Simple Linear Regression

Objectives

In this chapter, you learn:

- How to use regression analysis to predict the value of a dependent variable based on a value of an independent variable

- To understand the meaning of the regression coefficients b_0 and b_1

- To evaluate the assumptions of regression analysis and know what to do if the assumptions are violated

- To make inferences about the slope and correlation coefficient

- To estimate mean values and predict individual values
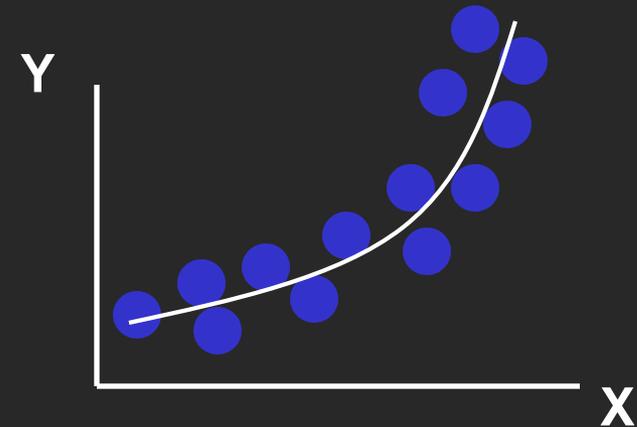
Correlation vs. Regression

- A scatter plot can be used to show the relationship between two variables

- Correlation analysis is used to measure the strength of the association (linear relationship) between two variables

  - Correlation is only concerned with strength of the relationship

  - No causal effect is implied with correlation

  - Scatter plots were first presented in Ch. 2

  - Correlation was first presented in Ch. 3

Types of Relationships

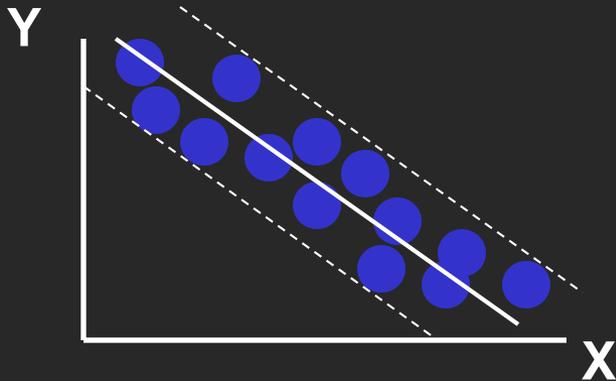**Linear relationships**
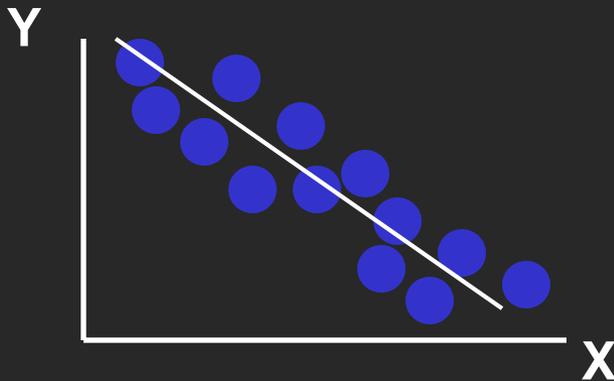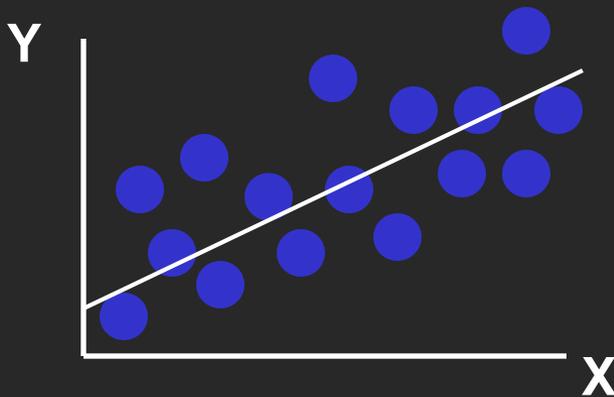
Curvilinear relationships
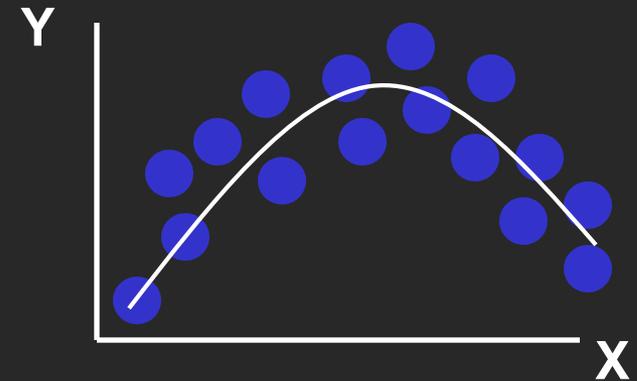
Types of Relationships

Strong relationships
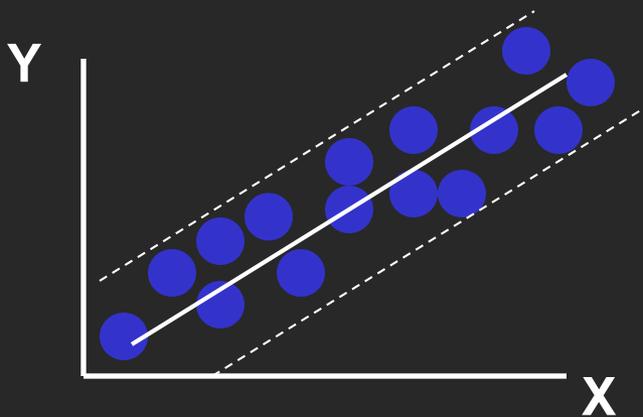
Weak relationships
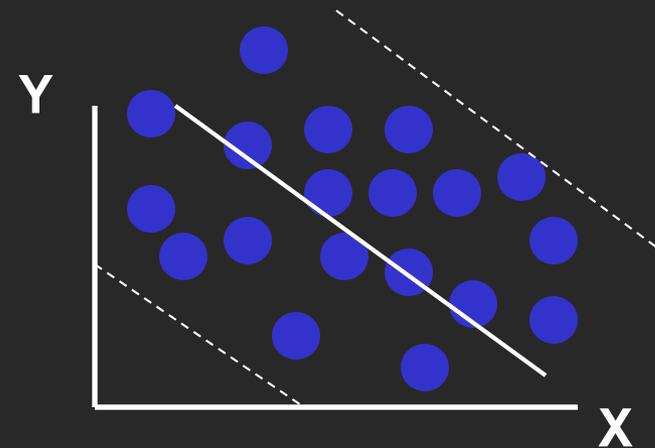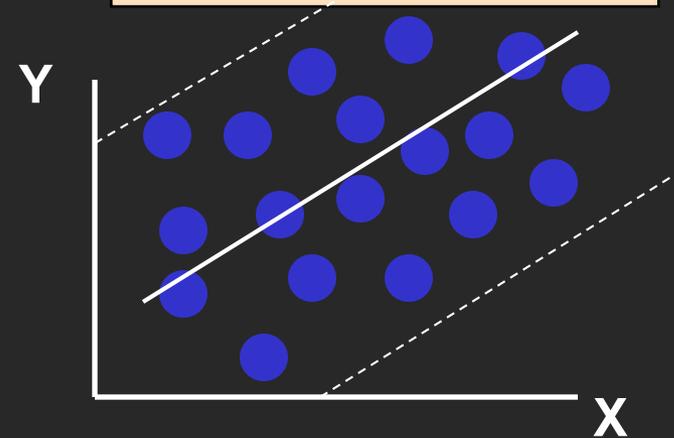
Types of Relationships

No relationship

PEARSON
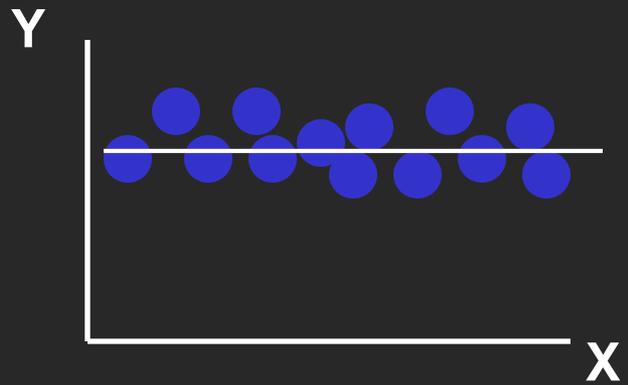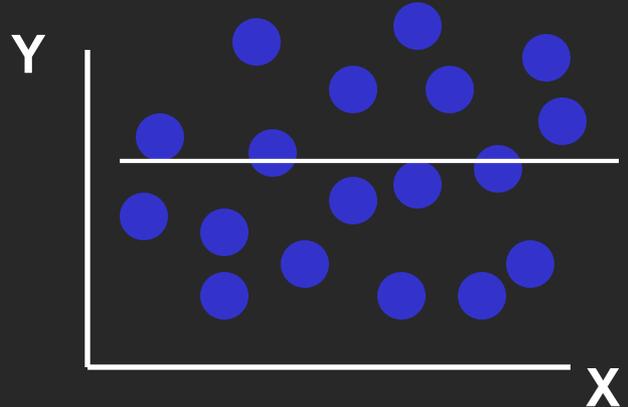
Introduction to Regression Analysis

- **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 predict or explain

- **Independent variable:** the variable used to predict or explain the dependent variable

PEARSON

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 related to changes in X

Simple Linear 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

Simple Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Observed Value of Y for X_i

