Assignments and Announcements

- Assignments:
  - Textbook Assignment #10 due Friday, April 19: Page 518, problems 6 and 8. Be sure to interpret the results in terms of the problem.
  - This week’s assignment is short to give you more time to work on Minitab Lab #5 (due April 26). The lab assignment is available on the ST204 webpage.

- Announcements:
  - Omit pages 446-455 from reading.
  - This week we will discuss a few details about multiple regression. However, no reading is required from Chapter 13.

Chapter Twelve - Part 3: Regression, continued

GOALS
When you have completed this chapter, you will be able to:

ONE
Interpret Minitab output for a regression analysis (simple and multiple regression)

TWO
Make predictions for a multiple regression model

THREE
Interpret the coefficient of determination for a variety of situations

Example 1 continued: Minitab output

Regression Analysis: PRICE versus PAGES

The regression equation is
PRICE = -4.41 + 0.121 PAGES

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-4.412</td>
<td>4.403</td>
<td>-1.00</td>
<td>0.355</td>
</tr>
<tr>
<td>PAGES</td>
<td>0.120672</td>
<td>0.007016</td>
<td>17.20</td>
<td>0.000</td>
</tr>
</tbody>
</table>

S = 2.706 R-Sq = 98.0% R-Sq(adj) = 97.7%

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>2166.1</td>
<td>2166.1</td>
<td>295.82</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>6</td>
<td>43.9</td>
<td>7.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td>2210.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Example 1, Questions on the Minitab output

1. What is value of the coefficient of determination?

2. What are the values of the slope and intercept?

3. What is the value of the correlation?

4. What is $s_{Y|X}$?
Example 1, questions continued

5. What is the interpretation of the p-value for “constant”?

6. What is the interpretation of the p-value for “PAGES”?

7. What is the interpretation of the p-value for the ANOVA?

Multiple Regression Analysis

- Multiple regression analysis can be useful when you want to use several independent variables to predict a dependent variable.
- The general multiple regression equation for \( k \) independent variables is given by:
  \[
  Y' = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k
  \]
- As before, the least squares criterion is used to determine the formulas for \( a, b_1, b_2, b_k \). We will use Minitab to perform these calculations.

Coefficient of Multiple Determination

- For multiple regression, the coefficient of (multiple) determination is called \( R^2 \).
- Properties of \( R^2 \):
  - \( R^2 \) is the proportion of variability in the response explained by the linear regression relationship between \( Y \) and \( X_1, X_2, \ldots, X_k \).
  - \( R^2 \) is between 0 and 1.
  - As the number of independent variables increases, \( R^2 \) increases.
- Details on the computation of \( R^2 \) for multiple regression will not be required for ST204. If you want more information, see the textbook.

Example 2

- A handout will be provided in class with the example.
- Questions:
  1. Which variable has the strongest correlation with the response?
  2. What is the regression equation?
  3. What is value of the coefficient of determination?
Example 2, continued

4. If we refit the model using only the first predictor, what can you say about the coefficient of determination for this new model?

5. What is the interpretation of the p-value for the ANOVA?

6. What is the predicted value of the response for

7. Do the residual plots suggest any violations of the assumptions?

Chapter 14: Nonparametric Methods: Chi-Square Applications

GOALS
When you have completed this chapter, you will be able to:

ONE
List the characteristics of the Chi-Square Distribution.

TWO
Conduct and interpret a goodness of fit test for equal cell frequencies

Hypothesis tests for categorical data

- Most techniques presented in previous chapters are for numerical data.
- What if the variable of interest is categorical?
- Examples:
  - Does the color of a car influence the chance that it will be stolen?
  - Is there a relationship between undergraduate class and illegal drug use?

Two Hypothesis tests for categorical data

- Goodness of fit tests: used to determine how well an observed set of data fits an assumed (expected) set of data.

- Contingency table analysis: used to investigate the association between two categorical variables in a single population.
EXAMPLE 1

- The following data on absenteeism was collected from a manufacturing plant. We wish to test at the .05 level of significance whether there is a difference in the absence rate by day of the week.

<table>
<thead>
<tr>
<th>Day</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monday</td>
<td>120</td>
</tr>
<tr>
<td>Tuesday</td>
<td>45</td>
</tr>
<tr>
<td>Wednesday</td>
<td>60</td>
</tr>
<tr>
<td>Thursday</td>
<td>90</td>
</tr>
<tr>
<td>Friday</td>
<td>130</td>
</tr>
</tbody>
</table>

Goodness-of-Fit Test: Equal Expected Frequencies

- Let $f_o$ be the frequency observed in the sample for each category.
- Let $f_e$ be the frequency that is expected if the categories are equally likely in the population.
- $H_0$: Categories are equally likely.
- $H_1$: Categories are not equally likely.
- The test statistic is:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

- The critical value is has a Chi-square distribution with $(k-1)$ degrees of freedom, where $k$ is the number of categories, provided expected frequencies are not too small.

Characteristics of the Chi-Square Distribution

- The major characteristics of the chi-square distribution are:
  - It is positively skewed.
  - Chi-square values are always $\geq 0$
  - The shape of the chi-square distribution is determined by the degrees of freedom.
EXAMPLE 1 continued

- Step 1:
  \( H_0: \)
  \( H_1: \)
- Step 2: \( \alpha = .05 \)
- Step 3: Test statistic:
  \[ X^2 = \sum \left( \frac{(f_o - f_e)^2}{f_e} \right) \]

Assuming that all five categories are equally likely, we should expect about 1/5 of the sample values to occur in each category. So the expected frequency for each of the categories is:

Example 2

- A group of department store buyers viewed a new line of shoes and gave their opinions of them. The results were:

<table>
<thead>
<tr>
<th>Opinion</th>
<th>Number of Buyers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>47</td>
</tr>
<tr>
<td>Good</td>
<td>45</td>
</tr>
<tr>
<td>Fair</td>
<td>40</td>
</tr>
<tr>
<td>Undesirable</td>
<td>39</td>
</tr>
</tbody>
</table>

Example 2, continued

- Because the largest number (47) of buyers indicated the new line is excellent, the head buyer thinks that this is a mandate to go into mass production of the shoes. A new buyer believes that slight differences among the various counts are probably due to chance.

- Test the null hypothesis that the buyers’ opinions are not distributed evenly among the four categories. Use a significance level of 0.01.
Example 2, continued

- Step 1:
- Step 2:
- Step 3:

Example 2, continued

- Step 4:
- Step 5: