Homeowners Ratemaking By Peril
— Data Issues —

Michael Nielsen, FCAS
United Services Automobile Association

2013 Ratemaking & Product Management Seminar
March 11-13, 2013

Agenda

- Basics
  - Response Variable Decisions
  - Predictor Variable Decisions
- Other Issues
  - Missing Data (Spatial Interpolation Example)
  - Principal Components Analysis

Response Variable Decisions

Frequency-Severity versus Pure Premium

Peril Group Definitions

- Limited by accuracy and detail of cause of loss codes
  - Water (weather vs non-weather)
  - Fire (environmental vs man-made)
  - Theft (on vs off premises)
  - Wind/Hail
  - Liability
  - Lightning
  - All Other
- Liability is both a coverage and a cause of loss
- A single claim may have multiple causes of loss

Claim exclusions & capping

Other adjustments to losses

Predictor Variable Decisions

Types of predictor variables:

- Structure characteristics
- Occupant characteristics
- Policy characteristics
- Location characteristics
  - Demographics
  - Weather
  - Topography
  - Proximity to other features

Consider purpose of modeling when selecting predictors

Which variables should be adjusted to current levels and which should be left at historical levels?

4

Dealing with missing values

Possible solutions:

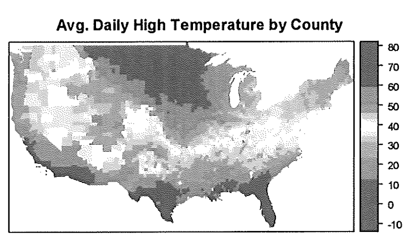
- Make no changes – leave it to the modelers
- Impute a new value
  - Use the mean
  - Interpolation
  - Build a model to predict the missing value

Good practice to create a new variable indicating an imputed value.

- Occasionally, the missingness of a variable is more predictive than the actual variable.

5

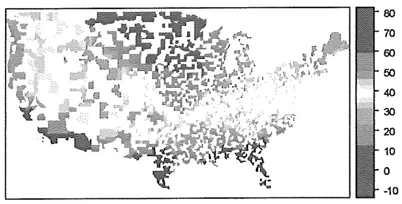
Spatial Data: You wish you had this ...



6

… but you've got this!

Avg. Daily High Temperature by County



7

Inverse Distance Weighted Interpolation

- A deterministic spatial interpolation method
- Key Assumption: Things that are close to one another are more alike than those that are farther apart.

$$\hat{Z}(s_0)=\frac{\sum_{i=1}^{n} w(s_i)Z(s_i)}{\sum_{i=1}^{n} w(s_i)} \qquad w(s_i)=\|s_i-s_0\|^{-p}$$

$\|\cdot\|$ indicates Euclidian distance

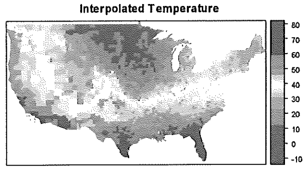
- Commonly available in GIS software.
- Also available in R.

8

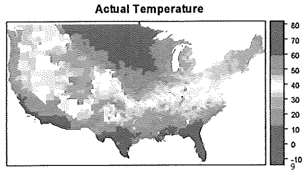
Interpolated Results

- Problems when there aren't many neighbors
  - Border counties
  - Islands (e.g., HI & AK)
- Interpolation can be slow
  - Many missing values
  - Many neighbors
- Considers proximity, but ignores other factors
  - Spatial correlation
  - Other predictors (e.g., elevation)

Interpolated Temperature



Actual Temperature



9

Spatial Interpolation in R

- `readShapeSpatial()` [package = maptools]
- `idw()` [package = gstat]
- `spplot()` [package = sp]
- `brewer.pal()` [package = RColorBrewer]

Great Resource:

- Bivand, Pebesma, and Gómez-Rubio. Applied Spatial Data Analysis with R

10

External Data – Too Much and Not Enough

Too Much Data:

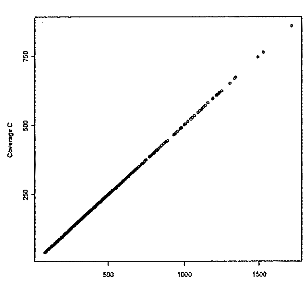
- Many geographic units:
  - 3,140 U.S. counties
  - 8.2 million census blocks
  - 211,267 census block groups
  - 74,002 census tracts
- High frequency of measurement
  - e.g., Weather data
- Large numbers of variables
  - American Community Survey, U.S. Census (over 21,000 variables)

We still want more!

11

Sometimes you have less data than you think

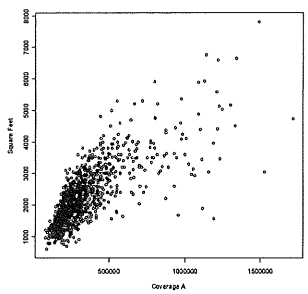
- Correlation = 100%
- Two Problems:
  - Unnecessary variable
  - Multicolinearity
- Two Solutions:
  - Throw out one variable
  - Rotate the axes



12

A more realistic example

- Correlation = 75.27%
  - Fairly high, but probably not problematic.
- Neither variable should be thrown out, but it's good to understand the relationship
- Correlations are more difficult to predict in higher dimensions.



