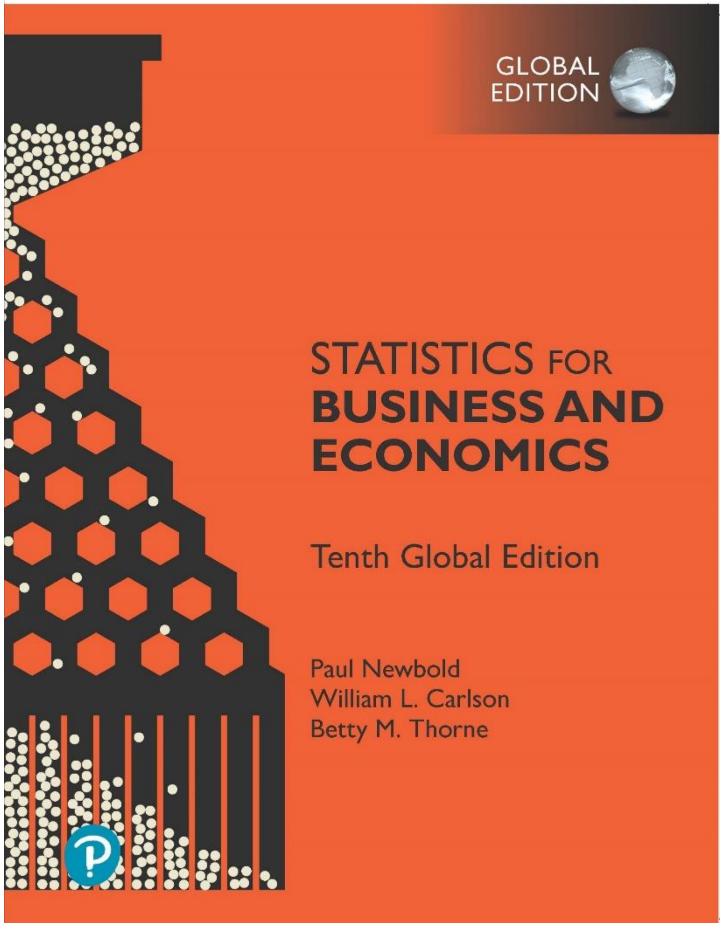
Statistics for Business and Economics

Tenth Edition, Global Edition



Chapter 11

Simple Regression

Section 11.1 Overview of Linear Models

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

$$Y = \beta_0 + \beta_1 X$$

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:

$$b_1 = \frac{Cov(x, y)}{s_x^2} = r\left(\frac{s_y}{s_x}\right) \qquad b_0 = \bar{y} - b_1\bar{x}$$

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)

Section 11.2 Linear Regression Model

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

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

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

Simple Linear Regression Model (1 of 2)

The population regression model:

Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

Dependent Variable

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Linear component

Random Error component

Linear Regression 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

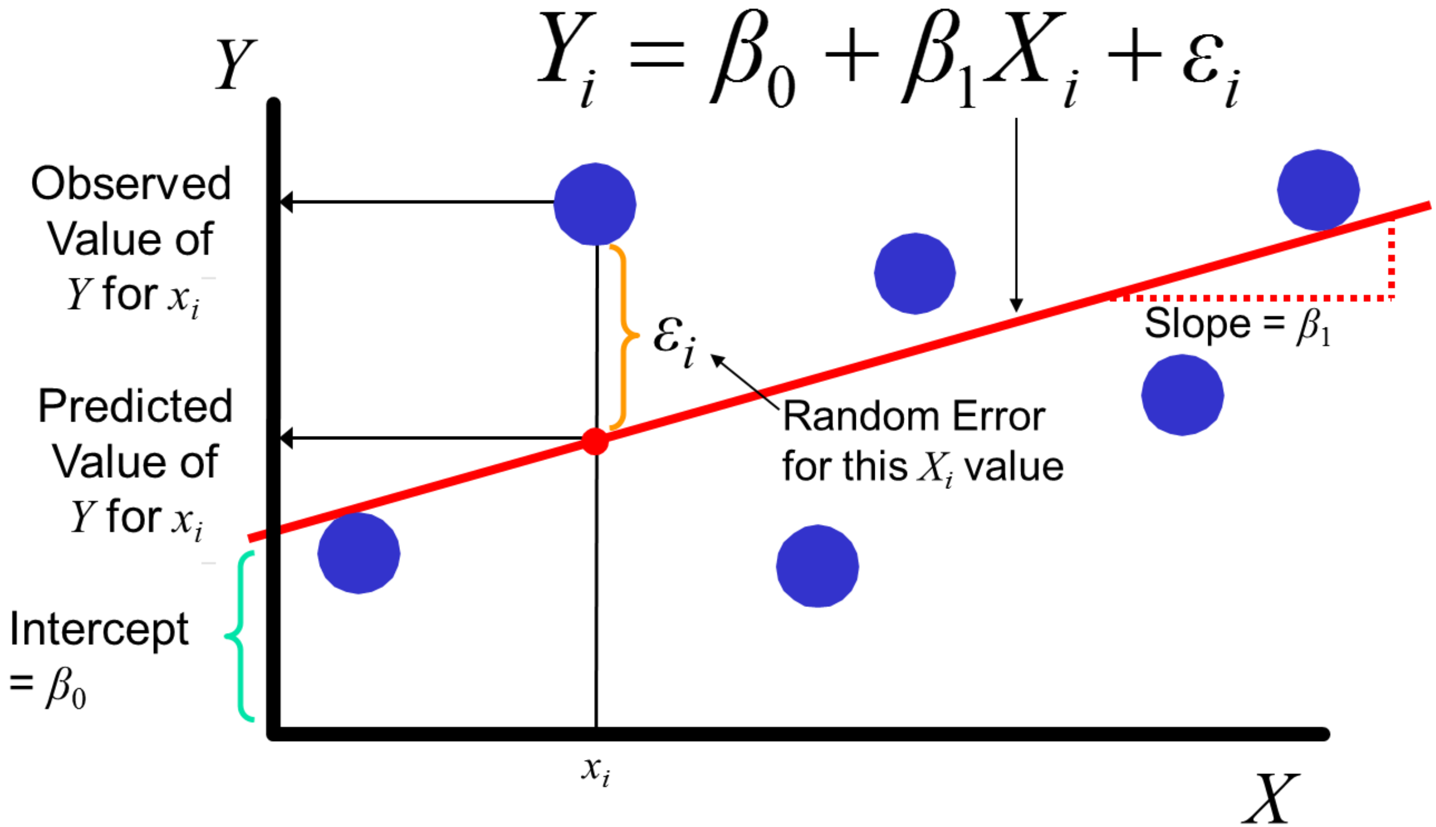
  (the uniform variance property is called homoscedasticity)

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

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

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

Simple Linear Regression Model (2 of 2)



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

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

Section 11.3 Least Squares Coefficient Estimators (1 of 2)

- 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 (errors), SSE:

$$\min \text{ SSE } = \min \sum_{i=1}^{n} e_i^2$$

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

$$= \min \sum \left[y_i - (b_0 + b_1 x_i) \right]^2$$

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

Least Squares Coefficient Estimators (2 of 2)

- The slope coefficient estimator is

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{Cov(x,y)}{s_x^2} = r\frac{s_y}{s_x}$$

- And the constant or *y*-intercept is

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

Simple Linear Regression Example

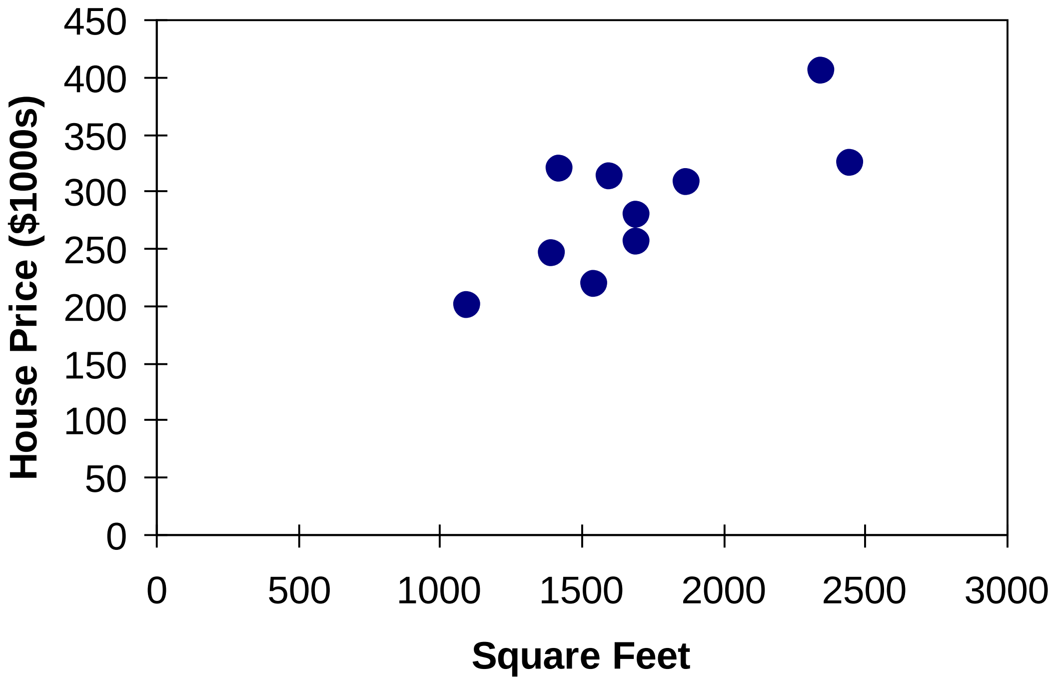
- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)
- A random sample of 10 houses is selected
  - Dependent variable (*Y*) = house price in $1000s
  - Independent variable (*X*) = square feet

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

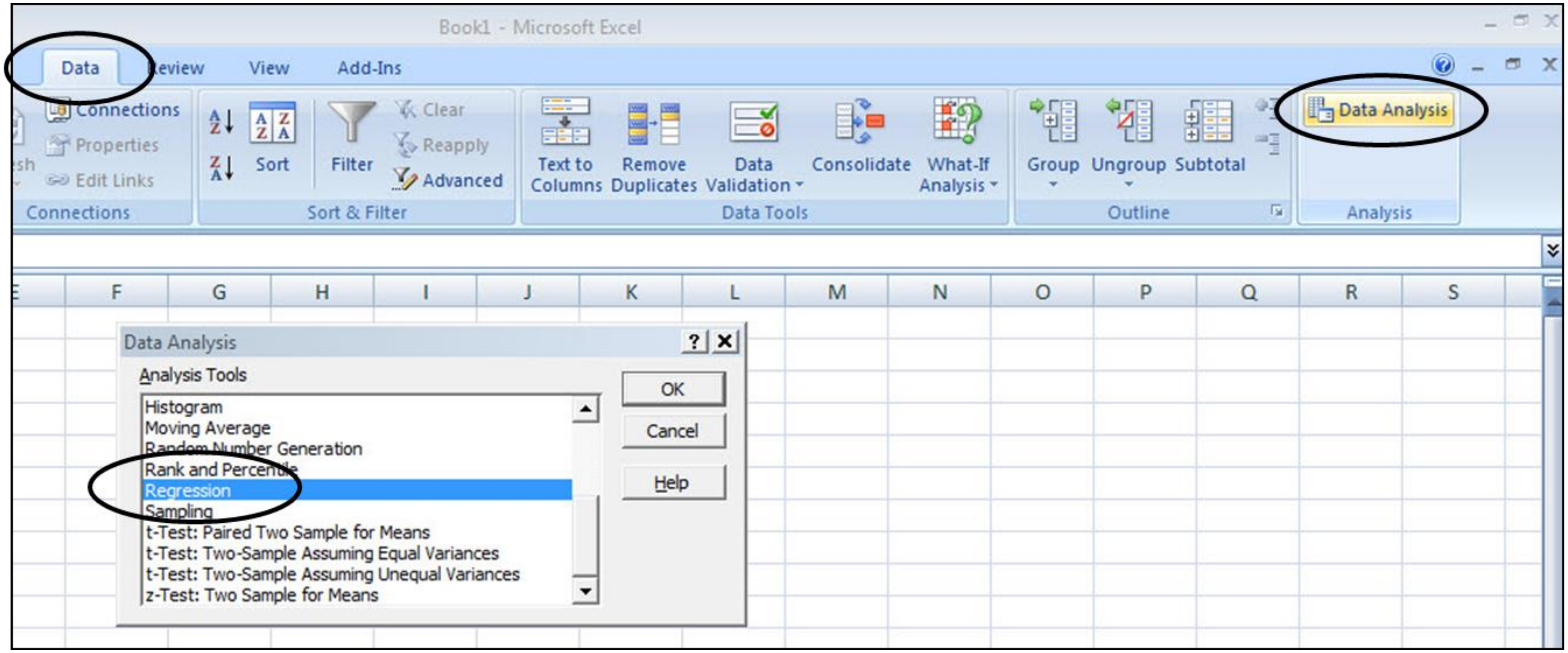
Graphical Presentation (1 of 2)

