Predictive Modeling of Healthcare Costs

Mike Ludkovski

Dept of Statistics & Applied Probability UC Santa Barbara

Actuarial Research Conference
based on joint work with Ian Duncan and Michael Loginov

Predictive Modeling in the Media

- NYT 6/28/2014: “When a Health Plan Knows How You Shop”
  - “Mail-order shoppers and Internet users, for example, were likelier than some other members to use more emergency services.”
  - Goes on to quote ”a chief analytics officer”, a ”law professor”, ”VP of client strategy”.
  - No mention of actuaries.

Predictive Modeling in the Media

- NYT 6/28/2014: “When a Health Plan Knows How You Shop”
  - “Mail-order shoppers and Internet users, for example, were likelier than some other members to use more emergency services.”
  - Goes on to quote ”a chief analytics officer”, a ”law professor”, ”VP of client strategy”.
  - No mention of actuaries.

(Academic) Actuaries should be **taking the lead** in this space

Modeling Healthcare Costs

- Forecasting healthcare expenditures is intrinsically difficult
- Huge variability year-over-year
- Many sub-categories (Rx, PCP, Inpatient, Outpatient, ER); multiple claims per year

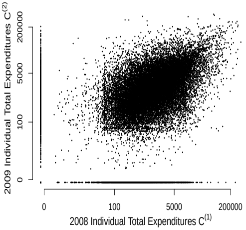
Modeling Healthcare Costs

- Forecasting healthcare expenditures is intrinsically difficult
- Huge variability year-over-year
- Many sub-categories (Rx, PCP, Inpatient, Outpatient, ER); multiple claims per year
- An enormous variety of potentially relevant risk factors:
  - Demographics
  - Lifestyle
  - Socioeconomic
  - Health history
  - Claims history

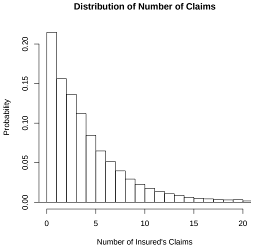
Encoding Health Information

- Ongoing challenge how to quantify health data into format suitable for risk management
- **H**ierarchical **C**oexisting **C**ondition codes: identify the presence of underlying medical conditions/diagnoses
- HCCs are derived from submitted claim descriptions
- Complement dollar-cost information and claim counts
- CMS-HCC variant used in Medicare Advantage
- Our dataset had 83 HCC flags (diabetes, hypertension, hip fracture, chemotherapy, etc.)
- Most are very rare

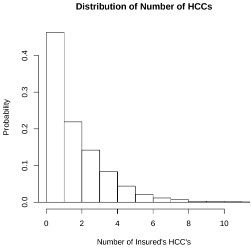
Data Features



COST



NUMCL



NUMHCC

Modeling Objectives

- Predict **individual aggregate next-year expenditures** given last year's data
- Classify insureds into risk groups
- Identify key predictive covariates, e.g
  - What is relationship between predictors/future costs?
  - How helpful is HCC information?
- Uncover new relationships (data mining)

Personal Objectives

- Expose students to a nontrivial dataset
- Introduce actuarial students to modern regression tools
- Test competing statistical approaches to understand their performance in this context
- Used as a case study for UCSB ActSci Masters students

Traditional Approach

- Linear Model: $COST = \sum_i \beta_i X^i + \varepsilon$
- Each predictive variable enters the model
- Coefficients β_i describe globally the effect on Y from changing X^i ceteris paribus
- Linear relationship between each predictor and the outcome
- Easy to understand + assess goodness-of-fit + test for significance

Linear Models Extended

- Because healthcare data have long tails and the noise is state-dependent, a LM is not appropriate
- Data Transformation:
  - $\log COST = \sum_i \beta_i X^i + \varepsilon$ (smearing)
  - $COST = \exp(\sum_i \beta_i X^i) + \varepsilon$ (log-link)
  - Truncate extreme losses
- Variable Transformation:
  - Add variable interactions (AGE*GENDER)
  - Add nonlinear predictors $1_{AGE \in [25,34]} \cdots$
- **Generalized Linear Model**