Principal Components

First Principal Component

$PC_1 = a_{11}x_1 + a_{12}x_2 + \cdots a_{1p}x_p$

- Choose $a_{11}, a_{12}, \ldots, a_{1p}$ such that the variance of PC_1 is maximized.
- One constraint: $\sum_i a_{1i}^2 = 1$

Second Principal Component

$PC_2 = a_{21}x_1 + a_{22}x_2 + \cdots a_{2p}x_p$

- Choose $a_{21}, a_{22}, \ldots, a_{2p}$ such that the variance of PC_2 is maximized.
- Two constraints: $\sum_i a_{2i}^2 = 1 \quad and \quad Cov(PC_1, PC_2) = 0$

Continue in this fashion for each additional principal component. The covariance with each of the preceding principal components is 0.

Principal Components Solution

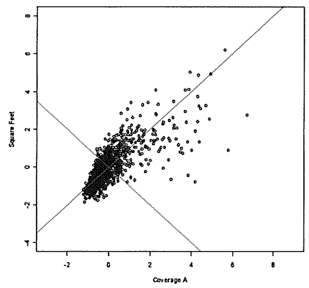
- The weights of the i^{th} principal component are given by the i^{th} eigenvector of the covariance matrix
- Principal components are affected by the scale of the underlying variables.
  - Best to obtain principal components from standardized variables
  - Equivalent to using the correlation matrix
- The variance of the i^{th} principle component is the i^{th} eigenvalue (λ_i) of the covariance matrix
- Total sample variance = $\sum_{i=1}^{p} \lambda_i$
- Use the eigenvalues to calculate the proportion of the total variance due to each principal component.

Our principal component solution

$PC_1 = 0.7071 \times x_1 + 0.7071 \times x_2$

$PC_2 = -0.7071 \times x_1 + 0.7071 \times x_2$

PC_1 explains 87.6% of the total variation



16

Variable Reduction Example

Census Variables:

- Total Population
- Civilian Employment
- Median Income
- Median Home Value
- Healthcare Employment
- College Graduates

```
> R<-cor(X);round(R,2)

      Pop  Emp  Inc  Home HEmp Col
Pop  1.00 0.95 0.26 0.19 0.69 0.59
Emp  0.95 1.00 0.40 0.31 0.76 0.70
Inc  0.26 0.40 1.00 0.85 0.53 0.73
Home 0.19 0.31 0.85 1.00 0.46 0.69
HEmp 0.69 0.76 0.53 0.46 1.00 0.73
Col  0.59 0.70 0.73 0.69 0.73 1.00

> round(eigen(R)$values,3)
[1] 3.978 1.363 0.290 0.185 0.146 0.038

> round(eigen(R)$vectors,3)
       [,1]   [,2]   [,3]   [,4]   [,5]   [,6]
[1,] -0.384  0.508 -0.364 -0.197  0.105  0.642
[2,] -0.428  0.405 -0.276 -0.095 -0.080 -0.749
[3,] -0.383 -0.486 -0.139 -0.254 -0.721  0.116
[4,] -0.353 -0.552 -0.176 -0.273  0.679 -0.069
[5,] -0.433  0.153  0.861 -0.213  0.041  0.036
[6,] -0.459 -0.117  0.002  0.876  0.028  0.084
```

Data Source:

http://www2.census.gov/acs2010_5yr/summaryfile/2006-2010_ACSSF_By_State_All_Tables/

17

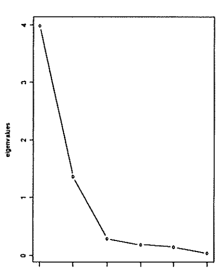
Variable Reduction Example

Choose the number of principal components by looking for an "elbow" in the scree plot.

Two or three principal components effectively summarize the total sample variance.

```
> round(cumsum(eigen(R)$values)
  /sum(eigen(R)$values),3)

[1] 0.663 0.890 0.938 0.969 0.994 1.000
```



18

Principal Components Analysis in R or SAS

SAS

- `proc princomp`
- `proc factor`

R

- `princomp()` [package = stats]
- `eigen()` [package = base]
- `prcomp()` [package = stats]
- `svd()` [package = base]

`prcomp()` calculates principal components using the singular value decomposition (preferred method for numerical accuracy)

19

Conclusions

- Data preparation usually takes more time and effort than the actual modeling
- Better data preparation leads to smoother modeling.
- Knowledge gained by preparing the data will improve the modeling process
- The person preparing the data needs to think like a modeler and the modeler needs to think like an actuary.

20