- House price model: scatter plot



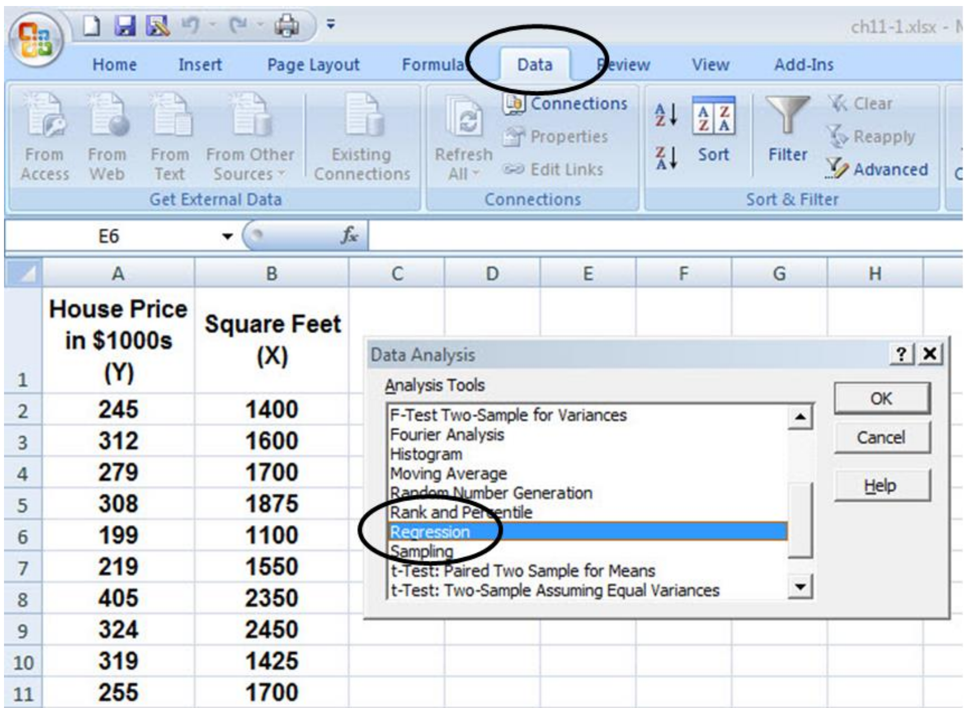
Regression Using Excel (1 of 2)

- Excel will be used to generate the coefficients and measures of goodness of fit for regression
  - Data / Data Analysis / Regression

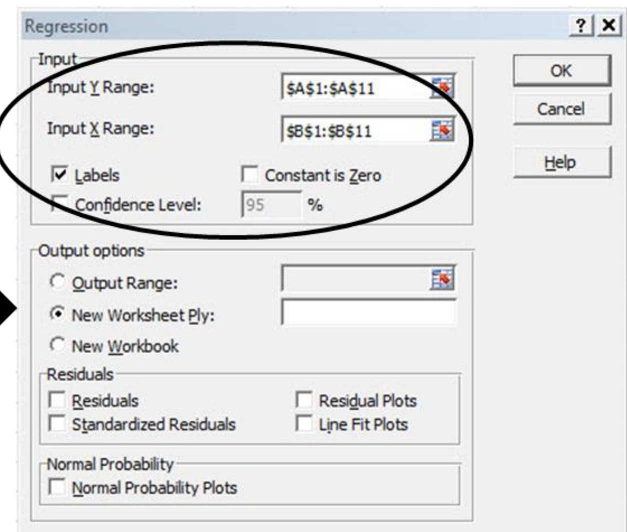


Regression Using Excel (2 of 2)

- Data / Data Analysis / Regression



Provide desired input:



Excel Output (1 of 6)

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.762113713					
5	R Square	0.580817312					
6	Adjusted R Square	0.528419476					
7	Standard Error	41.33032365					
8	Observations	10					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	18934.9348	18934.9348	11.0848	0.01039	
13	Residual	8	13665.5652	1708.1957			
14	Total	9	32600.5				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	98.24833	58.03348	1.69296	0.12892	-35.57711	232.07377
18	Square Feet (X)	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Excel Output (2 of 6)

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

The regression equation is:

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

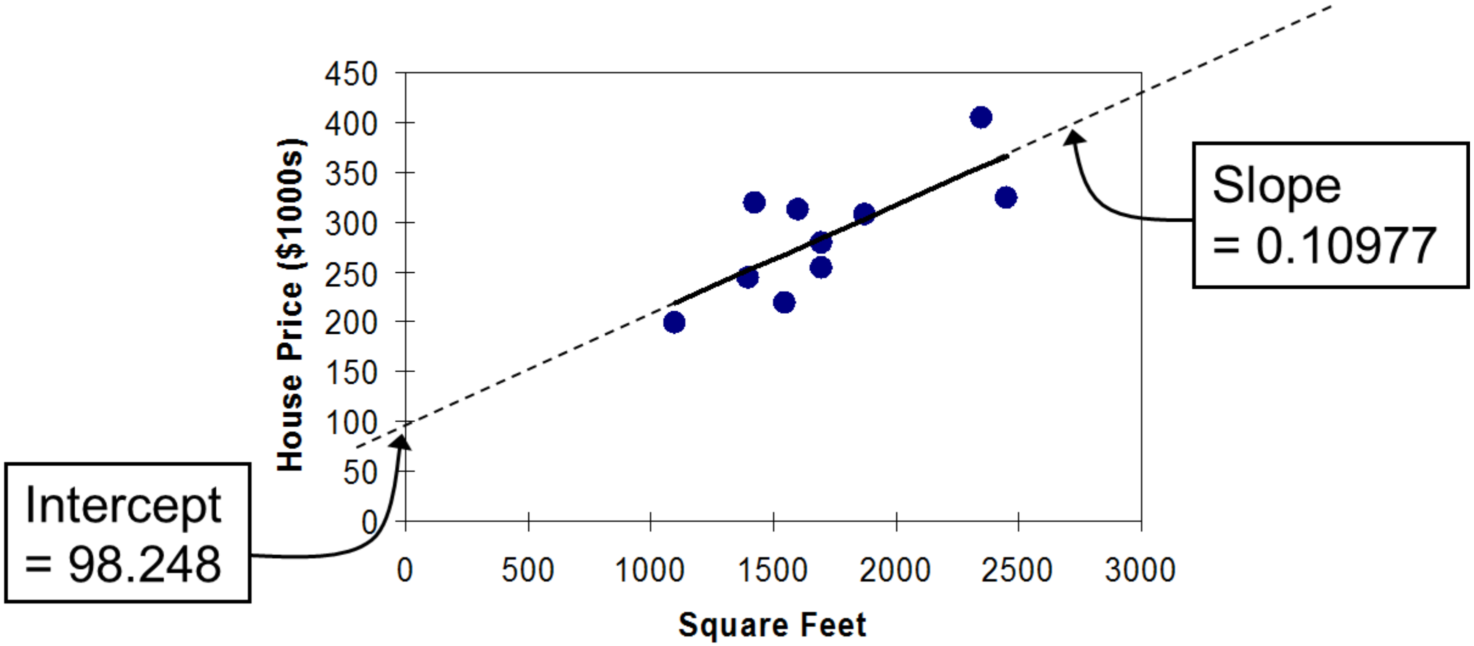
ANOVA

	df	*SS*	*MS*	*F*	Significance *F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	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

Graphical Presentation (2 of 2)

- House price model: scatter plot and regression line



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

Interpretation of the Intercept, *b* Sub 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)
  - Here, no houses had 0 square feet, so $\boxed{b_0 = 98.24833}$ 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* Sub 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

Section 11.4 Explanatory Power of a Linear Regression Equation

- Total variation is made up of two parts:

$$\text{SST} = \text{SSR} + \text{SSE}$$

Total Sum of Squares	Regression Sum of Squares	Error (residual) Sum of Squares

$$\text{SST} = \sum (y_i - \bar{y})^2 \quad \text{SSR} = \sum (\hat{y}_i - \bar{y})^2 \quad \text{SSE} = \sum (y_i - \hat{y}_i)^2$$

where:

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

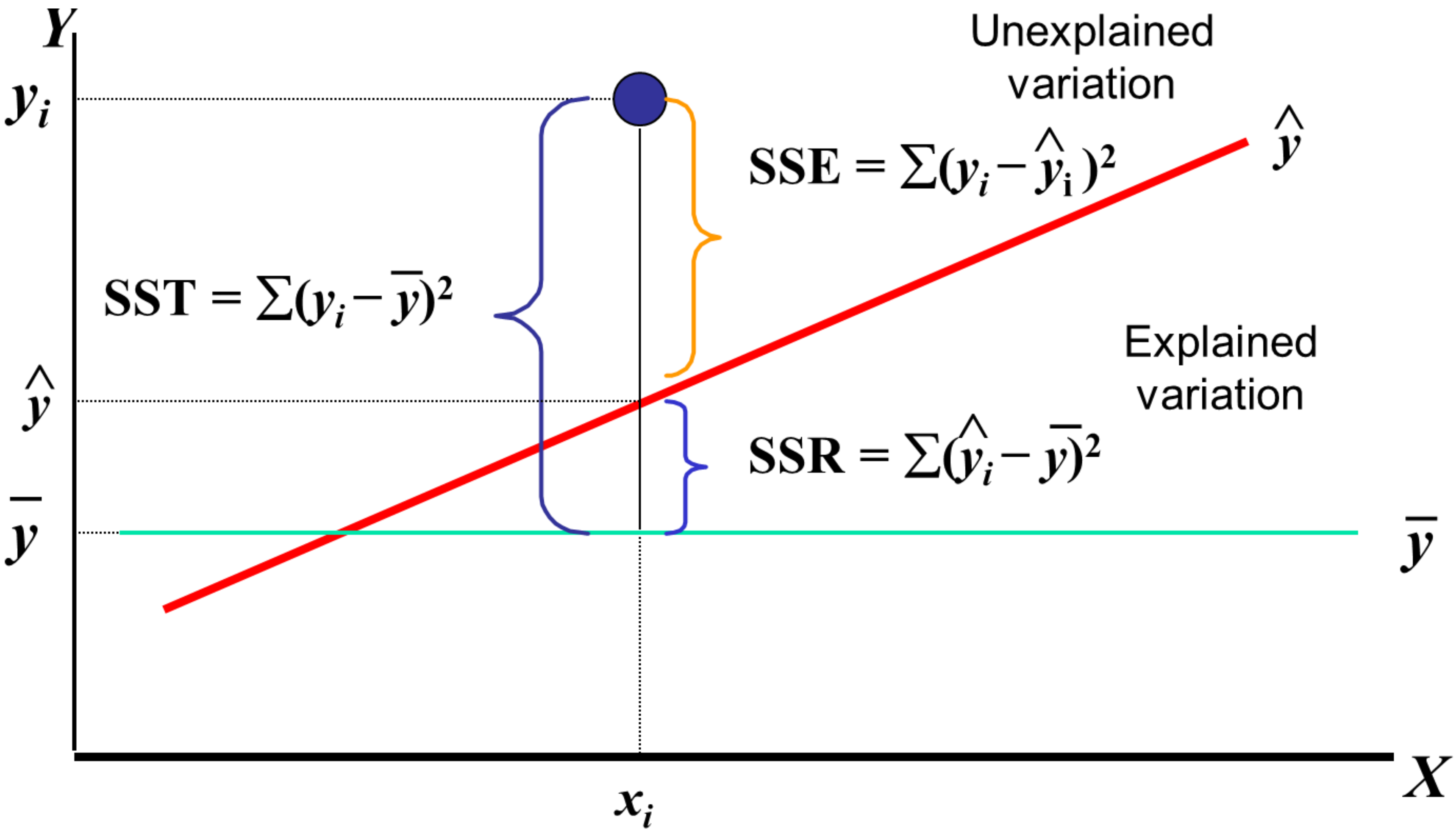
y_i = Observed values of the dependent variable

