Jared Campbell

Olson-Manning

Bio 291: Big Data in Medicine

21 Jan 2019

The Importance of Doctor Visits

Introduction

In 2016, the United States almost doubled the health care spending of other high-income countries (Papanicolas et al. 2018). With the Affordable Care Act(ACA), the uninsured healthcare rate has dropped, giving more people access to care (Obama 2016). With more access to healthcare in the U.S., it has become a difficult but important to find a way to minimize costs. A sizeable portion of these costs can be tied to inpatient care, or more simply, hospital visits. Each time a patient arrives to the hospital for care, a cost is incurred.

Adding to costs, current payment systems promote volume based care. Hospitals are incentivized to deliver more services to more people (Miller 2009). Healthcare providers gain more revenue for every service delivered, regardless of its need or outcome. Instead of promoting healthy patients with minimum intervention, the current system promotes the opposite.

Leading healthcare providers are beginning the switch to another payment system based on value (personal communication, Emily Griese of Director of Sanford Population Health). Patients pay a base charge for healthcare, regardless of the number of visits tendered over the time period. With this new system, healthcare providers are encouraged to maximum their efficiency. Being able to predict how often each patient will come in for a visit allows proper billing and minimizes excess costs.

The US healthcare system utilizing electronic health records combined with advances in big data techniques gives opportunities to reduce healthcare costs (Bates et al. 2014). Combining the Sanford Dataset with data analysis techniques, I will model the patient doctors visits based on hypertension, diabetes, vascular disease, age, and sex.

Study 1

Method

The topic of interest for this study was to model patient doctor visits using the variables presented in the given Sanford dataset. These variables included Sex, Age, Living Status, Hypertension, Vascular Disease, Payor, Diabetes, A1C Levels, BMI, Scheduled Visits, Missed Visits, Diastolic BP, Systolic BP, and Smoking Status. First, the BMI, Visits, Age, and A1C variables were converted to numeric values to better represent the variable. Similarly, the flag variables Hypertension, Vascular Disease, and Diabetes were converted to factors. Finally, rows with incomplete information or outlier BMI's (over 100) were removed from the dataset.

Using a random index, the cleaned dataset was split into two subsets: train and test. The training data set contains a randomized 60% of the original data, with the test containing the remaining 40%. Using the training set, I began to create linear models to predict Scheduled Clinic Visits. The first model used every other variable in the dataset for the prediction. The test data set is then used to compute the root-mean-square error to measure the accuracy of the predicted model. A smaller the root-mean-square error shows a higher accuracy of the model with 0 meaning a perfect fit.

Results and Discussion

The results for the first model is shown below in Table 1. While the model contained multiple variables with high significance (***), the R^2 value is only 0.2167. So the model using all variables only explains about 22% of the proportion of variance in Scheduled Clinic Visits. Since the variables Payor-Medicare and Smoking Status were not significant, removing those variables is a logical step to improve the next model. This model's root-mean-square error was 5.799, an important value to note but will not provide much knowledge until we compare it to another model.

Table 1

Regression coefficients, standard error, and significance of each predictor variable along with significance measurement(R^2) of the model.

```
[1] 5.79851

Call:
lm(formula = ScheduledClinicVisits ~ ., data = trainData)

Residuals:
    Min      1Q  Median      3Q     Max
-43.199  -3.285  -1.436   1.740  97.456

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            4.068893   0.193111  21.070  < 2e-16 ***
SexMale               -1.159955   0.039868 -29.095  < 2e-16 ***
Age                    0.013999   0.002127   6.581 4.71e-11 ***
Hypertension1          0.876240   0.041033  21.355  < 2e-16 ***
VascularDisease1       2.035801   0.070073  29.053  < 2e-16 ***
PayorMedicare          0.141931   0.111086   1.278    0.201
PayorPrivate Ins/Other -1.562858   0.102040 -15.316  < 2e-16 ***
Diabetes1              1.505460   0.046236  32.560  < 2e-16 ***
BMI                    0.025709   0.002905   8.849  < 2e-16 ***
MissedClinicVisits     2.719191   0.023434 116.034  < 2e-16 ***
SmokingStatus          0.018175   0.014979   1.213    0.225
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.663 on 83751 degrees of freedom
Multiple R-squared:  0.2168,    Adjusted R-squared:  0.2167
F-statistic:  2318 on 10 and 83751 DF,  p-value: < 2.2e-16
```

Study 2

Method

Building off of model one, I removed the insignificant variables Payor and Smoking Status. Since Payor is a factor variable type, and had one factor level insignificant, in order to remove the one level, the entire variable must be removed. This model will predict Scheduled Clinic Visits using only the significant variables from the first model.

