by variation in square feet.

```
> summary(model)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-49.388 -27.388  -6.388  29.577  64.333

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 98.24833   58.03348   1.693   0.1289
x            0.10977    0.03297   3.329   0.0104 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41.33 on 8 degrees of freedom
Multiple R-squared:  0.5808,    Adjusted R-squared:  0.5284
F-statistic: 11.08 on 1 and 8 DF,  p-value: 0.01039
```

Example - SPSS

$$R^2 = \frac{SSR}{SST} = \frac{18934.935}{32600.5000} = 0.581$$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.762[a]	.581	.528	41.3303

a. Predictors: (Constant), SQR_FT

58.1% of the variation in house prices is explained by variation in square feet

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	18934.935	1	18934.935	11.085	.010[a]
	Residual	13665.565	8	1708.196		
	Total	32600.500	9			

a. Predictors: (Constant), SQR_FT

b. Dependent Variable: H_PRICE

Standard Error of Estimate

- The standard deviation of the variation of observations around the regression line is estimated by

$$s_\varepsilon = \sqrt{\frac{SSE}{n-k-1}}$$

Where

SSE = Sum of squares error

n = Sample size

k = number of independent variables in the model

The Standard Deviation of the Regression Slope

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

$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{\sum (x - \bar{x})^2}} = \frac{s_\varepsilon}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

where:

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

$S_\varepsilon = \sqrt{\frac{SSE}{n-2}}$ = Sample standard error of the estimate

Example – R Programming

```
> summary(model)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-49.388 -27.388  -6.388  29.577  64.333

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 98.24833   58.03348   1.693   0.1289
x            0.10977    0.03297   3.329   0.0104 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41.33 on 8 degrees of freedom
Multiple R-squared:  0.5808,    Adjusted R-squared:  0.5284
F-statistic: 11.08 on 1 and 8 DF,  p-value: 0.01039
```

$s_\varepsilon = 41.33$

$s_{b_1} = 0.033$

Example – SPSS

$s_{\varepsilon} = 41.3303$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.762[a]	.581	.528	41.3303

a. Predictors: (Constant), SQR_FT

$s_{b_1} = 0.033$

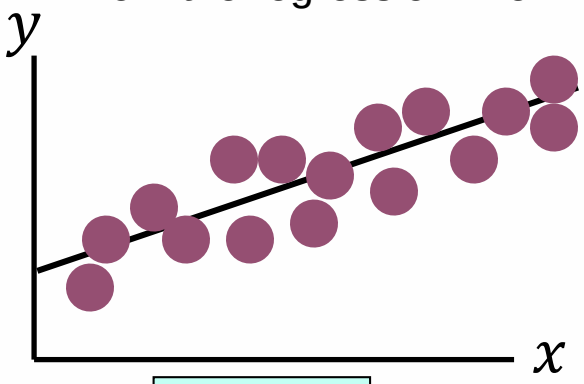
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	98.248	58.033		1.693	.129
	SQR_FT	.110	.033	.762	3.329	.010