Predicted Value of Y for X_i

ε_i

Slope = β_1

Random Error for this X_i value

Intercept = β_0

X_i

Y

X
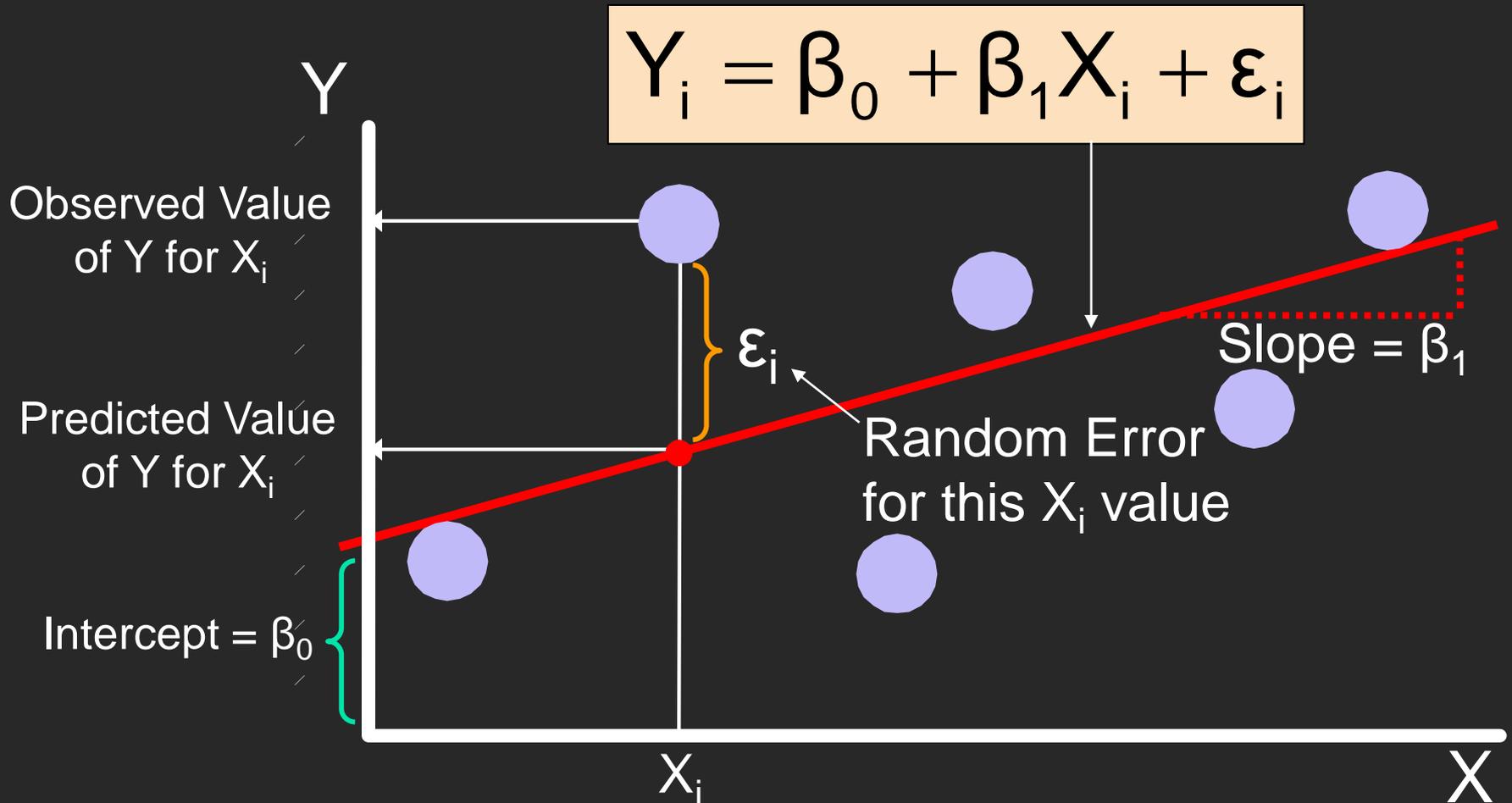
Simple Linear Regression Equation (Prediction Line)

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

$$\hat{Y}_i = b_0 + b_1 X_i$$

PEARSON

The Least Squares Method

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

$$\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (b_0 + b_1 X_i))^2$$

Finding the Least Squares Equation

- The coefficients b_0 and b_1, and other regression results in this chapter, will be found using Excel or Minitab

> Formulas are shown in the text for those who are interested

Interpretation of the Slope and the Intercept

- b_0 is the estimated mean value of Y when the value of X is zero

- b_1 is the estimated change in the mean value of Y as a result of a one-unit increase 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

Simple Linear Regression Example: Data

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

Simple Linear Regression Example: Scatter Plot

House price model: Scatter Plot
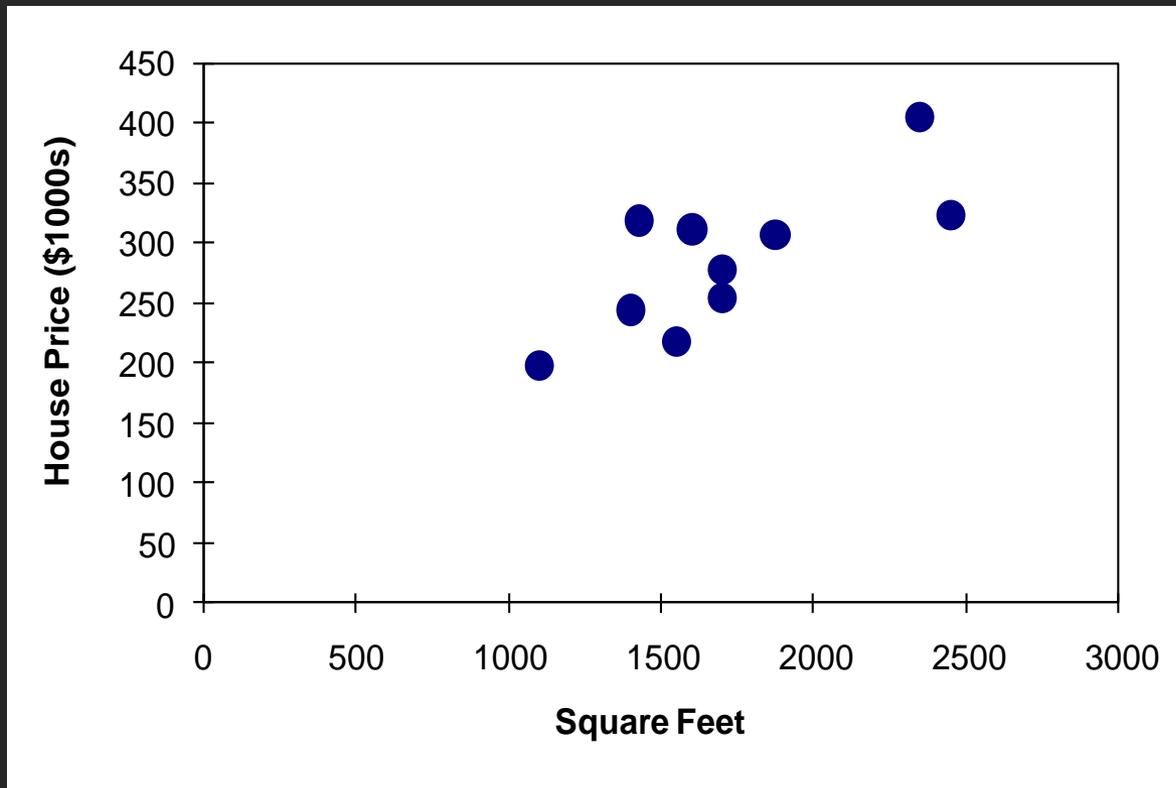
Simple Linear Regression Example: Using Excel Data Analysis Function

1. Choose Data

2. Choose Data Analysis

3. Choose Regression

Simple Linear Regression Example: Using Excel Data Analysis Function

(continued)

Enter Y range and X range and desired options

DCOVA

Simple Linear Regression Example: Using PHStat

Add-Ins: PHStat: Regression: Simple Linear Regression

Simple Linear Regression Example: Excel Output

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

Simple Linear Regression Example: Minitab Output

The regression equation is:

The regression equation is

Price = 98.2 + 0.110 Square Feet

Predictor	Coef	SE Coef	T	P
Constant	98.25	58.03	1.69	0.129
Square Feet	0.10977	0.03297	3.33	0.010

S = 41.3303   R-Sq = 58.1%   R-Sq(adj) = 52.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	18935	18935	11.08	0.010
Residual Error	8	13666	1708		
Total	9	32600			

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

Simple Linear Regression Example: Graphical Representation

House price model: Scatter Plot and Prediction Line
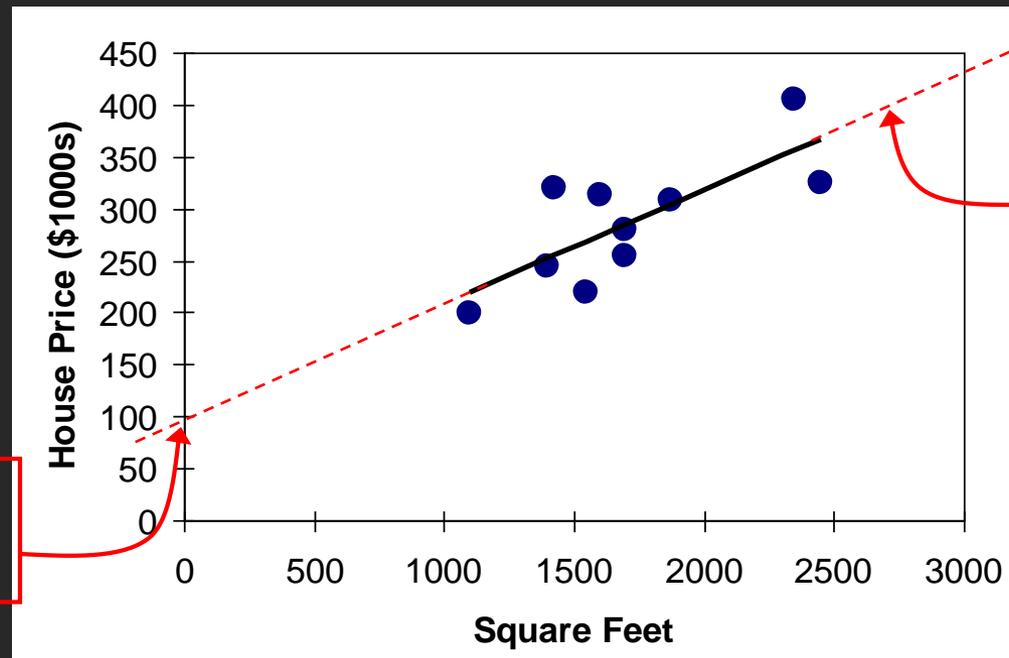
Slope = 0.10977

Intercept = 98.248

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

Simple Linear Regression Example: Interpretation of b_o

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

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

- Because a house cannot have a square footage of 0, b_0 has no practical application

Simple Linear Regression Example: Interpreting b_1

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

- b_1 estimates the change in the mean value of Y as a result of a one-unit increase in X

  - Here, $\boxed{b_1 = 0.10977}$ tells us that the mean value of a house increases by .10977($1000) = $109.77, on average, for each additional one square foot of size