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

Analysis of Variance (1 of 2)

- 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*

Analysis of Variance (2 of 2)



Pearson

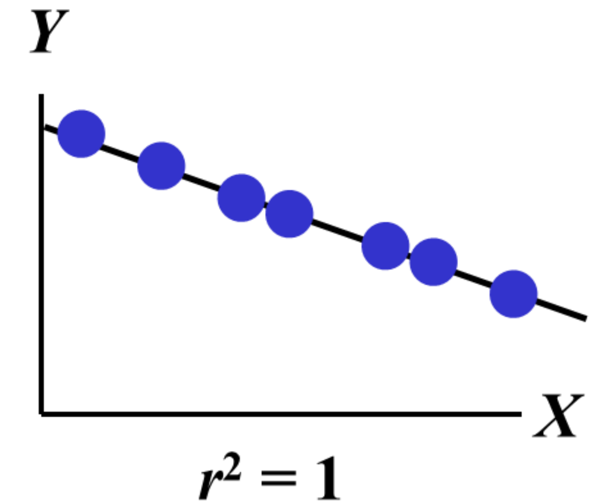
Coefficient of Determination, *R* Squared

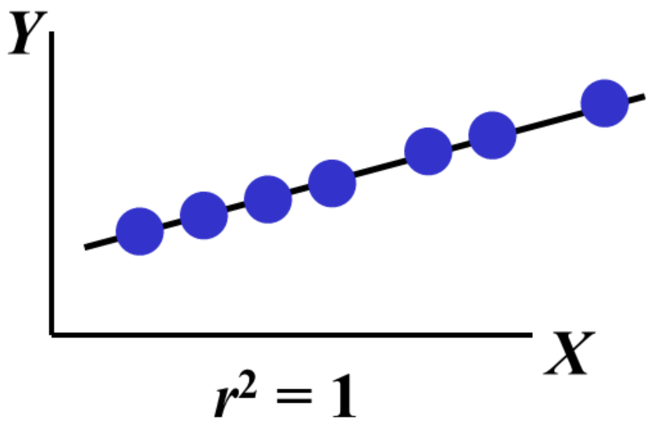
- The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable
- The coefficient of determination is also called *R*-squared and is denoted as R^2

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

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

Examples of Approximate *r* Squared Values (1 of 3)



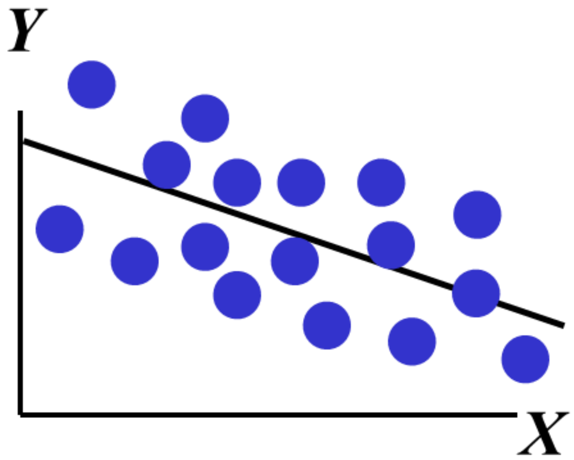


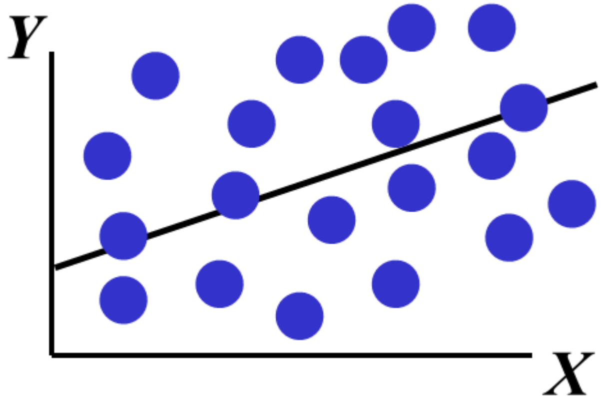
$$r^2 = 1$$

Perfect linear relationship between *X* and *Y*:

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

Examples of Approximate *r* Squared Values (2 of 3)



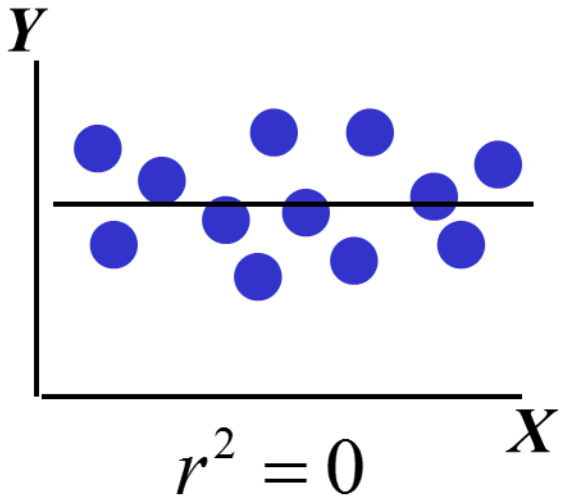


$$0 < r^2 < 1$$

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* Squared Values (3 of 3)



$$r^2 = 0$$

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*)

Excel Output (3 of 6)

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$$R^2 = \frac{SSR}{SST} = \frac{18934.9348}{32600.5000} = 0.58082$$

58.08% of the variation in house prices is explained by variation in square feet

ANOVA

	df	*SS*	*MS*	*F*	Significance *F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	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

Correlation and *R* Squared

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

$$R^2 = r^2$$

Estimation of Model Error Variance

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

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

- 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

Excel Output (4 of 6)

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

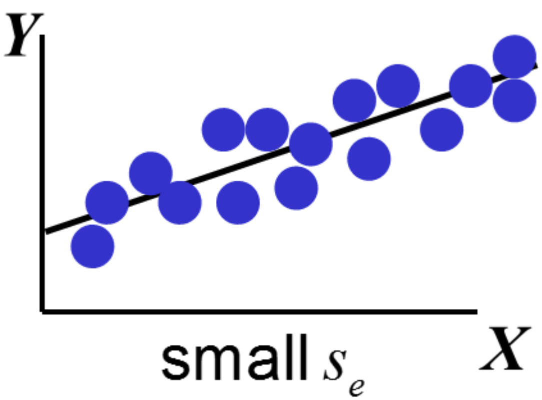
$$s_e = 41.33032$$

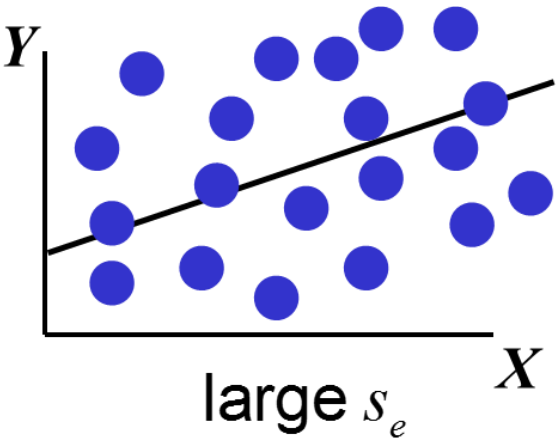
ANOVA	*df*	*SS*	*MS*	*F*	Significance *F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	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

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

Section 11.5 Statistical Inference: Hypothesis Tests and Confidence Intervals

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

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

where:

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

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

Excel Output (5 of 6)

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$$s_{b_1} = 0.03297$$

ANOVA	*df*	*SS*	*MS*	*F*	Significance *F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	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

Pearson

Inference About the Slope: *t* Test (1 of 2)

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

$$H_0 : \beta_1 = 0 \qquad \text{(no linear relationship)}$$

$$H_1 : \beta_1 \neq 0 \qquad \text{(linear relationship does exist)}$$

- Test statistic

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

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

where:

$b_1 =$ regression slope coefficient

$\beta_1 =$ hypothesized slope

$s_{b_1} =$ standard error of the slope

Inference About the Slope: *t* Test (2 of 2)

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 significantly affect its sales price?

Inferences About the Slope: *t* Test Example (1 of 3)

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

From Excel output:

b_1 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 (2 of 3)

Test Statistic: ***t* = 3.329**

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

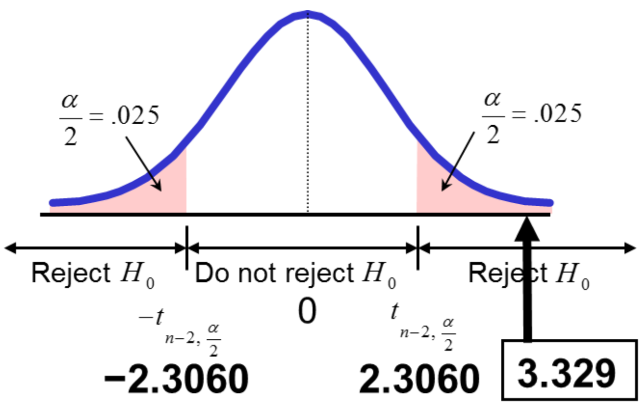
d.f. = 10 − 2 = 8

$t_{8,.025} = 2.3060$

From Excel output:

	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

(b_1 = 0.10977, S_{b_1} = 0.03297, t = 3.32938)



Decision:

Reject H_0

Conclusion:

There is sufficient evidence that square footage affects house price

Inferences About the Slope: *t* Test Example (3 of 3)

P-value = **0.01039**

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

From Excel output:

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 $< \alpha$ so
Reject H_0

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

Confidence Interval Estimate for the Slope (1 of 2)

Confidence Interval Estimate of the Slope:

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

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

Excel Printout for House Prices:

	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 (2 of 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

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

Hypothesis Test for Population Slope Using the *F* Distribution (1 of 2)

- *F* Test statistic:

$$F = \frac{\text{MSR}}{\text{MSE}}$$

where

$$\text{MSR} = \frac{\text{SSR}}{k}$$

$$\text{MSE} = \frac{\text{SSE}}{n-k-1}$$

where *F* follows an *F* distribution with *k* numerator and ($n - k - 1$) denominator degrees of freedom

(*k* = the number of independent variables in the regression model)

Hypothesis Test for Population Slope Using the *F* Distribution (2 of 2)

- An alternate test for the hypothesis that the slope is zero:

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

- Use the *F* statistic

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{\text{SSR}}{s_e^2}$$

- The decision rule is

reject H_0 if $F \geq F_{1,n-2,\alpha}$

Excel Output (6 of 6)

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$$F = \frac{MSR}{MSE} = \frac{18934.9348}{1708.1957} = 11.0848$$

With 1 and 8 degrees of freedom

P-value for the *F*-Test

ANOVA

	df	*SS*	*MS*	*F*	Significance *F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	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

F-Test for Significance

$$H_0: \beta_1 = 0$$

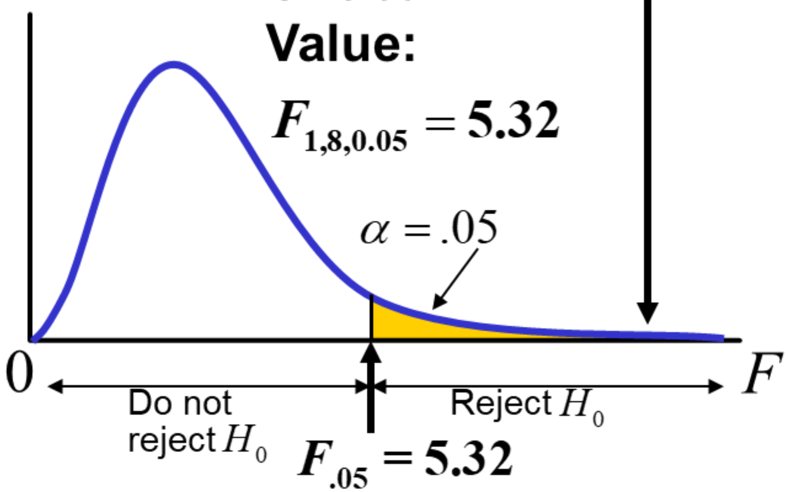
$$H_1: \beta_1 \neq 0$$

$$\alpha = .05$$

$$df_1 = 1 \qquad df_2 = 8$$

Critical Value:

$$F_{1,8,0.05} = 5.32$$



Test Statistic:

$$F = \frac{\text{MSR}}{\text{MSE}} = 11.08$$

Decision:

Reject H_0 at $\alpha = 0.05$

Conclusion:

