Simple Linear Regression

Lecture Outline

- Regression: A classical view
- Regression as an ANOVA
- Regression diagnostics
- Reading

Objectives of Regression

- The primary objective of regression is to develop a linear relationship between a response variable and a regressor for the purposes of prediction.

Objectives of Regression

- Regression is sometimes used to estimate the functional linear relationship between two variables. However, such an approach assumes that a functional linear relationship exists, and alternative approaches (functional regression) are superior.

Objectives of Regression

- Regression is sometimes interpreted as establishing causality. Only manipulative experiments can establish causality. Regression establishes a predictive relationship but does not on its own provide evidence of causality.

Example: Positive regression between Crime Rate and Number of Churches in cities.

Regression and Correlation

- Regression establishes a predictive relationship between response variable Y and regressor X. The regressor is fixed, that is, its values are under the control of the experimenter.
- Correlation provides a measure of how strongly two variables $X_1$ and $X_2$ vary together (covary). Both $X_1$ and $X_2$ are free to vary naturally.
Regression Terminology

- The **response variable** is sometimes called the **dependent variable**
- The **regressor** is sometimes called the **independent variable** or the **covariate**

Response Variable: Height (cm)
Regressor: Age (months)

Regression line is uniquely defined by its slope and its Y intercept

Statistical Regression

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Parameter</th>
<th>( \beta_0 )</th>
<th>( \beta_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slope</td>
<td>( B_1 )</td>
<td>( \beta_1 )</td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>( B_0 )</td>
<td>( \beta_0 )</td>
<td></td>
</tr>
<tr>
<td>Equation</td>
<td>( Y = B_0 + B_1x )</td>
<td>( Y = \beta_0 + \beta_1 )</td>
<td></td>
</tr>
</tbody>
</table>

The sample line is an estimate of a true but unknown parametric line

Regression as ANOVA

<table>
<thead>
<tr>
<th>Source</th>
<th>Df</th>
<th>SS</th>
<th>MS</th>
<th>F ratio</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Among Sites</td>
<td>9</td>
<td>99025.04</td>
<td>10992.73</td>
<td>31.76</td>
<td>0.0001</td>
</tr>
<tr>
<td>Regression</td>
<td>1</td>
<td>4812.70</td>
<td>4812.70</td>
<td>11.91</td>
<td>0.0005</td>
</tr>
<tr>
<td>Residual</td>
<td>88</td>
<td>1212.31</td>
<td>14.54</td>
<td>0.0001</td>
<td></td>
</tr>
<tr>
<td>Within</td>
<td>90</td>
<td>1143.00</td>
<td>12.7222</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>99</td>
<td>7176.44</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

More than one Y value for each X value
Regression as ANOVA

Mean Squares and what the estimate

<table>
<thead>
<tr>
<th>Source</th>
<th>MS</th>
<th>Df</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Among Sites</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regression</td>
<td>$\sigma^2 + \sigma_s^2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>$\sigma^2 + \sigma_r^2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within</td>
<td>$\sigma^2$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

F-test for variation Among Sites?
What does it mean?
F-test for Regression?
What does it mean?
F-test for Residual?
What does it mean?

One Y value for each X value

Regression Assumptions

- Heterogeneity of Variances
- Normality of errors

Residual Analysis

Homogeneity of Variances
Heterogeneity of Variances
Serial dependence
Influential aberrant outlier

Test of Normality
Curvilinear Regression

Pre-Transformation

Post-Transformation

Confidence Limits

Reading

- Module 6: Key concepts
- Regression as an ANOVA
- Worked Examples