Chapter 8
Trendlines and Regression Analysis
Contents

ício Trendlines
- Simple Linear Regression
- Multiple Linear Regression
- Systematic Model Building
- Practical Issues
  - Overfitting
  - Categorical Variables
  - Interaction Terms
  - Non-linear Terms
Common Mathematical Functions Used in Predictive Analytical Models

Linear \[ y = a + bx \]
Logarithmic \[ y = \ln(x) \]
Polynomial (2\text{nd} order) \[ y = ax^2 + bx + c \]
Polynomial (3\text{rd} order) \[ y = ax^3 + bx^2 + dx + e \]
Power \[ y = ax^b \]
Exponential \[ y = ab^x \]

(the base of natural logarithms, \(e = 2.71828\ldots\) is often used for the constant \(b\))
Excel Trendline Tool

- Right click on data series and choose *Add trendline* from pop-up menu
- Check the boxes *Display Equation on chart* and *Display R-squared value on chart*
$R^2$ is a measure of the “fit” of the line to the data.

- The value of $R^2$ will be between 0 and 1.
- A value of 1.0 indicates a perfect fit and all data points would lie on the line; the larger the value of $R^2$ the better the fit.
- $R^2$ is the squared correlation between the dependent variable and the prediction.

It is called the **coefficient of determination** and indicates the proportion of the variance in the dependent variable that is predictable from the independent variable.
Example 8.1: Modeling a Price-Demand Function

Linear demand function:
Sales = 20,512 - 9.5116(price)
Example 8.2: Predicting Crude Oil Prices

- Line chart of historical crude oil prices
Excel’s *Trendline* tool is used to fit various functions to the data.

<table>
<thead>
<tr>
<th>Function</th>
<th>Equation</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>$y = 50.49e^{0.021x}$</td>
<td>0.664</td>
</tr>
<tr>
<td>Logarithmic</td>
<td>$y = 13.02\ln(x) + 39.60$</td>
<td>0.382</td>
</tr>
<tr>
<td>Polynomial 2°</td>
<td>$y = 0.13x^2 - 2.399x + 68.01$</td>
<td>0.905</td>
</tr>
<tr>
<td>Polynomial 3°</td>
<td>$y = 0.005x^3 - 0.111x^2 + 0.648x + 59.497$</td>
<td>0.928 *</td>
</tr>
<tr>
<td>Power</td>
<td>$y = 45.96x^{0.0169}$</td>
<td>0.397</td>
</tr>
</tbody>
</table>
Example 8.2 Continued

- Third order polynomial trendline fit to the data

![Graph showing a trendline fitting to the data with a polynomial equation and R² value.](image)
Caution About Polynomials

- The $R^2$ value will continue to increase as the order of the polynomial increases; that is, a 4th order polynomial will provide a better fit than a 3rd order, and so on.
- Higher order polynomials will generally not be very smooth and will be difficult to interpret visually.
  - Thus, we don't recommend going beyond a third-order polynomial when fitting data.
- Use your eye to make a good judgment!
Regression Analysis

- **Regression analysis** is a tool for building mathematical and statistical models that characterize relationships between a dependent (ratio) variable and one or more independent, or explanatory variables (ratio or categorical), all of which are numerical.
- **Simple linear regression** involves a single independent variable.
- **Multiple regression** involves two or more independent variables.
Contents

- Trendlines
- **Simple Linear Regression**
- Multiple Linear Regression
- Systematic Model Building
- Practical Issues
  - Overfitting
  - Categorical Variables
  - Interaction Terms
  - Non-linear Terms
Simple Linear Regression

- Finds a linear relationship between:
  - one independent variable $X$ and
  - one dependent variable $Y$
- First prepare a scatter plot to verify the data has a linear trend.
- Use alternative approaches if the data is not linear.
Example 8.3: *Home Market Value Data*

Size of a house is typically related to its market value.

\[ X = \text{square footage} \]
\[ Y = \text{market value ($)} \]

The scatter plot of the full data set (42 homes) indicates a linear trend.
Finding the Best-Fitting Regression Line

- Market value = $a + b \times$ square feet
- Two possible lines are shown below.