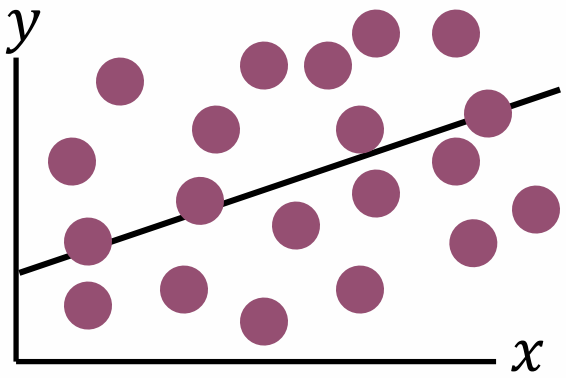
a. Dependent Variable: H_PRICE

Comparing Standard Errors

Variation of observed y values from the regression line

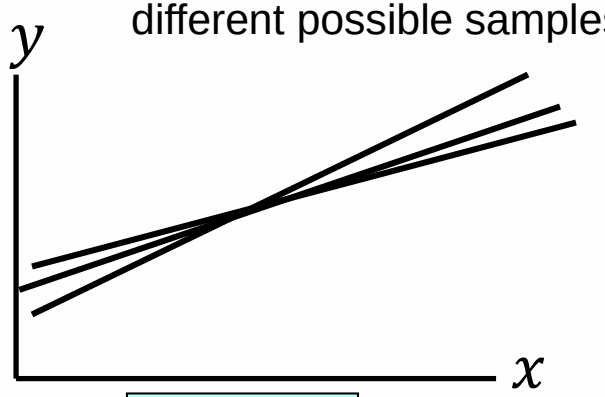


small s_ε

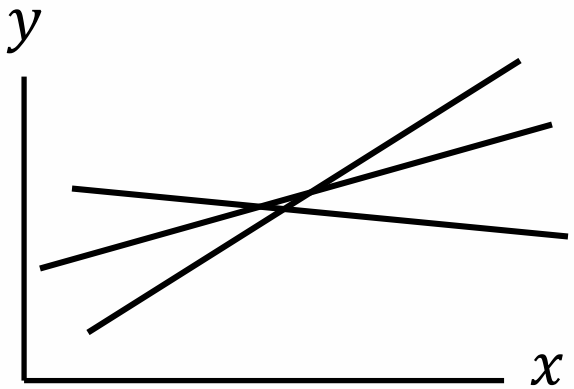


large s_ε

Variation in the slope of regression lines from different possible samples



small s_{b_1}



large s_{b_1}

Inference About The Slope: t -test

- 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

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

$$\text{d.f.} = n - 2$$

where:

b_1 = Sample regression slope coefficient

β_1 = Hypothesized slope

s_{b1} = Estimator of the standard error of the slope

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:

$$\hat{y} = 98.248 + 0.110\,x$$

The slope of this model is 0.110

Does square footage of the house affect its sales price?

Example – R Programming

$H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

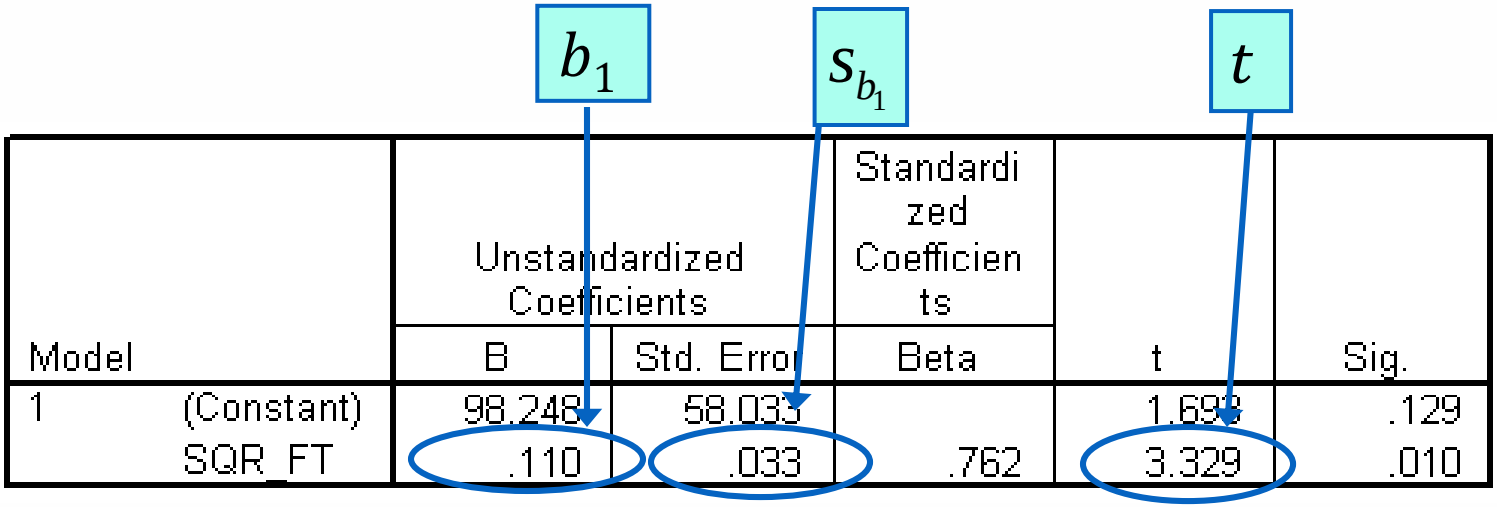
```
> summary(model)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-49.388 -27.388  -6.388  29.577  64.333

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 98.24833   58.03348   1.693   0.1289
x            0.10977    0.03297   3.329   0.0104 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41.33 on 8 degrees of freedom
Multiple R-squared:  0.5808,    Adjusted R-squared:  0.5284
F-statistic: 11.08 on 1 and 8 DF,  p-value: 0.01039
```

