Multiple Regression

- A single numerical response variable, Y.
- Multiple numerical explanatory variables, $X_1, X_2, ..., X_k$

Example

- Y, Response – Effectiveness score based on experienced teachers’ evaluations.
- Explanatory – Test 1, Test 2, Test 3, Test 4.

$Y = \mu_{Y|x_1,x_2,...,x_k} + \varepsilon$

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$
### Stat 401 B – Lecture 12

#### Table

<table>
<thead>
<tr>
<th>Student Teacher</th>
<th>Eval</th>
<th>Test1</th>
<th>Test2</th>
<th>Test3</th>
<th>Test4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>489</td>
<td>81</td>
<td>151</td>
<td>46</td>
<td>44</td>
</tr>
<tr>
<td>2</td>
<td>423</td>
<td>68</td>
<td>156</td>
<td>46</td>
<td>45</td>
</tr>
<tr>
<td>3</td>
<td>507</td>
<td>80</td>
<td>165</td>
<td>76</td>
<td>55</td>
</tr>
<tr>
<td>4</td>
<td>467</td>
<td>107</td>
<td>149</td>
<td>56</td>
<td>43</td>
</tr>
<tr>
<td>5</td>
<td>340</td>
<td>43</td>
<td>134</td>
<td>49</td>
<td>49</td>
</tr>
<tr>
<td>6</td>
<td>524</td>
<td>129</td>
<td>163</td>
<td>72</td>
<td>50</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>23</td>
<td>434</td>
<td>76</td>
<td>141</td>
<td>54</td>
<td>58</td>
</tr>
</tbody>
</table>

#### JMP

- Analyze – Fit Model
- Pick Role Variables
  - Y – EVAL
- Construct Model Effects
  - Add – Test1, Test2, Test3, Test4

- Analyze – Fit Model
  - Personality – Standard Least Squares
  - Emphasis – Minimal Report
Response EVAL

Summary of Fit

RSquare 0.802861
RSquare Adj 0.759052
Root Mean Square Error 37.53027
Mean of Response 444.4783
Observations (or Sum Wgts) 23

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4</td>
<td>103286.25</td>
<td>25821.6</td>
<td>16.3265</td>
<td>&lt;.0001*</td>
</tr>
<tr>
<td>Error</td>
<td>18</td>
<td>125361.49</td>
<td>7.02287</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>22</td>
<td>228647.74</td>
<td>12246.86</td>
<td>.307088</td>
<td></td>
</tr>
</tbody>
</table>

Parameter Estimates

| Term   | Estimate | Std Error | t Ratio | Prob>|t| |
|--------|----------|-----------|---------|-----|---|
| Intercept | -193.4994 | 1.540074 | -0.1399 | 0.8864 |
| Test1   | 1.1158539 | 3.489265 | 0.3026* | 0.7670 |
| Test2   | 2.243287 | 3.573303 | 0.628449 | 0.5334 |
| Test3   | -1.367001 | 2.422887 | 0.563965 | 0.5800 |
| Test4   | 6.0482367 | 5.029203 | 1.202281 | 0.2283 |

Prediction Equation

Predicted Evaluation = -193.50 + 1.116*Test1 + 2.243*Test2 - 1.367*Test3 + 6.048*Test4

Conditions

- The random error term, $\varepsilon$, is
  - Independent
  - Identically distributed
  - Normally distributed with standard deviation, $\sigma$. 
Estimate of Error Variance, $\sigma^2$

$$MS_{\text{Error}} = \frac{SS_{\text{Error}}}{df_{\text{Error}}}$$

$$MS_{\text{Error}} = \frac{\sum (y - \hat{y})^2}{n - (k + 1)}$$

$$MS_{\text{Error}} = \frac{25361.49}{18} = 1409.0$$

Estimate of Error Std Dev, $\sigma$

- Root Mean Square Error

$$RMSE = \sqrt{MS_{\text{Error}}}$$

$$RMSE = \sqrt{1409.0} = 37.54$$

Multiple $R^2$

$$R^2 = \frac{SS_{\text{Model}}}{SS_{\text{Total}}} = 1 - \frac{SS_{\text{Error}}}{SS_{\text{Total}}}$$

$$R^2 = \frac{103286.25}{128647.74} = 0.802861$$
Interpretation

- 80.3% of the variation in the evaluation scores can be explained by the model, i.e. the relationship with the explanatory variables.

Caution

- Including additional explanatory variables in a model can only increase the value of $R^2$, even if those explanatory variables have nothing to do with the response variable.

Adjusted $R^2$

$$adjR^2 = 1 - \frac{MS_{Error}}{MS_{Total}}$$

$$adjR^2 = 1 - \left( \frac{25361.49/18}{128647.74/22} \right) = 0.75905$$
Test of Model Utility

- Is there any explanatory variable in the model that is helping to explain significant amounts of variation in the response?

Step 1: Hypotheses

H₀ : \( \beta_1 = \beta_2 = \ldots = \beta_k = 0 \)
Hₐ : at least one parameter is not zero

Step 2: Test Statistic

\[
F = \frac{MS_{Model}}{MS_{Error}} = \frac{25821.6}{1409.0} = 18.3265
\]

P value < 0.0001
Step 3: Decision

- Reject the null hypothesis because the P-value is so small.

Step 4: Conclusion

- At least one of the tests is providing statistically significant information about the evaluation score.
- The model is useful. Maybe not the best, but useful.

Alternative Form

\[
F = \frac{\left( \frac{R^2}{k} \right)}{\left( \frac{1 - R^2}{n - (k + 1)} \right)}
\]

\[
F = \frac{0.802861}{4} = 0.197139
\]

\[
F = \frac{0.197139}{18} = 18.3265
\]

P-value < 0.0001