Learning Objectives
In this chapter, you learn:
- How to use regression analysis to predict the value of a dependent variable based on an independent variable
- The meaning of the regression coefficients $b_0$ and $b_1$
- How 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

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

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 $X$

Types of Relationships
- Linear relationships
- Curvilinear relationships
Types of Relationships

Strong relationships

Weak relationships

No relationship

Simple Linear Regression Model

\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \]

Simple Linear Regression Equation (Prediction Line)

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

\[ \hat{Y}_i = b_0 + b_1 X_i \]

The Least Squares Method

\[ b_0 \text{ and } b_1 \text{ are obtained by finding the values of} \]

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

<table>
<thead>
<tr>
<th>House Price in $1000s (Y)</th>
<th>Square Feet (X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>245</td>
<td>1400</td>
</tr>
<tr>
<td>312</td>
<td>1600</td>
</tr>
<tr>
<td>279</td>
<td>1700</td>
</tr>
<tr>
<td>308</td>
<td>1875</td>
</tr>
<tr>
<td>199</td>
<td>1100</td>
</tr>
<tr>
<td>219</td>
<td>1550</td>
</tr>
<tr>
<td>405</td>
<td>2350</td>
</tr>
<tr>
<td>324</td>
<td>2450</td>
</tr>
<tr>
<td>319</td>
<td>1425</td>
</tr>
<tr>
<td>255</td>
<td>1700</td>
</tr>
</tbody>
</table>

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's and X's and desired options

Add-Ins: PHStat2: Regression: Simple Linear Regression

Simple Linear Regression Example: Using PHStat2

Add-Ins: PHStat2: Regression: Simple Linear Regression

Simple Linear Regression Example: Excel Output

The regression equation is:

\[ \text{house price} = 98.24833 + 0.10977 \times \text{square feet} \]

Simple Linear Regression Example: Minitab Output

The regression equation is:

\[ \text{house price} = 98.2 + 0.110 \times \text{Square Feet} \]

Simple Linear Regression Example: Graphical Representation

House price model: Scatter Plot and Prediction Line

Simple Linear Regression Example: Interpretation of \( b_0 \)

\( b_0 \) 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 $b_1$

- $b_1$ estimates the change in the mean value of $Y$ as a result of a one-unit increase in $X$
- Here, $b_1 = 0.10977$ tells us that the mean value of a house increases by $0.10977 \times 1000 = 109.77$, on average, for each additional one square foot of size.

Predict the price for a house with 2000 square feet:

The predicted price for a house with 2000 square feet is $317.78(1000s) = 317,780$

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 = total sum of squares (Total Variation)
  - SSR = regression sum of squares (Explained Variation)
  - SSE = error sum of squares (Unexplained Variation)

\[
\begin{align*}
\text{SST} & = \sum (Y_i - \bar{Y})^2 \\
\text{SSR} & = \sum (\hat{Y}_i - \bar{Y})^2 \\
\text{SSE} & = \sum (Y_i - \hat{Y}_i)^2
\end{align*}
\]

where:

- $\bar{Y}$ = Mean value of the dependent variable
- $Y_i$ = Observed value of the dependent variable
- $\hat{Y}_i$ = Predicted value of $Y$ for the given $X$ value
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 $r^2$.

**Coefficient of Determination, $r^2$**

\[
0 \leq r^2 \leq 1
\]

**Examples of Approximate $r^2$ Values**

- $r^2 = 0$: 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.)
- $0 < r^2 < 1$: Weaker linear relationships between X and Y. Some but not all of the variation in Y is explained by variation in X.
- $r^2 = 1$: Perfect linear relationship between X and Y. 100% of the variation in Y is explained by variation in X.

**Simple Linear Regression Example: Coefficient of Determination, $r^2$ in Minitab**

The regression equation is

\[
\text{Price} = 98.2 + 0.110 \times \text{Square Feet}
\]

- Predictors: Constant, Square Feet
- Coefficients: Constant = 98.25, SE = 58.03, T = 1.69, P = 0.129; Square Feet = 0.110, SE = 0.03297, T = 3.33, P = 0.010
- $R^2$-Sqr = 58.1%; $R^2$ (adj) = 52.8%

**Analysis of Variance**

- Source: Regression, Residual Error, Total
- DF: 1, 8, 9
- SS: 18934.948, 13666, 32600
- MS: 18934.948, 1708.1957
- F: 11.0843
- Significance F: 0.01039

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

\[ S_{\text{YX}} = \sqrt{\frac{\sum (Y_i - \hat{Y}_i)^2}{n - 2}} \]

Where

- \( SSE \) = error sum of squares
- \( n \) = sample size

Simple Linear Regression Example: Standard Error of Estimate in Excel

The regression equation is

\[ \text{Price} = 98.2 + 0.110 \text{ Square Feet} \]

Comparing Standard Errors

\( S_{\text{YX}} \) is a measure of the variation of observed \( Y \) values from the regression line.