Simple Linear Regression Example: Making Predictions

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($1,000s) = $317,850

Simple Linear Regression Example: Making Predictions

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

Relevant range for interpolation
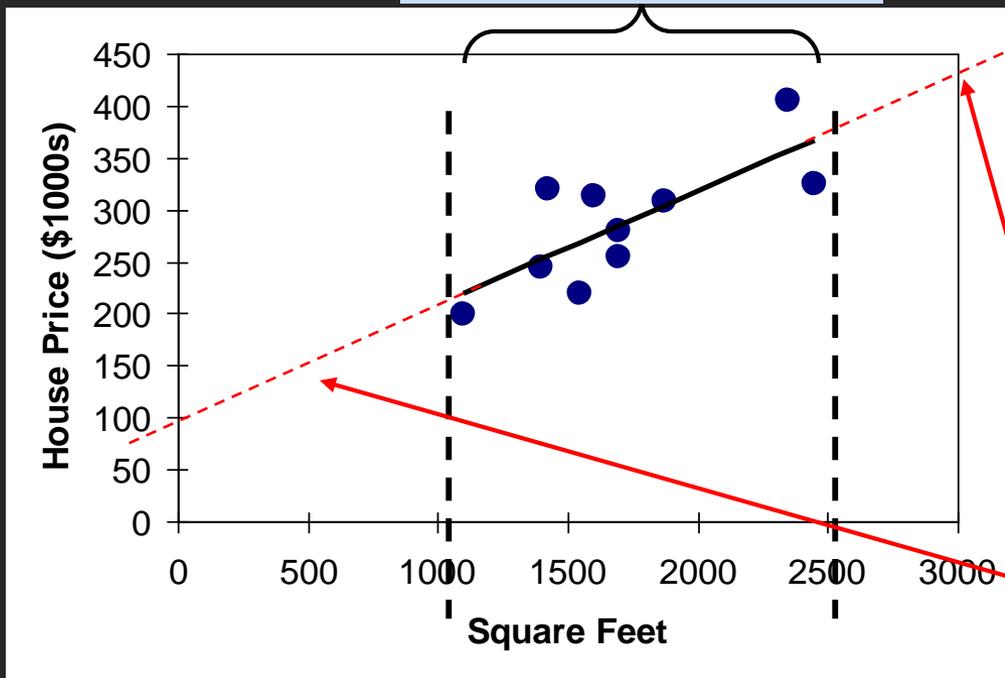
Do not try to extrapolate beyond the range of observed X's

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

$$SST = \sum(Y_i - \overline{Y})^2 \qquad SSR = \sum(\hat{Y}_i - \overline{Y})^2 \qquad SSE = \sum(Y_i - \hat{Y}_i)^2$$

where:

$\overline{Y}$ = Mean value of the dependent variable

Y_i = Observed value of the dependent variable

$\hat{Y}_i$ = Predicted value of Y for the given X_i value

Measures of Variation

DCOV<u>A</u>

- SST = total sum of squares   (Total Variation)
  - Measures the variation of the Y_i values around their mean $\bar{Y}$

- SSR = regression sum of squares (Explained Variation)
  - Variation attributable to the relationship between X and Y

- SSE = error sum of squares   (Unexplained Variation)
  - Variation in Y attributable to factors other than X

PEARSON

Measures of Variation

(continued)

$$SSE = \sum(Y_i - \hat{Y}_i)^2$$

$$SST = \sum(Y_i - \bar{Y})^2$$

$$SSR = \sum(\hat{Y}_i - \bar{Y})^2$$
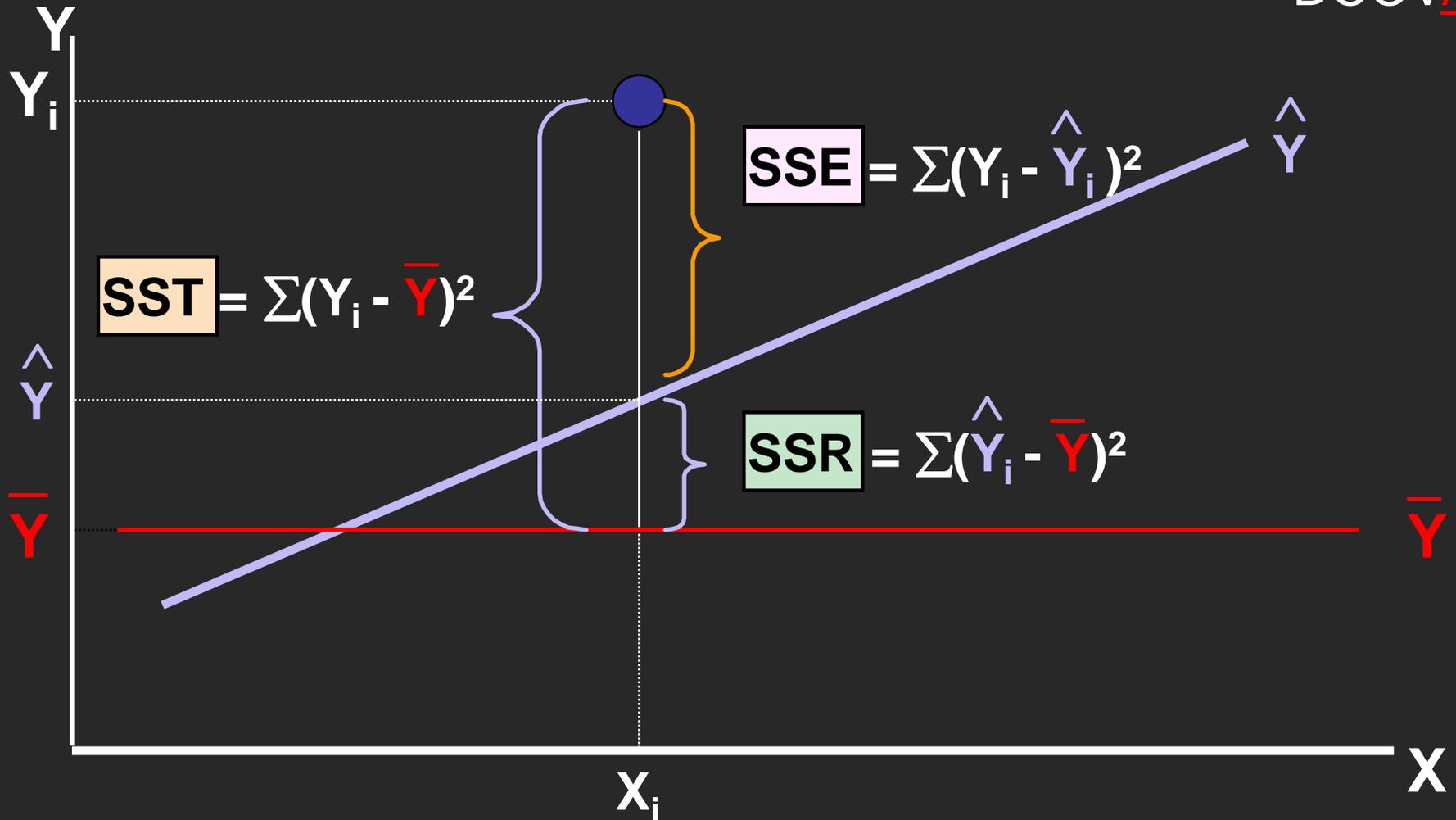
Coefficient of Determination, r²

- 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-square and is denoted as r²

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

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

Examples of Approximate r^2 Values

Y

$r^2 = 1$

X

Y

$r^2 = 1$

X
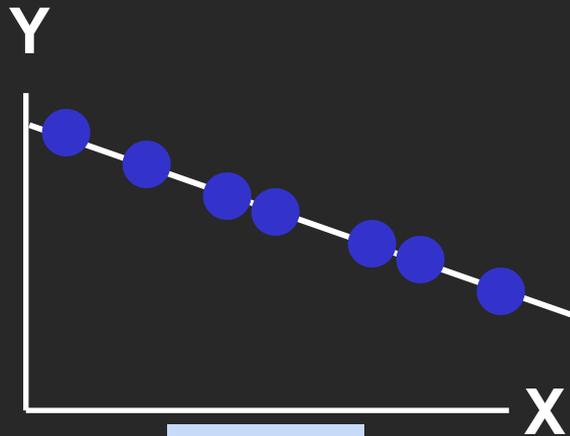
Perfect linear relationship between X and Y:
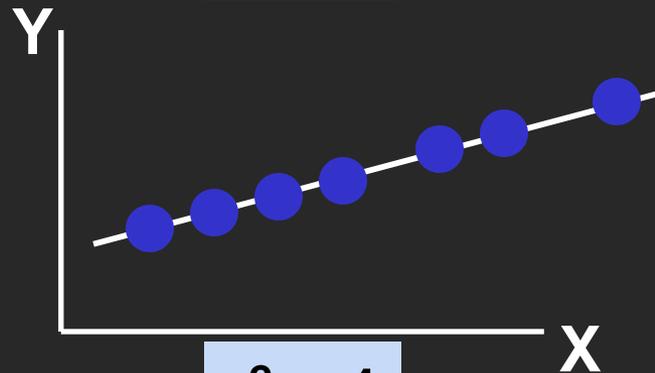
100% of the variation in Y is explained by variation in X

Examples of Approximate r^2 Values
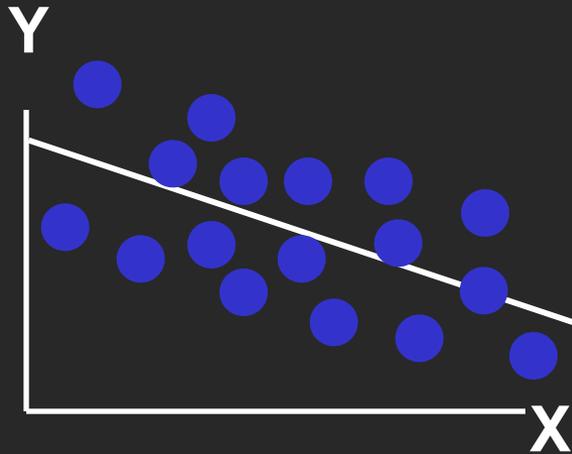
$$0 < r^2 < 1$$

Weaker linear relationships between X and Y:
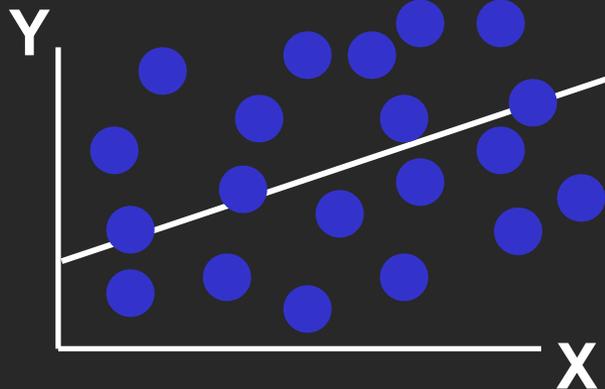
Some but not all of the variation in Y is explained by variation in X

ALWAYS LEARNING

Examples of Approximate r^2 Values

$r^2 = 0$

Y

X

$r^2 = 0$
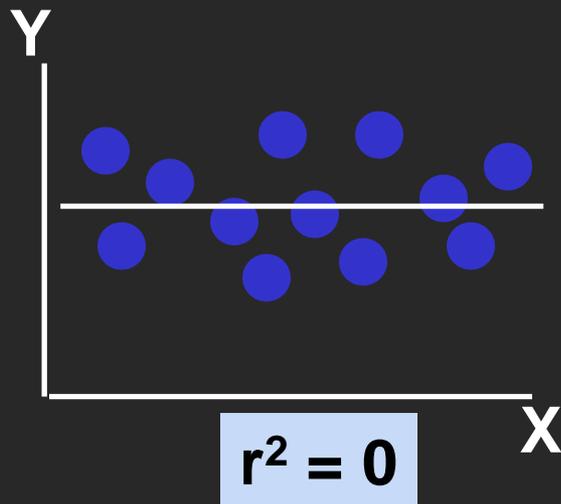
$r^2 = 0$

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)

