Population Segmentation in a Healthcare Environment

Enactment of the Affordable Care Act (ACA) impacted the United States health insurance industry in 2014. Previously, the insurer determined who their member population were based on actuarial practices, medical underwriting and product development. Now, member characteristics will vary and insurers will need to maintain a balance of high and low risk members to balance their portfolio of products to support business objectives.

MaryAnne DePesquio, BlueCross BlueShield of Arizona, USA

### Project Overview

- **Requirements Gathering**
  - Target Population
  - Project Plan
  - Resource Allocation
  - Stakeholders Engagement
  - Actionable Outcomes

- **Cluster Analysis**
  - PROC CLUSTER
  - PROC FASTCLUS

- **PROC CLUSTER**
  - Hierarchical methodology
  - Need to know underlying distribution
  - Three methods - Average, Centroid, Ward's
  - Used with survey data due to the ability to capture non-spherical clusters
  - Determine number of clusters best fits data

  ```
  PROC CLUSTER
  DATA = mydat.azpop
  METHOD = WARD
  SIMPLE
  CCC
  PSEUDO;
  VAR
  Insurvar1 Insurvar2 Channel
  Age Income Dwelling Tenure;
  RUN;
  ```

- **PROC FASTCLUS**
  - Non-Hierarchical methodology
  - K-Means
  - Use WARD’s method to determine # clusters

  ```
  PROC FASTCLUS
  DATA = mydat.azpop
  OUT = kmeans_out
  MAXC = 7
  MAXITER = 100
  MEANS = cluster_mns
  OUT = cluster_outall;
  VAR
  Insurvar1 Insurvar2 Channel
  Age Income Dwelling Tenure;
  RUN;
  ```

### Variable Reduction Methods

- **PROC VARCLUS**
  - PROC LOGISTIC

#### PROC VARCLUS

- MAXEIGEN = 0.7
- DATA = mydat.analytic;
- VAR
  - age
  - gender
  - tenure
  - education
  - ethnicity
  - income
  - dwelling
  - num_persons;
- RUN;

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Variable</th>
<th>Obs</th>
<th>Stat</th>
<th>P-val</th>
<th>R^2</th>
<th>R^2 Adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>Age</td>
<td>0.0397</td>
<td>0.0097</td>
<td>0.2623</td>
<td>0.2623</td>
<td></td>
</tr>
<tr>
<td>Cluster 1</td>
<td>Tenure</td>
<td>0.1276</td>
<td>0.0444</td>
<td>0.2623</td>
<td>0.2623</td>
<td></td>
</tr>
<tr>
<td>Cluster 2</td>
<td>Gender</td>
<td>0.3754</td>
<td>0.0188</td>
<td>0.1968</td>
<td>0.1968</td>
<td></td>
</tr>
<tr>
<td>Cluster 2</td>
<td>Education</td>
<td>0.3693</td>
<td>0.0138</td>
<td>0.1968</td>
<td>0.1968</td>
<td></td>
</tr>
<tr>
<td>Cluster 3</td>
<td>Dwelling</td>
<td>0.3977</td>
<td>0.0214</td>
<td>0.1968</td>
<td>0.1968</td>
<td></td>
</tr>
<tr>
<td>Cluster 3</td>
<td>Ethnicity</td>
<td>0.8909</td>
<td>0.0847</td>
<td>0.8909</td>
<td>0.8909</td>
<td></td>
</tr>
<tr>
<td>Cluster 4</td>
<td>Num_persons</td>
<td>0.8253</td>
<td>0.0338</td>
<td>0.8253</td>
<td>0.8253</td>
<td></td>
</tr>
<tr>
<td>Cluster 4</td>
<td>Income</td>
<td>0.7950</td>
<td>0.0556</td>
<td>0.7950</td>
<td>0.7950</td>
<td></td>
</tr>
</tbody>
</table>

#### PROC LOGISTIC

- DATA = mydat.analytic
- DESCENDING;
- MODEL
  - Target_LR = Num_Persons
  - Gender Dwelling Prev_Insur
  - Tenure Age Ethnicity Deductible
  - Marital_Status Education /
  - SELECTION = BACKWARD SLS = 0.05
- CORRB RSQ LACKFIT
- OUT = mydat.Logistic_out
- Prob = pred_prob;
- RUN;

### Cluster Results

- **7 Clusters**
- **Primary Drivers 10**
- **Target 3 Clusters**
- **Matched ‘Desired’ Characteristics**
- **Top 10% in Each Cluster Selected**
- **Direct Mail Campaigns**
- **Monitor Response to Conversion**

### Selecting & Transforming Data

- Database 150k
- 2,600 Fields
- Exploration
  - PROC FREQ
  - PROC MEANS
- Normality
  - Log Transform
  - Remove Outliers
  - Group to Deciles
  - Re-categorize
  - Missing Levels
  - Explode Levels
  - Combine Levels

### Campaign Implementation

- Target 3 Clusters
- Matched ‘Desired’ Characteristics
- Top 10% in Each Cluster Selected
- Direct Mail Campaigns
- Monitor Response to Conversion