Eileen S. Burns
Matthias Kollowatz
Predictive Models in Life and Annuities
May 18, 2016
Theory
Theory

• Questions of interest for life and annuity products
• Predictive model forms that are best suited to investigating them
• Associated theoretical concerns that may arise in the modeling process.
Questions of interest

• When will a policyholder...
  • Lapse?
  • Partially withdraw?
  • Die?

• How will a policyholder utilize the policy?
• What drives these “behaviors” and why?
Predictive model forms
Regression

• OLS, GLM, ridge, lasso

• Pros
  • Quick fitters
  • Interpretable coefficients and output
  • Harder to overfit

• Cons
  • Constrained by functional form
  • Multicollinearity issues
Tree-based models

• Decision trees, bagging, boosting

• Pros
  • Quick fitters
  • Non-parametric
  • Intuitive output

• Cons
  • Black-box formula
  • Constrained to rectangular regions
Clustering

• K-means, hierarchical, k-nearest neighbors

• Pros
  • Non-rectangular regions
  • Can be unsupervised

• Cons
  • Sensitive to outliers
  • Lacks proximity to other clusters (probabilities)
Agent-based modeling

• Pros
  • Causation is implied in the model
  • Outputs easily interpretable

• Cons
  • At high risk of modeler bias
  • Difficult to validate
Logistic GLM

• For predicting probabilities of binary outcomes
• Link function provides much needed flexibility
• Predictor variables can be quantitative or qualitative
Why a link function?

[Graph showing a link function with predictor and response axes.]
The logistic function

• $\hat{y} = g(L) = \frac{e^L}{1+e^L}$

$\triangledown L = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$

$\lim_{L \to \infty} g(L) = 1$ and $\lim_{L \to -\infty} g(L) = 0$

• $g^{-1}(\hat{y}) = \ln \left( \frac{\hat{y}}{1-\hat{y}} \right) = L$

  – Logit function (“logodds”)
Consequences of logit link

Between logodds of 0 and 1, the probability goes up from 50.0% to 73.1%, a change of 23.1% in probability.

Between logodds of -4 and -3, the probability goes up from 1.8% to 4.7%, a change of 2.9% in probability.
Generalized linear models

• Theoretical extras
  • Independent observations
  • The model is fit by maximizing the following:
    \[
    \text{loglikelihood} = \sum Y_i \ln(\hat{Y}_i) + (1 - Y_i)\ln(1 - \hat{Y}_i)
    \]
  • \(AIC = -2 \times \text{loglikelihood} + 2 \times \text{parameters}\)
  • \(BIC = -2 \times \text{loglikelihood} + \ln(N) \times \text{parameters}\)
Practical concerns
Predictive analytics process

Data Prep

Exploratory Analysis

Modeling

Validation

training/holdout  test
Practical concerns: Data

• Formatting variables (1)
• Identifying and dealing with outlier data values (2)
• Accounting for missing data (2)
• Derive new variables for modeling (3)
• Compile dataset into appropriate format (4)
Practical concerns: Modeling

• Holdout dataset (2A)
• Fitting a model (2C)
• Using the step function for variable selection (2D)
• Multicollinearity concerns (2E)
• Setting reference levels for factors (DataPrep 2)
• Piecewise terms (2F)
• Undersampling (3)
Data outliers

Histogram of tempAV

Histogram of log(tempAV)
Missing values

> summary(data.full$height)

<table>
<thead>
<tr>
<th></th>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
<th>NA's</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>47.00</td>
<td>64.00</td>
<td>67.00</td>
<td>66.82</td>
<td>70.00</td>
<td>83.00</td>
<td>10739</td>
</tr>
</tbody>
</table>

Histogram of data.full$height
## Missing values

<table>
<thead>
<tr>
<th>Model</th>
<th>NA treatment</th>
<th>Intercept</th>
<th>Height coefficient</th>
<th>Flag coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Death ~ height</td>
<td>Removed</td>
<td>-4.418</td>
<td>0.0100</td>
<td>N/A</td>
</tr>
<tr>
<td>Death ~ height + Ind</td>
<td>Set to 0</td>
<td>-3.580</td>
<td>0.0100</td>
<td>-0.838</td>
</tr>
<tr>
<td>Death ~ height + Ind</td>
<td>Set to mean</td>
<td>-4.245</td>
<td>0.0100</td>
<td>-0.173</td>
</tr>
<tr>
<td>Death ~ height</td>
<td>Set to 0</td>
<td>-3.589</td>
<td>-0.0024</td>
<td>N/A</td>
</tr>
<tr>
<td>Death ~ height</td>
<td>Set to mean</td>
<td>-4.343</td>
<td>0.0095</td>
<td>N/A</td>
</tr>
</tbody>
</table>

- The first three models are mathematically equivalent
- The second two are biased
Training versus holdout data

Exposure over time

[Bar chart showing exposure over time from 1999 to 2014]
Stepwise model building

\[
\text{logodds} = f(\text{attained age})
\]

\[
\text{logodds} = f(\text{attained age}) + f(\text{cad})
\]

\[
\text{logodds} = f(\text{attained age}) + f(\text{cad}) + f(\text{cognitive}) + \cdots
\]

\[
\text{step}(\text{model.intercept}, \text{scope} = \text{model.maximum}, \text{direction} = "\text{forward}", \text{steps} = 3, k = \log(\text{data.dim}[1]))
\]
Multicollinearity

- `pairs()`

- `cor()`

```
<table>
<thead>
<tr>
<th></th>
<th>height</th>
<th>weight</th>
<th>bmi</th>
</tr>
</thead>
<tbody>
<tr>
<td>height</td>
<td>1.000000</td>
<td>0.637640</td>
<td>0.052578</td>
</tr>
<tr>
<td>weight</td>
<td>0.637640</td>
<td>1.000000</td>
<td>0.795710</td>
</tr>
<tr>
<td>bmi</td>
<td>0.052578</td>
<td>0.795710</td>
<td>1.000000</td>
</tr>
</tbody>
</table>
```

- `vif()`
Reference levels
Piecewise linear effects

A/E by predictor before piecewise split

Piecewise impact of example predictor
Undersampling

• For logistic regression, undersampling can help improve runtimes:
  • All deaths (n) +
  • Randomly selected non-deaths (3n)

• Fitting the model Death ~ AttAge

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Records</th>
<th>Runtime</th>
<th>Intercept</th>
<th>AttAge coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>259,284</td>
<td>2.15</td>
<td>-14.13</td>
<td>0.129</td>
</tr>
<tr>
<td>Undersampled</td>
<td>25,152</td>
<td>0.12</td>
<td>-10.99</td>
<td>0.123</td>
</tr>
</tbody>
</table>
Validation
Validation and comparison

• Overall model fit (4A)
• Comparison between two candidate models (4B)
Model fit

- $R^2$
- AIC/BIC
- Actual-to-expected plots (4A-i)
- Confusion matrix (4A-ii)
- AUC (4A-iii)
Confusion matrix

- Select a threshold for predicting the outcome
- Build a 2x2 contingency table

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Death</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>65,815</td>
<td>66,650</td>
</tr>
<tr>
<td>1</td>
<td>18,500</td>
<td>19,813</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>84,315</td>
<td>86,463</td>
</tr>
</tbody>
</table>

True positive rate = 1,313/2,148 = 0.658
False positive rate = 18,500/84,315 = 0.301
Area under the curve (AUC)

- The curve here is the relationship of the true positive rate and false positive rate as the threshold moves from 0 to 1
Model comparison: Lift charts

- Actual to expected (4B)
- Two-way lift (4B)