Simple Linear Regression Example: Coefficient of Determination, r² in Excel

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

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

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

Simple Linear Regression Example: Coefficient of Determination, r² in Minitab

The regression equation is

Price = 98.2 + 0.110 Square Feet

Predictor	Coef	SE Coef	T	P
Constant	98.25	58.03	1.69	0.129
Square Feet	0.10977	0.03297	3.33	0.010

S = 41.3303   R-Sq = 58.1%   R-Sq(adj) = 52.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	18935	18935	11.08	0.010
Residual Error	8	13666	1708		
Total	9	32600			

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

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

Standard Error of Estimate

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

$$S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}}$$

Where

SSE = error sum of squares
n = sample size

Simple Linear Regression Example: Standard Error of Estimate in Excel

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

$$S_{YX} = 41.33032$$

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

Simple Linear Regression Example: Standard Error of Estimate in Minitab

The regression equation is

Price = 98.2 + 0.110 Square Feet

Predictor	Coef	SE Coef	T	P
Constant	98.25	58.03	1.69	0.129
Square Feet	0.10977	0.03297	3.33	0.010

S = 41.3303   R-Sq = 58.1%   R-Sq(adj) = 52.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	18935	18935	11.08	0.010
Residual Error	8	13666	1708		
Total	9	32600			

$$S_{YX} = 41.33032$$

Comparing Standard Errors

S_{YX} is a measure of the variation of observed Y values from the regression line

small S_{YX}

large S_{YX}
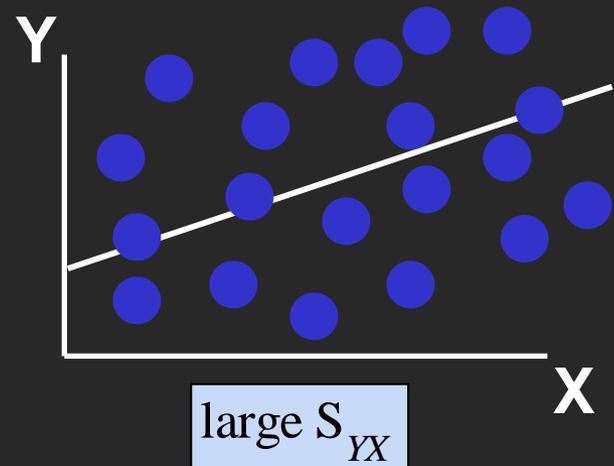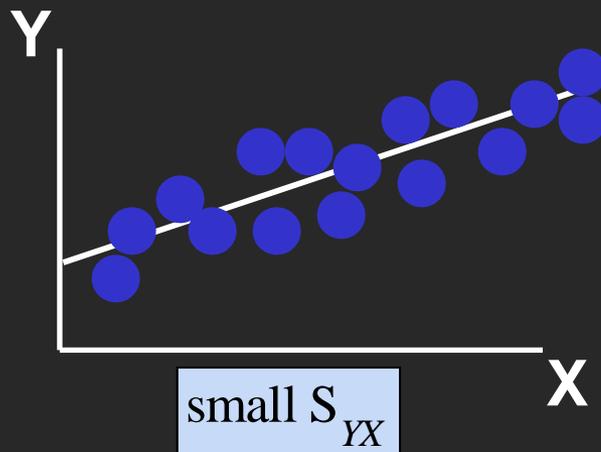
The magnitude of S_{YX} should always be judged relative to the size of the Y values in the sample data

i.e., S_{YX} = $41.33K is moderately small relative to house prices in the $200K - $400K range

Assumptions of Regression L.I.N.E

- Linearity
  - The relationship between X and Y is linear
- Independence of Errors
  - Error values are statistically independent
  - Particularly important when data are collected over a period of time
- Normality of Error
  - Error values are normally distributed for any given value of X
- Equal Variance (also called homoscedasticity)
  - The probability distribution of the errors has constant variance

Residual Analysis

$$e_i = Y_i - \hat{Y}_i$$

- The residual for observation i, e_i, is the difference between its observed and predicted value

- Check the assumptions of regression by examining the residuals
  - Examine for linearity assumption
  - Evaluate independence assumption
  - Evaluate normal distribution assumption
  - Examine for constant variance for all levels of X (homoscedasticity)

- Graphical Analysis of Residuals
  - Can plot residuals vs. X

PEARSON

Residual Analysis for Linearity
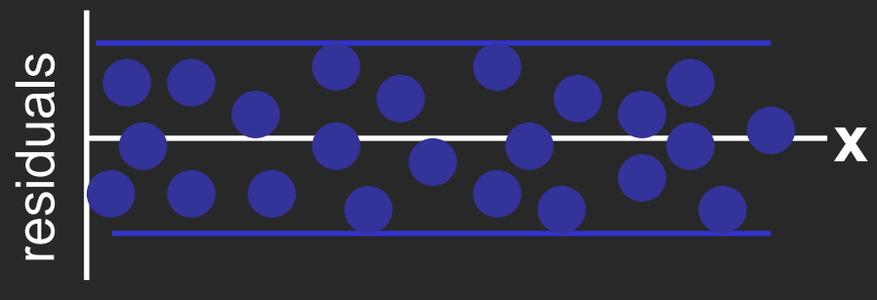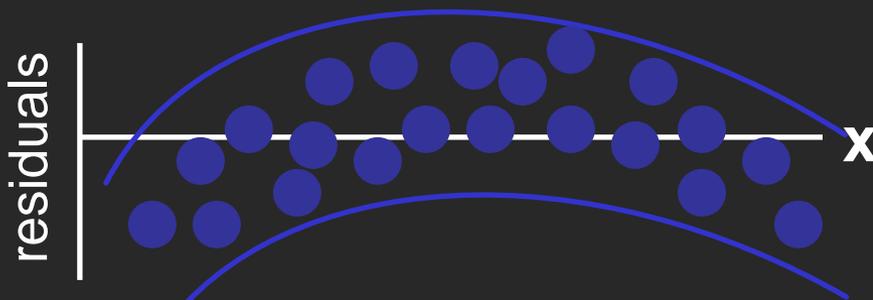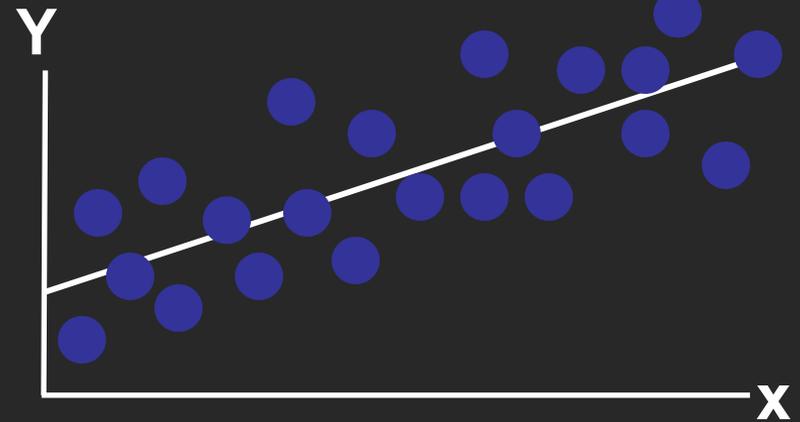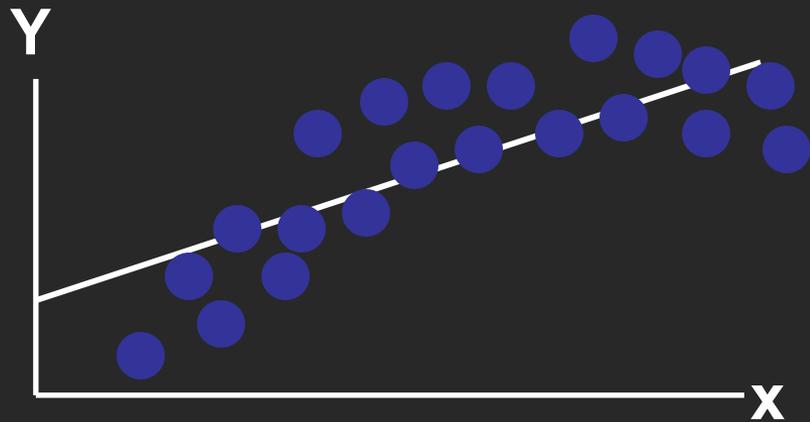
Not Linear

Linear

Residual Analysis for Independence

Cyclial Pattern:

Not Independent
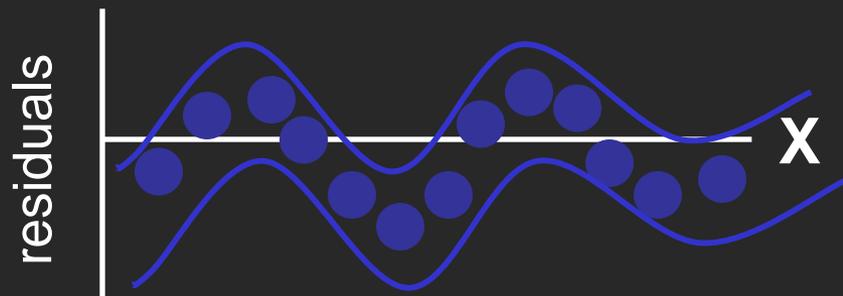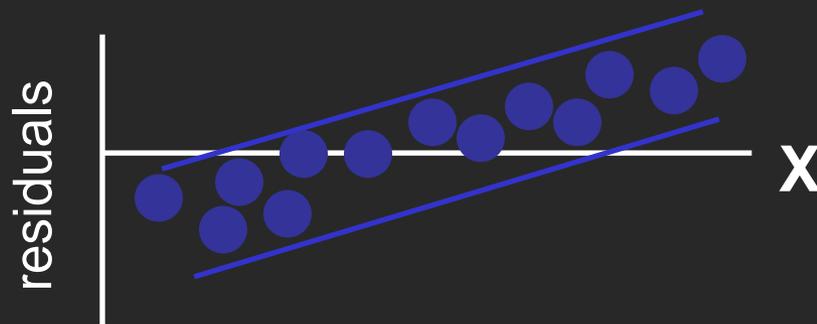
✓ **No Cyclial Pattern Independent**

