Example of Multivariate Regression (Cont.).
Use of the Chi-Square distribution (Multinomial Experiments and Contingency Tables).
Friday May 29th, 2015
Example: Corn production in Iowa

- Sec 9.5 # 6.

- $Y =$ Corn production (millions of bushels)
- $X_1 =$ Annual Precipitation (inches)
- $X_2 =$ Average annual temperature ($^\circ$F)
- $X_3 =$ Acres harvested (thousands of acres)
Linear Regression Analysis of Variance (using R)

- **Analysis of Variance Table**

- Response: Y

  - Df  | Sum Sq  | Mean Sq  | F      | P-value |
  - X1  | 1      | 116092   | 116092 | 1.045   | 0.3366  |
  - Residuals  | 8      | 888704   | 111088 |

*F test for the significance of the linear regression equation:*

\[
F = \frac{\text{explained variation} / \text{number of independent variables}}{\text{unexplained variation} / \text{df Residuals}} = \frac{116092/1}{888704/8} = 1.045
\]

- If P-value > alpha the regression equation does not have good overall significance. Since P-value > 0.05 the linear regression does not have a good overall significance.
Linear regression: Y vs. X3

- \( r = 0.8924660 \)

- Formal test of significance of the linear correlation coefficient
  - \( H_0: \rho = 0 \)
  - \( H_1: \rho \neq 0 \)

- The test statistic is (t Student with 10-2=8 d.f.)
  \[
  t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.892}{\sqrt{\frac{1-0.892^2}{10-2}}} = 5.580
  \]

- The critical values are found in Table A-3 using df = 10 – 2, and are
  \( t_{0.025} = \pm 2.306 \).
- Since the absolute value of the test statistic is greater than 2.306, we reject the null hypothesis. There is sufficient sample evidence to conclude that there is significant linear correlation between acres harvested and corn production amounts.
Regression equation (using R package)

- \( \hat{y} = -710.31 + 0.177 \times 3 \)
- Coefficient of Determination: \( r^2 = \frac{\text{Explained variation}}{\text{Total variation}} = 0.7965 \)
- Near 80% of the variation in Corn production is explained by this linear relationship with acres harvested.
- Best predicted amount of corn production for a year in which the amount of acres harvested is 13,300 thousands of acres is:
  \[ \hat{y} = -710.31 + 0.177 \times 13300 = 1644.989 \text{ (millions of bushels)} \]
Residuals diagnostic

- Residuals vs Fitted
- Normal Q-Q
- Scale-Location
- Residuals vs Leverage

Fitted values
Theoretical Quantiles
Leverage

Standardized residuals
Leverage
Cook's distance
Multiple Regression (Using R)

- Coefficients:
  - (Intercept)    -3102.7029  1749.7344  -1.773  0.126 55
  - X1             10.1878        7.3698    1.382     0.21611
  - X2             39.8253      32.4125    1.229     0.26519
  - X3              0.1859        0.0314    5.920     0.00103 **

- Residual standard error: 144.8 on 6 degrees of freedom
- Multiple R-squared: 0.8747, Adjusted R-squared: 0.8121
- F-statistic: 13.96 on 3 and 6 DF, p-value: 0.004094
Regression Equation

- \( \hat{y} = -3102.7 + 10.18X_1 + 39.82X_2 + 0.1859X_3 \)
  (3 independent variables)

- \( R^2 \): Multiple coefficient of determination: How well the multiple regression equation fits the sample data.

- Adjusted \( R^2 \) = is the multiple coefficient of determination modified to account for the number of variables and the sample size:

\[
\text{Adjusted } R^2 = 1 - \frac{(n-1)}{[n-(k+1)]} (1 - R^2)
\]

- In this example:

  Multiple R-squared: 0.8747,     Adjusted R-squared: 0.8121
  and P-value < 0.05. The regression equation has a good overall significance
Residuals diagnostic

- **Residuals vs Fitted**: Shows the residuals plotted against the fitted values. The plot indicates the residuals are generally around zero, suggesting no systematic pattern.
- **Normal Q-Q**: Displays the standardized residuals against theoretical quantiles. The points follow a straight line, indicating the residuals are normally distributed.
- **Scale-Location**: Plots the square root of the standardized residuals against the fitted values. The plot helps assess if the variance is constant across the fitted values.
- **Residuals vs Leverage**: Illustrates the residuals against leverage, with Cook's distance superimposed. Points with high leverage and high Cook's distance may indicate influential observations.
Best models and other extensions

• You would need other methods for model comparisons to select the best regression model and the more influential independent variables. One of these methods is called the stepwise regression.

