Chapter 10

Simple Linear Regression Analysis

What is Regression Analysis?
- A statistical technique that describes the relationship between a dependent variable and one or more independent variables.

Examples
- Consider the relationship between construction permits (x) and carpet sales (y) for a company.
  OR
- Relationship between advertising expenditures and sales
- There probably is a relationship...
  ...as number of permits increases, sales should increase.
  ...set advertising expenditure and we can predict sales
- But how would we measure and quantify this relationship?

Simple Linear Regression Model (SLR)
- Assume relationship to be linear
  \[ Y = a + bX + \varepsilon \]
- Where
  \( Y \) = dependent variable
  \( X \) = independent variable
  \( a \) = y-intercept
  \( b \) = slope
  \( \varepsilon \) = random error
Random Error Component ($\varepsilon$)

- Makes this a *probabilistic model*...
  - Represents uncertainty
- $\varepsilon \Rightarrow$ random variation not explained by $x$
- Deterministic Model = Exact relationship
- Example:
  - Temperature: $^\circ F = 9/5^\circ C + 32$
  - Assets = Liabilities + Equity
- Probabilistic Model = Det. Model + Error

Graphically, SLR line is displayed as...

---

Model Parameters

- $a$ and $b$
- Estimated from the data
- Data collected as a pair $(x, y)$

---

Process of Developing SLR Model

- Hypothesize the model: $E(Y) = a + bX$
- Estimate Coefficients
  $$\hat{y} = \hat{a} + \hat{b}x$$
- Specify distribution of error term
- How adequate is the model?
- When model is appropriate, use it for estimation and prediction
Fitting the Straight-Line Model
Ordinary Least Squares (OLS)

• Once it is assumed that the model is
  \[ Y = a + bX + \varepsilon \]
  Next we must collect the data

• Before estimating parameters, we must ensure that the data follows a linear trend
  • Use scatterplot, scattergram, scatter diagram

Assessing Fit

Assessing Fit (Deviations)

• aka errors or residuals \( r_i, e_i \)
  • Difference between the observed value of \( y \) and the predicted value of \( y \)

\[ e_i = r_i = y_i - \hat{y}_i \]

• Want \( r_i \) to be small
Assessing Fit (Cont.)

• NOTE: Sum of the residuals is 0
  \[ \sum \varepsilon_i = 0 = \sum (y_i - \hat{y}_i) \]

• Can fit many different lines; which one is best?
• Line that best fits the data is the one that minimizes the sum of squares of the errors (SSE).
• This is the least squares line.

Least Squares Line

• Find the line that minimizes
  \[ \sum (y_i - \hat{y}_i)^2 \]
  with respect to the parameters

• Recall that
  \[ \hat{y}_i = \hat{a} + \hat{b}x_i \]

• Minimize
  \[ \sum [y_i - (\hat{a} + \hat{b}x_i)]^2 \]

Least Squares Line (Cont.)

• Estimated parameters yield smallest SSE
• Estimated coefficients are given by:
  \[ \hat{b} = \frac{SS_{xy}}{SS_{xx}} \]
  \[ \hat{a} = \bar{y} - \hat{b}\bar{x} \]
  \[ SS_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) \]
  \[ SS_{xx} = \sum (x_i - \bar{x})^2 \]

Example 1

The Central Company manufactures a certain specialty item once a month in a batch production run. The number of items produced in each run varies from month to month as demand fluctuates. The company is interested in the relationship between the size of the production run (x) and the number of man-hours of labor (y) required for the run. The company has collected the following data for the 10 most recent runs:
Example 1 (Cont.)

<table>
<thead>
<tr>
<th>Run</th>
<th>Number of items</th>
<th>Labor (man-hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>40</td>
<td>83</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>60</td>
</tr>
<tr>
<td>3</td>
<td>70</td>
<td>138</td>
</tr>
<tr>
<td>4</td>
<td>90</td>
<td>180</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td>97</td>
</tr>
<tr>
<td>6</td>
<td>60</td>
<td>118</td>
</tr>
<tr>
<td>7</td>
<td>70</td>
<td>140</td>
</tr>
<tr>
<td>8</td>
<td>40</td>
<td>75</td>
</tr>
<tr>
<td>9</td>
<td>80</td>
<td>159</td>
</tr>
<tr>
<td>10</td>
<td>70</td>
<td>144</td>
</tr>
</tbody>
</table>

Example 1 (Cont.)