- Line A is clearly a better fit to the data.
- We want to determine the best regression line.
Example 8.4: Using Excel to Find the Best Regression Line

- Market value = $32,673 + $35.036 \times \text{square feet}
  - The estimated market value of a home with 2,200 square feet would be: market value = $32,673 + $35.036 \times 2,200 = $109,752

The regression model explains variation in market value due to size of the home. It provides better estimates of market value than simply using the average.
Simple linear regression model:

\[ Y = \beta_0 + \beta_1 X + \epsilon \]  \hspace{1cm} (8.1)

We estimate the parameters from the sample data:

\[ \hat{Y} = b_0 + b_1 X \]  \hspace{1cm} (8.2)

Let \( X_i \) be the value of the independent variable of the \( i^{th} \) observation. When the value of the independent variable is \( X_i \), then \( \hat{Y}_i = b_0 + b_1 X_i \) is the estimated value of \( Y \) for \( X_i \).
Residuals are the observed errors associated with estimating the value of the dependent variable using the regression line:

\[ e_i = Y_i - \hat{Y}_i \]  \hspace{1cm} (8.3)
Least Squares Regression

- The best-fitting line minimizes the sum of squares of the residuals.

\[
\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - [b_0 + b_1X_i])^2
\] (8.4)

\[
b_1 = \frac{\sum_{i=1}^{n} X_iY_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}
\] (8.5)

\[
b_0 = \bar{Y} - b_1\bar{X}
\] (8.6)

- Excel functions:
  - =INTERCEPT(known_y’s, known_x’s)
  - =SLOPE(known_y’s, known_x’s)
Example 8.5: Using Excel Functions to Find Least-Squares Coefficients

- **Slope** = \( b_1 = 35.036 \)
  
  \[ = \text{SLOPE(C4:C45, B4:B45)} \]

- **Intercept** = \( b_0 = 32,673 \)
  
  \[ = \text{INTERCEPT(C4:C45, B4:B45)} \]

- **Estimate** \( Y \) when \( X = 1750 \) square feet
  
  \[ \hat{Y} = 32,673 + 35.036(1750) = $93,986 \]
  
  \[ = \text{TREND(C4:C45, B4:B45, 1750)} \]
Simple Linear Regression With Excel

Data > Data Analysis > Regression
Input Y Range (with header)
Input X Range (with header)
Check Labels

Excel outputs a table with many useful regression statistics.
# Home Market Value Regression Results

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression Analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td><strong>Regression Statistics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Multiple R</td>
<td>0.731255223</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>R Square</td>
<td>0.534734202</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Adjusted R Square</td>
<td>0.523102557</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Standard Error</td>
<td>7287.722712</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Observations</td>
<td>42</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td><strong>ANOVA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>df</td>
<td>SS</td>
<td>MS</td>
<td>F</td>
<td>Significance F</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Regression</td>
<td>1</td>
<td>2441633669</td>
<td>2441633669</td>
<td>45.97236277</td>
<td>3.79802E-08</td>
</tr>
<tr>
<td>13</td>
<td>Residual</td>
<td>40</td>
<td>2124436093</td>
<td>53110902.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Total</td>
<td>41</td>
<td>4566069762</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td><strong>Coefficients</strong></td>
<td><strong>Standard Error</strong></td>
<td><strong>t Stat</strong></td>
<td><strong>P-value</strong></td>
<td><strong>Lower 95%</strong></td>
<td><strong>Upper 95%</strong></td>
</tr>
<tr>
<td>17</td>
<td>Intercept</td>
<td>32673.2199</td>
<td>8831.950745</td>
<td>3.699434116</td>
<td>0.000649604</td>
<td>14823.18178</td>
</tr>
<tr>
<td>18</td>
<td>Square Feet</td>
<td>35.03637258</td>
<td>5.16738385</td>
<td>6.780292234</td>
<td>3.79802E-08</td>
<td>24.59270036</td>
</tr>
</tbody>
</table>
Regression Statistics

