Final Report for MA678 Project

Yeyang Han

2022/12/11

Abstract:

In this essay, we focus on health and use All Payer Hospital Inpatient Discharges by Facility (SPARCS De-Identified) dataset to explore factors that influence patients' average length of stay. At first, we use EDA to analyze situations with raw data and convert some data into numbers which are easy for us to build models. Then, because of the difference in hospitals, we use the multilevel regression model to see how factors influent patients' average length of stay. Finally, we do some checks for the built model.

Introduction:

Health is a big topic that has received much attention. Here, we talk about all-payer hospital inpatient discharges by facilities. The collected data is from The Statewide Planning and Research Cooperative System which is a comprehensive data reporting system that collects patient-level detail on patient characteristics, diagnoses, treatments, services, and charges for every hospital discharge from an Article 28 facility; ambulatory surgery discharges from hospital-based ambulatory surgery centers and all other facilities providing ambulatory surgery services; and emergency department visits in New York State.

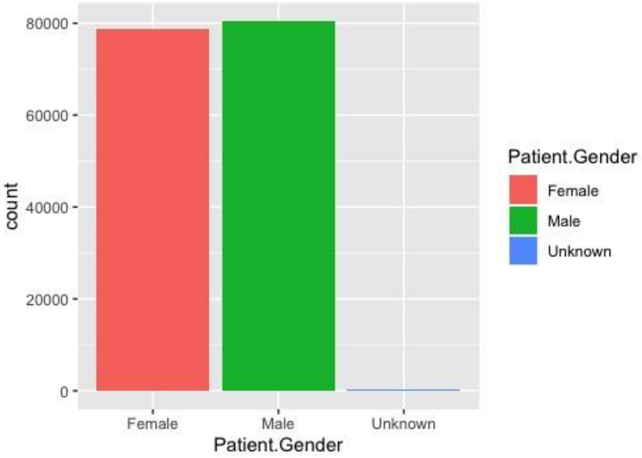
Data Processing:

For easy access to data in R Studio, we change some character variables into numeric ones. The details are below:

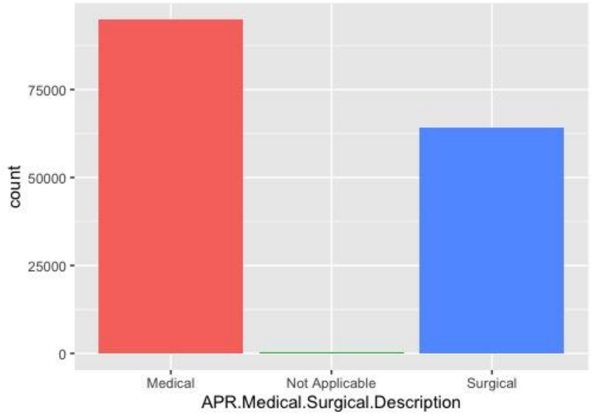
Column Name	Description
APR Medical Surgical Description	Medical(1), Surgical(2) or Not Applicable(3)
Type of Insurance	Commercial(1), Medicare(2), Medicaid(3), Other(4), Unknow(5)
Patient Gender	Male(1), Female(2), Unknown(3)
Discharged Dead or Alive	Discharged Alive(1), Discharged Dead(2)
Patient Age Group	0(0), 1-12(1), 13-17(2), 18-40(3), 41-64(4), 65-74(5), 75+(6)

EDA:

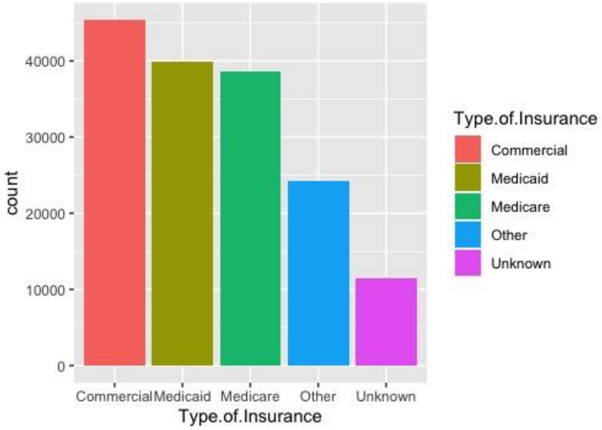
Before we make a model to analyze data, it is necessary to realize situations of data.