residuals

X

residuals

X

residuals

X
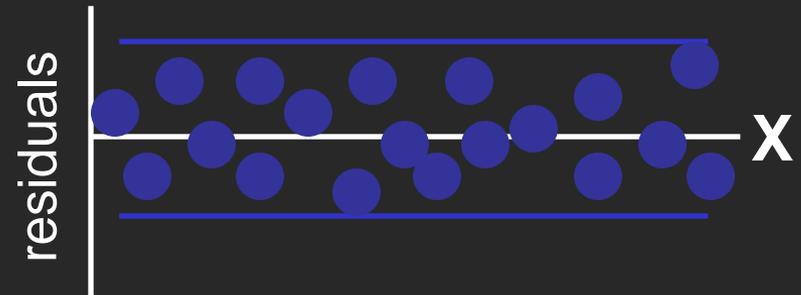
Checking for Normality

- Examine the Stem-and-Leaf Display of the Residuals

- Examine the Boxplot of the Residuals

- Examine the Histogram of the Residuals

- Construct a Normal Probability Plot of the Residuals

ALWAYS LEARNING

PEARSON

Residual Analysis for Normality

When using a normal probability plot, normal errors will approximately display in a straight line
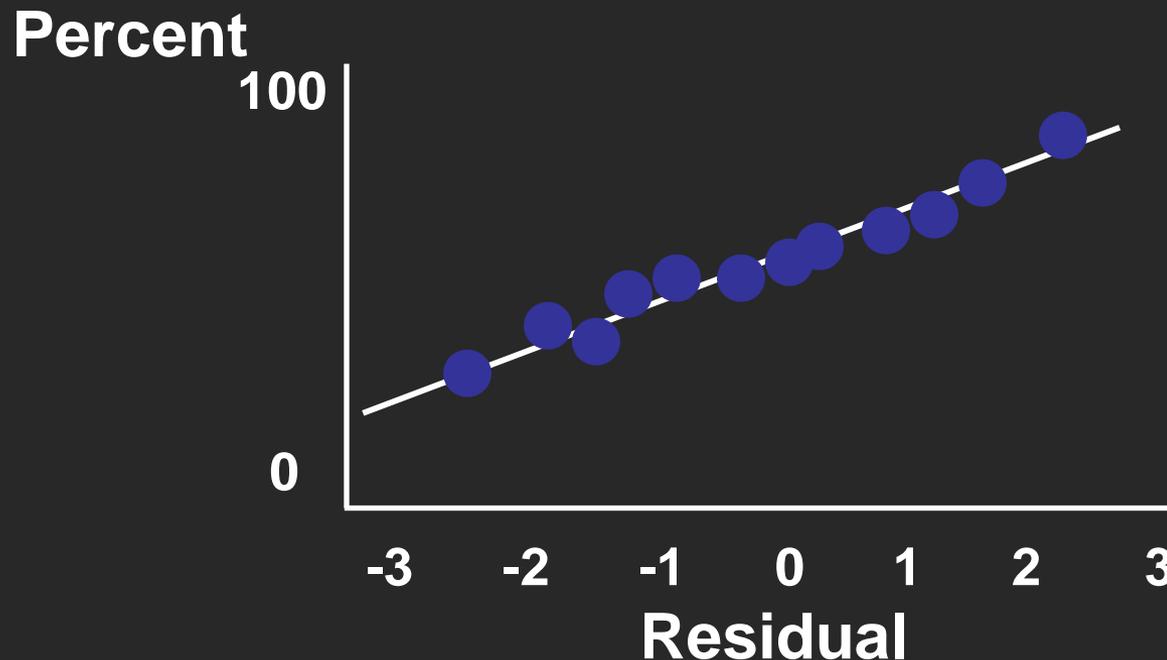
Residual Analysis for Equal Variance
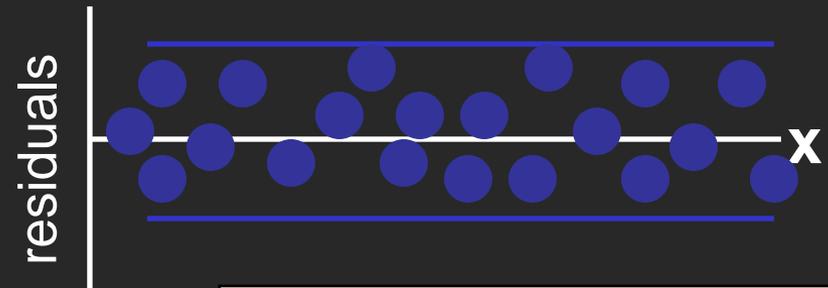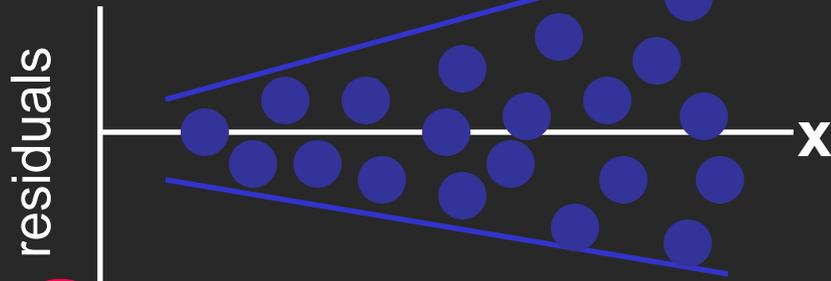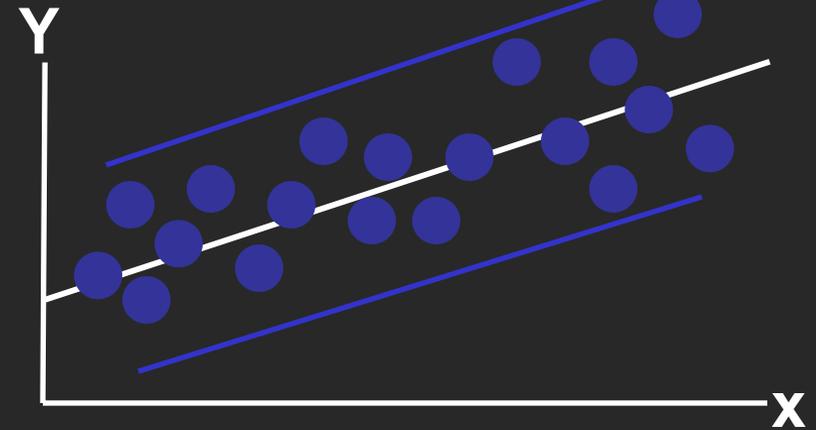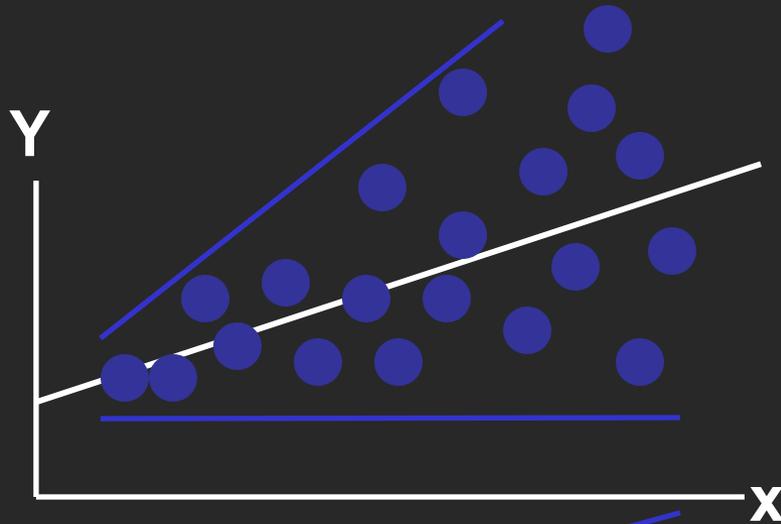
Non-constant variance

Constant variance

Simple Linear Regression Example: Excel Residual Output

RESIDUAL OUTPUT		
	Predicted House Price	*Residuals*
1	251.92316	-6.923162
2	273.87671	38.12329
3	284.85348	-5.853484
4	304.06284	3.937162
5	218.99284	-19.99284
6	268.38832	-49.38832
7	356.20251	48.79749
8	367.17929	-43.17929
9	254.6674	64.33264
10	284.85348	-29.85348

House Price Model Residual Plot
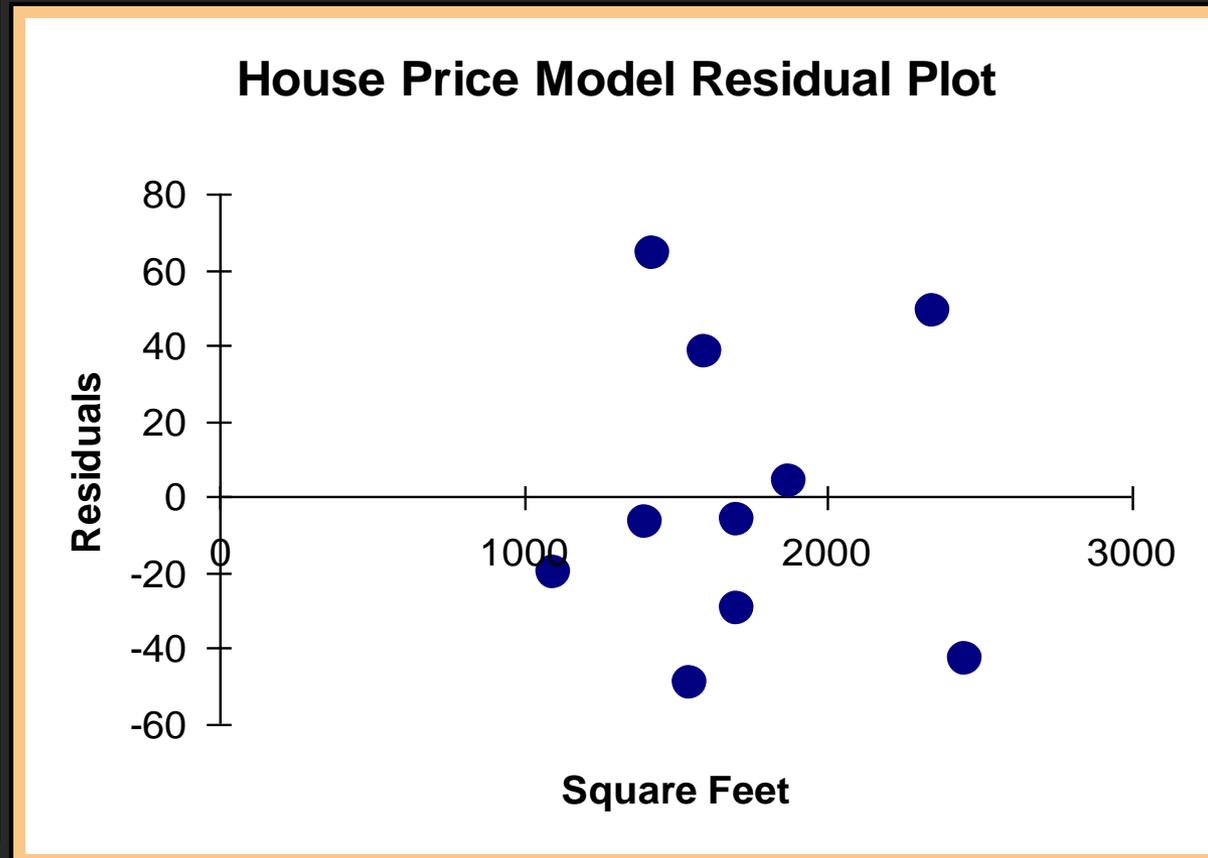
Does not appear to violate any regression assumptions