- **Multiple R** - $| r |$, where $r$ is the sample correlation coefficient. The value of $r$ varies from -1 to +1 ($r$ is negative if slope is negative)
- **R Square** - coefficient of determination, $R^2$, which varies from 0 (no fit) to 1 (perfect fit)
- **Adjusted R Square** - adjusts $R^2$ for sample size and number of $X$ variables
- **Standard Error** - variability between observed and predicted $Y$ values. This is formally called the standard error of the estimate, $S_{YX}$. 
Example 8.6: Interpreting Regression Statistics for Simple Linear Regression

53% of the variation in home market values can be explained by home size. The standard error of $7287 is less than standard deviation (not shown) of $10,553.
Regression as Analysis of Variance

ANOVA conducts an $F$-test to determine whether variation in $Y$ is due to varying levels of $X$.

ANOVA is used to test for *significance of regression*:

- $H_0$: population slope coefficient = 0
- $H_1$: population slope coefficient $\neq 0$

Excel reports the $p$-value (*Significance F*).
Rejecting $H_0$ indicates that $X$ explains variation in $Y$. 
P-values are small (<.01) 
Coefficients are significantly different from zero.
Confidence intervals \((\text{Lower 95\% and Upper 95\% values in the output})\) provide information about the unknown values of the true regression coefficients, accounting for sampling error.

- We may also use confidence intervals to test hypotheses about the regression coefficients.
  - To test the hypotheses
    \[
    H_0: \beta_1 = B_1 \\
    H_1: \beta_1 \neq B_1
    \]
    check whether \(B_1\) falls within the confidence interval for the slope. If it does, reject the null hypothesis.
CIs do not span zero!
Residual Analysis and Regression Assumptions

- **Residual** = Actual $Y$ value − Predicted $Y$ value
- **Standard residual** = residual / standard deviation
- Rule of thumb: Standard residuals outside of ±2 or ±3 are potential outliers.
- Excel provides a table and a plot of residuals.

This point has a standard residual of 4.53
Checking Assumptions

- **Linearity**
  - examine scatter diagram (should appear linear)
  - examine residual plot (should appear random)

- **Normality of Errors**
  - view a histogram of standard residuals
  - regression is robust to departures from normality

- **Homoscedasticity**: variation about the regression line is constant
  - examine the residual plot

- **Independence of Errors**: successive observations should not be related.
  - This is important when the independent variable is time.
Example 8.11: Checking Regression Assumptions for the *Home Market Value* Data

- **Linearity** - linear trend in scatterplot
  - no pattern in residual plot
Example 8.11 Continued

Normality of Errors – residual histogram appears slightly skewed but is not a serious departure
Example 8.11 Continued

- **Homoscedasticity** – residual plot shows no serious difference in the spread of the data for different $X$ values.
Example 8.11 Continued

- Independence of Errors – Because the data is cross-sectional, we can assume this assumption holds.
Contents

- Trendlines
- Simple Linear Regression
- Multiple Linear Regression
- Systematic Model Building
- Practical Issues
  - Overfitting
  - Categorical Variables
  - Interaction Terms
  - Non-linear Terms
A linear regression model with more than one independent variable is called a **multiple linear regression model**.