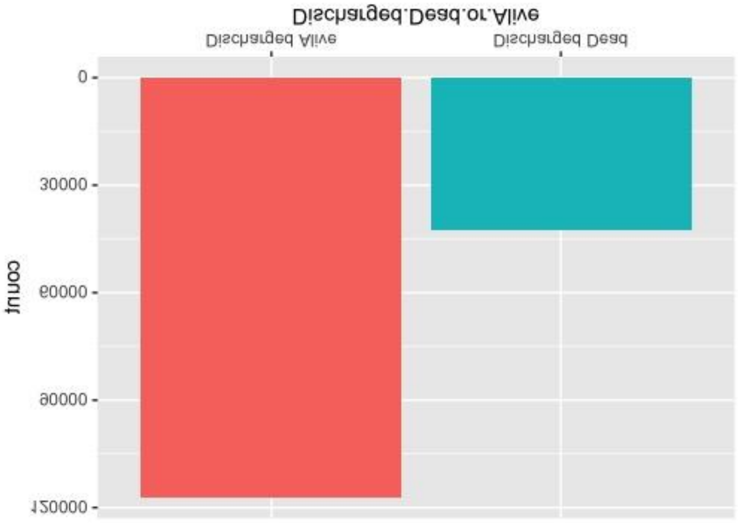
From figure 1, we know the man and women in this dataset are similar. Therefore, conclusions cannot be affected because of the difference percentage of gender.



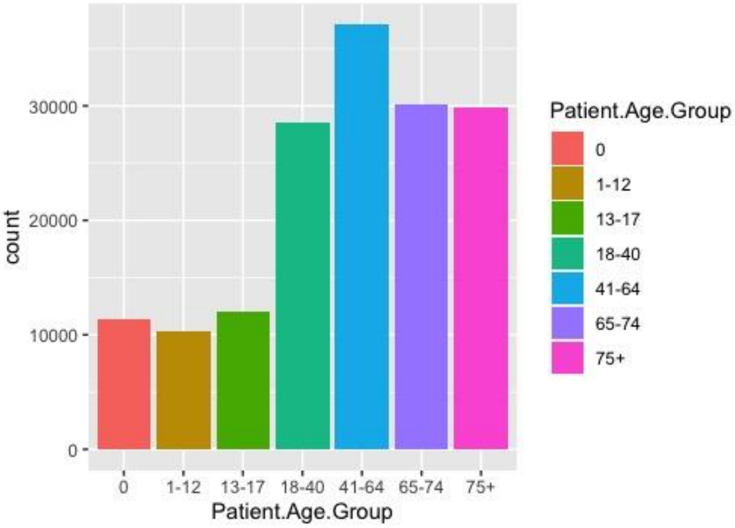
According to figure 2, we can see that most of the patients' APR-DRG-specific classification is medical, and a few patients' APR-DRG-specific classification is not applicable.



When we look at what kind of insurance(Figure 3) patients take when they are in the hospital, we can find that the number of people who take Commercial insurance makes up first place compared with others. The number of people who take Medicaid insurance ranks second place and the number of people with Medicare insurance is close to it. Only some patients take other insurance except those patients we do not know what kind of insurance they take.