There is sufficient evidence that house size affects selling price

Section 11.6 Prediction

- The regression equation can be used to predict a value for *y*, given a particular *x*
- For a specified value, x_{n+1}, the predicted value is

$$\hat{y}_{n+1} = b_0 + b_1 x_{n+1}$$

Predictions Using Regression Analysis

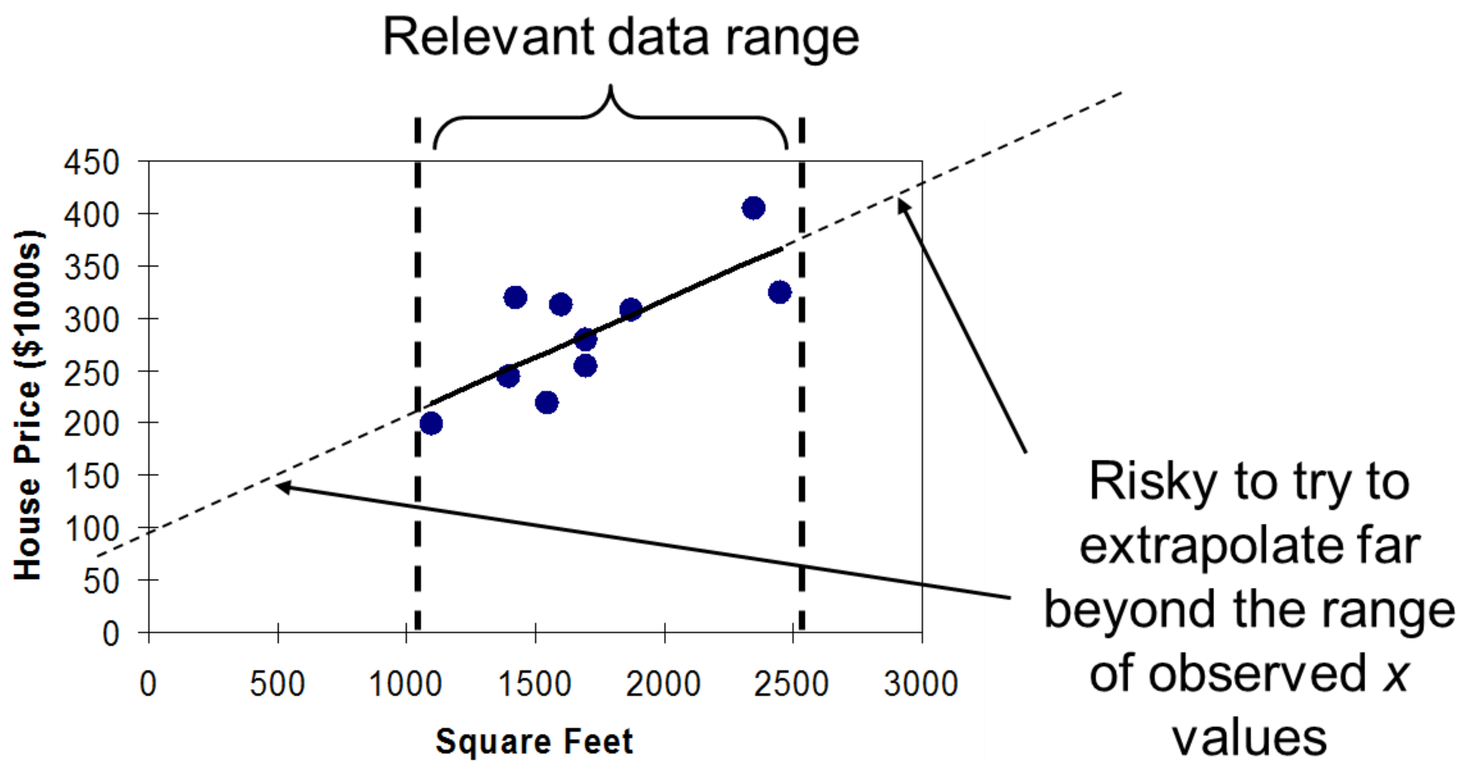
Predict the price for a house with 2000 square feet:

$$\widehat{\text{house price}} = 98.25 + 0.1098\ (\text{sq.ft.})$$
$$= 98.25 + 0.1098(2000)$$
$$= 317.85$$

The predicted price for a house with 2000 square feet is $317.85(\$1{,}000\text{s}) = \$317{,}850$

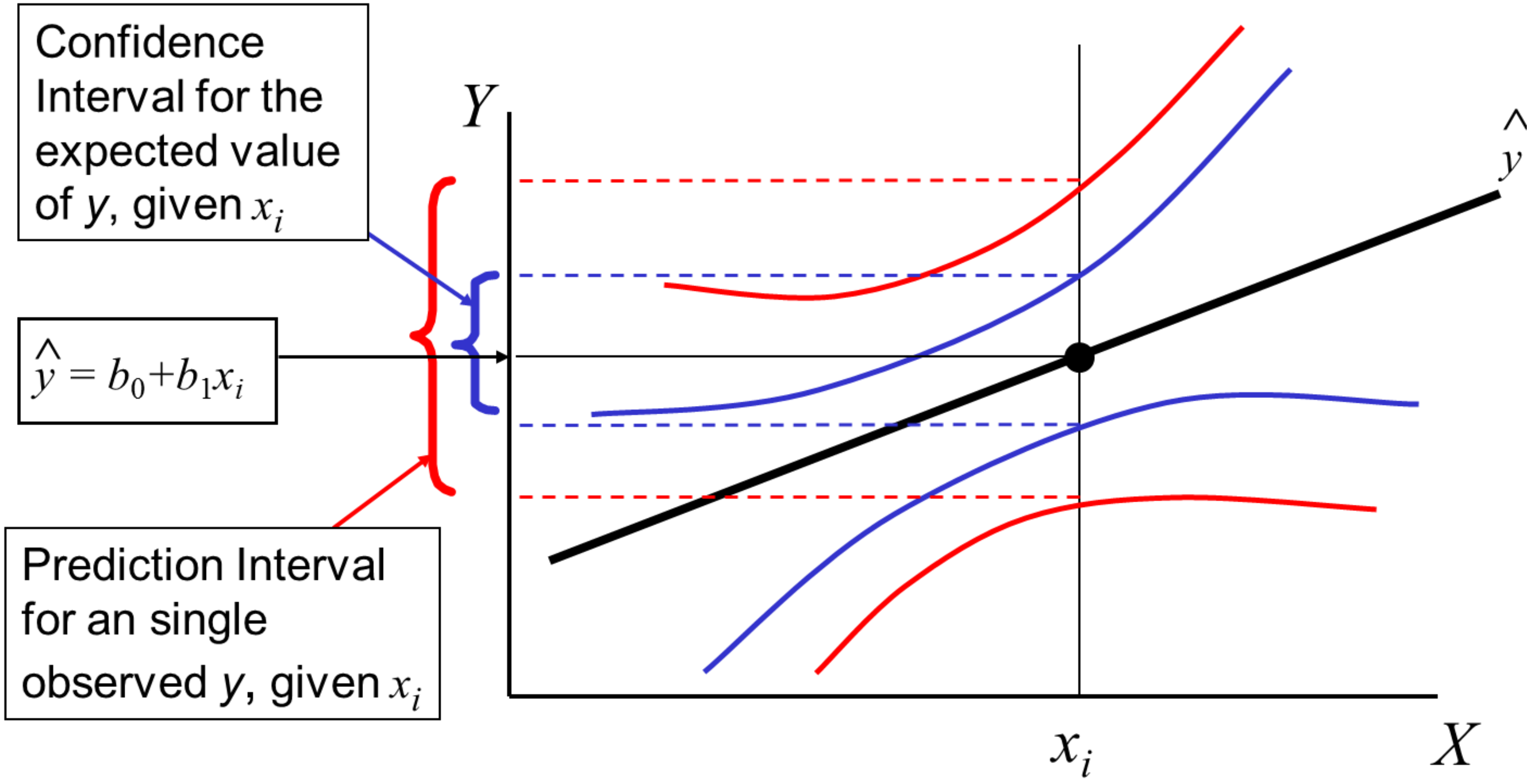
Relevant Data Range

- When using a regression model for prediction, only predict within the relevant range of data



Estimating Mean Values and Predicting Individual Values

Goal: Form intervals around *y* to express uncertainty about the value of *y* for a given x_i



Confidence Interval for the Average *Y*, Given *X*

Confidence interval estimate for the **expected value of *y*** given a particular x_i

Confidence interval for $E(Y_{n+1} | X_{n+1})$:

$$\hat{y}_{n+1} \pm t_{n-2,\frac{\alpha}{2}} s_e \sqrt{\left[\frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{\sum (x_i - \overline{x})^2}\right]}$$

Notice that the formula involves the term $(x_{n+1} - \overline{x})^2$

so the size of interval varies according to the distance x_{n+1} is from the mean, $\overline{x}$

Prediction Interval for an Individual *Y*, Given *X*

Confidence interval estimate for an **actual observed value of *y*** given a particular x_i

Confidence interval for $\hat{y}_{n+1}$:

$$\hat{y}_{n+1} \pm t_{n-2,\frac{\alpha}{2}} s_e \sqrt{\left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right]}$$

This extra term adds to the interval width to reflect the added uncertainty for an individual case

Example: Confidence Interval for the Average *Y*, Given *X* (1 of 2)

Confidence Interval Estimate for $E\left(Y_{n+1} \mid X_{n+1}\right)$

Find the 95% confidence interval for the mean price of 2,000 square-foot houses

Predicted Price $\hat{y}_i = 317.85$ ($1,000s)

$$\hat{y}_{n+1} \pm t_{n-2,\frac{\alpha}{2}} s_e \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}} = 317.85 \pm 37.12$$

The confidence interval endpoints are 280.73 and 354.97, or from $280,730 to $354,970

Example: Confidence Interval for the Average *Y*, Given *X* (2 of 2)

Confidence Interval Estimate for $\hat{y}_{n+1}$

Find the 95% confidence interval for an individual house with 2,000 square feet

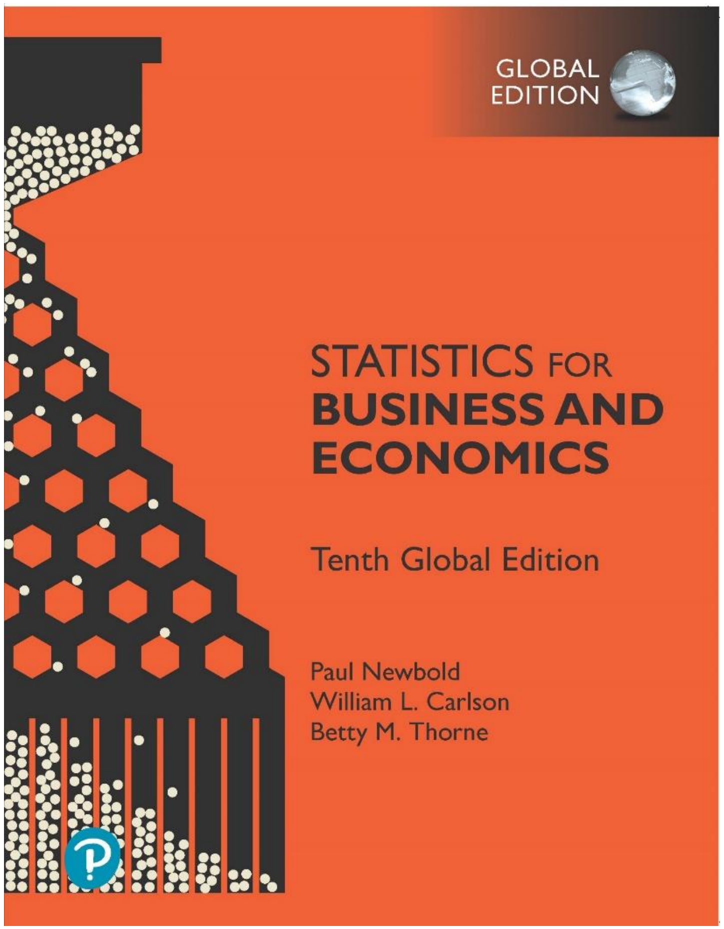
Predicted Price $\hat{y}_i = 317.85 \text{ (\$1,000s)}$

$$\hat{y}_{n+1} \pm t_{n-1,\frac{\alpha}{2}} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}} = 317.85 \pm 102.28$$

The confidence interval endpoints are 215.57 and 420.13, or from \$215,570 to \$420,130

