Healthcare Case Studies Project

Jeff Miller, Fernanda De La O, Mary Nevis
Department of Computer Science and Engineering
University of the Pacific
Stockton, CA

Abstract— **As medical insurance companies in the United States face stiff competition amongst themselves and pressures from the government, the need to cut costs and raise efficiency is as important as ever. In this study, an algorithm was developed to predict post-diagnosis costs for diabetic patients based a careful selection of factors that bear the most influence on these often expensive subsequent medical expenditures.**

Keywords—diabetes, healthcare

I. Introduction

According to the Center for Disease Control, 29 million Americans suffer from diabetes. With this disease affecting almost 10 percent of the American population, its impact on healthcare is phenomenal. People with diabetes are at increased risk of serious health complications including vision loss, heart disease, stroke, kidney failure, amputation of toes, feet or legs, and premature death. In 2012, diabetes and its related complications accounted for $245 billion in total medical costs and lost work and wages. This figure is up from $174 billion in 2007.[1]

The purpose of this study is to create a statistical model that identifies specific clinical and statistical variables capable of predicting the post-diagnosis costs to the insurance company of diabetic patients.

According to our industry expert, Dr. Rolin Wade, diabetic patients that may be at particular risk of emergency room visits and/or hospitalization if appropriate medication is not taken on a regular schedule. Although we did not specifically categorize diabetic patients on their adherence to medication behaviors, current scientific studies show a strong association between medication adherence and healthcare costs. In fact, a study on the impact of medication adherence on hospital risks and healthcare costs, showed that for diabetes, a high level of medication adherence was associated with lower disease-related medical costs. Additionally, higher medication costs were more than offset by medical cost reductions, producing a net reduction in overall healthcare costs. Also cost offsets were observed for all-cause medical costs at high levels of medication adherence, and hospitalization rates were significantly lower for patients with high medication adherence. (cost) A model that predicts the post-diagnosis costs of diabetic patients can assist the insurance company in identifying and providing intervention for those patient with non-adherence of medication behaviors. [2]

Another study of diabetic patients over a three-year period assessed such variables as the impact of baseline A1c (glucose levels), cardiovascular disease, and depression on subsequent health care costs among adults with diabetes. In this study, generalized linear models were used to analyze costs related to clinical predictors after adjusting for demographic and socioeconomic factors. This multivariate analysis of 1,694 adults with diabetes, found that the three-year costs in those with coronary heart disease (CHD) and hypertension were over 300% of those with diabetes only. Depression was associated with a 50% increase in costs. Relative to those with a baseline A1c of 6%, those with an A1c of 10% had three-year costs that were 11% higher. Higher A1c predicted higher costs only for those with baseline A1c >7.5%. The conclusion of this study was that in adults with diabetes, CHD, hypertension, and depression spectrum disorders more strongly predicted future costs than the A1c level, thus concurrent with aggressive efforts to control glucose, greater efforts to prevent or control CHD, hypertension, and depression are necessary to control health care costs in adults with diabetes. (Diabetes) In summary, although keeping those glucose levels under control with medication adherence, diet and exercise are important, it is also important to focus on other variables when trying to reduce the healthcare costs of diabetic patients. [3]

In yet another study, increased comorbidity severity and an emergency room visit during the year prior to enrollment in a Medicare HMO were independently associated with decreased antidiabetic medication possession ratios (MPRs) after enrollment. These researchers found that after controlling for type of medication therapy and other variables, increased antidiabetic MPR remained the strongest predictor of decreased total annual health care costs (8.6% to 28.9% decrease in annual costs with every 10% increase in MPR; $P < 0.001$). Thus, adherence to anti-diabetic medications proved to be a greater driver of cost reduction than other concurrent medications in this population. (predictors)

II. Dataset

The dataset was extracted from PharMetrics™, a legacy IMS claims data warehouse. The data was fully de-identified and HIPPA compliant. This analytic file was created to examine the relationship of drug adherence to overall medical costs and events (hospitalization/ER). The data originally was spread out over 15 drug categories but only two of them, antidiabetics and anticoagulants, were included in the dataset received for this study. The index dates ranged from February – October 2013 but the data also included six month pre- and post-index dates on either side. There were 3,552 patients represented in the dataset with 94 variables plus the patient key. The analytic file contained descriptive data as well as pre- and post-index morbidity, HC utilization and cost data. Outcome measures included PDC, and % with PDC>=0.8.