- Estimated Regression Equation

\[ \hat{y} = -1.836 + 2.021x \]

Interpretation of Regression Equation

- What does this mean?

\[ \hat{y} = -1.836 + 2.021x \]

- \( x \) = # of items produced
- \( y \) = # of man-hours of labor

State conclusions in terms of problem
- Intercept: when no items are produced, the est. # of hrs. is -1.836.
- Does this make sense?
- No!

Interpretation (Cont.)

- When using regression to predict a response, the value of the independent variable must fall in the range of the original data.
- Predictions made outside of the range of the data is called EXTRAPOLATION and may have little or no validity.
- In our example, our independent variable ranges from 30 to 90, and predictions should be made in this range.
Interpretation (Cont.)

• Slope: every unit change in x, the average value of y will change by the slope.
• In the example, 2.021 implies that for every item produced, the average # of man-hours is expected to increase by 2.021.

Example 2

The Tri-City Office Equipment Corporation sells an imported desk calculator on a franchise basis and performs preventive maintenance and repair service on this calculator. Data has been collected from 18 recent calls on users to perform routine preventive maintenance service; for each call, x is the number of machines serviced and y is the total number of minutes spent by the service person.

Example 2 (Cont.)

• Obtain Estimated Regression Equation

\[
\hat{b} = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{1098}{74.5} = 14.7383
\]

\[
\hat{a} = \bar{y} - \hat{b}\bar{x} = 64 - (14.7383)(4.5) = -2.3224
\]

\[
\hat{y} = -2.3224 + 14.7383x
\]

Example 2 (Cont.)

Interpretations

• Intercept: When no machines are serviced, the repairman spends an avg. of -2.3224 minutes; Note that x=0 is probably not in the range of the data, so intercept makes no sense.
• Slope: For each machine serviced, we would expect approx. 14.74 minutes of service time spent.
Model Assumptions

- $E(\varepsilon) = 0$
- $\text{Var}(\varepsilon) = \sigma^2$
- $\varepsilon$ is normally distributed
- $\varepsilon_i$ are independent

Before performing regression analysis, these assumptions should be validated.

Descriptive Measures of Association

- Coefficient of Determination ($R^2$)
Error Decomposition

\[ Y_i - \bar{Y} = \hat{Y}_i - \bar{Y} \]

\[ \hat{Y}_i = \hat{\alpha} + \hat{b} x_i \]

Coefficient of Determination (Cont.)

\[ 0 \leq \text{SSE} \leq \text{SS}_{yy} \]

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

Larger R^2, the more variability is explained by the regression model.
Correlation Coefficient ($r$)

- Positive square root of $R^2$
- aka Pearson product-moment correlation coefficient
- Unitless
- $-1 \leq r \leq 1$
- Describes the strength of the relationship between $x$ and $y$

Correlation Coefficient ($r$)

- Computational Formula
  \[ r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} \]
- $-1$ implies strong negative relationship
- $0$ implies no relationship
- $+1$ implies strong positive relationship

Measures of Association (Cont.)

- High correlation does not imply causation.
- What does this mean?

Estimation and Prediction

- Satisfied with the model, we can perform:
  - *Estimation* of the mean value of $y$ for a given value of $x$
  - *Prediction* of a new observation for a given value of $x$
- Where do we expect to have the most success?
Estimation & Prediction (Cont.)

- The fitted SLR model is 
  \[ \hat{y} = \hat{a} + \hat{b}x \]
- Estimating \( y \) at a given value of \( x \), say \( x_p \), yields the same value as predicting \( y \) at a given value of \( x_p \).
- Difference is in precision of the estimate... the sampling errors

Scatter Plot of Correct Model

- \( Y = 3.0 + 0.5X \)
- \( R^2 = 0.67 \)

Scatter Plot of Curvilinear Model

- \( Y = 3.0 + 0.5X \)
- \( R^2 = 0.67 \)

Scatter Plot of Outlier Model

- \( Y = 3.0 + 0.5X \)
- \( R^2 = 0.67 \)
Chapter 10

Scatter Plot of Influential Model

- \( Y = 3.0 + 0.5X \)
- \( R^2 = 0.67 \)

Verifying Assumptions

- \( r_i = Y_i - \hat{Y}_i \)

Examining Residual Plots

Regression and Excel

- Excel also has a built-in tool for performing regression that:
  - is easier to use
  - provides a lot more information about the problem
- To install the Regression tool,
  Tools → AddIns → Analysis ToolPak
- Then to perform the analysis
  Data → Data Analysis → Regression
The TREND() Function

TREND(Y-range, X-range, X-value for prediction)

where:
- **Y-range** is the spreadsheet range containing the dependent Y variable,
- **X-range** is the spreadsheet range containing the independent X variable(s),
- **X-value for prediction** is a cell (or cells) containing the values for the independent X variable(s) for which we want an estimated value of Y.

