Multiple Regression

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

$Y = \mu_{Y|x_1,x_2,\ldots,x_k} + \epsilon$

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon$

Example

- $Y$, Response - Effectiveness score based on experienced teachers' evaluations.
- Explanatory - Test 1, Test 2, Test 3, Test 4.
<table>
<thead>
<tr>
<th>Student</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>45</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>507</td>
<td>80</td>
<td>165</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>467</td>
<td>107</td>
<td>149</td>
<td>66</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>
</tbody>
</table>

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

JMP
- Analyze - Fit Model
  - Personality - Standard Least Squares
  - Emphasis - Minimal Report
Summary of Fit

- RSquare: 0.802861
- R(Square Adj): 0.759052
- Root Mean Square Error: 37.53627
- 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>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4</td>
<td>25821.60</td>
<td>18.3265</td>
<td>&lt;.0001*</td>
</tr>
<tr>
<td>Error</td>
<td>18</td>
<td>1409.00</td>
<td>1409.00</td>
<td>Prob &gt; F</td>
</tr>
<tr>
<td>C. Total</td>
<td>22</td>
<td>212864.74</td>
<td></td>
<td>&lt;.0001*</td>
</tr>
</tbody>
</table>

Parameter Estimates

| Term       | Estimate | Std Error | t Ratio | Prob>|t| |
|------------|----------|-----------|---------|-----|
| Intercept  | -193.4994| 129.3074  | -1.54   | 0.1394|
| Test1      | 1.114829 | 3.49      | 0.0026* | 0.9999* |
| Test2      | 2.243287 | 3.57      | 0.0022  | 0.9999* |
| Test3      | -1.367001| 0.623986  | -2.24   | 0.0261* |
| Test4      | 6.048236 | 1.023281  | 5.93    | <.0001* |

Prediction Equation

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

Conditions

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

$$MS_{Error} = \frac{SS_{Error}}{df_{Error}}$$
$$MS_{Error} = \frac{\sum (y - \hat{y})^2}{n - (k + 1)}$$
$$MS_{Error} = \frac{25361.49}{18} = 1409.0$$

Estimate of Error Std Dev, $\sigma$

Root Mean Square Error

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

Multiple $R^2$

$$R^2 = \frac{SS_{Model}}{SS_{Total}} = 1 - \frac{SS_{Error}}{SS_{Total}}$$
$$R^2 = \frac{103286.25}{128647.74} = 0.802861$$
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 $R^2$

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

$$adjR^2 = 1 - \frac{25361.49/18}{128647.74/22} = 0.75905$$
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

\[ H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0 \]
\[ H_A : \text{at least one parameter is not zero} \]

Step 2: Test Statistic

\[ F = \frac{\text{MS}_{\text{Model}}}{\text{MS}_{\text{Error}}} \]
\[ F = \frac{25821.6}{1409.0} = 18.3265 \]

\[ P - \text{value} < 0.0001 \]
Step 3: Decision

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

Step 4: Conclusion

- At least one of the Test 1, Test 2, Test 3 or Test 4 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