Michael Ruby

ECON484

6/1/2018

Introduction

As state run entities, hospitals frequently struggle to make their funding cover all their needs. This means that they may frequently be short staffed or running out of resources. This can become especially problematic during emergencies, causing a large influx of people in urgent need of attention with varying levels of severity. This issue quickly compounds; as patients badly in need of medical attention remain untreated their prospects continue to diminish causing not only an increased chance of death but also increasing the amount of medical attention the doctors are forced to pay to them. Instead, to most efficiently admit and aid patients, hospital administrators should be able to quickly predict based on admittance data, as well as subsequent data that comes in, how long a given patient will be in a bed before they are discharged or expire.

Further, if hospitals could reliably determine the amount of time before another bed was likely to be open they could quickly react to catastrophic events. In the event of a major local event, there is likely to be a shortage of beds in the immediate area, however, the ability to determine openings in your hospital or nearby hospitals could aid in the transferring of critical patients in need of specialist attention, or to help the government determine where the best allocation of resources and first responders is. We will attempt to solve these programs by doing an analysis of 2015 DGIS Hospital data to determine how different conditions and attributes determined during hospital intake and throughout the patients time at the hospital can help to predict the length of time a patient will spend at the hospital.

Methodology

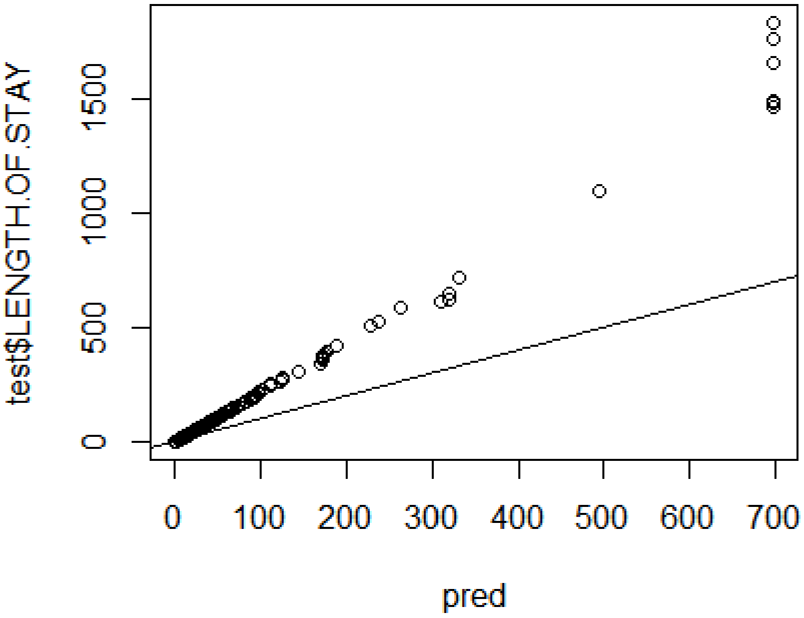
For our analysis we looked at DGIS, Mexico's governmental health administration and database. The data used for analysis is the *2015 Database of Hospital Discharges*, produced my Mexico's Ministry of health. Contained in the data received was anonymized data about public hospital patients throughout the entire country, including personal characteristics and background as well as information about any medical care received. While the data contained various files containing information about deaths, births and surgical procedures; for the purposes of this analysis we focused only on information contained in the "EGRESO" file. This database contained all data obtained by the hospital throughout their stay at the hospital as well as all the patient background and characteristics.

For the purposes of this analysis we did not use the "Productos", "Obstet", "Defunc", or "Afecciones" sets. Our reason for excluding them was since we were trying to predict based hospital stay length based on intake data; and this was all discharge data or events that occurred during their stay, they would not have been available at the time the data would have been analyzed. To clean the data, we began by removing all out of range values; these were generally 9, 99, 999, or 9999, these were all changed to "NA's". Next, we changed the appropriate data to factors and renamed the variables and levels accordingly. Finally, we removed psychiatric patients since they tended to have longer stays and have less impact on hospital bed use in any urgent or emergency rooms.

For analysis we will take a look at a handful of variables; we will run a boosted regression of age, sex, weight, height, insurance program, municipality (county) of residence, municipality of hospital, if the person is indigenous, the type of hospital service, the first and second diagnosis, the amount of times they have been diagnosed with that condition and if they were infected with an intra-hospital infection on length of stay. Since age was coded as a number and a character we first had to create a factor and an interaction between them, then we proceeded to complete the regression.

Results

The boosted regression is a very good fit for this model we have a residual-mean square error of 31.97999. Additionally, the predicted stay appears to be very accurate below roughly 300 days, when the data begins to act erratically until about 1000 days when there appears to be very little predictability. The most likely explanation for the outliers are that they are typos, or misclassifications, since it is extremely unlikely that someone is spending over 2.5 years in the hospital it is most likely that the fields were transposed for date of birth or another field. Although it appears that we have consist underestimated the length of stay, it does appear that there is a linear relationship between out predictors and the length of stay.



After running a surrogate model to determine the partial effects, we find that unsurprisingly, Intra-hospital infection has the greatest impact on increasing time at the hospital, at almost 7.5 days. While being a female has the largest decrease on you time in the hospital, decreasing it by nearly a full day. This is likely due to the number of expectant mothers that come in and lower the overall time. Of the variables Age clearly has the most diverse partial effect on hospital time increasing it early and late, while decreasing it in the middle.

Conclusion

With this knowledge hospital administrators will not only be able to efficiently distribute resources and assign hospital workers but will also be able to decrease waiting times. By implementing our analysis further to determine turnaround times in the emergency and urgent care centers. Administrators will be able to determine not only how long patients will be in the hospital, but further data will allow for determining how long the patient will be in the emergency or urgent care before they are discharged or transferred to a separate section of the hospital for more long-term care.