After the dataset was imported, all records in the anticoagulant drug category were removed. Next the data was randomly split into train and test sets with an 80/20 ratio respectively. All features that had low sample sizes (< 50) were removed as were any data with pvalues > 0.2. All post-index data was also removed.

III. Methodology

The variable selection process began by identifying the dependent variable – Post Total Cost – and testing it for linearity as shown in Fig. 1

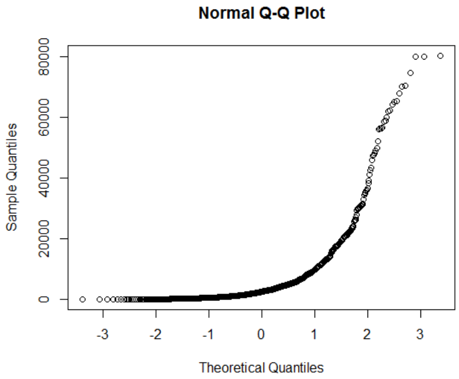

Figure 1. Pre-Normalization Linearity Plot of Post Total Cost

As the resulting q-q plot revealed it was non-linear, the variable was then normalized using the log transformation for both the train and test sets, and another q-q plot (Fig. 2) created for verification.

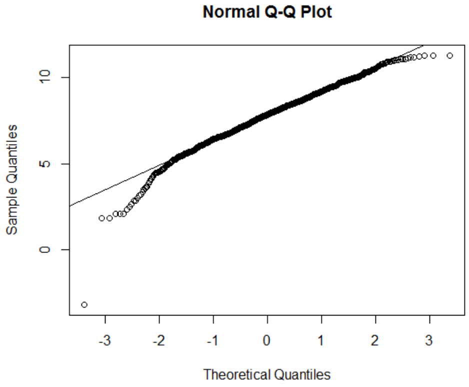

Figure 2. Post-Normalization Linearity Plot of Post Total Cost

A. *Pair Plots*

Pair plots were utilized to spot correlations between any of the continuous variables and remove one of any pair that were highly correlated. As a result of these tests, two variables – Pre Total Cost and Pre Medical Cost – were removed. A second round of pair plots revealed no visibly discernable independent correlation with the dependent variable, Post Total Cost. See Fig. 3.

B. *Correlation Coefficients*

In an effort to discover the optimum set of variables for the model, a test of the remaining continuous variables was run to indicate the correlation coefficients. The resultant variables were then used to create a model that was subsequently eliminated from the study.

C. *Box Plots*

Box plots were then employed as a means of checking for categorical features. The box plots revealed one group with a set of features with overlapping interquartile ranges and close medians, and a second group with dissimilar medians. No eliminations were made on any categorical features on the basis of box plot similarity because it was unclear if the differences were significant. See Figs. 4 and 5.