Challenges

- Have a lot of predictor variables; many are highly correlated
- Have nonlinear relationships that must be captured (eg impact of AGE)
- Much more than just main effects
- Impact on costs is state-dependent (eg having 10 HCC's at age 65 vs having 10 HCC's at age 30)
- Little *a priori* knowledge about all this

Consequently:

- Parametrizing assumed relationships is difficult
- Including all possible variables in the model leads to overfitting
- Approach must be flexible and localized

Recommendations

- Should utilize nonparametric statistical models (flexibility)

Recommendations

- Should utilize nonparametric statistical models (flexibility)
- Select which covariates to include (regularization)

Recommendations

- Should utilize nonparametric statistical models (flexibility)
- Select which covariates to include (regularization)
- Avoid making global statements (data partitioning)

Recommendations

- Should utilize nonparametric statistical models (flexibility)
- Select which covariates to include (regularization)
- Avoid making global statements (data partitioning)
- Strive for interpretability/computational efficiency

Recommendations

- Should utilize nonparametric statistical models (flexibility)
- Select which covariates to include (regularization)
- Avoid making global statements (data partitioning)
- Strive for interpretability/computational efficiency

Recommendations

- Should utilize nonparametric statistical models (flexibility)
- Select which covariates to include (regularization)
- Avoid making global statements (data partitioning)
- Strive for interpretability/computational efficiency

Further requirements:

- Able to deal with many predictive variables of different types (numerical, factor, 0/1)
- Distribution of costs is long-tailed: handle outliers

Approaches Tried

- **Lasso**/LARS: regularized linear model (M. Loginov, E. Marlow, V. Potruch, ARC 2012)

Approaches Tried

- **Lasso**/LARS: regularized linear model (M. Loginov, E. Marlow, V. Potruch, ARC 2012)
- **MARS**: multivariate adaptive splines (A. Mackenzie, T. Sun, R. Wu, ARC 2013)

Approaches Tried

- **Lasso**/LARS: regularized linear model (M. Loginov, E. Marlow, V. Potruch, ARC 2012)
- **MARS**: multivariate adaptive splines (A. Mackenzie, T. Sun, R. Wu, ARC 2013)
- CART/**Random Forest**: treed models (D. Mena, A. Moat, J. Wang, ARC 2013)

Approaches Tried

- **Lasso**/LARS: regularized linear model (M. Loginov, E. Marlow, V. Potruch, ARC 2012)
- **MARS**: multivariate adaptive splines (A. Mackenzie, T. Sun, R. Wu, ARC 2013)
- CART/**Random Forest**: treed models (D. Mena, A. Moat, J. Wang, ARC 2013)
- **Additive** Models: spline representation of main effects; still additive over covariates

Approaches Tried

- **Lasso**/LARS: regularized linear model (M. Loginov, E. Marlow, V. Potruch, ARC 2012)
- **MARS**: multivariate adaptive splines (A. Mackenzie, T. Sun, R. Wu, ARC 2013)
- CART/**Random Forest**: treed models (D. Mena, A. Moat, J. Wang, ARC 2013)
- **Additive** Models: spline representation of main effects; still additive over covariates
- **Boosted Trees**: ensemble model that sequentially fits (thousands) of small trees to the latest residuals

Approaches Tried

- **Lasso**/LARS: regularized linear model (M. Loginov, E. Marlow, V. Potruch, ARC 2012)
- **MARS**: multivariate adaptive splines (A. Mackenzie, T. Sun, R. Wu, ARC 2013)
- CART/**Random Forest**: treed models (D. Mena, A. Moat, J. Wang, ARC 2013)
- **Additive** Models: spline representation of main effects; still additive over covariates
- **Boosted Trees**: ensemble model that sequentially fits (thousands) of small trees to the latest residuals
- **Decision Rules** (Quinlan M5) – treed model with a linear fit at each node. Final answer is a weighted average of ancestral LMs.

