A note on using ‘forward prediction’ to assess precision and bias of genomic predictions

Christian Edel, Stefan Neuner, Reiner Emmerling und Kay-Uwe Götz
Motivation

- Investigation focuses on the property of reliabilities as measures of precision of estimates
  - How much will a future, more reliable estimate deviate from the current one?
  - How will selection affect the conditions in a validation sample?
Methods

- calculations based on model-derived reliabilities and related multivariate-normal distributions of TBV and BLUPs
Methods

- Calculations based on model-derived reliabilities and related multivariate BLUPs

\[
\begin{bmatrix}
TBV \\
PA \\
GEBV \\
EBV
\end{bmatrix} \sim N\{\mu, \Sigma\}
\]

\[
\begin{bmatrix}
TBV \\
PA \\
GEBV \\
EBV
\end{bmatrix} \sim N\left\{ \begin{bmatrix} 100 \\ 100 \\ 100 \\ 100 \end{bmatrix}, \begin{bmatrix}
R_{TBV}^2 & R_{PA}^2 & R_{GEBV}^2 & R_{EBV}^2 \\
R_{PA}^2 & R_{PA}^2 & R_{PA}^2 & R_{PA}^2 \\
R_{GEBV}^2 & R_{GEBV}^2 & R_{GEBV}^2 & R_{GEBV}^2 \\
R_{EBV}^2 & R_{EBV}^2 & R_{EBV}^2 & R_{EBV}^2
\end{bmatrix} \right\} \times 144
\]

\[R_{M1M2}^2 = R_{PA}^2 + (R_{GEBV}^2 - R_{PA}^2)(R_{EBV}^2 - R_{PA}^2)/(1 - R_{PA}^2)\]
Methods

- calculations based on model-derived reliabilities and related multivariate-normal distributions of TBV and BLUPs

- effects of selection:
  - explicit calculations in one step selection scenarios via conditional means and (co-)variances
Methods

- calculations based on model-derived reliabilities and related multivariate-normal distributions of TBV and BLUPs

- effects of selection:
  - explicit calculations in one step selection scenarios via conditional means and (co-)variances

\[
E(A|B) = \mu_A + V_{AB} V_B^{-1} (B - \mu_B)
\]

\[
Var(A|B) = \begin{bmatrix}
V_A - V_{AB} V_0 V_{BA} & V_{AB} V_B^{-1} V_{B_s} \\
V_{B_s} V_B^{-1} V_{BA} & V_{B_s}
\end{bmatrix}
\]

Henderson, 1975
Methods

- calculations based on model-derived reliabilities and related multivariate-normal distributions of TBV and BLUPs

- effects of selection:
  - explicit calculations in one step selection scenarios via conditional means and (co-)variances
  - via simulation in multistep scenarios
Correlation between true and estimated BV

Simulated squared correlation: 0.65
Reestimated squared correlation: 0.65
Correlation between true and estimated BV

Simulated squared correlation: 0.65
Reestimated squared correlation: 0.65

observed SD: 7.10
Correlation between true and estimated BV

- Simulated squared correlation: 0.65
- Reestimated squared correlation: 0.65

Prediction error distribution

- Observed SD: 7.10
- Expected SD: 7.10
Selection applied to correlated PA (R²=0.38), p = 0.25

Reestimated squared correlation: 0.51
Selection applied to correlated PA ($R^2=0.38$), $p = 0.25$

Reestimated squared correlation: 0.51

Prediction error distribution

Observed SD: 7.10
Illustration

Selection applied to correlated PA (R²=0.38), p = 0.25

Reestimated squared correlation: 0.51

Expected SD (CV): 8.42
Observed SD: 7.10
Illustration

Selection applied to correlated PA (R²=0.38), p = 0.25

Simulated squared correlation: 0.63
Reestimated squared correlation: 0.46

ebv (R²=0.89)
Illustration

Selection applied to correlated PA (R²=0.38), p = 0.25

Simulated squared correlation: 0.63
Reestimated squared correlation: 0.46

Distribution: gebv - ebv

observed SD: 6.95
expected SD: 6.95
Selection applied to correlated PA (R²=0.38), p = 0.25

