Math 1710 Class 23

Regression, Experimentation
Dr. Back

Oct. 21, 2009
Major Changes in Your Next Homework Assignment

Please check the website.
The slope $b_1$ of $\hat{\text{Fare}} = b_1 \text{Distance} + b_0$ is the average increase in fare per extra mile.

$\hat{\text{Fare}} = 177 + .079 \cdot \text{Distance}$ and $\hat{\text{Distance}} = -644 + 6.13 \cdot \text{Fare}$ are different lines!

(Note $\frac{1}{.079} \neq 6.13$.)

If you want to compute $r$ on a TI-83/84, the place to look is stat $\rightarrow$ calc $\rightarrow$ linreg. And ONCE, you need to set DiagnosticsOn in the Catalog.
\[ \text{Var}(d_i) = (1 - r^2) \text{Var}(y_i) \]

Sample Variance: \[ \text{Var}(x_i) = \frac{1}{n-1} \sum (x_i - \bar{x})^2 \]
Math 1710  
Class 23  

\[ \text{Var}(d_i) = (1 - r^2) \text{Var}(y_i) \]

Sample Variance: \( \text{Var}(x_i) = \frac{1}{n-1} \sum (x_i - \bar{x})^2 \)

For paired data \((x_i, y_i)\),

Sample Covariance: \( \text{Cov}(x_i, y_i) = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y}) \)
$Var(d_i) = (1 - r^2) Var(y_i)$

Sample Variance: $Var(x_i) = \frac{1}{n-1} \sum (x_i - \bar{x})^2$

For paired data $(x_i, y_i)$,

Sample Covariance: $Cov(x_i, y_i) = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})$

So $r = \frac{Cov(x_i, y_i)}{s_x s_y}$
\[
\text{Var}(d_i) = (1 - r^2) \text{Var}(y_i)
\]

Sample Variance: \( \text{Var}(x_i) = \frac{1}{n-1} \sum (x_i - \bar{x})^2 \)

For paired data \((x_i, y_i)\),

Sample Covariance: \( \text{Cov}(x_i, y_i) = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y}) \)

So \( r = \frac{\text{Cov}(x_i, y_i)}{s_x s_y} \)

Because \((a + b)^2 = a^2 + 2ab + b^2\)

\[
\text{Var}(a_i + b_i) = \text{Var}(a_i) + \text{Var}(b_i) + 2\text{Cov}(a_i, b_i).
\]
\[
\text{Var}(d_i) = (1 - r^2) \text{Var}(y_i)
\]

Because \((a + b)^2 = a^2 + 2ab + b^2\)

\[
\text{Var}(a_i + b_i) = \text{Var}(a_i) + \text{Var}(b_i) + 2\text{Cov}(a_i, b_i).
\]

Apply this to:

\[
\text{Var}(d_i) = \text{Var}(y_i - \hat{y}_i)
\]
\[ \text{Var}(d_i) = (1 - r^2) \text{Var}(y_i) \]

Apply this to:

\[ \text{Var}(d_i) = \text{Var}(y_i - \hat{y}_i) \]

\[ \text{Var}(y_i - (\bar{y} + b_1(x_i - \bar{x}))) = \text{Var}((y_i - \bar{y}) - b_1(x_i - \bar{x})) \]

\[ = s_y^2 + b_1^2 s_x^2 - 2b_1 \text{Cov}(y_i - \bar{y}, x_i - \bar{x}) \]

\[ = s_y^2 + r^2 \frac{s_y^2}{s_x^2} s_x^2 - 2r \frac{s_y}{s_x} (rs_x s_y) \]

\[ = s_y^2 - r^2 s_y^2 \]

\[ = (1 - r^2) s_y^2. \]
Crime Rates by Locality

Crime Rates vs Housing Prices in Philadelphia 1996
Crime Rate is Crimes Per 1000
Housing Prices in Dollars
Crime Rates by Locality

scatterplot
Crime Rates by Locality

with regression line

\[ \hat{HP} = -577 \cdot CR + 177K \quad r^2 = .06 \]
# Crime Rates by Locality

A regression display is shown below:

```
Dependent variable is: HousePrice
No Selector
111 total cases of which 12 are missing
R squared = 6.2%   R squared (adjusted) = 5.3%
\( s = 84.33e3 \) with \( 99 - 2 = 97 \) degrees of freedom

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>45.9673e9</td>
<td>1</td>
<td>45.9673e9</td>
<td>6.46</td>
</tr>
<tr>
<td>Residual</td>
<td>689.739e9</td>
<td>97</td>
<td>7.11071e9</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>s.e. of Coeff</th>
<th>t-ratio</th>
<th>prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>176629</td>
<td>11.25e3</td>
<td>15.7</td>
<td>≤ 0.0001</td>
</tr>
<tr>
<td>CrimeRate</td>
<td>-576.908</td>
<td>226.9</td>
<td>-2.54</td>
<td>0.0126</td>
</tr>
</tbody>
</table>
```
Crime Rates by Locality

Residuals

crime rate vs residuals
Now analyze without the Center City Outlier. The point is we now get a very different regression line with a much higher $r^2$. Outliers matter!
Crime Rates by Locality

scatterplot
Crime Rates by Locality

with regression line

\[ \hat{HP} = -2290 \cdot CR + 225K \quad r^2 = .18 \]
Crime Rates by Locality

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1.34831e11</td>
<td>1</td>
<td>1.34831e11</td>
<td>21.7</td>
</tr>
<tr>
<td>Residual</td>
<td>5.97038e11</td>
<td>96</td>
<td>6.21914e9</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>s.e. of Coeff</th>
<th>t-ratio</th>
<th>prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>225234</td>
<td>1.64e4</td>
<td>13.7</td>
<td>≤ 0.0001</td>
</tr>
<tr>
<td>CrimeRate</td>
<td>-2288.69</td>
<td>491.5</td>
<td>-4.66</td>
<td>≤ 0.0001</td>
</tr>
</tbody>
</table>
Crime Rates by Locality

Residuals

![Residuals Graph](image)

Philadelphia Crime Rates

Proof of $\text{Var}(d_i)$
Now transform from CR to $\frac{1}{CR}$. The point here is that transforming your data to get a form more appropriate to the assumptions of simple linear regression is an important activity. We don’t cover it, but chapter 10 of BVD is all about this. This is just 1 example to raise the possibility in your mind.
Crime Rates by Locality

scatterplot
Crime Rates by Locality

with regression line

\[ \hat{HP} = 1.3M \cdot \frac{1}{CR} + 97.9K \quad r^2 = .17 \]

But Center City included.
Crime Rates by Locality

regression display

Dependent variable is: No Selector
111 total cases of which 12 are missing
R squared = 17.4%  R squared (adjusted) = 16.5%
s = 79.16e3 with 99 - 2 = 97 degrees of freedom

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>127.948e9</td>
<td>1</td>
<td>127.948e9</td>
<td>20.4</td>
</tr>
<tr>
<td>Residual</td>
<td>607.758e9</td>
<td>97</td>
<td>6.26555e9</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>s.e. of Coeff</th>
<th>t-ratio</th>
<th>prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>97920.6</td>
<td>15.46e3</td>
<td>6.33</td>
<td>≤ 0.0001</td>
</tr>
<tr>
<td>1/CrR</td>
<td>1.30138e6</td>
<td>288e3</td>
<td>4.52</td>
<td>≤ 0.0001</td>
</tr>
</tbody>
</table>
Crime Rates by Locality

Residuals

![Residuals Graph](image)