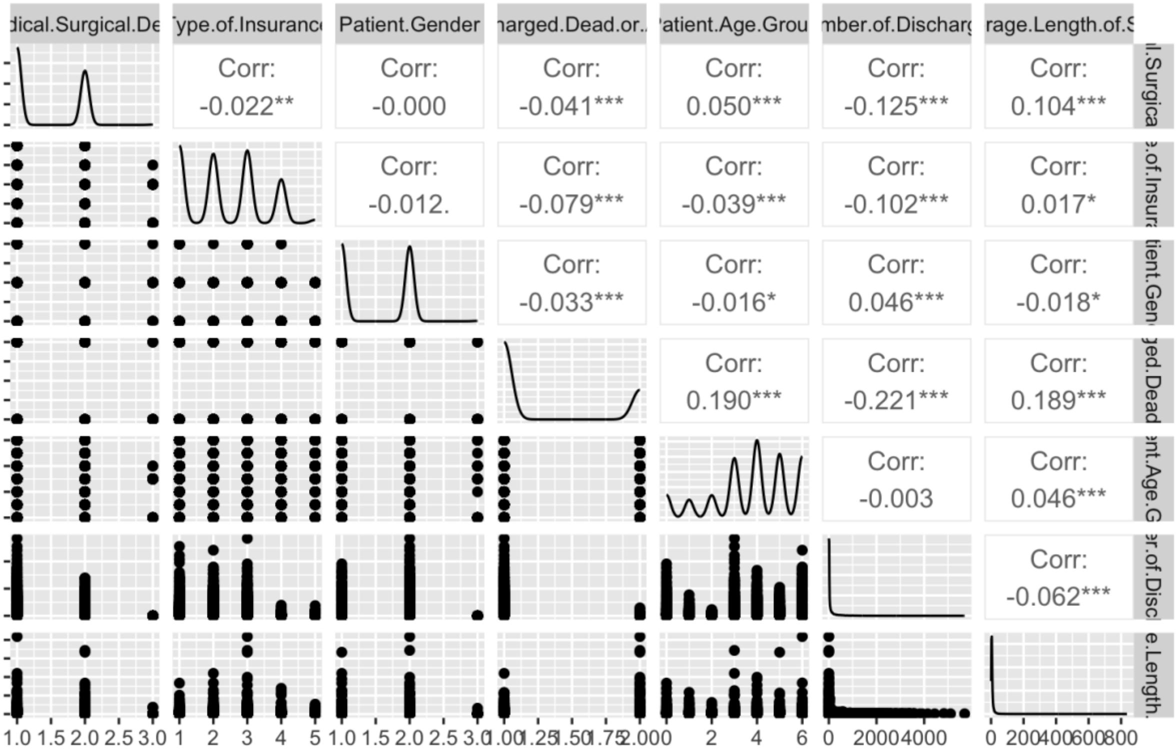
When identifying if the patient was discharged from the facility alive or dead (Figure 4), we can see most patients are alive even though some of them are dead.



According to the car chart, we find that most of the patients in this dataset are more than 18. People aged between 41-64 are the most, which means their bodies easily have problems and need to go to the hospital compared with others.

Model Build:

Different variables on different levels affect patients' average length of stay. But what exactly kind of variables influence patients' average length of stay, we need to explore it.



At there, patients' average length of stay is what we want to explore. So, this variable is very important and we cannot delete it. In addition, according to the plot, we see that the relationship coefficient between different variables is not big. I reckon the reason for that is most variables are discrete variables. Then, to select variables that we need in the model, we see the significance between different variables. Even though all factors are significant with patients' average length of stay, we know that Number.of.Discharges have no relationship with patients' average length of stay. Therefore, we delete it. Besides, the significance of the type of insurance and Patient Gender with patients' average length of stay is less compared with other variables.

Because the situation of different hospitals is different, so we cannot evaluate both hospitals together. To improve our model's accuracy, we treat different hospitals as different groups to calculate.