b_1 s_{b_1} t

Example – SPSS

$H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$



Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	98.248	58.033		1.693	.129
	SQR_FT	.110	.033	.762	3.329	.010

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	98.248	58.033		1.693	.129
	SQR_FT	.110	.033	.762	3.329	.010

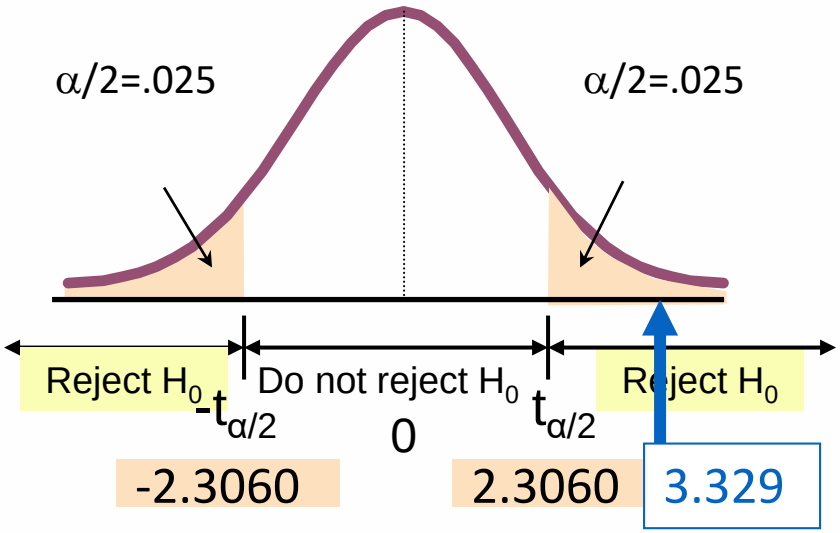
Test Statistic: $t = 3.329$

d.f. = 10-2 = 8; $\alpha = .05$

$t_{\alpha/2} = 2.3060$ (refer to table)

Decision: Reject H_0

Conclusion: There is sufficient evidence that square footage affects house price



Exercise #1

Representative data on x = carbonation depth (in millimeters) and y = strength (in megapascals) for a sample of concrete core specimens taken from a particular building were obtained from a particular survey.

Depth, x	8	20	20	30	35	40	50	55	65
Strength, y	22.8	17.1	21.1	16.1	13.4	12.4	11.4	9.7	6.8

i. Construct a scatterplot. Does the relationship between carbonation depth and strength appear to be linear?

ii. Find the equation of the least-square line.

iii. What would you predict for strength when carbonation depth is 25 mm?

iv. Explain why it would not be reasonable to use the least-square line to predict strength when carbonation depth is 100 mm.

Exercise #2

The following data (next page) on sale, size, and land-to-building ratio for 10 large industrial properties appeared in the paper "Using Multiple Regression Analysis in Real Estate Appraisal" (Appraisal Journal [2002]: 424-430):

i. Calculate and interpret the value of the correlation coefficient between sale price and size.

ii. Calculate and interpret the value of the correlation coefficient between sale price and land-to-building ratio.

iii. If you wanted to predict sale price and you could use either size or land-to-building ratio as the basis for making predictions, which would you use? Explain.

iv. Based on your choice (iii), find the equation of the least-square regression line you would use for predicting y = sale price.

Property	Sale Price (millions of dollars)	Size (thousands of sq. ft.)	Land-to-Building Ratio
1	10.6	2166	2.0
2	2.6	751	3.5
3	30.5	2422	3.6
4	1.8	224	4.7
5	20.0	3917	1.7
6	8.0	2866	2.3
7	10.0	1698	3.1
8	6.7	1046	4.8
9	5.8	1108	7.6
10	4.5	405	17.2