AMS7: WEEK 9. CLASS 1

More on Regression
Wednesday May 27th, 2015
Residual Analysis

- Residual difference: \( (y - \hat{y}) \): Observed - Predicted
This is the difference between an observed value \( y \) – the value \( \hat{y} \) (value of \( y \) using the regression equation)

**EXAMPLE:**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>1</th>
<th>3</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( y )</td>
<td>2</td>
<td>8</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>( \hat{y} )</td>
<td>5.273</td>
<td>5.273</td>
<td>4.909</td>
<td>4.545</td>
</tr>
<tr>
<td>( y - \hat{y} )</td>
<td>-3.273</td>
<td>2.727</td>
<td>1.091</td>
<td>-0.545</td>
</tr>
</tbody>
</table>
Residual Plot

Scatterplot Residuals \((y - \hat{y})\) vs. \(x\)
Points \((x, y - \hat{y})\)

Scatterplot of \(y\) vs. \(x\)
Points \((x, y)\)
Residual Sum of Squares (RSS)

• RSS = \( \sum (y - \hat{y})^2 \)
• In the example RSS = \((-3.273)^2 + (2.727)^2 + (1.091)^2 + (-0.545)^2 = 19.636\)

Least Square Property
The regression line satisfies the least-squares property. This is the line with the smallest Residual Sum of Squares (RSS). It is the straight line that fits data “best”.

Analysis of the Residuals

- Make a residual plot. This is a scattered plot with points \((x, y - \hat{y})\).

- There is a systematic pattern in the Residual Plot
  The regression equation is **not good**.

- No systematic pattern in the Residual Plot
  The regression equation **is good**.
Coefficient of Determination \((r^2)\)

- Amount of Variation in \(y\) explained by the regression line

\[
r^2 = \frac{\text{Explained Variation}}{\text{Total Variation}}
\]

- \(\text{Total Variation} = \text{Explained Var.} + \text{Unexplained Var.}\)

\[
\sum (y - \bar{y})^2 = \sum (\hat{y} - \bar{y})^2 + \sum (y - \hat{y})^2
\]

- We can also use the linear correlation coefficient and square it.
Prediction Interval

- It is a confidence interval for a predicted value of $y$

- We need to have a measure of the spread of the sample points about the regression line (unexplained deviation).

- This quantity is the standard error of estimate $(Se)$

- $Se = \sqrt{\frac{\sum (y-\hat{y})^2}{n-2}}$

- $Se = \sqrt{\frac{\sum y^2-b_0 \sum y-b_1 \sum x.y}{n-2}}$
Example: Length/weight of bears

- $n=8$
- $\sum y = 2176$
- $\sum y^2 = 728,520$
- $\sum x \cdot y = 151,879$
- $b_0 = -351.660$
- $b_1 = 9.6598$

$Se = \sqrt{\frac{728,520 - (-351.66)(2176) - (9.6598)(151,879)}{8-2}}$

= 66.6 (rounded)
Prediction interval for an individual $y$

- $\hat{y} - E < y < \hat{y} + E$

- **$E$: Margin of Error**

\[
E = t_{\alpha/2} \times Se \sqrt{1 + \frac{1}{n} + \frac{n(Xo - \bar{X})^2}{n (\sum X^2) - (\sum X)^2}}
\]

Where:

- **$Xo$:** A given value of $X$
- **$Se$:** Standard error of estimate
- **$t_{\alpha/2}$**: $t$ Student critical value with $n-2$ degrees of freedom
Example: Prediction Interval

• Find a 95% prediction interval for the weight of a bear corresponding to a length of 74 in.
• Point prediction was: $\hat{y} = 362.84$
• $n=8; \bar{X} = 64.562 ; \sum X = 516.5 ; \sum X^2 = 34,525.75$
• From Table A-3: $t_{\alpha/2} = 2.447$ (Use $8-2=6$ d.f.)

$$E = 2.447 \times 66.6 \times \sqrt{1 + \frac{1}{8} + \frac{8(74-64.5625)^2}{8 (34,525.75)- 516.5^2}} = 178.56$$

• $\hat{y} - E < y < \hat{y} + E$
  $362.84 - 178.56 < y < 362.84 + 178.56$
• 95% Confidence Prediction Interval
  $184.28 < y < 541.4$
Multiple Regression