Simulated squared correlation: 0.63
Reestimated squared correlation: 0.46

expected SD (CV): 8.88
observed sd: 6.95
expected sd: 6.95
Results

Table of expected values accounting for selection:

**squared correlations and deviations**

<table>
<thead>
<tr>
<th>selection applied to</th>
<th>proportion selected (%)</th>
<th>$p^2_{EBV_GEBV}$</th>
<th>peSD proxy (GEBV-EBV) “observed”</th>
<th>peSD CV „expected“ from $R^2$ in CV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EBV</td>
<td>75</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA/EBV</td>
<td>25/75</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Results

Table of expected values accounting for selection:

**squared correlations and deviations**

<table>
<thead>
<tr>
<th>selection applied to</th>
<th>proportion selected (%)</th>
<th>$R^2_{EBV_{-}GEBV}$</th>
<th>peSD proxy (GEBV-EBV) “observed”</th>
<th>peSD CV “expected“ from $R^2$ in CV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
<td>0.63</td>
<td>6.95</td>
<td>6.95</td>
</tr>
<tr>
<td>PA</td>
<td>25</td>
<td>0.45</td>
<td>6.95</td>
<td>8.88</td>
</tr>
<tr>
<td>EBV</td>
<td>75</td>
<td>0.47</td>
<td>6.49</td>
<td>8.71</td>
</tr>
<tr>
<td>PA/EBV</td>
<td>25/75</td>
<td>0.33</td>
<td>6.24</td>
<td>9.86</td>
</tr>
</tbody>
</table>
# Results

Table of expected values accounting for selection:

**means**

<table>
<thead>
<tr>
<th>selection applied to</th>
<th>proportion selected (%)</th>
<th>( \varnothing ) TBV</th>
<th>( \varnothing ) PA</th>
<th>( \varnothing ) GEBV</th>
<th>( \varnothing ) EBV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>PA</td>
<td>25</td>
<td>109</td>
<td>109</td>
<td>109</td>
<td>109</td>
</tr>
<tr>
<td>EBV</td>
<td>75</td>
<td>105</td>
<td>102 ( \rightarrow ) 103 ( \rightarrow ) 105</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA/EBV</td>
<td>25/75</td>
<td>113</td>
<td>110 ( \leftrightarrow ) 111 ( \leftrightarrow ) 113</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Results

Table of expected values accounting for selection:

**intercepts and slopes**

<table>
<thead>
<tr>
<th>selection applied to</th>
<th>proportion selected (%)</th>
<th>( b_{0_{TBV_GEBV}} )</th>
<th>( b_{1_{TBV_GEBV}} )</th>
<th>( b_{0_{EBV_GEBV}} )</th>
<th>( b_{1_{EBV_GEBV}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.00</td>
<td>1.00</td>
<td>7.37</td>
<td>0.93</td>
<td></td>
</tr>
<tr>
<td>PA</td>
<td>25</td>
<td>0.00</td>
<td>1.00</td>
<td>14.48</td>
<td>0.87</td>
</tr>
<tr>
<td>EBV</td>
<td>75</td>
<td>21.87</td>
<td>0.80</td>
<td>32.61</td>
<td>0.70</td>
</tr>
<tr>
<td>PA/EBV</td>
<td>25/75</td>
<td>28.30</td>
<td>0.76</td>
<td>46.15</td>
<td>0.60</td>
</tr>
</tbody>
</table>
General Conclusions

- Squared correlations derived from CV can be heavily influenced by effects of selection on validation sample.
  - Selected animals (preselection on PA, selective genotyping, etc.)
  - Data selection (best animals, most reliable animals, etc.)

- Effects of selection might be hard to characterize both in scope and nature.

- Conclusions from CV-correlations about true precision and bias of estimates might be limited.
General Conclusions

- distribution of differences ("peSD proxy")
  - relatively robust: effects of selection are weak
  - could be a helpful extension to GEBV-test

- approach capable for calculation of "expected b1" under selection
  - using $b_{1_{TBV_GEBV}}$: everything depends on deregression
  - using $b_{1_{EBV_GEBV}}$: tests simultaneously for coherence of estimates and assigned reliabilities
In the near future....
Selection applied to correlated GEBV (R^2=0.65), p = 0.10

Simulated squared correlation: 0.63
Reestimated squared correlation: 0.24
Thank you for your attention