Note: The TREND() function is dynamically updated whenever any inputs to the function change. However, it does not provide the statistical information provided by the regression tool. It is best to use these two different approaches to regression in conjunction with one another.

Important Software Note

When using more than one independent variable, all variables for the X-range must be in one contiguous block of cells (that is, in adjacent columns).
Regression Plot

Multiple Regression & Model Building

Multiple Regression Analysis

- Most regression problems involve more than one independent variable.
- If each independent variable varies in a linear manner with $y$, the estimated regression function in this case is:
  \[ \hat{y}_i = \hat{a} + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \cdots + \hat{b}_k X_k \]
- The optimal values for the $\hat{b}_i$ can again be found by minimizing the ESS.
- The resulting function fits a hyperplane to our sample data.
Multiple Regression Analysis

Example

Admissions data
In SLR, we had
\[ x_1 = \text{Entrance test score} \]
\[ y = \text{End of year GPA} \]
Suppose other factors involved
\[ x_2 = \text{HS GPA} \]
\[ x_3 = \text{SAT score} \]
Model becomes
\[ y = a + b_1x_1 + b_2x_2 + b_3x_3 + \varepsilon \]

Multiple Regression Analysis

Independent Variables

- May represent higher-order terms
  - \( x_1 = \text{age} \)
  - \( x_2 = \text{age}^2 \)
- May be dummy/indicator variables
  \[ x_3 = \begin{cases} 0, & \text{if female} \\ 1, & \text{if male} \end{cases} \]
- May be functions of independent variables
  - \( x_4 = \text{price} \)
  - \( x_5 = \text{industry average price} \)
  - \( x_6 = \text{price difference} = x_5 - x_4 \)

Steps to Developing Multiple Regression Model

1. Hypothesize the model:
   \[ y = a + b_1x_1 + \ldots + b_kx_k + \varepsilon \]
2. Estimate coefficients
3. Specify distribution of \( \varepsilon \) and estimate \( \sigma^2 \)
4. Validate model assumptions
5. Evaluate model adequacy
6. Use for estimation and prediction

Multiple Regression Analysis

Fitting the Model

- \( b_i \) represents the change in \( y \) with respect to each unit change in \( x \) when ALL other \( x \)'s are held constant
- Method of fitting is the same as in SLR
- Estimate \( b_i \)'s to minimize SSE
- Computationally intensive
- Use MS Excel
Multiple Regression
Salsberry Realty

Salsberry Realty sells homes along the east coast of the United States. One of the questions frequently asked by prospective buyers is: If we purchase this home, how much can we expect to pay to heat it during the winter? The research department at Salsberry has been asked to develop some guidelines regarding heating costs for single family homes. Three variables are thought to relate to the heating costs: (1) the mean daily outside temperature, (2) the number of inches of insulation in the attic, and (3) the age of the furnace. To investigate, Salsberry's research department selected a random sample of 20 recently sold homes. They determined the cost to heat the home last January, as well as the mean outside temperature during January in the region, the number of inches of insulation in the attic, and the age of the furnace. The sample information is given below.

<table>
<thead>
<tr>
<th>Home</th>
<th>Heating Cost ($)</th>
<th>Mean Outside Temperature (°F)</th>
<th>Attic Insulation (inches)</th>
<th>Age of Furnace (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>250</td>
<td>35</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>360</td>
<td>29</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>165</td>
<td>36</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>43</td>
<td>60</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>90</td>
<td>65</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>200</td>
<td>32</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>355</td>
<td>10</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>290</td>
<td>7</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>9</td>
<td>230</td>
<td>21</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>10</td>
<td>120</td>
<td>55</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>11</td>
<td>73</td>
<td>54</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>12</td>
<td>205</td>
<td>48</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>400</td>
<td>20</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>14</td>
<td>320</td>
<td>39</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>15</td>
<td>72</td>
<td>60</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>16</td>
<td>272</td>
<td>20</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>17</td>
<td>94</td>
<td>58</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>18</td>
<td>190</td>
<td>40</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>19</td>
<td>235</td>
<td>27</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>20</td>
<td>135</td>
<td>30</td>
<td>7</td>
<td>5</td>
</tr>
</tbody>
</table>