Statistics for Business and Economics

Tenth Edition, Global Edition



Chapter 12

Multiple Regression

Section 12.1 The Multiple Regression Model

Idea: Examine the linear relationship between
1 dependent (Y) & 2 or more independent variables (X_i)

Multiple Regression Model with *K* Independent Variables:

Y-intercept — Population slopes — Random Error

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K + \varepsilon$$

Multiple Regression Equation

The coefficients of the multiple regression model are estimated using sample data

Multiple regression equation with *K* independent variables:

Estimated (or predicted) value of *y*

Estimated intercept

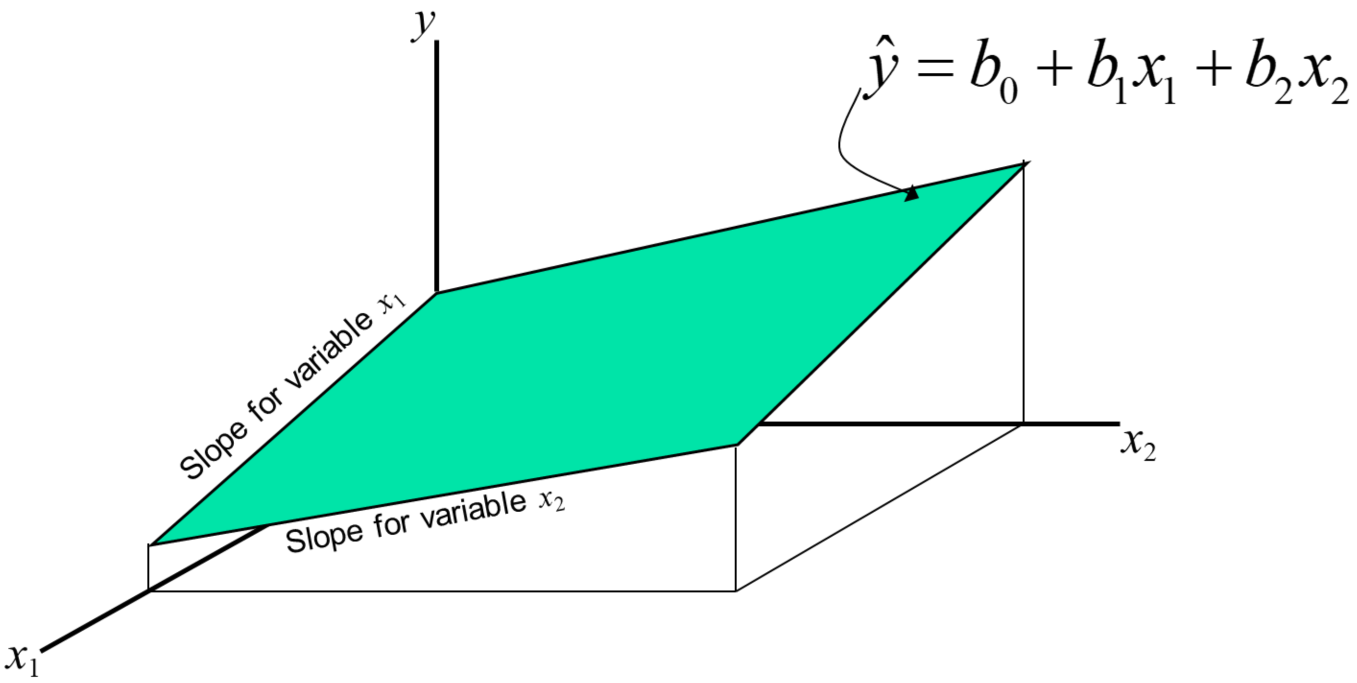
Estimated slope coefficients

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_K x_{Ki}$$

In this chapter we will always use a computer to obtain the regression slope coefficients and other regression summary measures.

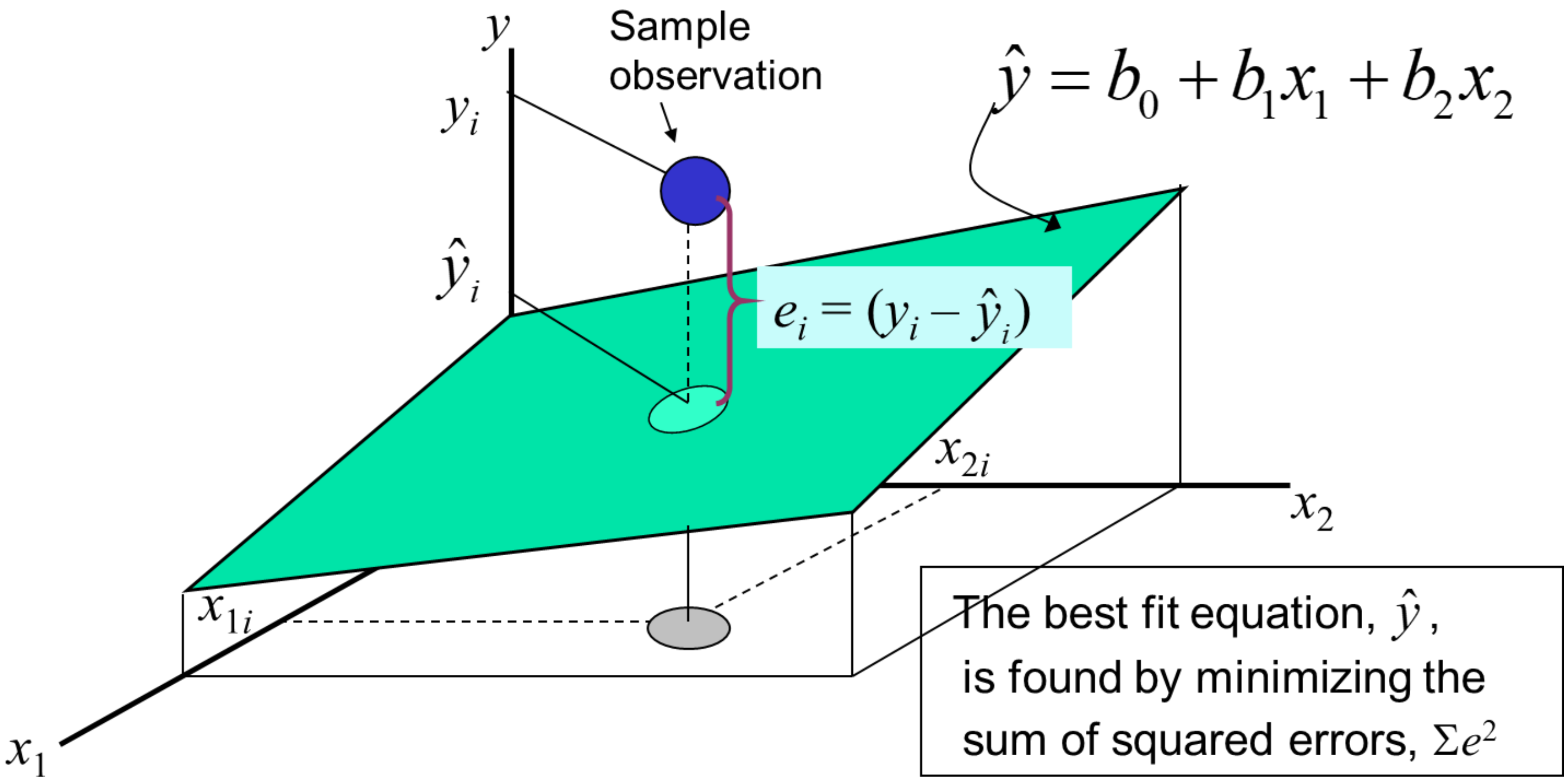
Three Dimensional Graphing (1 of 2)

Two variable model



Three Dimensional Graphing (2 of 2)

Two variable model



Section 12.2 Estimation of Coefficients

Standard Multiple Regression Assumptions

- 1. The x_{ji} terms are fixed numbers, or they are realizations of random variables X_j that are independent of the error terms, ε_i
- 2. The expected value of the random variable *Y* is a linear function of the independent X_j variables.
- 3. The error terms are normally distributed random variables with mean 0 and a constant variance, σ^2.

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

(The constant variance property is called homoscedasticity)

Standard Multiple Regression Assumptions

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

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

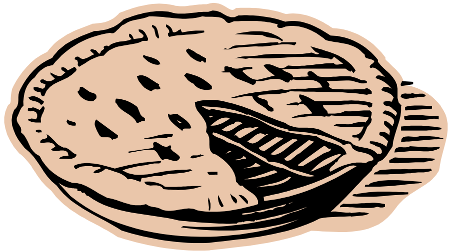
- 5. It is not possible to find a set of numbers, $c_0, c_1, \ldots, c_k$, such that

$$c_0 + c_1 x_{1i} + c_2 x_{2i} + \ldots + c_K x_{Ki} = 0$$

(This is the property of no linear relation for the $X_j s$)

Example 1: 2 Independent Variables

- A distributor of frozen desert pies wants to evaluate factors thought to influence demand
  - Dependent variable: Pie sales (units per week)
  - Independent variables: { Price (in $), Advertising ($100's) }
- Data are collected for 15 weeks



Pie Sales Example

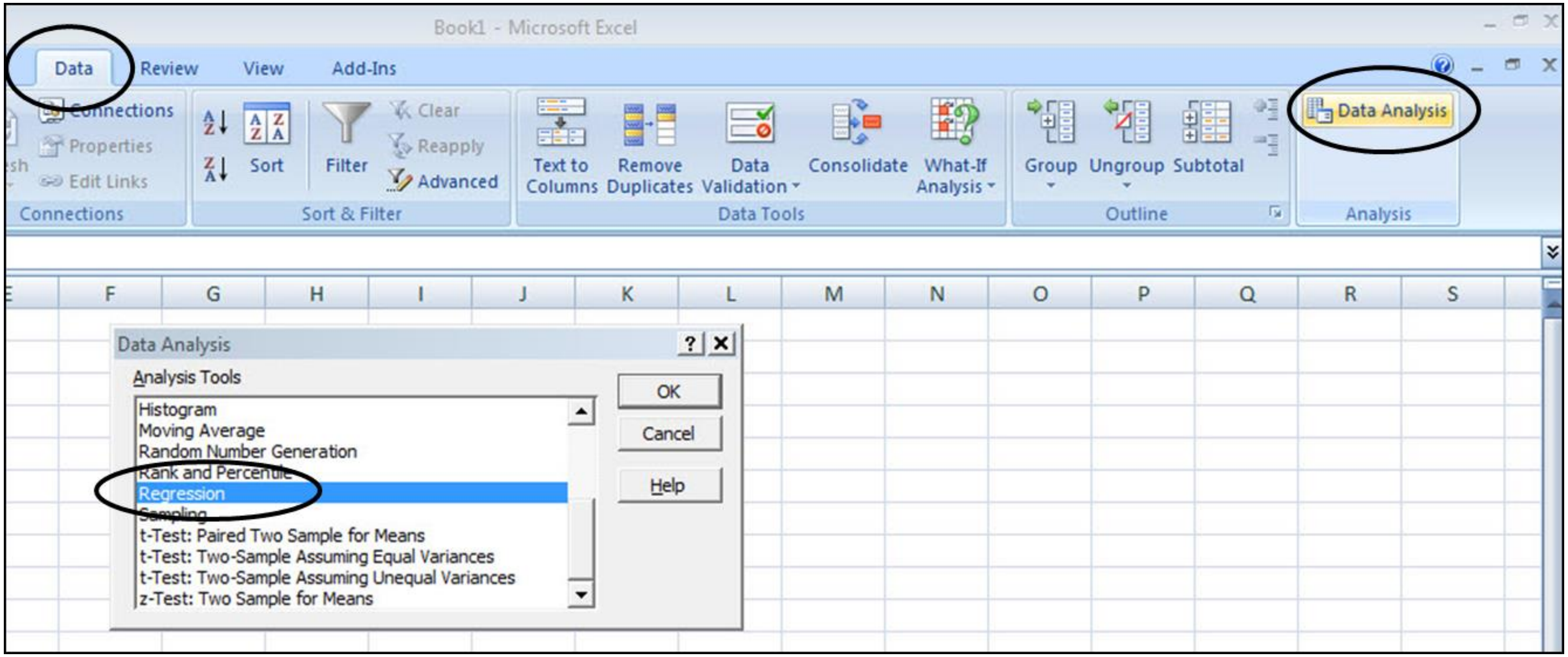
Week	Pie Sales	Price ($)	Advertising ($100s)
1	350	5.50	3.3
2	460	7.50	3.3
3	350	8.00	3.0
4	430	8.00	4.5
5	350	6.80	3.0
6	380	7.50	4.0
7	430	4.50	3.0
8	470	6.40	3.7
9	450	7.00	3.5
10	490	5.00	4.0
11	340	7.20	3.5
12	300	7.90	3.2
13	440	5.90	4.0
14	450	5.00	3.5
15	300	7.00	2.7

Multiple regression equation:

$$\widehat{\text{Sales}} = b_0 + b_1 \text{ (Price)} + b_2 \text{ (Advertising)}$$

Estimating a Multiple Linear Regression Equation

- Excel can be used to generate the coefficients and measures of goodness of fit for multiple regression
  - Data / Data Analysis / Regression



Multiple Regression Output

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$\widehat{Sales} = 306.526 - 24.975(Price) + 74.131(Advertising)$$

ANOVA	*df*	*SS*	*MS*	*F*	Significance *F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	Standard Error	*t* Stat	*P*-value	Lower 95%	Upper 95%
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	−24.97509	10.83213	−2.30565	0.03979	−48.57626	−1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

The Multiple Regression Equation

$$\widehat{Sales} = 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising})$$

where
Sales is in number of pies per week
Price is in $
Advertising is in $100's.