- Linear association between a dependent variable $y$ and two or more independent variables ($X_1, X_2, \ldots, X_k$)

\[
\hat{y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k
\]

- $b_0$: y-intercept
- $b_1, \ldots, b_k$: regression coefficients (sample estimates)
- $\beta_0$: y-intercept (from population)
- $\beta_1, \ldots, \beta_k$: Population parameters
Example: Corn production prediction

- \( Y \) = Corn production in a given year (millions of bushels)
- \( X_1 \) = Annual precipitation (in.)
- \( X_2 \) = Average annual temperature (°F)

- \( \hat{y} = b_0 + b_1 X_1 + b_2 X_2 \)
- \( b_0 \) = y intercept
- \( b_1 \) = increase in crop production per 1 in. increase of mean annual rainfall
- \( b_2 \) = increase in crop production per 1 °F increase in mean annual temperature
- We normally use a computer program to get the values of \( b_0, b_1, \ldots, b_k \)
### DATA (Page 484)

<table>
<thead>
<tr>
<th>Year</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precipitation</td>
<td>32.4</td>
<td>41.8</td>
<td>36.5</td>
<td>37.5</td>
<td>31.6</td>
<td>40.3</td>
<td>33.3</td>
<td>21.6</td>
<td>24.7</td>
<td>39.4</td>
</tr>
<tr>
<td>Average Temperature</td>
<td>47.80</td>
<td>46.93</td>
<td>48.47</td>
<td>48.28</td>
<td>46.69</td>
<td>49.31</td>
<td>51.87</td>
<td>49.34</td>
<td>47.06</td>
<td>49.88</td>
</tr>
<tr>
<td>Corn Production</td>
<td>1731</td>
<td>1578</td>
<td>744</td>
<td>1445</td>
<td>1707</td>
<td>1627</td>
<td>1320</td>
<td>899</td>
<td>1446</td>
<td>1562</td>
</tr>
<tr>
<td>Acres harvested</td>
<td>13,850</td>
<td>13,150</td>
<td>8,550</td>
<td>12,900</td>
<td>13,500</td>
<td>12,050</td>
<td>10,150</td>
<td>10,700</td>
<td>12,250</td>
<td>12,400</td>
</tr>
</tbody>
</table>
The strongest linear association seems to be between Y and X3. No strong association can be observed between the independent variables.
Pairwise Coefficients of linear correlation (Correlation Matrix)

<table>
<thead>
<tr>
<th></th>
<th>Y</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>1.00000000</td>
<td>0.33990910</td>
<td>-0.15641293</td>
<td>0.8924660</td>
</tr>
<tr>
<td>X1</td>
<td>0.33990910</td>
<td>1.00000000</td>
<td>0.02673135</td>
<td>0.1411452</td>
</tr>
<tr>
<td>X2</td>
<td>-0.15641293</td>
<td>0.02673135</td>
<td>1.00000000</td>
<td>-0.3783065</td>
</tr>
<tr>
<td>X3</td>
<td>0.8924660</td>
<td>0.14114519</td>
<td>-0.37830647</td>
<td>1.00000000</td>
</tr>
</tbody>
</table>
Linear regression: Y vs. X1

- $r = 0.34$
- 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.)
  $$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.34}{\sqrt{\frac{1-0.34^2}{10-2}}} = 1.022$$

- 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 less than 2.306, we fail to reject the null hypothesis. There is not sufficient sample evidence to conclude that there is significant linear correlation between annual precipitation amounts and corn production amounts.
Regression equation (using R package)

- \( \hat{y} = 826.15 + 17.10 \times X_1 \)
- Coefficient of Determination: \( r^2 = \frac{\text{Explained variation}}{\text{Total variation}} = 0.1155 \)
- Only 11.55% of the variation in Corn production is explained by this linear relationship with rainfall.
- Best predicted amount of corn production for a year in which the precipitation amount is 29.3 inches is the sample mean of \( Y \) (\( \bar{y} \)). This value is 1405.9. In this case the regression line is not a good model.

Remember:
- Total Variation = Explained variation + Unexplained variation
- \( \sum (y - \bar{y})^2 = \sum (\hat{y} - \bar{y})^2 + \sum (y - \hat{y})^2 \)