Forward Selection

- The Forward selection procedure looks to add variables to the model.
- Once added, those variables stay in the model even if they become insignificant at a later step.

Backward Selection

- The Backward selection procedure looks to remove variables from the model.
- Once removed, those variables cannot reenter the model even if they would add significantly at a later step.

Mixed Selection

- A combination of the Forward and Backward selection procedures.
- Starts out like Forward selection but looks to see if an added variable can be removed at a later step.
Mixed – Set up

Stepwise Fit
Stepwise Regression Control
Prob to Enter: 0.250
Prob to Leave: 0.250
Direction: Mixed

Stepwise Regression Control

- Direction – Mixed
- Prob to Enter – controls what variables are added.
- Prob to Leave – controls what variables are removed.
- Prob to Enter = Prob to Leave

Current Estimates
- The current estimates are exactly the same as with the Forward selection procedure.
- Clicking on Step will initiate the Mixed procedure that starts like the Forward procedure.
Stepwise Fit

Response: MDBH

Stepwise Regression Control

Prob to Enter: 0.250
Prob to Leave: 0.250

Direction: Mixed

Current Estimates

<table>
<thead>
<tr>
<th>SSE</th>
<th>DFE</th>
<th>MSE</th>
<th>RSquare</th>
<th>RSquare Adj</th>
<th>Cp</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.0902449</td>
<td>18</td>
<td>0.1694581</td>
<td>0.7063</td>
<td>0.6990</td>
<td>19.387747</td>
<td>-33.6102</td>
</tr>
</tbody>
</table>

Lock Entered Parameter Estimate NSP  SS "F Ratio" "Prob>F"

- Intercept: 3.8956688
- X1: 0 1 1.000159 8.294 0.0104
- X2: 0 1 0.403296 0.403296 0.1259
- X3: 32.9371533 1 7.335255 43.287 0.0000

Step History

<table>
<thead>
<tr>
<th>Step</th>
<th>Parameter</th>
<th>Action</th>
<th>&quot;Sig Prob&quot;</th>
<th>Seq SS</th>
<th>RSquare</th>
<th>Cp</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X3</td>
<td>Entered</td>
<td>0.0000</td>
<td>7.335255</td>
<td>0.7063</td>
<td>19.388</td>
<td>2</td>
</tr>
</tbody>
</table>

---

Current Estimates – Step 1

- X3 is added to the model
- Predicted MDBH = 3.896 + 32.937*X3
- R^2 = 0.7063
- RMSE = \sqrt{MSE} = \sqrt{0.1694581} = 0.4117

---

Current Estimates – Step 1

- By clicking on Step you will invoke the Backward part of the Mixed procedure.
- Because X3 is statistically significant and is the only variable in the model, clicking on Step will not do anything.
Current Estimates – Step 1
- Of the remaining variables not in the model, \( X_1 \) will add the largest sum of squares if added to the model.
  - \( SS = 1.000 \)
  - "F Ratio" = 8.294
  - "Prob>F" = 0.0104

JMP Mixed – Step 2
- Because \( X_1 \) will add the largest sum of squares and that addition is statistically significant, by clicking on Step, JMP will add \( X_1 \) to the model with \( X_3 \).
Current Estimates – Step 2

- $X_1$ is added to the model
- Predicted MDBH = 3.143 + 0.0314*$X_1$ + 22.954*$X_3$
- $R^2 = 0.8026$
- $RMSE = \sqrt{MSE} = \sqrt{0.1205933} = 0.3473$

Current Estimates – Step 2

- By clicking on Step you will invoke the Backward part of the Mixed procedure.
- Because $X_3$ and $X_1$ are statistically significant, clicking on Step will not do anything.

Current Estimates – Step 2

- Of the remaining variables not in the model $X_2$ will add the largest sum of squares if added to the model.
  - $SS = 0.671$
  - "F Ratio" = 7.784
  - "Prob>F" = 0.0131
JMP Mixed – Step 3

- Because $X_2$ will add the largest sum of squares and that addition is statistically significant, by clicking on Step, JMP will add $X_2$ to the model with $X_3$ and $X_1$.