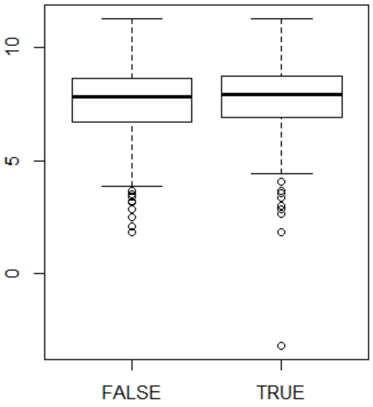

Figure 4. Typical Box Plot of Features with Q1-Q3, Overlap, Close Medians

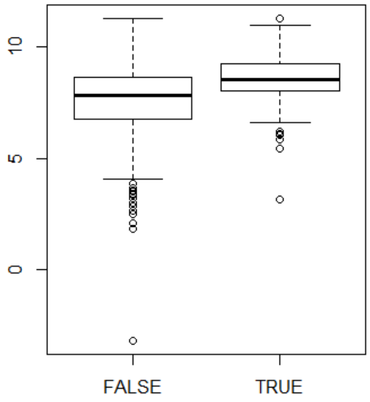

Figure 5. Typical Box Plot of Features with Dissimilar Medians

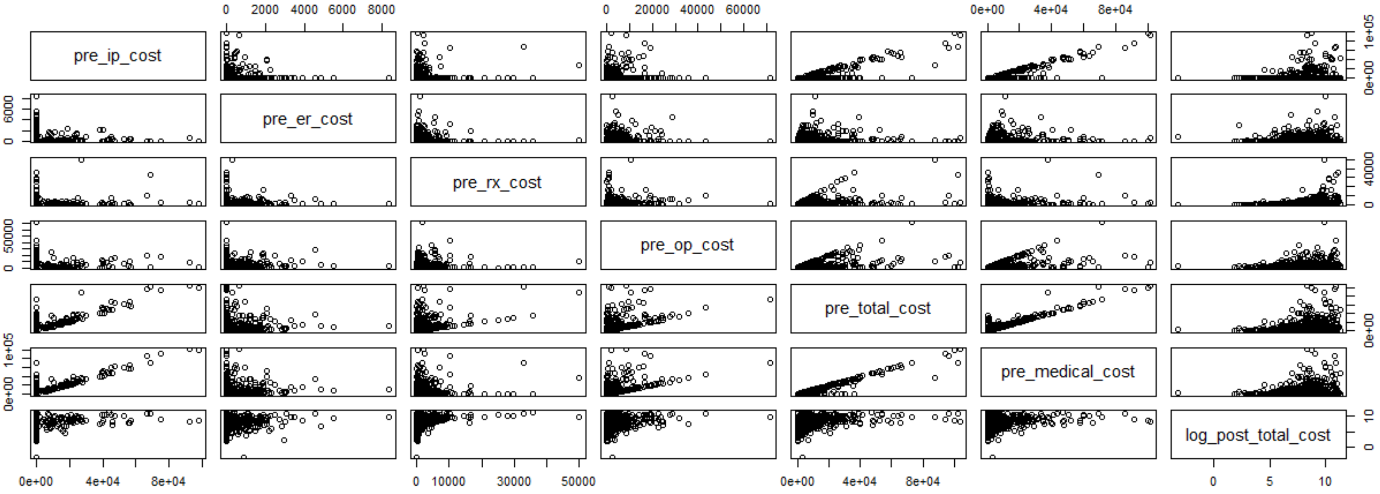

Figure 3. Pair Plots of Continuous Features

D. *Leaps Information Criteria*

The continuous features were then run through the Leaps Information Criteria (IC). The leaps command carries out an exhaustive search for the best subsets of the explanatory variables for predicting a response variable. The Bayesian information criterion (BIC) method was chosen. It is a criterion for model selection among a finite set of models; the model with the lowest BIC is preferred. It is based, in part, on the likelihood function and it is closely related to the Akaike information criterion (AIC).

When fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in overfitting. Both BIC and AIC resolve this problem by introducing a penalty term for the number of parameters in the model; the penalty term is larger in BIC than in AIC. (https://en.)

Fig. 6 shows that the BIC method suggests an optimality of four features without overfitting before the forced inclusion of Gender and Age variables.

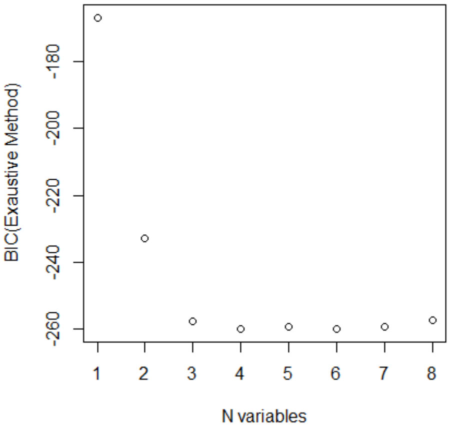

Figure 6. Leaps BIC vs. N Variables

Table 1 lists the features and their respective coefficients as suggested by the BIC Exhaustive Method. These were used for a model that was later eliminated.