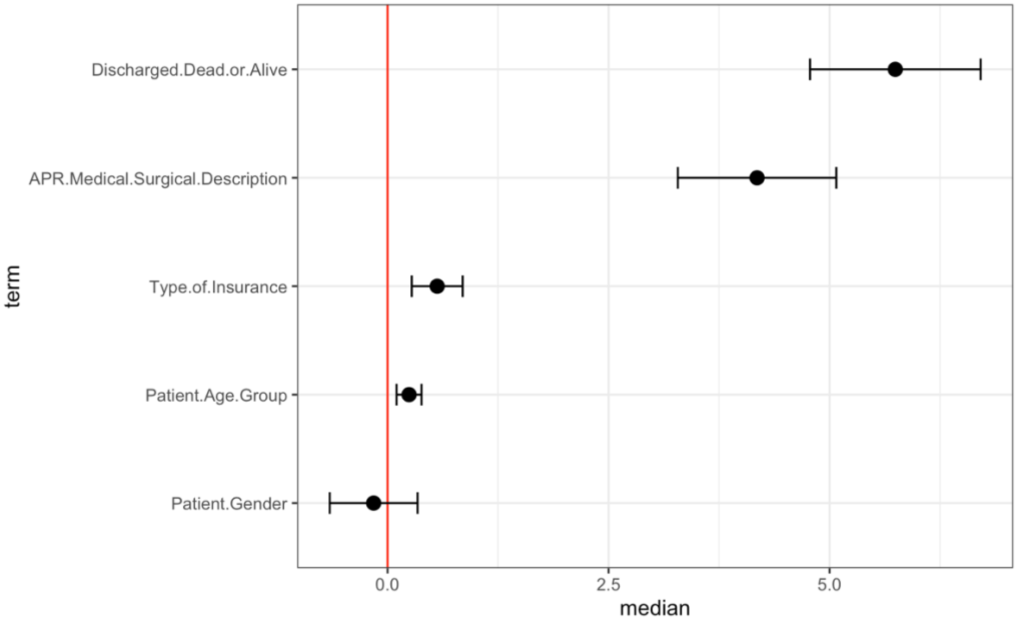
Formula: Average.Length.of.Stay ~ APR.Medical.Surgical.Description + Type.of.Insurance + Patient.Gender + Discharged.Dead.or.Alive + Patient.Age.Group + (1 + APR.Medical.Surgical.Description + Type.of.Insurance + Patient.Gender + Discharged.Dead.or.Alive + Patient.Age.Group | Facility.ID)

According to the results, we can look at the estimated model averaging over hospitals by fixed effect and the hospital-level errors by random effects.

Fixed effect:

```
                         Estimate Std. Error       df t value Pr(>|t|)
(Intercept)                 -6.71211    1.15103 112.45478  -5.831 5.36e-08 ***
APR.Medical.Surgical.Description   4.20270    0.51685  28.46017   8.131 6.62e-09 ***
Type.of.Insurance               0.58651    0.17434 107.36693   3.364  0.00107 **
Patient.Gender                -0.17938    0.28399 133.09713  -0.632  0.52870
Discharged.Dead.or.Alive          5.73628    0.58227 181.84849   9.852  < 2e-16 ***
Patient.Age.Group               0.25790    0.08324 212.21836   3.098  0.00221 **
```

From the above table, we can see most variables' p-vale are below 0.05 which is significant, except Patient.Gender. The reason for this situation maybe Patient.Gender cannot effect patients' average length of stay. More details about these variables can look at following plot.



From the fixed effect, we can get the estimated regression line in every hospital is

$$y_{\text{Average.Length.of.Stay}} = -6.71211 + 4.20270 * \text{x}_{\text{APR.Medical.Surgical.Description}} + 0.58651 * \text{x}_{\text{Type.of.Insurance}} - 0.17938 * \text{x}_{Patient.Gender} + 5.73628 * \text{x}_{Discharged.Dead.or.Alive} + 0.25790 * \text{x}_{Patient.Age.Group}$$

The coefficient of APR.Medical.Surgical.Description, Type.of.Insurance, Discharged.Dead.or.Alive, Patient.Age.Group is 4.20270, 0.58651, 5.73628, 0.25790 separately. Besides, all of them are positive which that means with every increase of these variables, patients' average length of stay also increases. Patient.Gender has a negative effect on patients' average length of stay. Compared with males, female patients' average length of stay is lower.