Results and Discussion

The results for the first model is shown below in Table 2. As expected, the model only contains significant predictor variables, but the R^2 value is only 0.2117. This model only explains about 21% of the proportion of variance in Scheduled Clinic Visits. Even though the insignificant variables were removed, the model has a lower significance than the original model with all variables used. This result was unexpected, as removing insignificant data should improve the model's significance. Furthermore, the root-square-mean error for the second model (5.730) is roughly the same as the first (5.799). Based on the R^2 and root-square-mean error it is evident that neither model predicts scheduled visits well. The true predictor variables for Scheduled Clinic Visits are unknown or at least not given in this dataset.

Table 2
Regression coefficients, standard error, and significance of each predictor variable along with significance measurement(R^2) of the model.

```
[1] 5.734087

Call:
lm(formula = ScheduledClinicVisits ~ Sex + Age + Hypertension +
    VascularDisease + Diabetes + BMI + MissedClinicVisits, data = trainData)

Residuals:
    Min      1Q  Median      3Q     Max
-46.665  -3.404  -1.512   1.691 125.926

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.801923   0.157806   5.082 3.75e-07 ***
SexMale            -1.155724   0.040413 -28.598  < 2e-16 ***
Age                 0.056303   0.001531  36.786  < 2e-16 ***
Hypertension1       0.873630   0.041739  20.931  < 2e-16 ***
VascularDisease1    2.117529   0.071301  29.698  < 2e-16 ***
Diabetes1           1.528335   0.046933  32.565  < 2e-16 ***
BMI                 0.025111   0.002931   8.568  < 2e-16 ***
MissedClinicVisits  2.881085   0.022807 126.325  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.764 on 83754 degrees of freedom
Multiple R-squared:  0.2118,    Adjusted R-squared:  0.2117
F-statistic:  3215 on 7 and 83754 DF,  p-value: < 2.2e-16
```

Study 3

Method

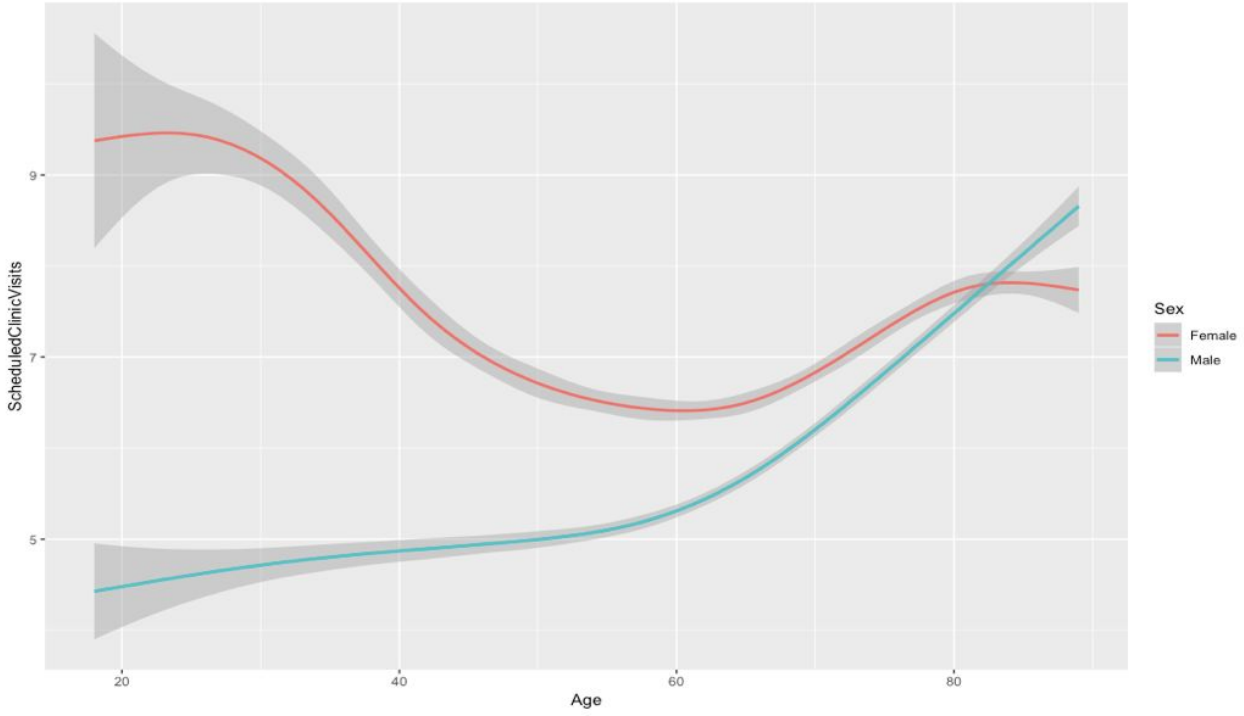
The aim of this study is to visualize the relationship between Sex, Age, and Scheduled Clinic Visits. Sex and Age are two of the most basic pieces of information about a person, and most often the first two known. Since both variables were significant when modeling in the previous studies, there is a correlation between Age and Sex with Scheduled Clinic Visits. Using the cleaned dataset, I generated a smooth correlation plot of Age vs. Scheduled Clinic Visits grouped by Sex, with the uncertainty shown on the shaded regions.