TABLE 1. Leaps BIC Coefficients

VARIABLE	COEFFICIENT
Intercept	7.1277581359
Age Group 2: 19-44 years (age_grp2)	-0.2377351084
Gender (sexN2)	0.2185798526
Sleep Disorders (Sleep_Disorders1)	0.3748578471
Charlson Comorbidity Index (CCI) Score, pre-index (pre_CCI)	0.1780233479
RX Cost, pre-index (pre_rx_cost)	0.0001374982
Total Outpatient Visits, pre-index (num_op)	0.0304412418

E. *Ridge Regression*

Ridge Regression (alpha = 0) was used next on both the continuous and categorical features. The results are shown in Fig. 7. The best Lambda from the Ridge Regression is 0.6445485. These results were used for a model that was also later eliminated.

F. *Lasso Regression*

Lasso Regression (alpha = 1) was then used on both the continuous and categorical variables. The results are in Fig. 8. The best Lambda that resulted from the Lasso method is 0.02211104. These results were used for yet another model that was eliminated.

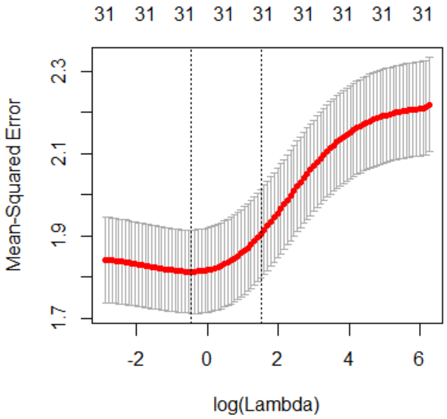

Figure 7. Ridge Regression MSE Vs Log Lambda

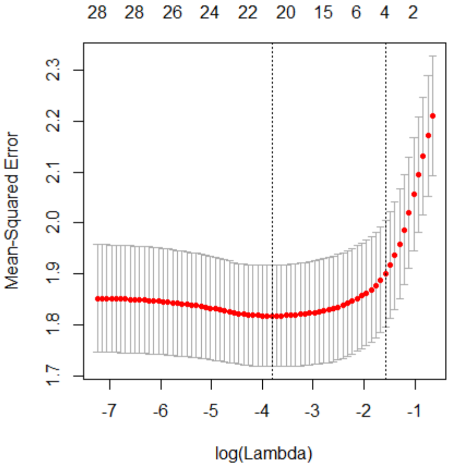

Figure 8. Ridge Regression MSE Vs Log Lambda

Based on observed heteroscedasticity in the residual vs fitted plot for our initial four sets of selected features (models), we eliminated all four of our original models and repeated the selection process a second time.

G. Final Model Selection

Our first model of series II was the result of our employment of Step Wise regression. After being asked to choose core features from the entirety of the dataset, the model suggested a rather large collection of variables which included recommendations of the following features: Age Group 2: 19-44 years (age_grp2), Age Group 3: 45-65 years (age_grp3), Gender (sexN), Osteoarthritis, Sleep Disorders, Pre-Index Charlson Comorbidity Index (CCI) Score (pre_CCI), Pre-Index RX cost (pre_rx), Pre-Index Total Outpatient Costs (pre_op_cost), Pre-Index Inpatient Hospital Stays (num_ip) and Pre-Index Total Outpatient Visits (num_op).

A model comprised of 10 features is prone to overfitting, however since this is the model with the lowest AIC, it was decided to build the second model based on the first using Ridge Regression to further cut down on the number of predictors in our model. In plotting the BIC it was concluded that the best model would be comprised of three to five variables (Fig. 9); as such, we used the Ridge Regression's suggested five variable model which was comprised of Depression, Sleep Disorders, Pre-Index Charlson Comorbidity Index (CCI) Score (pre_CCI), Pre-Index RX cost (pre_rx), and Pre-Index Total Outpatient Visits (num_op).