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \]  

(8.10)

where

- \( Y \) is the dependent variable,
- \( X_1, \ldots, X_k \) are the independent (explanatory) variables,
- \( \beta_0 \) is the intercept term,
- \( \beta_1, \ldots, \beta_k \) are the regression coefficients for the independent variables,
- \( \varepsilon \) is the error term
Estimated Multiple Regression Equation

- We estimate the regression coefficients—called **partial regression coefficients** — \( b_0, b_1, b_2, \ldots b_k \), then use the model:

\[
\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k
\]  

(8.11)

- The partial regression coefficients represent the expected change in the dependent variable when the associated independent variable is increased by one unit *while the values of all other independent variables are held constant.*
The independent variables in the spreadsheet must be in contiguous columns.
  ◦ So, you may have to manually move the columns of data around before applying the tool.

Key differences:

Multiple R and R Square are called the **multiple correlation coefficient** and the **coefficient of multiple determination**, respectively, in the context of multiple regression.

ANOVA tests for significance of the entire model. That is, it computes an F-statistic for testing the hypotheses:

\[ H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \]
\[ H_1: \text{at least one } \beta_j \text{ is not 0} \]
ANOVA for Multiple Regression

- ANOVA tests for significance of the entire model. That is, it computes an F-statistic for testing the hypotheses:

\[ H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \]
\[ H_1: \text{at least one } \beta_j \text{ is not 0} \]

- The multiple linear regression output also provides information to test hypotheses about each of the individual regression coefficients.
  - If we reject the null hypothesis that the slope associated with independent variable \(i\) is 0, then the independent variable \(i\) is significant and improves the ability of the model to better predict the dependent variable. If we cannot reject H0, then that independent variable is not significant and probably should not be included in the model.
A good regression model should include only significant independent variables.

However, it is not always clear exactly what will happen when we add or remove variables from a model; variables that are (or are not) significant in one model may (or may not) be significant in another.

Therefore, you should not consider dropping all insignificant variables at one time, but rather take a more structured approach.

Adding an independent variable to a regression model will always result in $R^2$ equal to or greater than the $R^2$ of the original model.

Adjusted $R^2$ reflects both the number of independent variables and the sample size and may either increase or decrease when an independent variable is added or dropped. An increase in adjusted $R^2$ indicates that the model has improved.
Contents

- Trendlines
- Simple Linear Regression
- Multiple Linear Regression
- Systematic Model Building
- Practical Issues
  - Overfitting
  - Categorical Variables
  - Interaction Terms
  - Non-linear Terms
1. Construct a model with all available independent variables. Check for significance of the independent variables by examining the p-values.

2. Identify the independent variable having the largest p-value that exceeds the chosen level of significance.

3. Remove the variable identified in step 2 from the model and evaluate adjusted $R^2$.
   (Don’t remove all variables with p-values that exceed a at the same time, but remove only one at a time.)

4. Continue until all variables are significant.
Example 8.13: Identifying the Best Regression Model

- **Banking Data**

Home value has the largest p-value; drop and re-run the regression.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SUMMARY OUTPUT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>Regression Statistics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Multiple R</td>
<td>0.97309221</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>R Square</td>
<td>0.946908448</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Adjusted R Square</td>
<td>0.944143263</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Standard Error</td>
<td>2055.64333</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Observations</td>
<td>102</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>ANOVA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>df</td>
<td>SS</td>
<td>MS</td>
<td>F</td>
<td>Significance F</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Regression</td>
<td>5</td>
<td>7235179873</td>
<td>1447035975</td>
<td>342.4394584</td>
<td>1.5184E-59</td>
</tr>
<tr>
<td>13</td>
<td>Residual</td>
<td>96</td>
<td>405664271.9</td>
<td>4225669.499</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Total</td>
<td>101</td>
<td>764084145</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Coefficients</td>
<td>Standard Error</td>
<td>t Stat</td>
<td>P-value</td>
<td>Lower 95%</td>
<td>Upper 95%</td>
</tr>
<tr>
<td>17</td>
<td>Intercept</td>
<td>-10710.64278</td>
<td>4260.976308</td>
<td>-2.513659314</td>
<td>0.013613179</td>
<td>-19168.61391</td>
</tr>
<tr>
<td>18</td>
<td>Age</td>
<td>318.6649626</td>
<td>60.98611242</td>
<td>5.225205378</td>
<td>1.01152E-06</td>
<td>197.6084862</td>
</tr>
<tr>
<td>19</td>
<td>Education</td>
<td>621.8603472</td>
<td>318.9595184</td>
<td>1.949652891</td>
<td>0.054135377</td>
<td>-11.26929279</td>
</tr>
<tr>
<td>20</td>
<td>Income</td>
<td>0.146323453</td>
<td>0.040781001</td>
<td>3.588029937</td>
<td>0.000526666</td>
<td>0.085373806</td>
</tr>
<tr>
<td>21</td>
<td>Home Value</td>
<td>0.009183087</td>
<td>0.011038075</td>
<td>0.831944635</td>
<td>0.407504891</td>
<td>-0.012727338</td>
</tr>
<tr>
<td>22</td>
<td>Wealth</td>
<td>0.074331533</td>
<td>0.011189265</td>
<td>6.643111131</td>
<td>1.84838E-09</td>
<td>0.052121017</td>
</tr>
</tbody>
</table>
Example 8.13 Continued