$b_1 = -24.975$: sales will decrease, on average, by 24.975 pies per week for each $1 increase in selling price, net of the effects of changes due to advertising

$b_2 = 74.131$: sales will increase, on average, by 74.131 pies per week for each $100 increase in advertising, net of the effects of changes due to price

Section 12.3 Explanatory Power of a Multiple Regression Equation

Coefficient of Determination, R^2

- Reports the proportion of total variation in *y* explained by all *x* variables taken together

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

- This is the ratio of the explained variability to total sample variability

Coefficient of Determination, *R* Squared

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$R^2 = \frac{SSR}{SST} = \frac{29460.0}{56493.3} = .52148$$

52.1% of the variation in pie sales is explained by the variation in price and advertising

ANOVA	*df*	*SS*	*MS*	*F*	Significance *F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	Standard Error	*t* Stat	*P*-value	Lower 95%	Upper 95%
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	−24.97509	10.83213	−2.30565	0.03979	−48.57626	−1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Estimation of Error Variance

- Consider the population regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$$

- The unbiased estimate of the variance of the errors is

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

where $e_i = y_i - \hat{y}_i$

- The square root of the variance, s_e, is called the standard error of the estimate

Standard Error, *s* Sub Epsilon

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$s_e = 47.463$$

The magnitude of this value can be compared to the average y value

ANOVA	*df*	*SS*	*MS*	*F*	Significance *F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	Standard Error	*t* Stat	*P*-value	Lower 95%	Upper 95%
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	−24.97509	10.83213	−2.30565	0.03979	−48.57626	−1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Adjusted Coefficient of Determination, *R* Bar Squared (1 of 2)

- R^2 never decreases when a new *X* variable is added to the model, even if the new variable is not an important predictor variable
  - This can be a disadvantage when comparing models
- What is the **net effect** of adding a new variable?
  - We lose a degree of freedom when a new *X* variable is added
  - Did the new *X* variable add enough explanatory power to offset the loss of one degree of freedom?

Adjusted Coefficient of Determination, *R* Bar Squared (2 of 2)

- Used to correct for the fact that adding non-relevant independent variables will still reduce the error sum of squares

$$\bar{R}^2 = 1 - \frac{\text{SSE}/(n-K-1)}{\text{SST}/(n-1)}$$

(where *n* = sample size, *K* = number of independent variables)

  - Adjusted R^2 provides a better comparison between multiple regression models with different numbers of independent variables
  - Penalize excessive use of unimportant independent variables
  - Value is less than R^2

R Bar Squared

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$\bar{R}^2 = .44172$$

44.2% of the variation in pie sales is explained by the variation in price and advertising, taking into account the sample size and number of independent variables

ANOVA	*df*	*SS*	*MS*	*F*	Significance *F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	Standard Error	*t* Stat	*P*-value	Lower 95%	Upper 95%
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	−24.97509	10.83213	−2.30565	0.03979	−48.57626	−1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Section 12.4 Conf. Intervals and Hypothesis Tests for Regression Coefficients

The variance of a coefficient estimate is affected by:

- the sample size
- the spread of the *X* variables
- the correlations between the independent variables, and
- the model error term

We are typically more interested in the regression coefficients b_j than in the constant or intercept b_0

Confidence Intervals (1 of 2)

Confidence interval limits for the population slope β_j

$$b_j \pm t_{n-K-1,\frac{\alpha}{2}} S_{b_j}$$

where t has $(n - K - 1)$ d.f.

	Coefficients	Standard Error
Intercept	306.52619	114.25389
Price	−24.97509	10.83213
Advertising	74.13096	25.96732

Here, t has

$(15 - 2 - 1) = 12$ d.f.

Example: Form a 95% confidence interval for the effect of changes in price (x_1) on pie sales:

$$-24.975 \pm (2.1788)(10.832)$$

So the interval is $-48.576 < \beta_1 < -1.374$

Confidence Intervals (2 of 2)

Confidence interval for the population slope β_i

	Coefficients	Standard Error	…	Lower 95%	Upper 95%
Intercept	306.52619	114.25389	…	57.58835	555.46404
Price	−24.97509	10.83213	…	−48.57626	−1.37392
Advertising	74.13096	25.96732	…	17.55303	130.70888

Example: Excel output also reports these interval endpoints:

Weekly sales are estimated to be reduced by between 1.37 to 48.58 pies for each increase of $1 in the selling price

Hypothesis Tests

- Use *t*-tests for individual coefficients
- Shows if a specific independent variable is conditionally important
- Hypotheses:
  - $H_0: \beta_j = 0$ (no linear relationship)
  - $H_1: \beta_j \neq 0$ (linear relationship does exist between x_j and *y*)

Evaluating Individual Regression Coefficients (1 of 3)

$H_0 : \beta_j = 0$ (no linear relationship)

$H_1 : \beta_j \neq 0$ (linear relationship does exist between x_i and y)

Test Statistic:

$$t = \frac{b_j - 0}{S_{b_j}} \qquad (\text{df} = n - k - 1)$$

Evaluating Individual Regression Coefficients (2 of 3)

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

***t*-value for Price is $t = -2.306$, with *p*-value .0398**

***t*-value for Advertising is $t = 2.855$, with *p*-value .0145**

ANOVA	*df*	*SS*	*MS*	*F*	Significance *F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	Standard Error	*t* Stat	*P*-value	Lower 95%	Upper 95%
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	−24.97509	10.83213	−2.30565	0.03979	−48.57626	−1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Example 2: Evaluating Individual Regression Coefficients

$H_0 : \beta_j = 0$

$H_1 : \beta_j \neq 0$

$\text{d.f.} = 15 - 2 - 1 = 12$

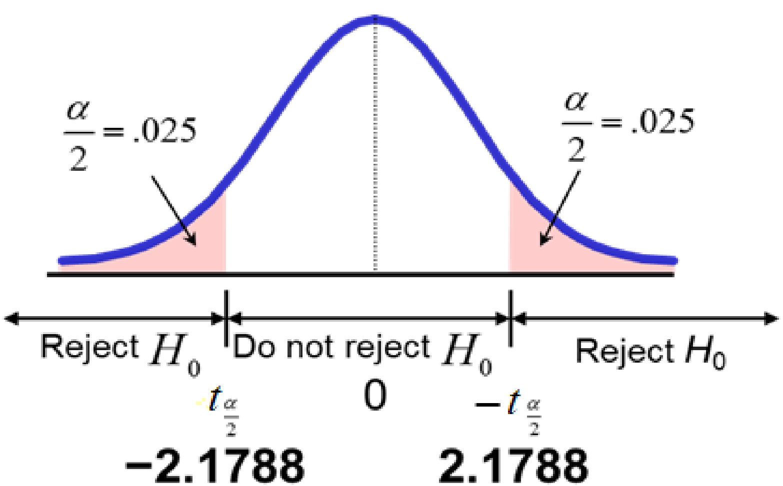
$\alpha = .05$

$t_{12,.025} = 2.1788$

From Excel output:

	Coefficients	Standard Error	*t* Stat	*P*-value
Price	−24.97509	10.83213	−2.30565	0.03979
Advertising	74.13096	25.96732	2.85478	0.01449

The test statistic for each variable falls in the rejection region (*p*-values < .05)



Decision:

Reject H_0 for each variable

Conclusion:

There is evidence that both Price and Advertising affect pie sales at $\alpha = .05$

Section 12.5 Tests on Regression Coefficients

Tests on All Coefficients

- *F*-Test for Overall Significance of the Model
- Shows if there is a linear relationship between all of the *X* variables considered together and *Y*
- Use *F* test statistic
- Hypotheses:

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_K = 0$ (no linear relationship)

H_1 : at least one $\beta_i \neq 0$ (at least one independent variable affects *Y*)

F-Test for Overall Significance (1 of 3)

- Test statistic:

$$F = \frac{\text{MSR}}{s_e^2} = \frac{\text{SSR}/K}{\text{SSE}/(n-K-1)}$$

where F has K (numerator) and
$(n - K - 1)$ (denominator)
degrees of freedom

- The decision rule is

$$\text{Reject } H_0 \text{ if } F = \frac{\text{MSR}}{s_e^2} > F_{K,n-K-1,\alpha}$$

F-Test for Overall Significance (2 of 3)

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{14730.0}{2252.8} = 6.5386$$

With 2 and 12 degrees of freedom

P-value for the F-Test

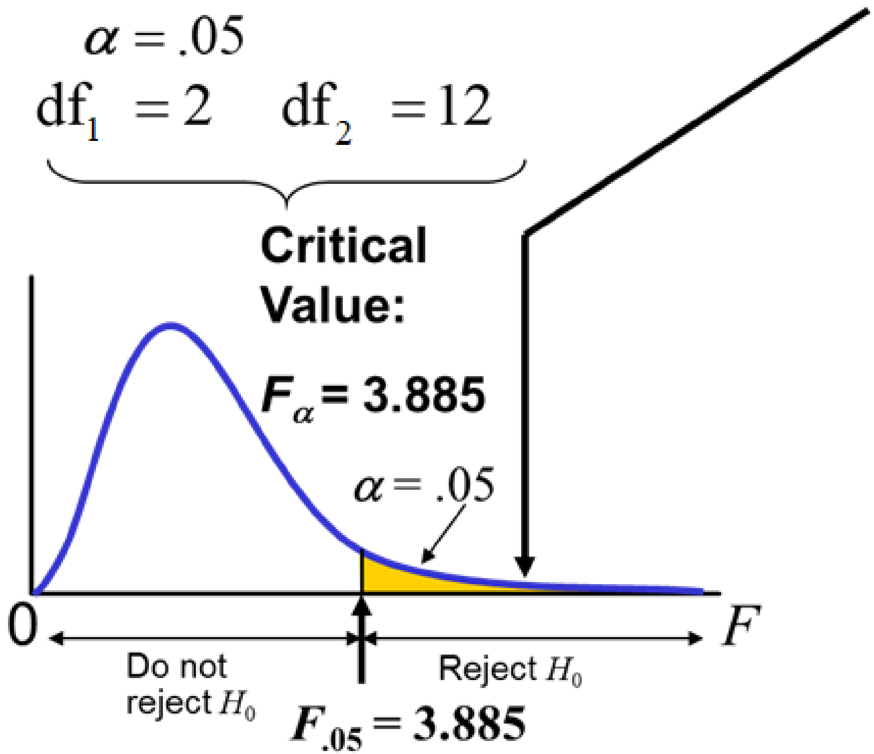
ANOVA	*df*	*SS*	*MS*	*F*	Significance *F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	Standard Error	*t* Stat	*P*-value	Lower 95%	Upper 95%
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	−24.97509	10.83213	−2.30565	0.03979	−48.57626	−1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

F-Test for Overall Significance (3 of 3)

$H_0 : \beta_1 = \beta_2 = 0$

$H_1 : \beta_1$ and β_2 not both zero



Test Statistic:

$$F = \frac{\text{MSR}}{\text{MSE}} = 6.5386$$

Decision:

Since *F* test statistic is in the rejection region (*p*-value < .05), reject H_0

Conclusion:

There is evidence that at least one independent variable affects *Y*

Test on a Subset of Regression Coefficients (1 of 2)

- Consider a multiple regression model involving variables X_j and Z_j, and the null hypothesis that the *Z* variable coefficients are all zero:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K + \alpha_1 z_1 + \cdots \alpha_R z_R + \varepsilon$$

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_R = 0$$

H_1 : at least one of $\alpha_j \neq 0 \ \left(j = 1, \ldots, R \right)$

Test on a Subset of Regression Coefficients (2 of 2)

- Goal: compare the error sum of squares for the complete model with the error sum of squares for the restricted model
  - First run a regression for the complete model and obtain SSE
  - Next run a restricted regression that excludes the *Z* variables (the number of variables excluded is *R*) and obtain the restricted error sum of squares SSE(*R*)
  - Compute the *F* statistic and apply the decision rule for a significance level α

$$\text{Reject } H_0 \text{ if } F = \frac{\left(\text{SSE}(\text{R}) - \text{SSE}\right)/\text{R}}{s_e^2} > F_{R,n-K-R-1,\alpha}$$

Section 12.6 Prediction

- Given a population regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i \quad (i = 1, 2, \ldots, n)$$

- then given a new observation of a data point

$$\left(x_{1,n+1},\ x_{2,n+1}, \ldots,\ x_{K,n+1}\right)$$

the best linear unbiased forecast of $\hat{y}_{n+1}$ is

$$\hat{y}_{n+1} = b_0 + b_1 x_{1,n+1} + b_2 x_{2,n+1} + \cdots + b_K x_{K,n+1}$$

- It is risky to forecast for new *X* values outside the range of the data used to estimate the model coefficients, because we do not have data to support that the linear model extends beyond the observed range.