• If your dependent variable is a dichotomous variable (e.g. positive/negative; cured/not cured and so on) we must use a different method known as logistic regression.
Use of the $\chi^2$-distribution

- **Goodness-of-fit test:**
  Determine whether a distribution of a sample data agrees with or fits some claimed distribution

- **Independence test:**
  Test independence between the row variable and the column variable in a contingency table
Example of a Contingency Table

- Gender and Handedness: 100 individuals randomly sampled from a population

<table>
<thead>
<tr>
<th></th>
<th>Right-handed</th>
<th>Left-handed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td>43</td>
<td>9</td>
<td>52</td>
</tr>
<tr>
<td>Females</td>
<td>44</td>
<td>4</td>
<td>48</td>
</tr>
<tr>
<td>Totals</td>
<td>87</td>
<td>13</td>
<td>100</td>
</tr>
</tbody>
</table>
Multinomial Experiment

1) The number of trials is fixed.
2) The trials are independent.
3) All outcomes of each trial must be classified into exactly one of several categories.
4) The probabilities for the different categories remain constant for each trial.

Example: Drivers who had a car crash in the last year categorized by age. Total no. of Drivers: 88

<table>
<thead>
<tr>
<th>AGE</th>
<th>&lt;25</th>
<th>25-44</th>
<th>45-64</th>
<th>&gt;64</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRIVERS</td>
<td>36</td>
<td>21</td>
<td>12</td>
<td>19</td>
</tr>
</tbody>
</table>
Claim: The observed frequencies fit some given distribution

Example: Since all the ages have the same crash rate, we expect these categories to have a number of subjects which match the age distribution of licensed drivers:

<table>
<thead>
<tr>
<th>AGES</th>
<th>&lt;25</th>
<th>25-44</th>
<th>45-64</th>
<th>&gt;64</th>
</tr>
</thead>
<tbody>
<tr>
<td>%Licensed Drivers</td>
<td>16%</td>
<td>44%</td>
<td>27%</td>
<td>13%</td>
</tr>
</tbody>
</table>

We will perform a goodness-of-fit test to check if the observed frequencies fits this distribution.
Notation

- O: Observed frequency of an outcome
- E: Expected frequency of an outcome
- k: # of different categories
- n: Total # of trials

In this example: k=4
n=88

**Expected frequencies:**
E=n×p, where p is the claimed proportion for each category. n is the number of trials.
\( \chi^2 \) Goodness-of-fit test

- Hypothesis test:
  Ho: \( p_1=0.16, \ p_2=0.44, \ p_3=0.27, \ p_4=0.13 \) (Claim)
  H1: At least one of the above proportions is different from the claim value

- Calculation of Expected Frequencies

<table>
<thead>
<tr>
<th>( k )</th>
<th>( E )</th>
<th>( O )</th>
<th>( O-E )</th>
<th>( (O-E)^2/E )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( 88 \times 0.16 = 14.08 )</td>
<td>36</td>
<td>21.92</td>
<td>34.12545</td>
</tr>
<tr>
<td>2</td>
<td>( 88 \times 0.44 = 38.72 )</td>
<td>21</td>
<td>-17.74</td>
<td>14.9523</td>
</tr>
<tr>
<td>3</td>
<td>( 88 \times 0.27 = 23.76 )</td>
<td>12</td>
<td>-11.76</td>
<td>5.820606</td>
</tr>
<tr>
<td>4</td>
<td>( 88 \times 0.13 = 11.44 )</td>
<td>19</td>
<td>7.56</td>
<td>4.995944</td>
</tr>
</tbody>
</table>
**χ² Goodness-of-fit test (Cont.)**

- Test Statistic:

  \[ \chi^2 = \sum \frac{(O-E)^2}{E} = 34.12545 + 14.9523 + 5.820606 + 4.995944 \]
  \[ = 59.89431 \text{ (with k-1 degrees of freedom)} \]

- Using \( \alpha = 0.05 \), Critical value is \( \chi^2 = 7.815 \).