Multiple Regression
Salsberry Realty

The hypothesized model is given by

\[ y = a + b_1x_1 + b_2x_2 + b_3x_3 + \varepsilon \]

where

- \( y \) = heating cost
- \( x_1 \) = mean outside temp.
- \( x_2 \) = attic insulation
- \( x_3 \) = age of furnace
- \( \varepsilon \) = random error
Chapter 10

Scatterplot 1
Mean Temp. vs. Heating Cost

Scatterplot 2
Attic Insulation vs. Heating Cost

Scatterplot 3
Age of Furnace vs. Heating Cost

Multiple Regression
Salsberry Realty

**SUMMARY OUTPUT**

**Regression Statistics**

- Multiple R: 0.896755299
- R Square: 0.804170066
- Adjusted R Square: 0.767451954
- Standard Error: 51.04855358
- Observations: 20

**ANOVA**

- df: 3, 16, 19
- SS: 171220.4728, 41695.27717, 212915.75
- MS: 57073.49094, 2605.954823, Total
- F: 21.90118203, 6.56178E-06

**Coefficients**

- Intercept: 427.1938033, 59.60142931, 7.167509374, 2.23764E-06
- Mean Outside Temperature (F): -4.582662626, 0.772319353, -5.933636915, 2.10035E-05
- Attic Insulation (inches): -14.83086269, 4.754412281, -3.119389277, 0.00605963
- Age of Furnace (years): 6.101032061, 4.012120166, 1.520650381, 0.1478624

- Lower 95%: 300.844446, -6.219906146, -24.9097642, -2.40428082
- Upper 95%: 553.5431606, -2.945419105, 14.60634494, 8.79491175

- Significance F: 2.23764E-06, 2.10035E-05, 0.1478624
Multiple Regression
Salsberry Realty

- Estimated regression equation:
  \[ \hat{y} = 427.19 - 4.58x_1 - 14.83x_2 + 6.10x_3 \]
- Discussion
  - Meaningful interpretations of coefficients
  - Check range of each independent variable
- Estimate the heating cost for a mean outside temp. of 30°F, there are 5 in. of insulation, and the furnace is 10 years old.

Multiple Regression Analysis
Estimation and Prediction

- Model Assumptions:
  \[ \varepsilon \sim \text{iid } \mathcal{N}(\mu = 0, \sigma^2) \]
- Estimation of the variance, \(\sigma^2\):
  \[ s^2 = \frac{\text{MSE}}{n - (k + 1)} = \frac{\text{SSE}}{n - (k + 1)} \]

Multiple Regression Analysis
Using Dummy/Indicator Variables

- Qualitative variables can also be used in the regression model
- Dummy/indicator or binary (0, 1) variables denote the presence or absence of the variable of interest

Multiple Regression Analysis
Using Dummy/Indicator Variables

- A qualitative variable with \(c\) classes will be represented by \((c-1)\) dummy/indicator variables in the model, with each taking on the values of 0 and 1.
- Example: Suppose we have an independent var. that represents type of diet: Weight Watchers, Atkins, Body for Life, and Protein.
  - Note we have 4 classes \((c = 4)\)
  - We will need \((c-1) = 3\) variables in the model
Multiple Regression Analysis
Using Dummy/Indicator Variables

Types of Diet could be modeled as:

\[ X_1 = \begin{cases} 
1, & \text{if WW} \\
0, & \text{otherwise} 
\end{cases} \]

\[ X_2 = \begin{cases} 
1, & \text{if Atkins} \\
0, & \text{otherwise} 
\end{cases} \]

\[ X_3 = \begin{cases} 
1, & \text{if BFL} \\
0, & \text{otherwise} 
\end{cases} \]

The model could resemble:

\[ y = a + b_1x_1 + b_2x_2 + b_3x_3 + \varepsilon \]

Multiple Regression Analysis
Motion Picture Industry Example

A motion picture industry analyst wants to estimate the gross earnings generated by a movie. The estimate will be based on different variables involved in the film's production. The independent variables considered are \( X_1 = \) production cost of the movie and \( X_2 = \) total cost of all promotional activities. A third variable (\( X_3 \)) that the analyst wants to consider is whether or not the movie is based on a book published before the release of the movie. The analyst obtains information on a random sample of 20 Hollywood movies made within the last five years. The data is given in the following table.