- Bank regression after removing *Home Value*

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SUMMARY OUTPUT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Regression Statistics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Multiple R</td>
<td>0.97289551</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>R Square</td>
<td>0.946525674</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Adjusted R Square</td>
<td>0.944320547</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Standard Error</td>
<td>2052.378536</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Observations</td>
<td>102</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>ANOVA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>df</td>
<td>SS</td>
<td>MS</td>
<td>F</td>
<td>Significance F</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Regression</td>
<td>4</td>
<td>7232255152</td>
<td>1808063788</td>
<td>429.2386497</td>
<td>9.68905E-61</td>
</tr>
<tr>
<td>11</td>
<td>Residual</td>
<td>97</td>
<td>408588992.5</td>
<td>4212257.655</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Total</td>
<td>101</td>
<td>7640844145</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Coefficients</td>
<td>Standard Error</td>
<td>t Stat</td>
<td>P-value</td>
<td>Lower 95%</td>
<td>Upper 95%</td>
</tr>
<tr>
<td>14</td>
<td>Intercept</td>
<td>-12432.4567</td>
<td>3718.674319</td>
<td>-3.343249681</td>
<td>0.001177705</td>
<td>-19812.99587</td>
</tr>
<tr>
<td>15</td>
<td>Age</td>
<td>325.0652837</td>
<td>60.40284488</td>
<td>5.381622098</td>
<td>5.1267E-07</td>
<td>205.1823574</td>
</tr>
<tr>
<td>16</td>
<td>Education</td>
<td>773.38040148</td>
<td>261.4330936</td>
<td>2.958233142</td>
<td>0.003886994</td>
<td>254.5077194</td>
</tr>
<tr>
<td>17</td>
<td>Income</td>
<td>0.159747379</td>
<td>0.037393587</td>
<td>4.272052794</td>
<td>4.52422E-05</td>
<td>0.085531459</td>
</tr>
<tr>
<td>18</td>
<td>Wealth</td>
<td>0.072988791</td>
<td>0.011054865</td>
<td>6.602532898</td>
<td>2.16051E-09</td>
<td>0.051048341</td>
</tr>
</tbody>
</table>

Adjusted $R^2$ improves slightly. All $X$ variables are significant.
Multicollinearity

- **Multicollinearity** occurs when there are strong correlations among the independent variables, and they can predict each other better than the dependent variable.
  - When significant multicollinearity is present, it becomes difficult to isolate the effect of one independent variable on the dependent variable, the signs of coefficients may be the opposite of what they should be, making it difficult to interpret regression coefficients, and $p$-values can be inflated.
- Correlations exceeding $\pm 0.7$ may indicate multicollinearity
- The **variance inflation factor** is a better indicator, but not computed in Excel.
Example 8.14: Identifying Potential Multicollinearity