Predictions from a Multiple Regression Model

Predict sales for a week in which the selling price is $5.50 and advertising is $350:

$$\begin{aligned}\widehat{\text{Sales}} &= 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising}) \\ &= 306.526 - 24.975(5.50) + 74.131(3.5) \\ &= 428.62\end{aligned}$$

Predicted sales is 428.62 pies

Note that Advertising is in $100's, so $350 means that $X_2 = 3.5$

Section 12.7 Transformations for Nonlinear Regression Models

- The relationship between the dependent variable and an independent variable may not be linear
- Can review the scatter diagram to check for non-linear relationships
- Example: Quadratic model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon$$

  – The second independent variable is the square of the first variable

Quadratic Model Transformations

Quadratic model form:

Let $z_1 = x_1$ and $z_2 = x_1^2$

And specify the model as

$$y_i = \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} + \varepsilon_i$$

- where:

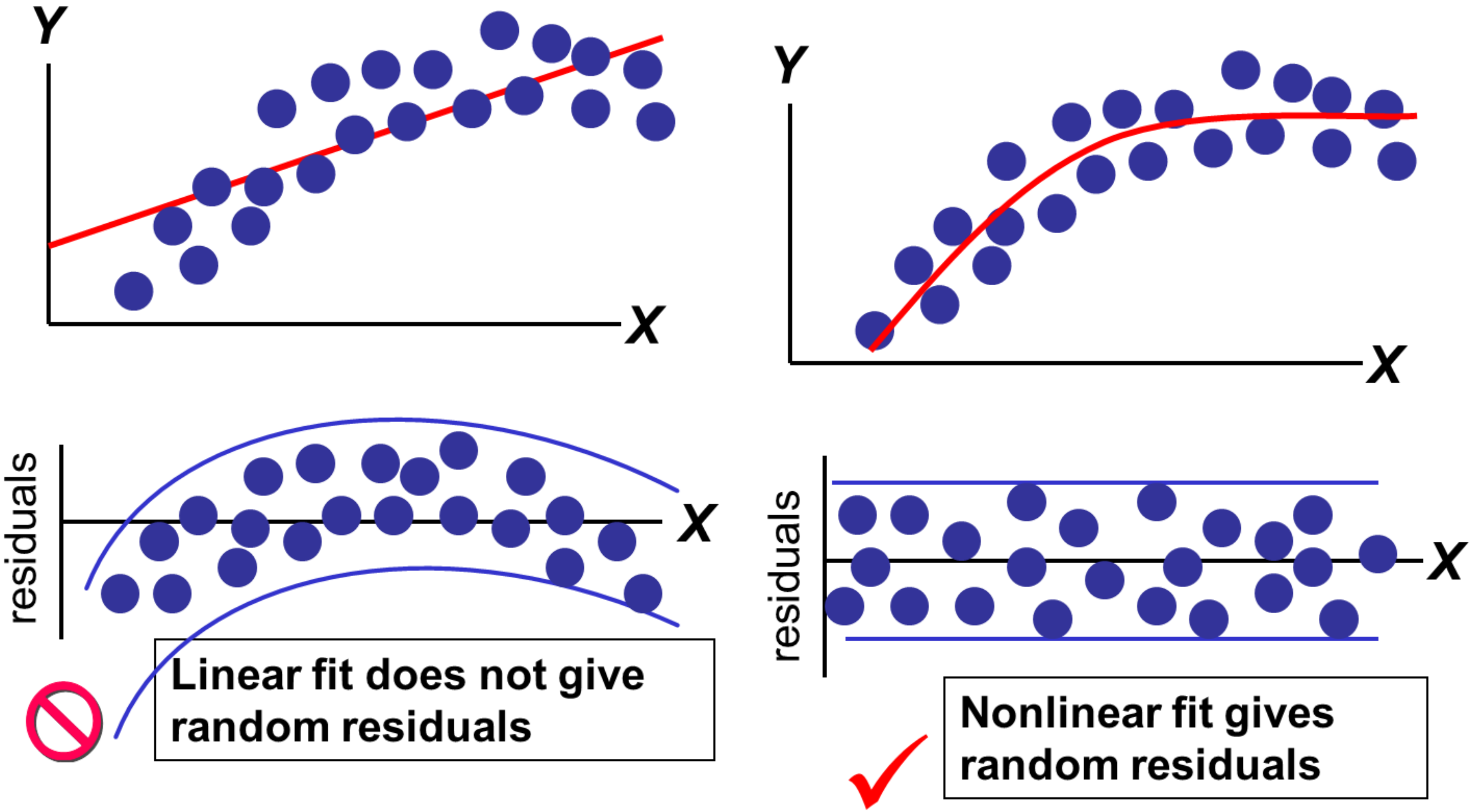
  $\beta_0 =$ *Y* intercept

  $\beta_1 =$ regression coefficient for linear effect of *X* on *Y*

  $\beta_2 =$ regression coefficient for quadratic effect on *Y*

  $\varepsilon_i =$ random error in *Y* for observation *i*

Linear vs. Nonlinear Fit



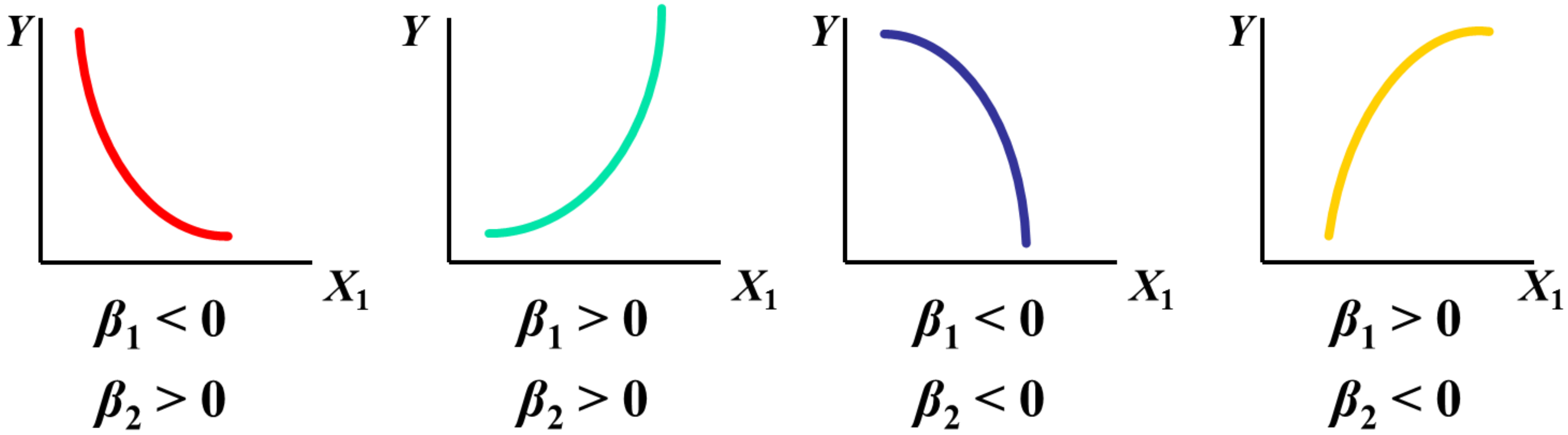
Linear fit does not give random residuals

Nonlinear fit gives random residuals

Quadratic Regression Model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \varepsilon_i$$

Quadratic models may be considered when the scatter diagram takes on one of the following shapes:



β_1 = the coefficient of the linear term

β_2 = the coefficient of the squared term

Testing for Significance: Quadratic Effect (1 of 3)

- Testing the Quadratic Effect
  - Compare the linear regression estimate

$$\hat{y} = b_0 + b_1 x_1$$

  - with quadratic regression estimate

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_1^2$$

  - Hypotheses
    - $H_0 : \beta_2 = 0$ (The quadratic term does not improve the model)
    - $H_1 : \beta_2 \neq 0$ (The quadratic term improves the model)

Testing for Significance: Quadratic Effect (2 of 3)

- Testing the Quadratic Effect

  Hypotheses

  - $H_0 : \beta_2 = 0$ (The quadratic term does not improve the model)
  - $H_1 : \beta_2 \neq 0$ (The quadratic term improves the model)

- The test statistic is

$$t = \frac{b_2 - \beta_2}{S_{b_2}}$$

$$\text{d.f} = n - 3$$

where:

$b_2 =$ squared term slope coefficient

$\beta_2 =$ hypothesized slope (zero)

$S_{b_2} =$ standard error of the slope

Testing for Significance: Quadratic Effect (3 of 3)

- Testing the Quadratic Effect

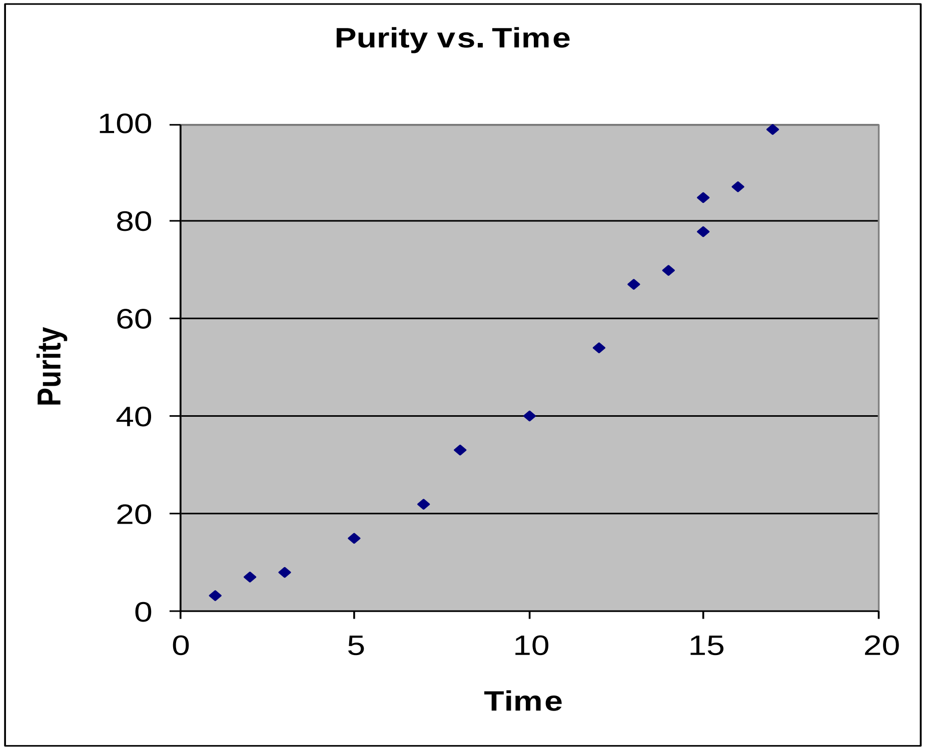
  Compare R^2 from simple regression to $\bar{R}^2$ from the quadratic model

- If $\bar{R}^2$ from the quadratic model is larger than R^2 from the simple model, then the quadratic model is a better model