<table>
<thead>
<tr>
<th>Movie</th>
<th>Gross Earnings, Millions $</th>
<th>Production Cost, Millions $</th>
<th>Promotion Cost, Millions $</th>
<th>Book</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>28</td>
<td>4.2</td>
<td>1</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>35</td>
<td>6.0</td>
<td>3</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td>5.5</td>
<td>6</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>3.3</td>
<td>1</td>
<td>Yes</td>
</tr>
<tr>
<td>5</td>
<td>75</td>
<td>12.5</td>
<td>11</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>60</td>
<td>9.6</td>
<td>8</td>
<td>Yes</td>
</tr>
<tr>
<td>7</td>
<td>15</td>
<td>2.5</td>
<td>0.5</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>45</td>
<td>10.8</td>
<td>5</td>
<td>No</td>
</tr>
<tr>
<td>9</td>
<td>50</td>
<td>8.4</td>
<td>3</td>
<td>Yes</td>
</tr>
<tr>
<td>10</td>
<td>34</td>
<td>6.6</td>
<td>2</td>
<td>No</td>
</tr>
<tr>
<td>11</td>
<td>48</td>
<td>10.7</td>
<td>1</td>
<td>Yes</td>
</tr>
<tr>
<td>12</td>
<td>82</td>
<td>11.0</td>
<td>15</td>
<td>Yes</td>
</tr>
<tr>
<td>13</td>
<td>24</td>
<td>3.5</td>
<td>4</td>
<td>No</td>
</tr>
<tr>
<td>14</td>
<td>50</td>
<td>6.9</td>
<td>10</td>
<td>No</td>
</tr>
<tr>
<td>15</td>
<td>58</td>
<td>7.8</td>
<td>9</td>
<td>Yes</td>
</tr>
<tr>
<td>16</td>
<td>63</td>
<td>10.1</td>
<td>10</td>
<td>No</td>
</tr>
<tr>
<td>17</td>
<td>30</td>
<td>5.0</td>
<td>1</td>
<td>Yes</td>
</tr>
<tr>
<td>18</td>
<td>37</td>
<td>7.5</td>
<td>8</td>
<td>No</td>
</tr>
<tr>
<td>19</td>
<td>45</td>
<td>6.4</td>
<td>8</td>
<td>Yes</td>
</tr>
<tr>
<td>20</td>
<td>72</td>
<td>10.0</td>
<td>12</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Prepare a scatter plot of gross earnings versus production cost and promotion cost. Does there appear to be a linear relationship between gross earnings and either production cost or promotion cost? If the analyst were to use a simple linear regression model to predict gross earnings, which variable should be used? Explain. Determine the parameter estimates for the model given by

\[ \hat{Y} = \hat{a} + \hat{b}_1X_1 + \hat{b}_2X_2 \]

Analyze the results. Determine the parameter estimates for the model given by

\[ \hat{Y} = \hat{a} + \hat{b}_1X_1 + \hat{b}_2X_2 + \hat{b}_3X_3 \]

Does \( X_3 \) help explain the gross earnings when \( X_1 \) and \( X_2 \) are also in the model? Explain.
### Multiple Regression Analysis

**Motion Picture Industry Example**

- With simplicity in mind, suppose we fit three simple linear regression functions:
  \[
  \hat{y}_1 = \hat{a} + \hat{b}_1 x_1 \\
  \hat{y}_2 = \hat{a} + \hat{b}_2 x_2 \\
  \hat{y}_3 = \hat{a} + \hat{b}_3 x_3
  \]

- Key regression results are:

<table>
<thead>
<tr>
<th>Variables in the Model</th>
<th>Adjusted R²</th>
<th>R²</th>
<th>S_i</th>
<th>Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>X₁</td>
<td>0.751</td>
<td>0.738</td>
<td>9.506</td>
<td>( a=5.071 ), ( b_1=5.527 )</td>
</tr>
<tr>
<td>X₂</td>
<td>0.779</td>
<td>0.766</td>
<td>8.970</td>
<td>( a=24.332 ), ( b_2=3.761 )</td>
</tr>
<tr>
<td>X₃</td>
<td>0.299</td>
<td>0.260</td>
<td>15.960</td>
<td>( a=35.111 ), ( b_3=19.889 )</td>
</tr>
</tbody>
</table>

- The model using \( X_2 \) accounts for 77.9% of the variation in \( y \), leaving approx. 22% unaccounted for.
Multiple Regression Analysis
Motion Picture Industry Example

Modeling earnings using production and promotion costs yields:

\[ \hat{y} = 8.15 + 3.27x_1 + 2.37x_2 \]

\[ R^2 = 0.9344, \] which implies 93.44% of the variation in earnings can be explained by prod. and prom. costs.

\[ s = 5.025, \] which is significantly less than either of the SLR models.