- **Colleges and Universities** correlation matrix; none exceed the recommend threshold of ±0.7

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Median SAT</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Median SAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Acceptance Rate</td>
<td>-0.601901959</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Expenditures/Student</td>
<td>0.572741729</td>
<td>-0.284254415</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Top 10% HS</td>
<td>0.503467995</td>
<td>-0.609720972</td>
<td>0.505782049</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Graduation %</td>
<td>0.564146627</td>
<td>-0.55037751</td>
<td>0.042503514</td>
<td>0.138612687</td>
<td>1</td>
</tr>
</tbody>
</table>

- **Banking Data** correlation matrix; large correlations exist

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Age</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Education</td>
<td>0.173407147</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Income</td>
<td>0.4771474</td>
<td>0.57539402</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Home Value</td>
<td>0.386493114</td>
<td>0.753521067</td>
<td>0.795355158</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Wealth</td>
<td>0.468091791</td>
<td>0.469413035</td>
<td>0.946665447</td>
<td>0.698477789</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Balance</td>
<td>0.565466834</td>
<td>0.55488066</td>
<td>0.951684494</td>
<td>0.766387128</td>
<td>0.948711734</td>
<td>1</td>
</tr>
</tbody>
</table>
Example 8.14 Continued

- If we remove Wealth from the model, the adjusted $R^2$ drops to 0.9201, but we discover that Education is no longer significant.
- Dropping Education and leaving only Age and Income in the model results in an adjusted $R^2$ of 0.9202.
- However, if we remove Income from the model instead of Wealth, the Adjusted $R^2$ drops to only 0.9345, and all remaining variables (Age, Education, and Wealth) are significant.
Contents

- Trendlines
- Simple Linear Regression
- Multiple Linear Regression
- Systematic Model Building
- Practical Issues
  - Overfitting
  - Categorical Variables
  - Interaction Terms
  - Non-linear Terms
Practical Issues in Trendline and Regression Modeling

- Identifying the best regression model often requires experimentation and trial and error.
- The independent variables selected should make sense in attempting to explain the dependent variable
  - Logic should guide your model development. In many applications, behavioral, economic, or physical theory might suggest that certain variables should belong in a model.
- Additional variables increase $R^2$ and, therefore, help to explain a larger proportion of the variation.
  - Even though a variable with a large p-value is not statistically significant, it could simply be the result of sampling error and a modeler might wish to keep it.
- Good models are as simple as possible (the principle of parsimony).
Overfitting means fitting a model too closely to the sample data at the risk of not fitting it well to the population in which we are interested.

- In fitting the crude oil prices in Example 8.2, we noted that the $R^2$-value will increase if we fit higher-order polynomial functions to the data. While this might provide a better mathematical fit to the sample data, doing so can make it difficult to explain the phenomena rationally.

- In multiple regression, if we add too many terms to the model, then the model may not adequately predict other values from the population.

- Overfitting can be mitigated by using good logic, intuition, theory, and parsimony.
Regression with Categorical Variables

- Regression analysis requires numerical data.
- Categorical data can be included as independent variables, but must be coded numeric using *dummy variables*.
- For variables with 2 categories, code as 0 and 1.
Employee Salaries provides data for 35 employees

Predict Salary using Age and MBA (code as yes=1, no=0)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

where

- $Y$ = salary
- $X_1$ = age
- $X_2$ = MBA indicator (0 or 1)
Example 8.15 Continued