Decision Rule (1 Tree out of 25)

```
Rule 1/4: [5060 cases, mean 3472.2988, range 0 to 40000, est err 3140.2654]
    if
        age > 51
        allow_current_total <= 5764.29
    then
        outcome = 2351.8117 + 1.04 allow_current_rx + 0.3 allow_current_prof
                  + 234 pcp_visit_cnt_current + 391 no_of_HCCs - 933 HCC11
                  - 55 total_count + 0.04 allow_current_op - 402 male
                  + 950 HCC144 - 256 HCC91 - 201 HCC22 - 6 age - 375 HCC13

  Rule 1/5: [792 cases, mean 4481.6519, range 0 to 40000, est err 3826.8042]
    if
        age <= 55
        allow_current_prof <= 8855.27
        allow_current_total > 5764.29
        pcp_visit_cnt_current <= 6
    then
        outcome = 6516.2337 + 6835 pcp_visit_cnt_current - 6392 total_count
                  + 1.37 allow_current_rx + 0.36 allow_current_op
                  + 5727 admit_cnt_current - 162 age + 0.16 allow_current_ip
                  + 124 no_of_HCCs - 225 HCC22 + 0.02 allow_current_prof
                  - 256 HCC11 - 142 male
```

Marginal Dependence

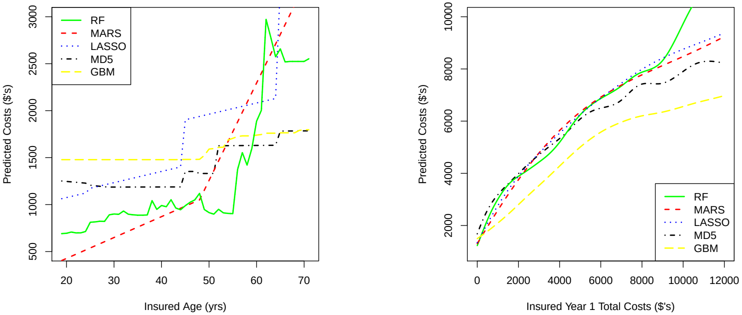


Figure : Partial dependence plots. Left panel: predicted 2009 costs for a Male with zero 2008 expenditures $C^{(1)} = 0$ and no HCC codes, as a function of **AGE**. Right panel: predicted 2009 costs as a function of **2008 costs** averaged out over the testing population.

HCC Effects

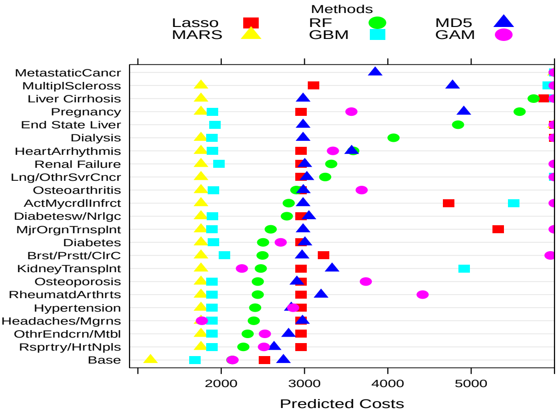


Figure : Predicted costs for a randomly picked **55-year old male** with a single given medical condition. Year-1 expenditures were fixed at $C^{(1)} = 1620$ with $NUMCL = 2$. We show the results for 22 most influential/common HCC's. The last row shows the base case where no HCC flags are given.