Using the full model:

\[ \hat{y} = 7.84 + 2.85x_1 + 2.28x_2 + 7.17x_3 \]

\[ R^2 \] increases to 96.67% and the std. error is reduced to 3.6895.

Indicator variables – revisited

Note that \( x_3 \) takes on the values of 0 and 1.
Multiple Regression Analysis
Selecting the Model

• We want to identify the simplest model that adequately accounts for the systematic variation in the dependent variable, y.
• Arbitrarily using all of the independent variables may result in overfitting.

Multiple Regression Analysis
Adjusted R² Statistic

• As additional independent variables are added to a model:
  – The $R^2$ statistic can only increase.
  – The Adjusted-$R^2$ statistic can increase or decrease.

$$R^2_a = 1 - \frac{\text{SSE}}{\text{SS}_y} \left( \frac{n - 1}{n - k - 1} \right)$$

• Adjusted $R^2 \leq R^2$
• The $R^2$ statistic can be artificially inflated by adding any independent variable to the model.
• We can compare adjusted-$R^2$ values as a heuristic to tell whether adding an additional independent variable really helps to improve a regression model.

Multiple Regression Analysis
Motion Picture Industry Example

• Key regression results are:

<table>
<thead>
<tr>
<th>Variables in the Model</th>
<th>Adjusted $R^2$</th>
<th>$R^2$</th>
<th>Parameter Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>0.751</td>
<td>0.738</td>
<td>$a=5.071, b=5.527$</td>
</tr>
<tr>
<td>$x_1$ &amp; $x_2$</td>
<td>0.934</td>
<td>0.927</td>
<td>$a=8.152, b_1=-3.267, b_2=2.367$</td>
</tr>
<tr>
<td>$x_1, x_2$ &amp; $x_3$</td>
<td>0.967</td>
<td>0.961</td>
<td>$a=7.836, b_1=2.848, b_2=2.278, b_3=7.166$</td>
</tr>
</tbody>
</table>

• The model using $x_1$, $x_2$, and $x_3$ appears to be best:
  – Highest adjusted-$R^2$ and highest $R^2$
  – Lowest $s$ (most precise prediction intervals)

Multiple Regression Analysis
Estimation and Prediction

• Same as in SLR
• Like SLR, difference lies in the error of estimation and prediction errors
• In multiple regression, these standard errors are complex and beyond the scope of this class
• Will rely on MS Excel output
Multiple Regression Analysis

Concerns

- Parameter Estimability
  - inability of the model to estimate parameters because data is concentrated in one area
  - data must include at least one more level of x than the highest order of the x-variable that is included in the model
- Multicollinearity
  - relationship between two or more independent variables
  - variables contributing the same information
  - if two or more variables are highly correlated, then we only need one in the model
- Extrapolation (already discussed in SLR)
- Correlated Errors
  - measurements on the dependent variable are correlated
  - time series analysis

Polynomial Regression

Sometimes the relationship between a dependent and independent variable is not linear.

This graph suggests a quadratic relationship between square footage (X) and selling price (Y).

An appropriate regression function in this case might be,

\[ \hat{y}_i = \hat{a} + \hat{b}_1 x_1 + \hat{b}_2 x_1^2 \]

or equivalently,

\[ \hat{y}_i = \hat{a} + \hat{b}_1 x_1 + \hat{b}_2 x_2 \]

where,

\[ x_2 = x_1^2 \]
Multiple Regression Analysis
Fitting a Third Order Polynomial Model

We could also fit a third order polynomial model,
\[ \hat{Y}_i = a + \hat{b}_1 X_{i1} + \hat{b}_2 X_{i1}^2 + \hat{b}_3 X_{i1}^3 \]
or equivalently,
\[ \hat{Y}_i = a + \hat{b}_1 X_{i1} + \hat{b}_2 X_{i2} + \hat{b}_3 X_{i3} \]
where,
\[ X_{i2} = X_{i1}^2 \]
\[ X_{i3} = X_{i1}^3 \]

Graph of Estimated Third Order Polynomial Regression Function

Multiple Regression Analysis
Polynomial Regression
Overfitting

- When fitting polynomial models, care must be taken to avoid overfitting.
- The adj.-R² statistic can also be used for building/fitting polynomial regression models.
- We can gauge the amount of overfitting by Validating the fit, or using a training sample to build the model and a validation sample to examine its estimation or prediction accuracy.