Example 3: Quadratic Model (1 of 3)

Purity	Filter Time
3	1
7	2
8	3
15	5
22	7
33	8
40	10
54	12
67	13
70	14
78	15
85	15
87	16
99	17

- Purity increases as filter time increases:



Example 3: Quadratic Model (2 of 3)

- Simple regression results:

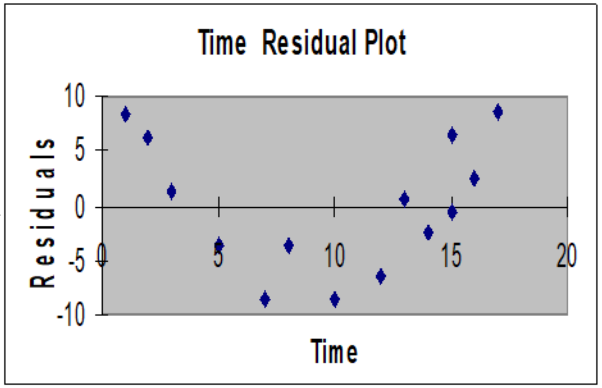
$$\hat{y} = -11.283 + 5.985 \text{ Time}$$

	Coefficients	Standard Error	*t* Stat	*P*-value
Intercept	−11.28267	3.46805	−3.25332	0.00691
Time	5.98520	0.30966	**19.32819**	2.078E-10

Regression Statistics	
R Square	**0.96888**
Adjusted R Square	0.96628
Standard Error	**6.15997**

F	Significance *F*
373.57904	2.0778E-10

t statistic, *F* statistic, and R^2 are all high, but the residuals are not random:



Example 3: Quadratic Model (3 of 3)

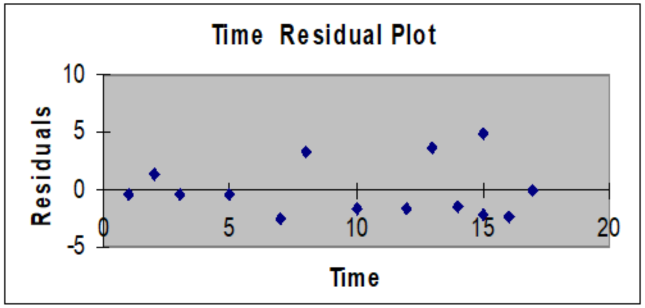
- Quadratic regression results:

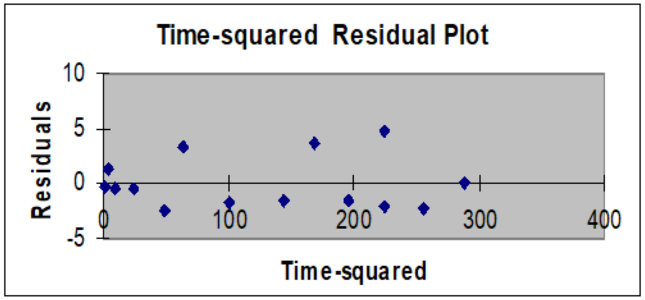
$$\hat{y} = 1.539 + 1.565\ \text{Time} + 0.245\ (\text{Time})^2$$

	Coefficients	Standard Error	t Stat	P-value
Intercept	1.53870	2.24465	0.68550	0.50722
Time	1.56496	0.60179	2.60052	0.02467
Time-squared	0.24516	0.03258	**7.52406**	1.165E-05

Regression Statistics	
R Square	0.99494
Adjusted R Square	**0.99402**
Standard Error	**2.59513**

F	Significance *F*
1080.7330	2.368E-13





The quadratic term is significant and improves the model: $\bar{R}^2$ is higher and s_e is lower, residuals are now random

Logarithmic Transformations

The Exponential Model:

- Original exponential model

$$Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} \varepsilon$$

- Transformed logarithmic model

$$\log(Y) = \log(\beta_0) + \beta_1 \log(X_1) + \beta_2 \log(X_2) + \log(\varepsilon)$$

Interpretation of coefficients

For the logarithmic model:

$$\log Y_i = \log \beta_0 + \beta_1 \log X_{1i} + \log \varepsilon_i$$

- When both dependent and independent variables are logged:
  - The estimated coefficient b_k of the independent variable X_k can be interpreted as

    a 1 percent change in X_k leads to an estimated b_k percentage change in the average value of *Y*
  - b_k is the elasticity of *Y* with respect to a change in X_k

Section 12.8 Dummy Variables for Regression Models

- A dummy variable is a categorical independent variable with two levels:
  - yes or no, on or off, male or female
  - recorded as 0 or 1
- Regression intercepts are different if the variable is significant
- Assumes equal slopes for other variables
- If more than two levels, the number of dummy variables needed is (number of levels - 1)

Dummy Variable Example (1 of 2)

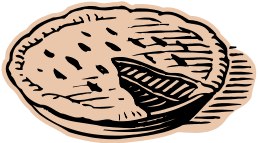
$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

Let:

y = Pie Sales

x_1 = Price

x_2 = Holiday ($x_2 = 1$ if a holiday occurred during the week)
($x_2 = 0$ if there was no holiday that week)

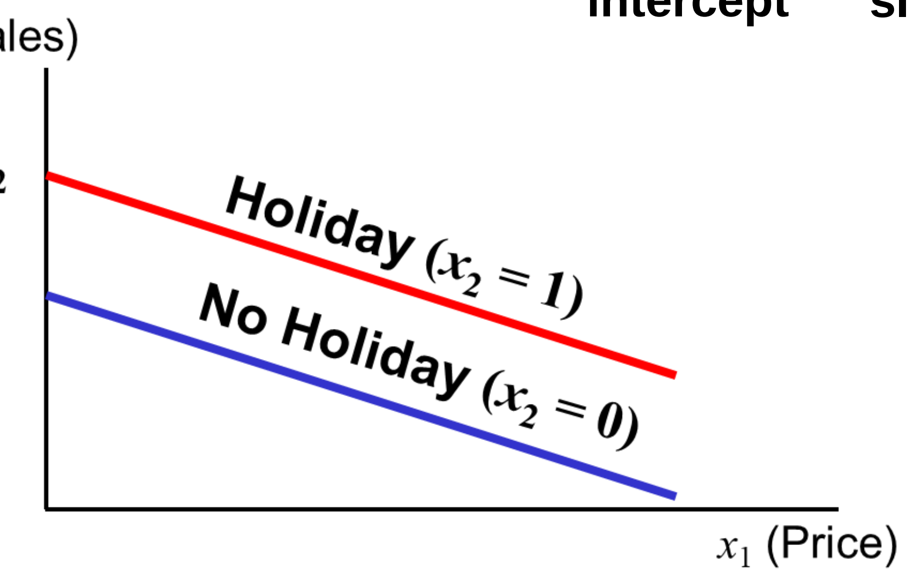


Dummy Variable Example (2 of 2)

$$\hat{y} = b_0 + b_1 x_1 + b_2(1) = (b_0 + b_2) + b_1 x_1 \quad \text{Holiday}$$

$$\hat{y} = b_0 + b_1 x_1 + b_2(0) = b_0 + b_1 x_1 \quad \text{No Holiday}$$

Different intercept — **Same slope**



If $H_0: \beta_2 = 0$ is rejected, then "Holiday" has a significant effect on pie sales

Interpreting the Dummy Variable Coefficient

Example: Sales = 300 − 30(Price) + 15(Holiday)

Sales: number of pies sold per week

Price: pie price in $

$$\text{Holiday}: \begin{cases} 1 \text{ If a holiday occurred during the week} \\ 0 \text{ If no holiday occurred} \end{cases}$$

$b_2 = 15$: on average, sales were 15 pies greater in weeks with a holiday than in weeks without a holiday, given the same price

Differences in Slope

- Hypothesizes interaction between pairs of *x* variables
  - Response to one *x* variable may vary at different levels of another *x* variable

- Contains two-way cross product terms
  - $$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$$
    $$= b_0 + b_1 x_1 + b_2 x_2 + b_3 \left(x_1 x_2 \right)$$

Effect of Interaction

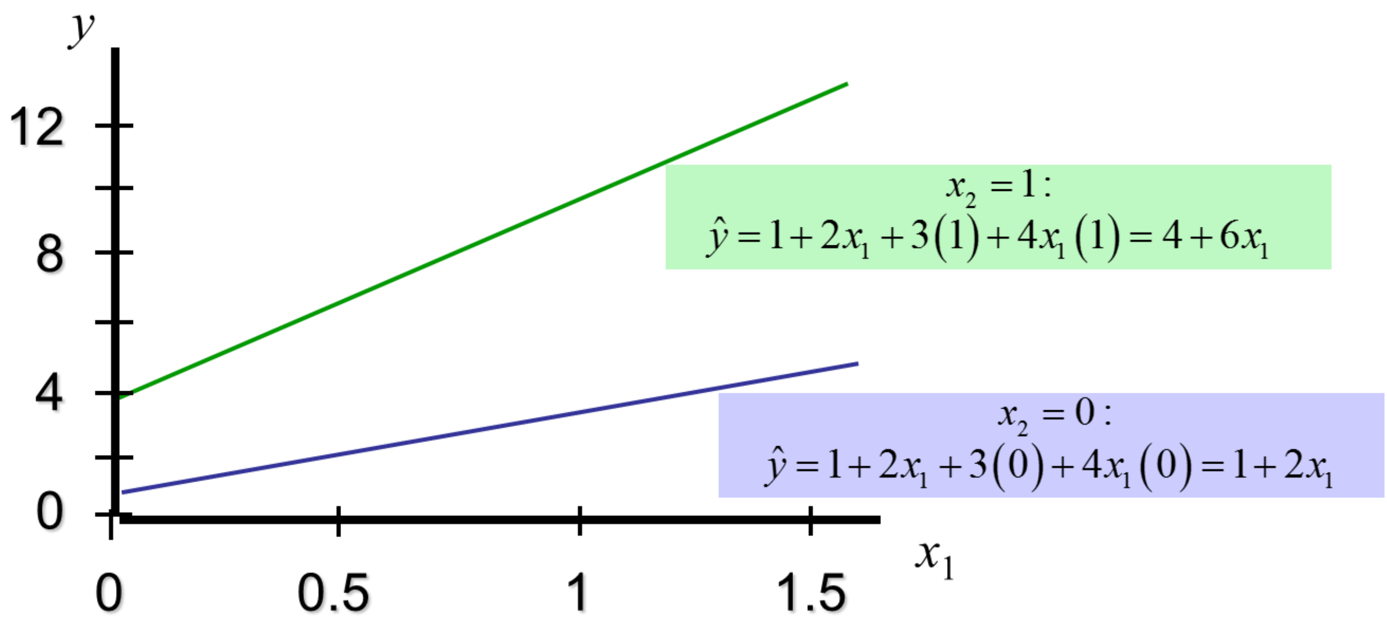
- Given:

$$Y = \beta_0 + \beta_2 X_2 + \left(\beta_1 + \beta_3 X_2\right) X_1$$
$$= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

- Without interaction term, effect of X_1 on *Y* is measured by β_1
- With interaction term, effect of X_1 on *Y* is measured by $\beta_1 + \beta_3 X_2$
- Effect changes as X_2 changes

Interaction Example

Suppose x_2 is a dummy variable and the estimated regression equation is $\hat{y} = 1 + 2x_1 + 3x_2 + 4x_1x_2$



Slopes are different if the effect of x_1 on *y* depends on x_2 value

Significance of Interaction Term

- The coefficient b_3 is an estimate of the difference in the coefficient of x_1 when $x_2 = 1$ compared to when $x_2 = 0$

- The *t* statistic for b_3 can be used to test the hypothesis

$$H_0: \beta_3 = 0 \mid \beta_1 \neq 0,\ \beta_2 \neq 0$$

$$H_1: \beta_3 \neq 0 \mid \beta_1 \neq 0,\ \beta_2 \neq 0$$

- If we reject the null hypothesis we conclude that there is a difference in the slope coefficient for the two subgroups

Section 12.9 Multiple Regression Analysis Application Procedure

Errors (residuals) from the regression model:

$$e_i = (y_i - \hat{y}_i)$$

Assumptions:

- The errors are normally distributed
- Errors have a constant variance
- The model errors are independent

Analysis of Residuals

- These residual plots are used in multiple regression:
  - Residuals vs. $\hat{y}_i$
  - Residuals vs. x_{1i}
  - Residuals vs. x_{2i}
  - Residuals vs. time (if time series data)

Use the residual plots to check for violations of regression assumptions