- Salary = 893.59 + 1044.15 \times \text{Age} + 14767.23 \times \text{MBA}
  - If MBA = 0, salary = 893.59 + 1044 \times \text{Age}
  - If MBA = 1, salary = 15,660.82 + 1044 \times \text{Age}

\[
\begin{array}{|c|c|c|c|c|c|c|}
\hline
\text{A} & \text{B} & \text{C} & \text{D} & \text{E} & \text{F} & \text{G} \\
\hline
\text{SUMMARY OUTPUT} & & & & & & \\
\hline
\text{Regression Statistics} & & & & & & \\
\hline
\text{Multiple R} & 0.976118476 & & & & & \\
\hline
\text{R Square} & 0.952807278 & & & & & \\
\hline
\text{Adjusted R Square} & 0.949857733 & & & & & \\
\hline
\text{Standard Error} & 2941.914352 & & & & & \\
\hline
\text{Observations} & 35 & & & & & \\
\hline
\text{ANOVA} & & & & & & \\
\hline
\text{df} & \text{SS} & \text{MS} & \text{F} & \text{Significance F} & & & \\
\hline
\text{Regression} & 2 & 5591651177 & 2795825589 & 323.0353318 & 6.05341E-22 & \\
\hline
\text{Residual} & 32 & 2769555521.7 & 8654860.054 & & & \\
\hline
\text{Total} & 34 & 5868606699 & & & & \\
\hline
\text{Coefficients} & \text{Standard Error} & \text{t Stat} & \text{P-value} & \text{Lower 95%} & \text{Upper 95%} & \\
\hline
\text{Intercept} & 893.5875971 & 1824.575283 & 0.489751015 & 0.627650922 & -2822.950634 & 4610.125828 \\
\hline
\text{Age} & 1044.146043 & 42.14128238 & 24.77727265 & 1.8878E-22 & 958.3070599 & 1129.985026 \\
\hline
\text{MBA} & 14767.23159 & 1351.801764 & 10.92411031 & 2.49752E-12 & 12013.7015 & 17520.76168 \\
\hline
\end{array}
\]
An interaction occurs when the effect of one variable is dependent on another variable.

We can test for interactions by defining a new variable as the product of the two variables, \( X_3 = X_1 \times X_2 \), and testing whether this variable is significant, leading to an alternative model.

\[
Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon
\]
Example 8.16: Incorporating Interaction Terms in a Regression Model

- Define an interaction between Age and MBA and re-run the regression.

The MBA indicator is not significant; drop and re-run.
Example 8.16 Continued

- Adjusted $R^2$ increased slightly, and both age and the interaction term are significant. The final model is:

$$\text{salary} = 3,323.11 + 984.25 \times \text{age} + 425.58 \times \text{MBA} \times \text{age}$$
Categorical Variables with More Than Two Levels

- When a categorical variable has $k > 2$ levels, we need to add $k - 1$ additional variables to the model.
Example 8.17: A Regression Model with Multiple Levels of Categorical Variables

- The Excel file *Surface Finish* provides measurements of the surface finish of 35 parts produced on a lathe, along with the revolutions per minute (RPM) of the spindle and one of four types of cutting tools used.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surface Finish Data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Part</td>
<td>Surface Finish</td>
<td>RPM</td>
<td>Cutting Tool</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>45.44</td>
<td>225</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>42.03</td>
<td>200</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>50.10</td>
<td>250</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>48.75</td>
<td>245</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>47.92</td>
<td>235</td>
</tr>
<tr>
<td>9</td>
<td>6</td>
<td>47.79</td>
<td>237</td>
</tr>
<tr>
<td>10</td>
<td>7</td>
<td>52.26</td>
<td>265</td>
</tr>
<tr>
<td>11</td>
<td>8</td>
<td>50.52</td>
<td>259</td>
</tr>
<tr>
<td>12</td>
<td>9</td>
<td>45.58</td>
<td>221</td>
</tr>
<tr>
<td>13</td>
<td>10</td>
<td>44.78</td>
<td>218</td>
</tr>
<tr>
<td>14</td>
<td>11</td>
<td>33.50</td>
<td>224</td>
</tr>
<tr>
<td>15</td>
<td>12</td>
<td>31.23</td>
<td>212</td>
</tr>
<tr>
<td>16</td>
<td>13</td>
<td>37.52</td>
<td>248</td>
</tr>
<tr>
<td>17</td>
<td>14</td>
<td>37.13</td>
<td>260</td>
</tr>
<tr>
<td>18</td>
<td>15</td>
<td>34.70</td>
<td>243</td>
</tr>
</tbody>
</table>
Because we have $k = 4$ levels of tool type, we will define a regression model of the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