**Stepwise Fit**

<table>
<thead>
<tr>
<th>Response: MDBH</th>
<th>Prob to Enter</th>
<th>Prob to Leave</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.250</td>
<td>0.250</td>
</tr>
</tbody>
</table>

**Stepwise Regression Control**

Current Estimates

<table>
<thead>
<tr>
<th>Step</th>
<th>Parameter</th>
<th>Action</th>
<th>&quot;Sig Prob&quot;</th>
<th>Seq SS</th>
<th>R^2</th>
<th>Cp</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Intercept</td>
<td>Entered</td>
<td>0.0000</td>
<td>7.33</td>
<td>0.75</td>
<td>19.38</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>X1</td>
<td>Entered</td>
<td>0.0104</td>
<td>1.00</td>
<td>0.80</td>
<td>7.78</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>X2</td>
<td>Entered</td>
<td>0.0131</td>
<td>0.80</td>
<td>0.87</td>
<td>0.87</td>
<td>4</td>
</tr>
</tbody>
</table>

**Step History**

- Predicted MDBH = 3.236 + 0.0974*X1 - 0.000169*X2 + 3.467*X3
- $R^2=0.8672$
- $RMSE = \sqrt{MSE} = \sqrt{0.861949} = 0.2936$
Current Estimates – Step 3

- By clicking on Step you will invoke the Backward part of the Mixed procedure.
- Note that variable $X_3$ is no longer statistically significant and so it will be removed from the model when you click on Step.

---

Current Estimates – Step 4

- $X_3$ is removed from the model
- Predicted $\text{MDBH} = 3.2605 + 0.1069X_1 - 0.0001898X_2$
- $R^2 = 0.8658$
- $\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{0.0819937} = 0.2863$
Current Estimates - Step 4

- Because $X_1$ and $X_2$ add significantly to the model they cannot be removed.
- Because $X_3$ will not add significantly to the model it cannot be added.
- The Mixed procedure stops.

Response MDBH

Summary of Fit
- RSquare: 0.865785
- RSquare Adj: 0.849995
- Root Mean Square Error: 0.203045
- Mean of Response: 6.265
- Observations (or Sum Wgts): 20

Analysis of Variance
- Source: DF, Sum of Squares, Mean Square, F Ratio, Prob > F
  - Model: 2, 8.991607, 54.8311, <.0001*
  - Error: 17, 0.08199, 0.00480, <.0001*
  - C. Total: 19, 10.38550, <.0001*

Parameter Estimates
- Term: Estimate, Std Error, t Ratio, Prob>|t|
  - Intercept: 3.26051, 0.05000, 65.20, <.0001*
  - X1: 0.10691, 0.01057, 10.11, <.0001*
  - X2: -0.00019, 3.256e-5, 0.00, <.0001*

Effect Tests
- Source: DF, Sum of Squares, F Ratio, Prob > F
  - X1: 1, 4.875, 102.14, <.0001*
  - X2: 1, 2.795, 33.96, <.0001*

Finding the “best” model

- For this example, the Forward selection procedure did not find the “best” model.
- The Backward and Mixed selection procedures came up with the “best” model.
Finding the “best” model

- None of the automatic selection procedures are guaranteed to find the “best” model.
- The only way to be sure, is to look at all possible models.

All Possible Models

- For $k$ explanatory variables there are $2^k - 1$ possible models.
- There are $k$ 1-variable models.
- There are $\binom{k}{2}$ 2-variable models.
- There are $\binom{k}{3}$ 3-variable models.

All Possible Models

- When confronted with all possible models, we often rely on summary statistics to describe features of the models.
  - $R^2$
  - $adjR^2$
  - RMSE
Another summary statistic used to assess the fit of a model is Mallows $C_p$.

$$C_p = \left( \frac{SSE_p}{MSE_{Full}} \right) - (n - 2p)$$

$p = k + 1$

The smaller $C_p$ is the “better” the fit of the model.

The full model will have $C_p = p$.

Personality – Stepwise
Red triangle pull down – All Possible Models
Right click on table – Columns
Check $C_p$
### All Possible Models

- Lists all 7 models.
- 3-variable (full) model first.
- 2-variable models – listed in order of the $R^2$ value.
- 1-variable models – listed in order of the $R^2$ value.

### Model with $X_1$, $X_2$, $X_3$ – Highest $R^2$ value.

### Model with $X_1$, $X_2$ – Lowest RMSE and lowest $C_p$.

### Which is “best”? Can’t tell until you look at significance of the variables in the model.