Predicting Healthcare Costs Using Regression
Andrew Mackenzie, Daniel Mena, Allie Moat, Tiffany Sun, Jessie Wang, Roger Wu
Faculty Advisor: Ian Duncan, Michael Ludkovski
Department of Statistics and Applied Probability, University of California Santa Barbara

Objectives
- Use hierarchical regression to predict healthcare costs:
  - Classification and Regression Trees (CART)
  - Multivariate Adaptive Regression Splines (MARS)
  - Identify important predictor variables, and compare accuracy of the methods

Data
- 2 years of private health insurance claims data: 10,000 individuals with 133 predictor variables each
- 110 0/1 flags indicating presence of certain health conditions (diabetes, leukemia, pregnancy, etc.); chronic conditions could indicate high probability of expenses
- Costs by categories (ex: total costs, pharmacy cost etc.)
- Numeric (ex: age) & demographic variables (ex: location)
- Counts (ex: # of visits to the hospital)

Approach
- Challenges in predicting year 2 cost with year 1 data:
  - Data is highly nonlinear; costs range from $0 to $300,000; fluctuate from year to year; many variables; future costs of those with zero costs currently are hard to predict

Results
- Comparing $\hat{MSE}$ (mean squared error on Testing data set) and $R^2$ (coefficient of determination on Training data set):

<table>
<thead>
<tr>
<th>Model</th>
<th>$\hat{MSE}$</th>
<th>$R^2$</th>
<th>$\hat{MSE}$</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>4900 ± 190</td>
<td>0.3851</td>
<td>4800 ± 170</td>
<td>0.3798</td>
</tr>
<tr>
<td>Male</td>
<td>4960 ± 290</td>
<td>0.4073</td>
<td>4700 ± 310</td>
<td>0.5809</td>
</tr>
<tr>
<td>Female</td>
<td>5100 ± 260</td>
<td>0.3958</td>
<td>5000 ± 240</td>
<td>0.5724</td>
</tr>
</tbody>
</table>

- Standard error of the MSE computed using 20 bootstrap runs
- $R^2$ only have standard error of 0.001
- Full Model is the best for both RF CART and MARS
- RF CART is significantly better because of its higher $R^2$ for all 3 models

Variable Importance (most influential variables for predicting future costs):

<table>
<thead>
<tr>
<th>Variables</th>
<th>MARS</th>
<th>RF CART</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year 1 Cost</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td># of PCP Visits</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Pharmacy Cost</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Inpatient Cost</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>Age</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Low Risk Pregnancy</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Professional Cost</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Renal Failure</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

- Current Cost is the most important variable when predicting future cost of healthcare, several other categorical costs are also crucial
- Low Risk Pregnancy is likely to require high cost of females in the data set
- Renal (kidney) Failure can be chronic & may require continual medical care cost

Decrease in MSE by Variable, Random Forest

Citations, and Acknowledgments

[6] In order to conduct this research we used the open source statistical software R with the package earth for MARS and rpart for CART, available at http://www.r-project.org/
[7] We would like to thank our faculty advisors, Ian Duncan, Raya Feldman, and Mike Ludkovski for their assistance, their guidance, and their enthusiasm for this research.