where

- $Y = \text{surface finish}$
- $X_1 = \text{RPM}$
- $X_2 = 1$ if tool type is B and 0 if not
- $X_3 = 1$ if tool type is C and 0 if not
- $X_4 = 1$ if tool type is D and 0 if not
Example 8.17 Continued

- Add 3 columns to the data, one for each of the tool type variables
Example 8.17 Continued

- Regression results

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SUMMARY OUTPUT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Regression Statistics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Multiple R</td>
<td>0.994447053</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>R Square</td>
<td>0.888924942</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Adjusted R Square</td>
<td>0.987448267</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Standard Error</td>
<td>1.089163115</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Observations</td>
<td>86</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>ANOVA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>df</td>
<td>SS</td>
<td>MS</td>
<td>F</td>
<td>Significance F</td>
</tr>
<tr>
<td>12</td>
<td>Regression</td>
<td>4</td>
<td>3177.784271</td>
<td>794.4460678</td>
<td>669.6973322</td>
<td>7.324E-29</td>
</tr>
<tr>
<td>13</td>
<td>Residual</td>
<td>30</td>
<td>35.58828875</td>
<td>1.196276292</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Total</td>
<td>34</td>
<td>3213.37256</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Surface finish = 24.49 + 0.098 RPM - 13.31 type B - 20.49 type C - 26.04 type D
Regression Models with Nonlinear Terms

- Curvilinear models may be appropriate when scatter charts or residual plots show nonlinear relationships.
- A second order polynomial might be used
  \[ Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon \]
- Here \( \beta_1 \) represents the linear effect of \( X \) on \( Y \) and \( \beta_2 \) represents the curvilinear effect.
- This model is linear in the \( \beta \) parameters so we can use linear regression methods.
The U-shape of the residual plot (a second-order polynomial trendline was fit to the residual data) suggests that a linear relationship is not appropriate.
Example 8.18 Continued

- Add a variable for temperature squared.
- The model is:

\[ \text{sales} = 142,850 - 3,643.17 \times \text{temperature} + 23.3 \times \text{temperature}^2 \]

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SUMMARY OUTPUT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Regression Statistics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Multiple R</td>
<td>0.973326989</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>R Square</td>
<td>0.947365428</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Adjusted R Square</td>
<td>0.941517142</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Standard Error</td>
<td>635.1365123</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Observations</td>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>ANOVA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>df</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Regression</td>
<td>2</td>
<td>130693232.2</td>
<td>65346616.12</td>
<td>161.9902753</td>
<td>3.10056E-12</td>
</tr>
<tr>
<td>13</td>
<td>Residual</td>
<td>18</td>
<td>7261171.007</td>
<td>403398.3893</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Total</td>
<td>20</td>
<td>137954403.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Coefficients</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Intercept</td>
<td>142850.3406</td>
<td>30575.70155</td>
<td>4.67021683</td>
<td>0.000189738</td>
<td>78613.17532</td>
</tr>
<tr>
<td>18</td>
<td>Temperature</td>
<td>-3643.171723</td>
<td>705.2304165</td>
<td>-5.165931075</td>
<td>6.492E-05</td>
<td>-5124.805849</td>
</tr>
<tr>
<td>19</td>
<td>Temp^2</td>
<td>23.30035581</td>
<td>4.053196314</td>
<td>5.748637374</td>
<td>1.89343E-05</td>
<td>14.78490634</td>
</tr>
</tbody>
</table>
Additional Information

- Interaction effects
- Subset selection
- LASSO (least absolute shrinkage and selection operator)
- Generalized linear models and logistic regression