Results and Discussion

The results of the Age vs. Scheduled Clinic Visits grouped by Sex is shown below in Figure 1. The results are largely as expected. Women during reproductive age have the highest number of visits and once people age beyond 60, their number of visits start to increase. While these results do not generate new ideas, it is evidence to support the existing conceptions of how age and sex correlate to doctor visits.

Figure 1

Smooth plot of patient Age vs. Scheduled Clinic Visits color coded by sex (blue = male, red = female). The shaded region represents the 95% confidence interval of the data.



Conclusion

While Study 1 and 2 failed to show a model that accurately predicts doctor visits by having R^2 values of 0.2167 and 0.2117, respectively, they did show some predictor variables that must be considered in further studies with more variables. Study 3 showed the expected correlations between Age, Sex and Scheduled Doctor Visits by highlighting the increase in visits for reproductive age women and the elderly. With a new US healthcare system focused around

value, being able to predict doctor visits can provide a key opportunity to smooth the transition. The studies here are another small step towards understanding what factors lead people to the hospital more.

Reference List

Asaria M, Doran T, Cookson R. The costs of inequality: whole-population modelling study of lifetime inpatient hospital costs in the English National Health Service by level of neighbourhood deprivation. *J Epidemiol Community Health* 2016;**70:**990-996.

Bates, David W., et al. “Big Data In Health Care: Using Analytics To Identify And Manage High-Risk And High-Cost Patients.” *Health Affairs*, vol. 33, no. 7, 2014, pp. 1123–1131.

Lotterhos, Kathleen E., et al. “Analysis Validation Has Been Neglected in the Age of Reproducibility.” *PLOS Biology*, vol. 16, no. 12, 2018.

Malik, M. M., et al. “Data Mining and Predictive Analytics Applications for the Delivery of Healthcare Services: a Systematic Literature Review.” *Annals of Operations Research*, vol. 270, no. 1-2, 2016, pp. 287–312.

Miller, Harold D. “From Volume To Value: Better Ways To Pay For Health Care.” *Health Affairs*, vol. 28, no. 5, 2009, pp. 1418–1428.

Obama B. United States Health Care Reform: Progress to Date and Next Steps. *JAMA.* 2016;316(5):525–532.

Papanicolas I, Woskie LR, Jha AK. Health Care Spending in the United States and Other High-Income Countries. *JAMA.* 2018;319(10):1024–1039.

R J Rubin, W M Altman, D N Mendelson; Health care expenditures for people with diabetes mellitus, 1992, *The Journal of Clinical Endocrinology & Metabolism*, Volume 78, Issue 4, 1 April 1994, Pages 809A–809F.

Sjoding, Michael W., et al. “Rising Billing for Intermediate Intensive Care among Hospitalized Medicare Beneficiaries between 1996 and 2010.” *American Journal of Respiratory and Critical Care Medicine*, vol. 193, no. 2, 2016, pp. 163–170.

Teuscher A, Egger M, Herman JB. Diabetes and Hypertension: Blood Pressure in Clinical Diabetic Patients and a Control Population. *Arch Intern Med.* 1989;149(9):1942–1945