Model Selection

- In multiple regression we often have many explanatory variables.
- How do we find the “best” model?

Model Selection

- How can we select the set of explanatory variables that will explain the most variation in the response and have each variable adding significantly to the model?

Cruising Timber

- Response: Mean Diameter at Breast Height (MDBH) of a tree.
- Explanatory:
  - $X_1$ = Mean Height of Pines
  - $X_2$ = Age of Tract times the Number of Pines
  - $X_3$ = Mean Height of Pines divided by the Number of Pines
Forward Selection

- Begin with no variables in the model.
- At each step check to see if you can add a variable to the model.
  - If you can, add the variable.
  - If not, stop.

Forward Selection – Step 1

- Select the variable that has the highest correlation with the response.
- If this correlation is statistically significant, add the variable to the model.

JMP

- Multivariate Methods
- Multivariate
- Put MDBH, X₁, X₂, and X₃ in the Y, Columns box.
Comment

- The explanatory variable $X_3$ has the highest correlation with the response MDBH.
  - $r = 0.8404$
- The correlation between $X_3$ and MDBH is statistically significant.
  - Signif Prob < 0.0001, small P-value.
Step 1 - Action

- Fit the simple linear regression of MDBH on $X_3$.
- Predicted MDBH = 3.896 + 32.937*$X_3$
- $R^2 = 0.7063$
- RMSE = 0.4117

SLR of MDBH on $X_3$

- Test of Model Utility
  - $F = 43.2886$, P-value < 0.0001
- Statistical Significance of $X_3$
  - $t = 6.58$, P-value < 0.0001
- Exactly the same as the test for significant correlation.

Can we do better?

- Can we explain more variation in MDBH by adding one of the other variables to the model with $X_3$?
- Will that addition be statistically significant?
Forward Selection - Step 2
- Which variable should we add, $X_1$ or $X_2$?
- How can we decide?

Correlation among explanatory variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>MDBH</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>MDBH</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDBH</td>
<td>1.0</td>
<td>0.7731</td>
<td>0.7546</td>
<td>0.8404</td>
<td>0.6345</td>
<td>0.0557</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X1</td>
<td>0.7731</td>
<td>1.0</td>
<td>0.2435</td>
<td>0.7546</td>
<td>0.6345</td>
<td>0.0557</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X2</td>
<td>0.7546</td>
<td>0.2435</td>
<td>1.0</td>
<td>0.8404</td>
<td>0.0557</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X3</td>
<td>0.8404</td>
<td>0.7546</td>
<td>0.8404</td>
<td>1.0</td>
<td>0.0557</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Signif Prob: <.0001* 0.3008 0.0001* 0.0027* 0.8156

Multicollinearity
- Because some explanatory variables are correlated, they may carry overlapping information about the response.
- You can't rely on the simple correlations between explanatory and response to tell you which variable to add.
Forward selection - Step 2
- Look at partial residual plots.
- Determine statistical significance.

Partial Residual Plots
- Look at the residuals from the SLR of $Y$ on $X_3$ plotted against the other variables once the overlapping information with $X_3$ has been removed.

How is this done?
- Fit MDBH versus $X_3$ and obtain residuals - Resid($Y$ on $X_3$)
- Fit $X_1$ versus $X_3$ and obtain residuals - Resid($X_1$ on $X_3$)
- Fit $X_2$ versus $X_3$ and obtain residuals - Resid($X_2$ on $X_3$)
Correlations

<table>
<thead>
<tr>
<th></th>
<th>Resid(YonX3)</th>
<th>Resid(X1onX3)</th>
<th>Resid(X2onX3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resid(YonX3)</td>
<td>1.0000</td>
<td>0.5726</td>
<td>0.3636</td>
</tr>
<tr>
<td>Resid(X1onX3)</td>
<td>0.5726</td>
<td>1.0000</td>
<td>0.9320</td>
</tr>
<tr>
<td>Resid(X2onX3)</td>
<td>0.3636</td>
<td>0.9320</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Comment

- The residuals (unexplained variation in the response) from the SLR of MDBH on $X_3$ have the highest correlation with $X_1$ once we have adjusted for the overlapping information with $X_3$. 
Statistical Significance

- Does $X_1$ add significantly to the model that already contains $X_3$?
  - $t = 2.88$, $P$-value $= 0.0104$
  - $F = 8.29$, $P$-value $= 0.0104$
  - Because the $P$-value is small, $X_1$ adds significantly to the model with $X_3$.

Summary

- Step 1 - add $X_3$
  - $R^2 = 0.706$
- Step 2 - add $X_1$ to $X_3$
  - $R^2 = 0.803$
  - Can we do better?

Forward Selection - Step 3

- Does $X_2$ add significantly to the model that already contains $X_3$ and $X_1$?
  - $t = -2.79$, $P$-value $= 0.0131$
  - $F = 7.78$, $P$-value $= 0.0131$
  - Because the $P$-value is small, $X_2$ adds significantly to the model with $X_2$ and $X_1$. 
Summary

- Step 1 – add $X_3$
  - $R^2 = 0.706$
- Step 2 – add $X_1$ to $X_3$
  - $R^2 = 0.803$
- Step 3 – add $X_2$ to $X_1$ and $X_3$
  - $R^2 = 0.867$

Summary

- At each step the variable being added is statistically significant.
- Has the forward selection procedure found the "best" model?

“Best” Model?

- The model with all three variables is useful.
  - $F = 34.83$, P-value < 0.0001
- The variable $X_3$ does not add significantly to the model with just $X_1$ and $X_2$.
  - $t = 0.41$, P-value = 0.6844