Ranking Risks

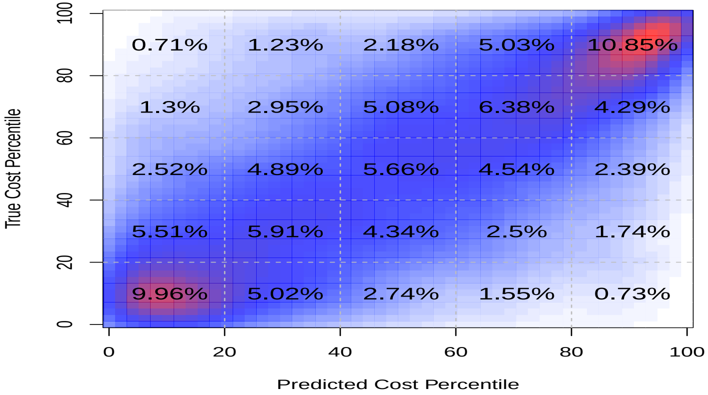


Figure : Classifying risks into quintiles. The numbers indicate percentage of test population ($M = 10000$ individuals) that fall into the particular cell (grouped by predicted quintile versus the actual quintile) so that the total sum is 100% and each row/column adds up to 20%. Colors provide a more granular description of the same information, with dark red indicating highest density.

What is the Loss Function?

How to assess accuracy/goodness of fit?

- Mean-squared error/R^2 is too sensitive to outliers
- Median absolute deviation will have a strong bias (washes out left-tail)
- Rank dependence is a simple way to standardize (but probably too crude)
- Predictive power is never high, so individual predictions are not too meaningful; what is the right group size to think about?
- Cross-validation is important to trust the results

$\Rightarrow$ strongly affects how models are estimated

Model Complexity/Variable Importance

How to benchmark/compare models?

- No simple way to compare different model complexity (what is “size”?)

Model Complexity/Variable Importance

How to benchmark/compare models?

- No simple way to compare different model complexity (what is “size”?)
- No simple way to define parsimony (eg Paraplegy: rare but obviously influential)

Model Complexity/Variable Importance

How to benchmark/compare models?

- No simple way to compare different model complexity (what is “size”?)
- No simple way to define parsimony (eg Paraplegy: rare but obviously influential)
- No simple way to rank covariates (eg testing for significance of HCCs)

Model Complexity/Variable Importance

How to benchmark/compare models?

- No simple way to compare different model complexity (what is “size”?)
- No simple way to define parsimony (eg Paraplegy: rare but obviously influential)
- No simple way to rank covariates (eg testing for significance of HCCs)
- No simple way to respect hierarchies when doing model selection

Model Complexity/Variable Importance

How to benchmark/compare models?

- No simple way to compare different model complexity (what is “size”?)
- No simple way to define parsimony (eg Paraplegy: rare but obviously influential)
- No simple way to rank covariates (eg testing for significance of HCCs)
- No simple way to respect hierarchies when doing model selection
- There are many knobs to finetune

Endless Possibilities

- Many of the mentioned methods have multiple implementations (eg lasso2, lars, grplasso, elasticnet, glmpath)
- There are currently 71 `R` packages in the Machine Learning directory `http://cran.r-project.org/web/views/MachineLearning.html`
- Great source of **pedagogic** examples!

Summary of Findings

- None of the above off-the-shelf methods are perfect
- Clear that LM/GLM is very *far from optimal*
- Ideally, should design a custom tool (eg treed GLM models are still rudimentary)
- **Human experts** are indispensable
- Visualizing/explaining the models is KEY to make them acceptable to decision-makers. Need to understand the pitfalls of each method.
- Properly framing the objectives is as important as building the actual model
- Actuaries are best positioned to take charge of the entire modeling process.
- Need both new theory and new applied studies.

More Takeways

- Costs are more predictive than pure HCCs
- HCCs are difficult to incorporate into a model
- Need a bespoke approach to deal with the right-tail
- Impossible to model zero-cost risks ($R^2 < 0.1$)

Future Directions

- Risk Management: don't care only about the expected cost but also about its distribution $\rightarrow$ quantile regression
- Modeling the full predictive distribution (Bayesian methods)
- Classification: run an ordinal classification method to directly assign risk groups
- Frequency + Severity modeling (bottom-up approaches)

Future Directions

- Risk Management: don't care only about the expected cost but also about its distribution $\rightarrow$ quantile regression
- Modeling the full predictive distribution (Bayesian methods)
- Classification: run an ordinal classification method to directly assign risk groups
- Frequency + Severity modeling (bottom-up approaches)

THANK YOU!