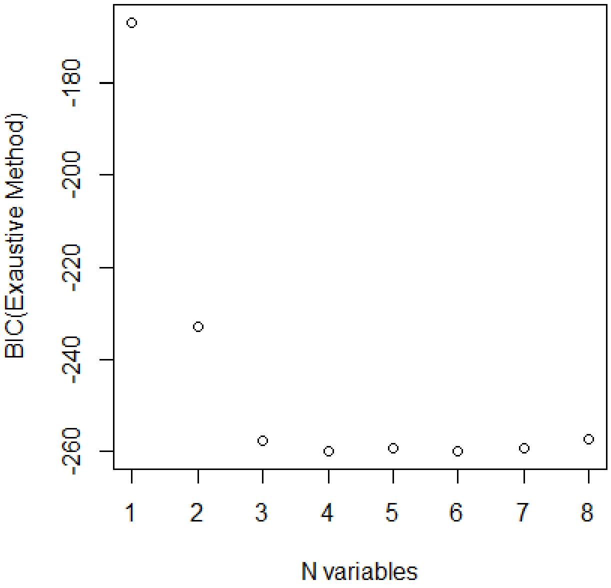

Figure 9. Ridge Regression #2 MSE Vs Log Lambda

For the third model, the initial model of series II was revisited. While paying close attention was paid to the p-values of each feature, it was decided to eliminate those with greater than a .05 p-value. This resulted in a model with a combination of the following features: Age Group 2: 19-44 years (age_grp2), Age Group 3: 45-65 years Gender (sexN)+ Sleep Disorders, Pre-Index Charlson Comorbidity Index (CCI) Score (pre_CCI), Pre-Index Total Emergency Department Costs (pre_er_cost) + Pre-Index RX cost (pre_rx), Pre-Index Total Outpatient Costs (pre_op_cost) and Pre-Index Total Outpatient Visits (num_op).

All three of the models have an adjusted R squared value close to .2, decent enough to begin the residual analysis process. Unfortunately, the similarities in the models do not stop at the adjusted R squared value; when looking at the fitted plot a clear pattern to our errors is evident, thus violating the linearity assumption. The funnel shape, seen below indicates that as the predicted values get larger there is less error. The plot of a good model would instead show error randomly scattered around zero. The resulting funnel shape suggests that the homoscedasticity assumption is also violated.

IV. Results

The ANOVA test was used to compare the three models. It returned small p-values suggesting that most of the features are significant. This also allowed the rejection of the null hypothesis which stated that the models' variation is due to sample variation.

Next, the F-Statistic was utilized to garner an idea of the efficacy of the models. The F-Statistics provided by ANOVA, compares the variation among the different models to the variation within each model. This reveals whether or not the variation among sample means dominates over the variation within groups. The findings from the F-Statistic shows that a good amount of variation among the different models can be owed to variation within the models themselves All in all, the ANOVA shows specifics about the models in comparison to each other, something that simple summaries for each model do not allow us to do.

The ANOVA sum of squares for model 1 is 1565 which is the least of all the three models, thus model 1 was chosen as the best of the three.

References

[1] Diabetes Latest. Centers for Disease Control and Preventions. http://www.cdc.gov/features/diabetesfactsheet/

[2] MC Sokol, KA McGuigan, RR Verbrugge, RS Epstein, Impact of medication adherence on hospitalization risk and healthcare cost. Med Care. 2005 Jun;43(6):521-30.

[3] Todd P. Gilmer, PHD, Patrick J. O'Connor, MD, MPH, William A. Rush, PHD, A. Lauren Crain, PHD, Robin R. Whitebird, PHD, Ann M. Hanson, BA and Leif I. Solberg, MD, Predictors of Health Care Costs in Adults With Diabetes. Diabetes Care, January 2005, vol. 28 no. 1 59-64.