Simple Linear Regression Example: Minitab Residual Output

Residual Plots for House Price (Y)
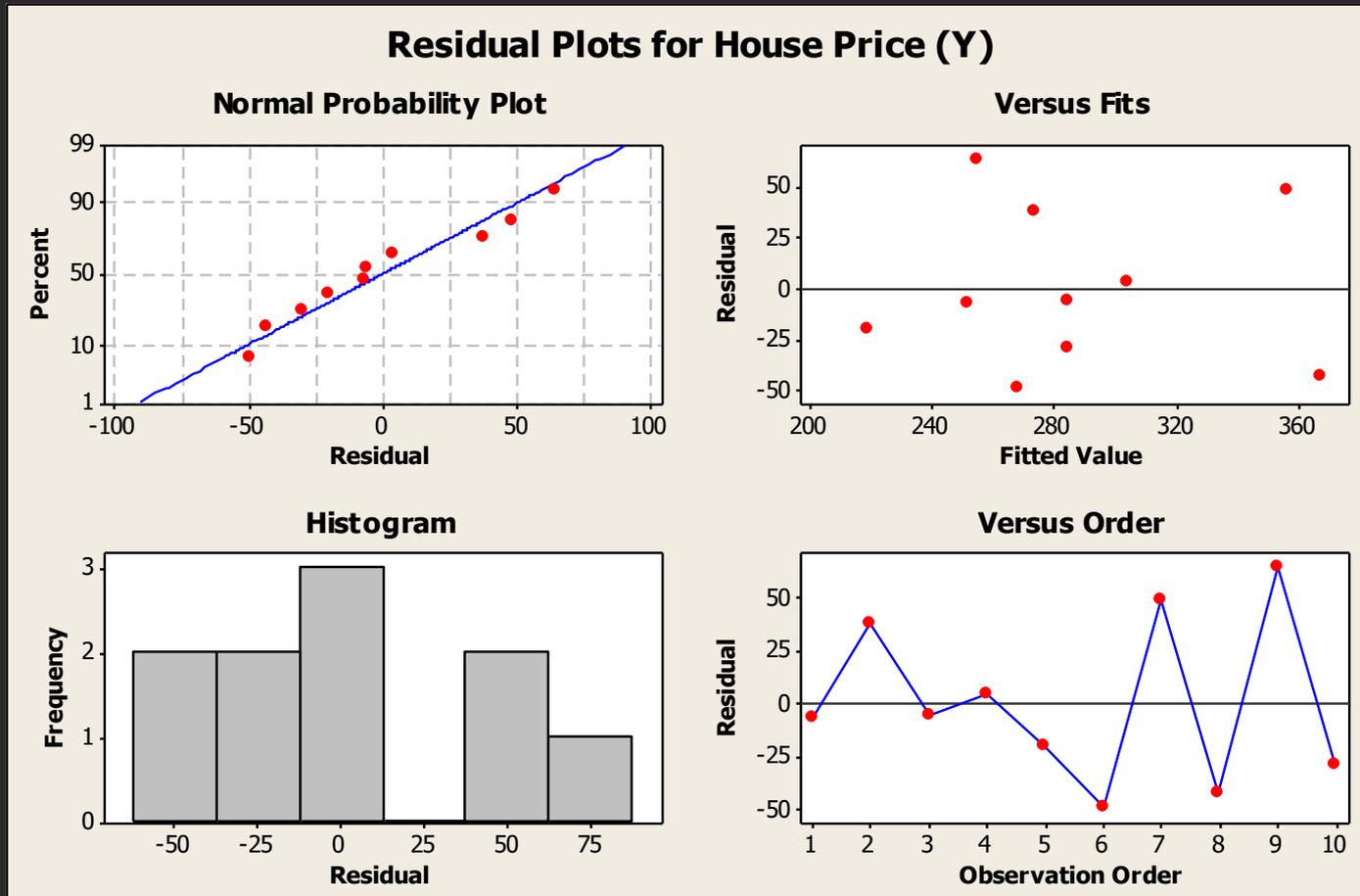
Does not appear to violate any regression assumptions

Measuring Autocorrelation: The Durbin-Watson Statistic

- Used when data are collected over time to detect if autocorrelation is present

- Autocorrelation exists if residuals in one time period are related to residuals in another period

Autocorrelation

- **Autocorrelation is correlation of the errors (residuals) over time**

- Here, residuals show a cyclic pattern, not random. Cyclical patterns are a sign of positive autocorrelation

Time (t) Residual Plot

Residuals vs. Time (t)
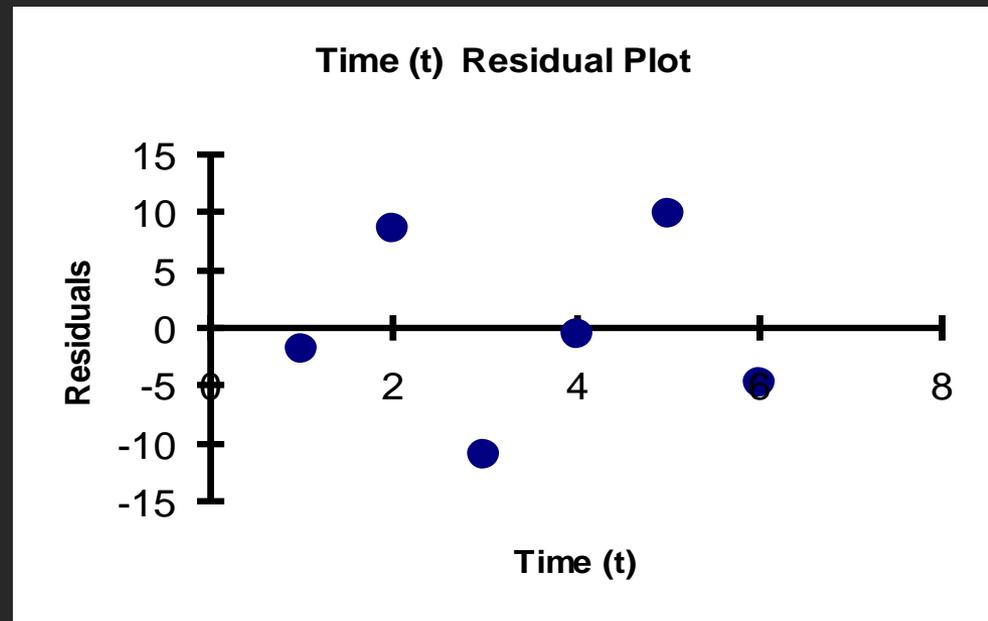
- Violates the regression assumption that residuals are random and independent

The Durbin-Watson Statistic

- The Durbin-Watson statistic is used to test for autocorrelation

H_0: positive autocorrelation doe not exist
H_1: positive autocorrelation is present

$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n}e_i^2}$$

- The possible range is $0 \le D \le 4$

- D should be close to 2 if H_0 is true

- D less than 2 may signal positive autocorrelation, D greater than 2 may signal negative autocorrelation

Testing for Positive Autocorrelation

H_0: positive autocorrelation does not exist

H_1: positive autocorrelation is present

- Calculate the Durbin-Watson test statistic = D
  (The Durbin-Watson Statistic can be found using Excel or Minitab)

- Find the values d_L and d_U from the Durbin-Watson table
  (for sample size **n** and number of independent variables **k**)

Decision rule: reject H_0 if $D < d_L$

| Reject H_0 | Inconclusive | Do not reject H_0 |

0   d_L d_U 2
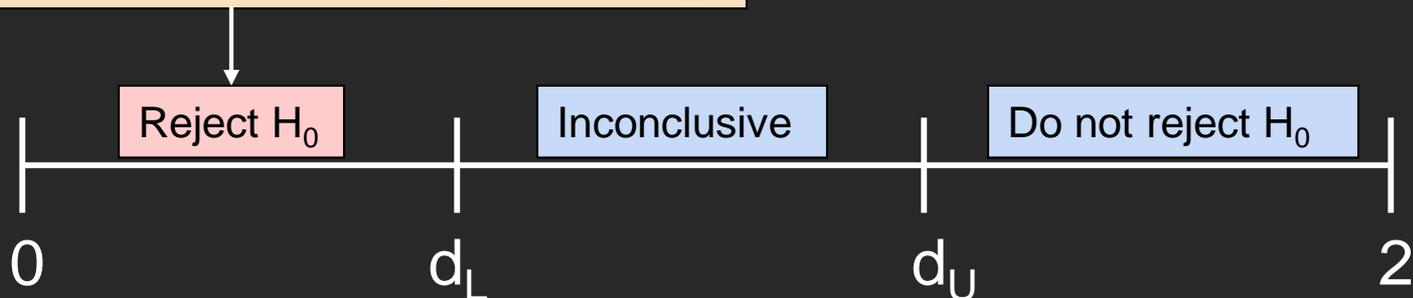
Testing for Positive Autocorrelation

(continued)

- Suppose we have the following time series data:

$$y = 30.65 + 4.7038x$$
$$R^2 = 0.8976$$
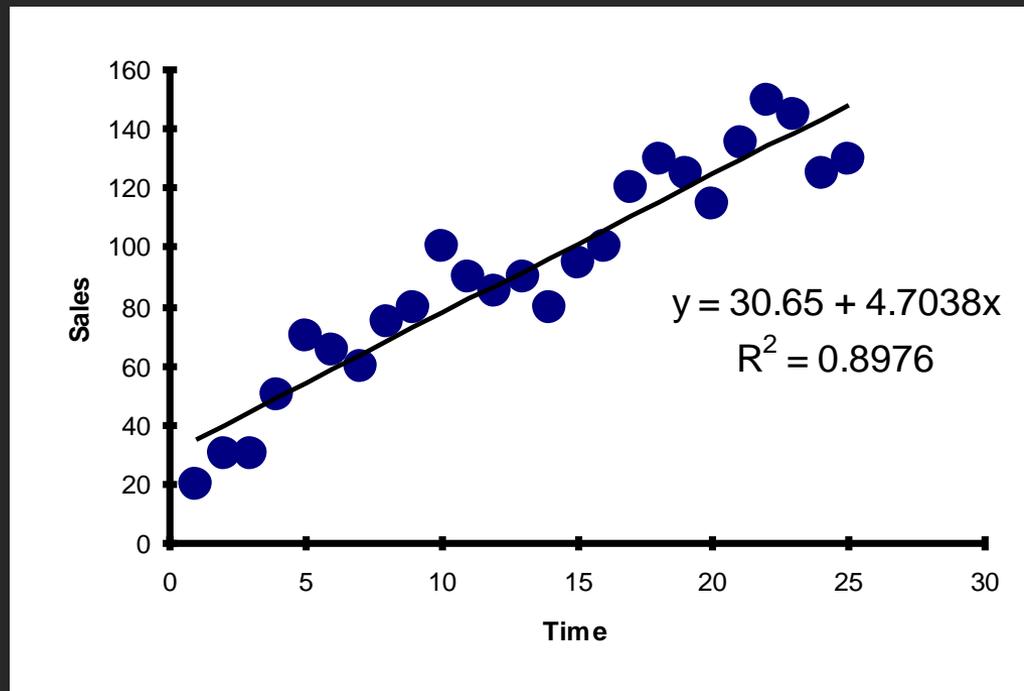
- Is there autocorrelation?