Since \( 59.89431 > 7.815 \) **Reject Ho**. (We can also use the p-Value)

**Conclusion**: Not a good fit with the assumed distribution
\( \chi^2 \) test for Independence

- Data comes in the form of a contingency table: Two-way frequency table with 2 variables:
  - One variable has categories in the rows of the table
  - Other variable has categories in the columns of the table
- Example (2\( \times \)2 Table): Effectiveness of Bicycle Helmets

<table>
<thead>
<tr>
<th></th>
<th>HELMET WORN</th>
<th>NO HELMET</th>
</tr>
</thead>
<tbody>
<tr>
<td>FACIAL INJURIES</td>
<td>30</td>
<td>182</td>
</tr>
<tr>
<td>ALL OTHER INJURES (NON-FACIAL)</td>
<td>83</td>
<td>236</td>
</tr>
</tbody>
</table>

- Total: 531
χ² test for Independence (Cont.)

- Ho: There is no association between the row variable and the column variable (They are Independent) (Claim)
- H1: There is an association between row and column variables

**Important requirement:**
- For every cell in the contingency table, the expected frequency E is at least 5.
- **Test Statistic:**
  \[ \chi^2 = \sum \frac{(O-E)^2}{E} \]
  with \((r-1)\times(c-1)\) degrees of freedom.  \( r \) is the number of rows. \( c \) is the number of columns.
\( \chi^2 \) test for Independence (Cont.)

- If the type of injury is independent of the Helmet use we had:
- \( P(ri & cj) = P(ri) \times P(cj) \) where \( ri = \text{Row } i; cj = \text{Column } j \)

**For example:**
\[
P(\text{Facial Injury} & \text{No Helmet}) =
\]
\[
P(\text{Facial Injury}) \times P(\text{no Helmet}) \text{ (if we assumed ind.)} =
\]
\[
\frac{30+182}{531} \times \frac{182+236}{531} = 0.31
\]

- Expected Frequency for that cell (Row 1, Column 2) =
\[
531 \times 0.31 = 166.88
\]
\( \chi^2 \) test for Independence (Cont.)

- In General: Expected Frequency for a Cell is

\[
E = \frac{(Row \ Total) \times (Column \ Total)}{(Grand \ Total)}
\]

<table>
<thead>
<tr>
<th></th>
<th>Helmet</th>
<th>No Helmet</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facial Injury</td>
<td>30 (45.11)</td>
<td>182 (166.88)</td>
<td>212</td>
</tr>
<tr>
<td>Non-Facial Injury</td>
<td>83 (67.88)</td>
<td>236 (251.11)</td>
<td>319</td>
</tr>
<tr>
<td>Col. Total</td>
<td>113</td>
<td>418</td>
<td>531</td>
</tr>
</tbody>
</table>

Numbers in parentheses are Expected Frequencies
\( \chi^2 \) test for Independence (Cont.)

Calculation of Expected Frequencies:

- \( \frac{113 \times 212}{531} = 45.11 \)
- \( \frac{113 \times 319}{531} = 67.88 \)
- \( \frac{418 \times 212}{531} = 166.88 \)
- \( \frac{418 \times 319}{531} = 251.11 \)

\[
\chi^2 = \frac{(30 - 45.11)^2}{45.11} + \frac{(83 - 67.88)^2}{67.88} + \frac{(182 - 166.88)^2}{166.88} + \frac{(236 - 251.11)^2}{251.11} = 10.708
\]
The number of degrees of freedom is:
\[(2-1) \times (2-1)=1\]

For \( \alpha=0.05 \) the critical value is \( \chi^2_{0.05}=3.841 \)

Since \( 10.708>3.841 \), we Reject Ho, i.e., we reject the hypothesis of independence.

Sample data does not provide enough evidence to assure that getting or not a facial injury is independent of wearing or not wearing a helmet. Wearing a Helmet seems to be effective in helping to prevent facial injuries.

Note that the expected frequencies for each cell are >5.

If this condition is not satisfied, statisticians use the Fisher exact test (see page 513 of Textbook).