Assumptions of Regression

L.I.N.E

- Linearity: The relationship between \( X \) and \( Y \) is linear
- Independence of Errors: Error values are statistically independent
- 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

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. \( X \)
Residual Analysis for Linearity
- Linear: Data points follow a straight line
- Not Linear: Data points do not follow a straight line

Residual Analysis for Independence
- Independent: Data points are equally distributed across the range of X values
- Not Independent: Data points are not equally distributed across the range of X values

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

Residual Analysis for Normality
- When using a normal probability plot, normal errors will approximately display in a straight line

Residual Analysis for Equal Variance
- Constant variance: Residuals are equally distributed across the range of Y values
- Non-constant variance: Residuals are not equally distributed across the range of Y values

Simple Linear Regression Example: Excel Residual Output
- Residuals do not appear to violate any regression assumptions
Decision rule: reject $H_0$ if $D < d_l$; do not reject $H_0$ if $D > d_u$. Is there autocorrelation?
Example with n = 25:

Durbin-Watson Calculations

\[ D = \frac{\sum (e_i - e_{i-1})^2}{\sum e_i^2} \]

where:

- \( n \) = number of data points
- \( e_i \) = residuals

Here, \( n = 25 \) and there is \( k = 1 \) one independent variable

Using the Durbin-Watson table, \( d_0 = 1.29 \) and \( d_1 = 1.45 \)

\[ D = 1.00494 < d_0 = 1.29 \]

Decision: reject \( H_0 \) since

\[ D = 1.00494 < d_0 \]

\[ 0.0 \]

Inferences About the Slope

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

\[ S_{b_1} = \frac{S_{SYX}}{\sqrt{SSX}} = \frac{S_{SYX}}{\sqrt{n - 2}} \]

where:

- \( S_{b_1} \) = estimate of the standard error of the slope
- \( S_{SYX} \) = 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

\[ t_{STAT} = \frac{b_1 - \beta_1}{S_{b_1}} \]

where:

- \( b_1 \) = regression slope coefficient
- \( \beta_1 \) = hypothesized slope
- \( S_{b_1} \) = standard error of the slope

\[ \]

Inferences About the Slope: \( t \) Test Example

Estimated Regression Equation:

\[ y = 30.65 + 4.7038x \]

The slope of this model is 0.1098

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

From Excel output:

- \( H_0: \beta_1 = 0 \)
- \( H_1: \beta_1 \neq 0 \)

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef.</th>
<th>SE Coef.</th>
<th>T-Value</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>30.65</td>
<td>4.7038</td>
<td>6.5395</td>
<td>0.0000</td>
</tr>
<tr>
<td>Square Feet</td>
<td>4.7038</td>
<td>0.2238</td>
<td>20.9452</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

From Minitab output:

- \( b_1 = 4.7038 \)
- \( S_{b_1} = 0.2238 \)
- \( t_{STAT} = \frac{4.7038 - 0}{0.2238} = 20.9452 \)
- \( \text{P-Value} = 0.0000 \)
Inferences About the Slope: t Test Example

H₀: β₁ = 0
H₁: β₁ ≠ 0

Test Statistic: t = 3.329

Decision: Reject H₀

There is sufficient evidence that square footage affects house price.

F-Test for Significance

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 a 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

ANOVA

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>18934.9348</td>
<td>18934.9348</td>
<td>11.0848</td>
<td>0.01039</td>
</tr>
<tr>
<td>Residual</td>
<td>8</td>
<td>13665.5652</td>
<td>1708.1957</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>9</td>
<td>32600.5000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Decision: Reject H₀ at α = 0.05

There is sufficient evidence that house size affects selling price.

F-Test for Significance

Minitab Output

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>18934.9348</td>
<td>18934.9348</td>
<td>11.0848</td>
<td>0.01039</td>
</tr>
<tr>
<td>Residual</td>
<td>8</td>
<td>13665.5652</td>
<td>1708.1957</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>9</td>
<td>32600.5000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Decision: Reject H₀ at α = 0.05

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

Excel printout for House Prices:

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

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.

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

R = \sqrt{r} if bi > 0
R = -\sqrt{r} if bi < 0

Decision: Reject H0

Conclusion: There is evidence of a linear association at the 5% level of significance.
Chapter 13

13-13

Basic Business Statistics, 10/e © 2006 Prentice Hall, Inc.

Estimating Mean Values and Predicting Individual Values

Goal: Form intervals around $Y$ to express uncertainty about the value of $Y$ for a given $X_i$.

$Y = b_0 + b_1 X_i$

Confidence Interval for the mean of $Y$, given $X_i$

$\bar{Y} = \hat{Y} \pm t_{\alpha/2} \hat{S}_{Y|X}$

Prediction Interval for an individual $Y$, given $X_i$

$\hat{Y} = \hat{Y} \pm t_{\alpha/2} \hat{S}_{Y|X} \sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}}$

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

$\hat{Y} = \hat{Y} \pm t_{\alpha/2} \hat{S}_{Y|X} \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}}$

Estimation of Mean Values: Example

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

Predicted Price $\hat{Y}_i = 317.78$ ($1,000s)

$\bar{Y} \pm t_{0.025} \hat{S}_{Y|X} \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}} = 317.78 \pm 37.12$

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

Estimation of Individual Values: Example

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

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

$\hat{Y} \pm t_{0.025} \hat{S}_{Y|X} \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}} = 317.85 \pm 102.28$

The prediction interval endpoints 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

Input values

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

Prediction Interval Estimate for \( Y_{X=X_i} \)

Finding Confidence and Prediction Intervals in Minitab

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
Obs  Feet
1  2000

Pitfalls of Regression Analysis

- Lacking an awareness of the assumptions underlying least-squares regression
- Not knowing how to evaluate the assumptions
- 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

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

(continued)

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
- Avoid making predictions or forecasts outside the relevant range

Chapter Summary

- Introduced types of regression models
- Reviewed assumptions of regression and correlation
- Discussed determining the simple linear regression equation
- Described measures of variation
- Discussed residual analysis
- Addressed measuring autocorrelation
Chapter Summary (continued)

- Described inference about the slope
- Discussed correlation – measuring the strength of the association
- Addressed estimation of mean values and prediction of individual values
- Discussed possible pitfalls in regression and recommended strategies to avoid them