Testing for Positive Autocorrelation

DCOV**A**

- Example with n = 25:

Excel/PHStat output:

Durbin-Watson Calculations	
Sum of Squared Difference of Residuals	3296.18
Sum of Squared Residuals	3279.98
Durbin-Watson Statistic	**1.00494**
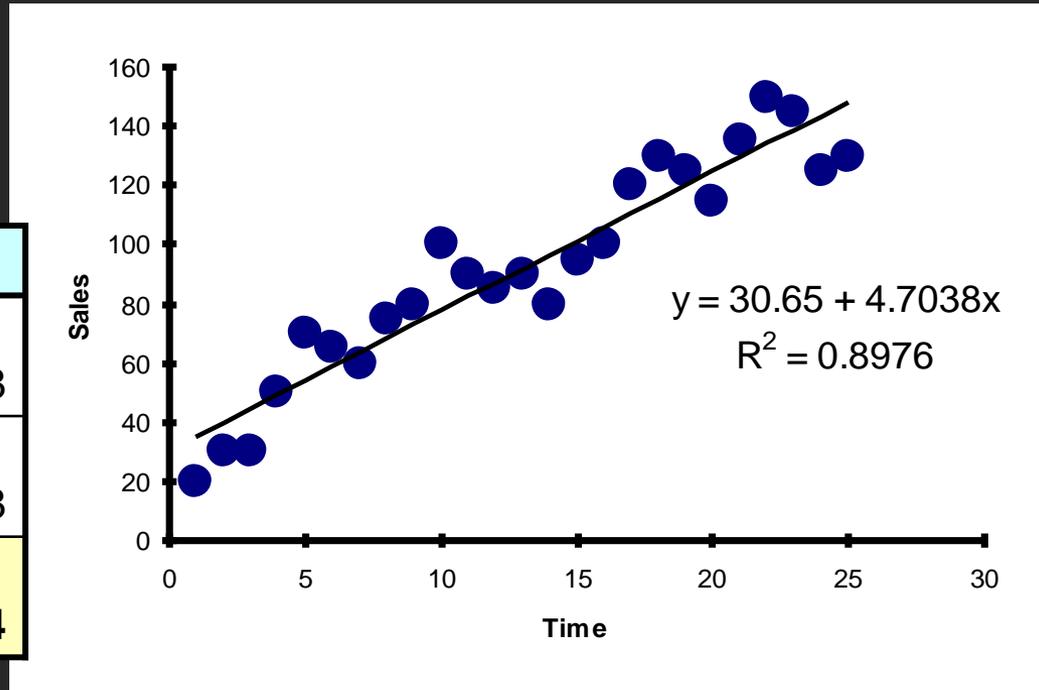
$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} = \frac{3296.18}{3279.98} = 1.00494$$

$y = 30.65 + 4.7038x$

$R^2 = 0.8976$

Testing for Positive Autocorrelation

- Here, n = 25 and there is k = 1 one independent variable

- Using the Durbin-Watson table, $d_L = 1.29$ and $d_U = 1.45$

- $D = 1.00494 < d_L = 1.29$, so reject H_0 and conclude that significant positive autocorrelation exists

Decision: **reject H_0** since

$D = 1.00494 < d_L$

Reject H_0 Inconclusive   Do not reject H_0

0      $d_L=1.29$ $d_U=1.45$ 2
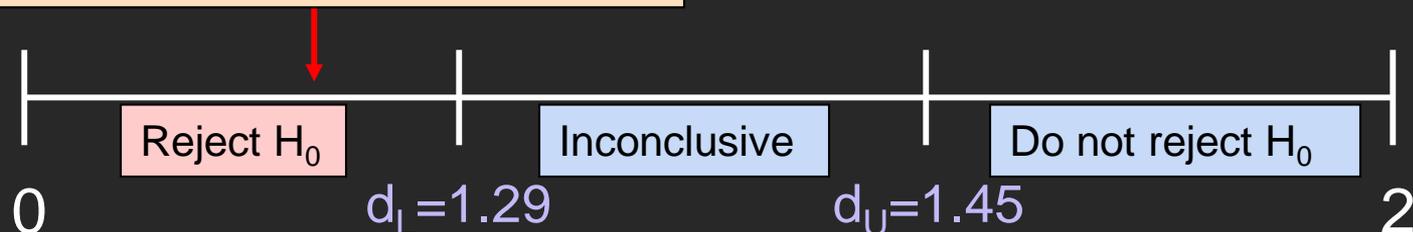
Inferences About the Slope

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

$$S_{b_1} = \frac{S_{YX}}{\sqrt{SSX}} = \frac{S_{YX}}{\sqrt{\sum(X_i - \overline{X})^2}}$$

where:

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

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

Inferences 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

$$t_{STAT} = \frac{b_1 - \beta_1}{S_{b_1}}$$

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

where:

b_1 = regression slope coefficient

β_1 = hypothesized slope

S_{b1} = standard error of the slope

Inferences About the Slope:
t Test Example

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:

house price = 98.25 + 0.1098 (sq. ft.)

The slope of this model is 0.1098

Is there a relationship between the square footage of the house and its sales price?

Inferences About the Slope: t Test Example

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

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 S_{b_1}

From Minitab output:

```
Predictor      Coef      SE Coef      T      P
Constant       98.25     58.03        1.69   0.129
Square Feet    0.10977   0.03297      3.33   0.010
```

b_1 S_{b_1}

$$t_{STAT} = \frac{b_1 - \beta_1}{S_{b_1}} = \frac{0.10977 - 0}{0.03297} = 3.32938$$

Inferences About the Slope: t Test Example

Test Statistic: $t_{STAT} = 3.329$

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

d.f. = 10 - 2 = 8

$\alpha/2 = .025$ $\alpha/2 = .025$

Reject H_0 $-t_{\alpha/2}$ Do not reject H_0 $t_{\alpha/2}$ Reject H_0

0

-2.3060   2.3060   3.329
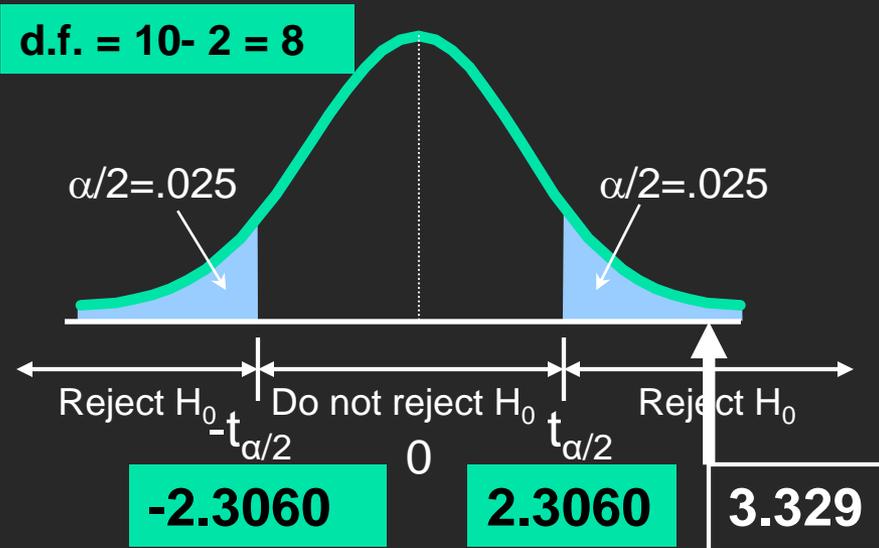
Decision: Reject H_0

There is sufficient evidence that square footage affects house price

Inferences About the Slope: t Test Example

$H_0: \beta_1 = 0$

From Excel output: $H_1: \beta_1 \neq 0$

	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

From Minitab output:

```
Predictor      Coef      SE Coef     T      P
Constant       98.25     58.03       1.69   0.129
Square Feet    0.10977   0.03297     3.33   0.010
```

p-value

Decision: Reject H_0, since p-value $< \alpha$

There is sufficient evidence that square footage affects house price.

F Test for Significance

■ F Test statistic:

$$F_{STAT} = \frac{MSR}{MSE}$$

where

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

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

where F_{STAT} 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)

F-Test for Significance Excel Output

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

$$F_{STAT} = \frac{MSR}{MSE} = \frac{18934.9348}{1708.1957} = 11.0848$$

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			

F-Test for Significance Minitab Output

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	18935	18935	11.08	0.010
Residual Error	8	13666	1708		
Total	9	32600			

p-value for the F-Test

With 1 and 8 degrees of freedom

$$F_{STAT} = \frac{MSR}{MSE} = \frac{18934.9348}{1708.1957} = 11.0848$$

F Test for Significance

(continued)

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

$\alpha = .05$

$df_1 = 1$ $df_2 = 8$

Critical Value:

$F_\alpha = 5.32$

$\alpha = .05$

0

Do not reject H_0

Reject H_0

F

$F_{.05} = 5.32$

Test Statistic:

$$F_{STAT} = \frac{MSR}{MSE} = 11.08$$

Decision:

Reject H_0 at $\alpha = 0.05$

Conclusion:

There is sufficient evidence that house size affects selling price
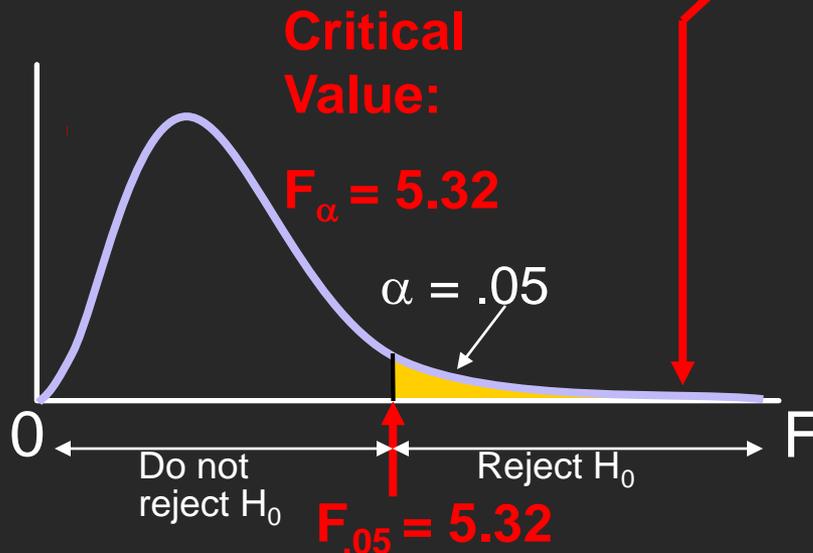
Confidence Interval Estimate for the Slope

Confidence Interval Estimate of the Slope:

$$b_1 \pm t_{\alpha/2} S_{b_1}$$

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

(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 $33.74 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

Confidence Interval Estimate for the Slope from Minitab

DCOVA

Minitab does not automatically calculate a confidence interval for the slope but provides the quantities necessary to use the confidence interval formula.

