Session 2 – Model Evaluation

Comparing Model Predictions With Data

Graphical first, then statistical
scatterplots can be quite revealing
new data, old data, yield curves

Re-measured plots
grow to cruise years option in FPS
actual vs projected plots by species and component

Statistics
statistical significance? biological relevance?
MAE & RMSE by species and component
hypothesis tests by species and component
New Stand Inv Vol x Age
New Stand Inv & Grown Vol x Age
Old Inv Added
Yield Curve Added
Trend Line of Grown Inv Added
Trend line added

Grown vs Actual TPA

\[ y = 0.9364x + 35.411 \]
\[ R^2 = 0.6857 \]
Grown vs Actual Qdbh
Trend line added

Grown vs Actual QDbh

\[ y = 0.7808x + 1.4708 \]

\[ R^2 = 0.9113 \]
Trend line added

Grown vs Actual HT

\[ y = 0.8972x + 1.3711 \]

\[ R^2 = 0.9524 \]
Proj v Act BA

Grown vs Actual BA

Grown BA

Actual BA
1:1 line added
Grown vs Actual BA

\[ y = 0.9454x - 6.7159 \]

\[ R^2 = 0.9153 \]
Grown v Actual Gmbf
Grown vs Actual Gmbf

Actual Gmbf vs Grown Gmbf
Trend line added

Grown vs Actual Gmbf

\[ y = 0.7512x + 0.3808 \]

\[ R^2 = 0.9486 \]
### Grown v Act Regression Coefficient CIs

<table>
<thead>
<tr>
<th></th>
<th>Coefficients</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TPA</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>35.41</td>
<td>-110.67</td>
<td>181.49</td>
</tr>
<tr>
<td>X Variable 1</td>
<td>0.94</td>
<td>0.69</td>
<td>1.18</td>
</tr>
<tr>
<td><strong>QD Dbh</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>1.47</td>
<td>0.54</td>
<td>2.40</td>
</tr>
<tr>
<td>X Variable 1</td>
<td>0.78</td>
<td>0.69</td>
<td>0.88</td>
</tr>
<tr>
<td><strong>BA</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-6.72</td>
<td>-37.11</td>
<td>23.68</td>
</tr>
<tr>
<td>X Variable 1</td>
<td>0.95</td>
<td>0.83</td>
<td>1.06</td>
</tr>
<tr>
<td><strong>HT</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>1.37</td>
<td>-5.07</td>
<td>7.81</td>
</tr>
<tr>
<td>X Variable 1</td>
<td>0.90</td>
<td>0.82</td>
<td>0.97</td>
</tr>
<tr>
<td><strong>Gmbf</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>0.38</td>
<td>-1.70</td>
<td>2.46</td>
</tr>
<tr>
<td>X Variable 1</td>
<td>0.75</td>
<td>0.68</td>
<td>0.82</td>
</tr>
</tbody>
</table>
Accuracy Stats

Root Mean Square Error (RMSE) sensitive to “outliers”

\[
RMSE_{\text{Errors}} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}
\]

Mean Absolute Error (MAE)

\[\text{RMSE} \geq \text{MAE}\]

“Smaller is better”

useful in comparing and refining models

before and after calibration

both estimate “absolute bias”

plots or means indicate bias direction
On Statistical Testing

Yang, Monserud & Huang, 2004: CJFR: 34:619-629. An evaluation of diagnostic tests and their roles in validating forest biometric models

- Compared 5 parametric and 5 non-parametric tests

“It was shown that the usefulness of statistical tests in model validation is very limited. None of the tests seems to be generic enough to work well across a wide range of models and data. Each model passed one or more tests, but not all of them. Because of this, caution should be exercised when choosing a statistical test or several tests together to try to validate a model. It is important to reduce and remove any potential personal bias in selecting a favorite test, which can influence the outcome of the results.”
Hypothesis Testing

Dbh & Height & BA

tempting to use a standard paired-t test
check distribution assumptions first
log ratio of Grown over Actual is better
\[ \ln(G/A) \]: H0: stat = 0, [+]: G>A; [-]:G<A

Tpa (survival and/or mortality)
chi-square or K-S goodness-of-fit test

Equivalence testing
set “practical or acceptable difference”
TOST (two one-sided tests)
Significance level to use?
**Technique Summary**

**Graphical first, then statistical**

proj vs actual scatterplots can be quite revealing
add 1:1 line (perfect agreement)
add linear trend line
slope (1)
intercept (0)

**Statistics**

RMSE and MAE --> smaller is better
distribution assumption for standard paired-t
Dbh, Ht, BA: ln(G/A):
  H0: stat = 0; [+]: G>A; [-]: G<A
Tpa: chi-squared or K-S goodness-of-fit test
Equivalence tests --> incorporate “practical est”
Re-Cap Model Evaluation

- Evaluation/Validation is relative to Utility and Purpose
- Methods parallel the methods used to develop models
- Evaluation with Data
  - Predicted – Actual
- Evaluation without data
  - Patterns and principles
- Techniques
  - Bias, Precision