PolyanNA, a Novel, Prediction-Oriented R Package for Missing Values

Norm Matloff
University of California at Davis

BARUG, at Databricks, SF
November 13, 2018 (updated Nov. 14)

coauthor: Pete Mohanty, Stanford University
Overview

Missing values (MVs):

- A perennial headache.
- Vast, VAST literature.
- Major R packages, e.g. mice and Amelia.
- New CRAN Task View, already quite extensive.
Estimation vs. Prediction

- Almost all (all?) of the MV literature is on *estimation*, e.g. estimation of treatment effects.
- Almost all of those methods are based on *imputation*. Requires extra assumptions beyond usual MAR.
- We’re interested in *prediction*.
- We’ll present a novel new technique we call the Tower Method.
- Non-imputational.
- Available at http://github.com/matloff/polyanNA.
**Theorem from Probability Theory**

> Please be patient; R code and real-data examples soon. :-)

Famous formula in probability theory:

\[ EY = E[E(Y|X)] \]

More general version, known as the Tower Property:

\[ E[E(Y|U, V)|U] = E(Y|U) \]

Why is this relevant to us?

- **Y**: variable to be predicted
- **U**: vector of known predictor values
- **V**: vector of unknown predictor values
Example: Census Data

- Programmer/engineer data, Silicon Valley, 2000 (prgeng in pkg).
- Predict $Y =$ wage income. In one particular case to be predicted, we might have
  - $U = (\text{education,occupation,weeks worked})$
  - $V = (\text{age,gender})$

  In another case, maybe $U = (\text{age,gender,education,weeks worked})$ and $V = (\text{occupation})$. Etc.
- Wish we had $U,V$, for prediction $E(Y|U,V)$, but forced to use $E(Y|U)$.
- But then must estimate many $E(Y \mid U)$, since many different patterns for MVs ($2^5$ here).
- Hard enough to fit one good model, let alone dozens or more.
- With Tower, need only one.
Basic idea:

- Fit full regression model to the complete cases.
- Use Tower to get the marginal models from the full one:

\[
\hat{E}(Y \mid U = s) = \text{avg. } \hat{E}(Y \mid U = s, V) \text{ full model}
\]

over all complete cases with \( U = s \)

- In practice, use \( U \approx s \) instead of \( U = s \), using \( k \) nearest neighbors.
Census Example (cont’d.)

(a) Use, say, `lm()` on the complete cases, predicting wage income from (age,gender,education,occupation,weeks worked).

(b) Save the fitted values, e.g. `fitted.values` from `lm()` output.

(c) Say need to predict case with education = MS, occupation = 102, weeks worked = 52 but with age and gender missing.

(d) Find the complete cases for which (education,occupation,weeks worked) = (MS,102,52).

(e) Predicted value for this case is average of the fitted values for the cases in (d).
polyanNA Package API

- `toweranNA(x,fittedReg,k,newx,scaleX=TRUE)`
  - **x**: Data frame of complete cases.
  - **fittedReg**: Estimated values of full regress. ftn. at those cases (from `lm()`, `glm()`, neural nets, whatever).
  - **k**: Number of nearest neighbors.
  - **newx**: Data frame of new cases to be predicted.
  - Return value: Vector of predictions.
Structure of Examples

- 3 real datasets.
- Break into random training and test sets.
- Predict all test-set cases with at least one MV.
Example: WordBank Data

- Kids’ vocabulary growth trajectories.
- About 5500 cases, 6 variables. About 29% MVs.

Mean Absolute Prediction Errors:

<table>
<thead>
<tr>
<th>Amelia</th>
<th>Tower</th>
</tr>
</thead>
<tbody>
<tr>
<td>102.7</td>
<td>96.2</td>
</tr>
<tr>
<td>122.9</td>
<td>119.9</td>
</tr>
<tr>
<td>89.4</td>
<td>88.1</td>
</tr>
<tr>
<td>115.3</td>
<td>107.0</td>
</tr>
<tr>
<td>111.1</td>
<td>102.5</td>
</tr>
</tbody>
</table>

- Times about 6s each.
- The `mice` package crashed.
UCI Bank Data

- About 50K cases.
- Only about 2% MVs. Not much need for MV methods, but let’s make sure Tower doesn’t bring harm. :-)
- Tower run 8.3s, **mice** 442.2s.
- Too long to do multiple runs. About the same accuracy, 0.92 or 0.93.
- **Amelia** crashed.
World Values Study

- World political survey.
- 48 countries, sample 500-3500 from each.
- MVs artificially added.
- Tower outperformed **mice** in 39 of 48 countries.

<table>
<thead>
<tr>
<th></th>
<th>Tower</th>
<th>Mice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Absolute Predictive Error</td>
<td>1.7603</td>
<td>1.8270</td>
</tr>
<tr>
<td>Elapsed Time (seconds)</td>
<td>0.1825</td>
<td>14.0822</td>
</tr>
</tbody>
</table>
Concerning Assumptions

- Most MV methods assum MAR, Missing at Random.
- Tower assumptions similar, but assumptions matter much less in prediction than in estimation.
Next for Us

- Package is called \texttt{polyanNA} because we want to make use of our \texttt{polyreg} package.
- Better regression models through polynomials (NOT machine learning!).
- \url{https://arxiv.org/abs/1806.06850}