Predictor	Coef	SE Coef	T	P
Constant	98.25	58.03	1.69	0.129
Square Feet	0.10977	0.03297	3.33	0.010

$$b_1 \pm t_{\alpha/2} S_{b_1}$$

t Test for a Correlation Coefficient

- ## Hypotheses

$H_0: \rho = 0$ (no correlation between X and Y)

$H_1: \rho \neq 0$ (correlation exists)

- ## Test statistic

$$t_{STAT} = \frac{r - \rho}{\sqrt{\dfrac{1 - r^2}{n - 2}}}$$

(with n – 2 degrees of freedom)

where

$r = +\sqrt{r^2}$ if $b_1 > 0$

$r = -\sqrt{r^2}$ if $b_1 < 0$

t-test For A Correlation Coefficient *(continued)*

Is there evidence of a linear relationship between square feet and house price at the .05 level of significance?

H_0: $\rho = 0$ (No correlation)

H_1: $\rho \neq 0$ (correlation exists)

$\alpha = .05$, df = 10 - 2 = 8

$$t_{STAT} = \frac{r - \rho}{\sqrt{\dfrac{1 - r^2}{n - 2}}} = \frac{.762 - 0}{\sqrt{\dfrac{1 - .762^2}{10 - 2}}} = 3.329$$

t-test For A Correlation Coefficient *(continued)*

$$t_{STAT} = \frac{r - \rho}{\sqrt{\dfrac{1 - r^2}{n - 2}}} = \frac{.762 - 0}{\sqrt{\dfrac{1 - .762^2}{10 - 2}}} = 3.329$$

Decision: Reject H_0

Conclusion: There **is evidence** of a linear association at the 5% level of significance

d.f. = 10-2 = 8

$\alpha/2 = .025$ $\alpha/2 = .025$

Reject H_0 Do not reject H_0 Reject H_0

$-t_{\alpha/2}$ 0 $t_{\alpha/2}$

-2.3060   **2.3060**

3.329

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 mean of Y, given X_i

$\hat{Y} = b_0 + b_1 X_i$

Prediction Interval for an individual Y, given X_i
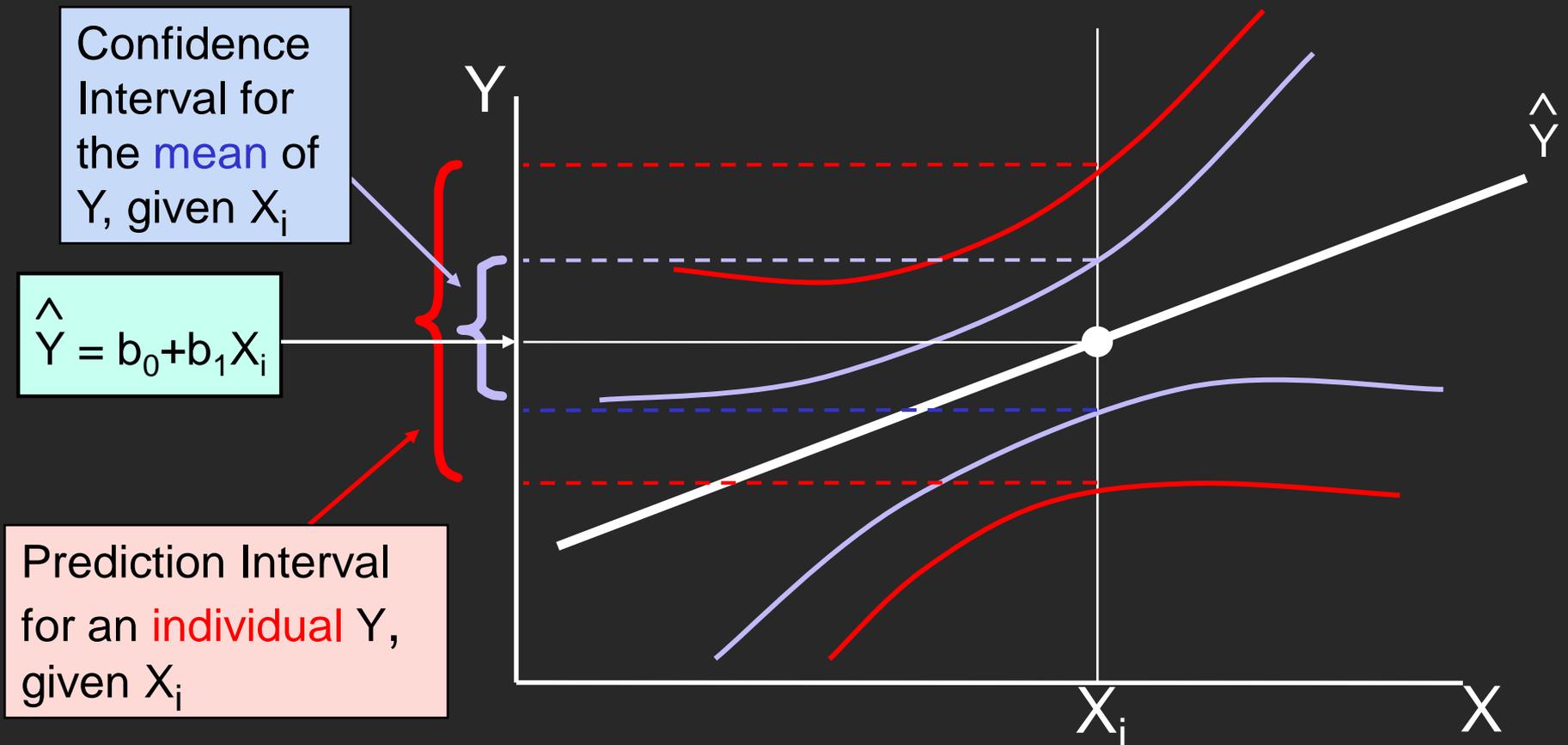
Confidence Interval for the Average Y, Given X

Confidence interval estimate for the **mean value of Y** given a particular X_i

Confidence interval for $\mu_{Y|X=X_i}$:

$$\hat{Y} \pm t_{\alpha/2} S_{YX} \sqrt{h_i}$$

Size of interval varies according to distance away from mean, $\overline{X}$

$$h_i = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{SSX} = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum (X_i - \overline{X})^2}$$

Prediction Interval for an Individual Y, Given X

Confidence interval estimate for an **Individual value of Y** given a particular X_i

Confidence interval for $Y_{X=X_i}$:

$$\hat{Y} \pm t_{\alpha/2} S_{YX} \sqrt{1 + h_i}$$

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

Estimation of Mean Values: Example

Confidence Interval Estimate for $\mu_{Y|X=X_i}$

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

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

$$\hat{Y} \pm t_{0.025} S_{YX} \sqrt{\frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum (X_i - \overline{X})^2}} = 317.85 \pm 37.12$$

The confidence interval endpoints (from Excel) are 280.66 and 354.90, or from $280,660 to $354,900

Estimation of Individual Values: Example

Prediction Interval Estimate for $Y_{X=X_i}$

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

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

$$\hat{Y} \pm t_{0.025} S_{YX} \sqrt{1 + \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum (X_i - \overline{X})^2}} = 317.85 \pm 102.28$$

The prediction interval endpoints from Excel are 215.50 and 420.07, or from $215,500 to $420,070

Finding Confidence and Prediction Intervals in Excel

- From Excel, use

  PHStat | regression | simple linear regression …

    - Check the

      "confidence and prediction interval for X="

      box and enter the X-value and confidence level desired

Finding Confidence and Prediction Intervals in Excel

DCOV<u>A</u>

	A	B
1	**Confidence Interval Estimate**	
2		
3	**Data**	
4	**X Value**	**2000**
5	**Confidence Level**	**95%**
6		
7	Intermediate Calculations	
8	Sample Size	10
9	Degrees of Freedom	8
10	t Value	2.306006
11	Sample Mean	1715
12	Sum of Squared Difference	1571500
13	Standard Error of the Estimate	41.33032
14	h Statistic	0.151686
15	Average Predicted Y (YHat)	317.7838
16		
17	**For Average Predicted Y (YHat)**	
18	Interval Half Width	37.11952
19	**Confidence Interval Lower Limit**	**280.6643**
20	**Confidence Interval Upper Limit**	**354.9033**
21		
22	**For Individual Response Y**	
23	Interval Half Width	102.2813
24	**Prediction Interval Lower Limit**	**215.5025**
25	**Prediction Interval Upper Limit**	**420.0651**

$\hat{Y}$

Confidence Interval Estimate for $\mu_{Y|X=X_i}$

Prediction Interval Estimate for $Y_{X=X_i}$

PEARSON

Finding Confidence and Prediction Intervals in Minitab

Confidence Interval Estimate for $\mu_{Y|X=Xi}$

Predicted Values for New Observations

New
Obs   Fit   SE Fit   95% CI   95% PI
 1 317.8   16.1   (280.7, 354.9) (215.5, 420.1)

Values of Predictors for New Observations

New Square
Obs   Feet
 1   2000

$\hat{Y}$

Input value(s)

Prediction Interval Estimate for $Y_{X=Xi}$

Pitfalls of Regression Analysis

- Lacking an awareness of the assumptions of least-squares regression

- Not knowing how to evaluate the assumptions of least-squares regression

- Not knowing the alternatives to least-squares regression if a particular assumption is violated

- Using a regression model without knowledge of the subject matter

- Extrapolating outside the relevant range

- Concluding that a significant relationship identified in an observational study is due to a cause-and-effect relationship

Strategies for Avoiding the Pitfalls of Regression

- Start with a scatter plot of X vs. Y to observe possible relationship

- Perform residual analysis to check the assumptions

  - Plot the residuals vs. X to check for violations of assumptions such as homoscedasticity

  - Use a histogram, stem-and-leaf display, boxplot, or normal probability plot of the residuals to uncover possible non-normality

Strategies for Avoiding
the Pitfalls of Regression

- If there is violation of any assumption, use alternative methods or models

- If there is no evidence of assumption violation, then test for the significance of the regression coefficients and construct confidence intervals and prediction intervals

- Refrain from making predictions or forecasts outside the relevant range

- Remember that the relationships identified in observational studies may or may not be due to cause-and-effect relationships.

Chapter Summary

In this chapter we discussed:

- How to use regression analysis to predict the value of a dependent variable based on a value of an independent variable

- To understand the meaning of the regression coefficients b_0 and b_1

- To evaluate the assumptions of regression analysis and know what to do if the assumptions are violated

- To make inferences about the slope and correlation coefficient

- To estimate mean values and predict individual values