Random effects:

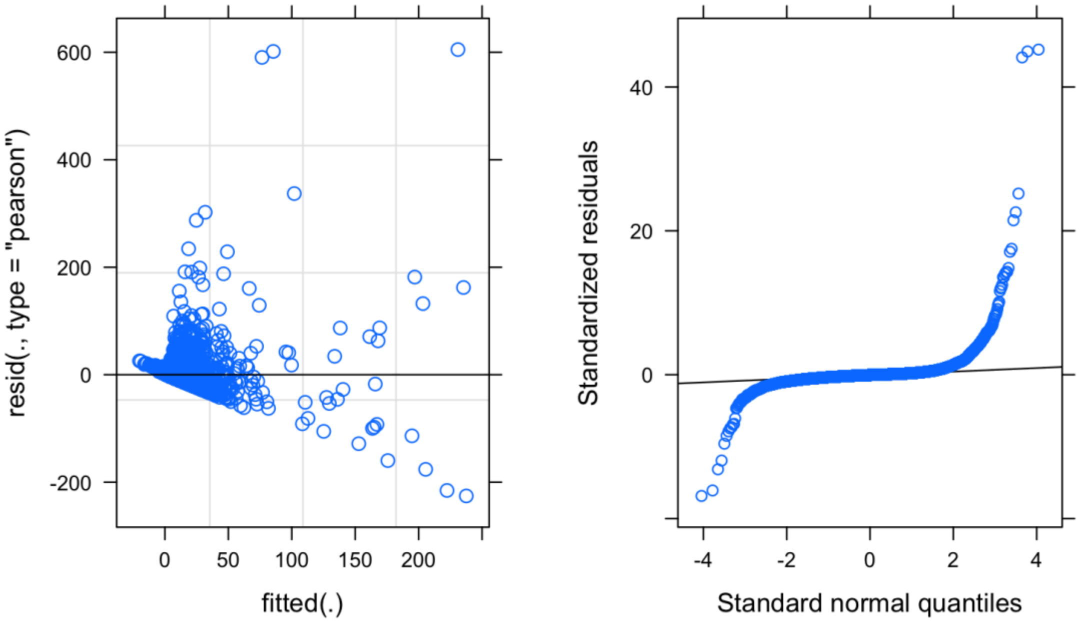
Facility.ID	(Intercept)	APR.Medical.Surgical.Description	Type.of.Insurance	Patient.Gender	Discharged.Dead.or.Alive	Patient.Age.Group
1	2.03	-0.49	-0.28	-0.40	-0.28	-0.28
2	4.38	-2.54	-0.93	-1.32	-2.20	0.08
4	5.66	-2.48	-0.59	-0.63	-3.06	0.23
5	2.34	-1.38	-0.39	-0.50	-1.48	0.18
12	4.16	-1.18	0.20	0.68	-2.78	0.44
37	3.75	-1.96	-0.62	-0.83	-1.96	0.11

There are more than 200 hospitals in this dataset, I only show the head of them. These data tell us how much the intercept is shifted up or down in the different hospitals. Take South Nassau Communities Hospital as an example, Facility.ID is 1, the estimated intercept is 2.03 higher than average and the estimated APR.Medical.Surgical.Description, Type.of.Insurance, Patient.Gender, Discharged.Dead.or.Alive, Patient.Age.Group is 0.49, 0.28, 0.40, 0.28, 0.28 separately lower than average.

Thus, add the fixed effects to South Nassau Communities Hospital which Facility.ID is 1getting the final intercept and slope.

$$y_{\text{Average.Length.of.Stay}} = -4.681 + 3.709 * \mathrm{x}_{\text{APR.Medical.Surgical.Description}} + 0.310 * \mathrm{x}_{\text{Type.of.Insurance}} - 0.580 * \mathrm{x}_{Patient.Gender} + 5.456 * \mathrm{x}_{Discharged.Dead.or.Alive} + 3.533 * \mathrm{x}_{Patient.Age.Group}$$

Model Check:



When we check this model, we use residual plot and Q-Q plot. As for the residual plot, we use it to assess the model's overall accuracy and the accuracy of individual predictions. However, we see that residuals not distribute equally around 0 and have some trend. So, this model’s accuracy is not good. Besides, we also can see that the model’s accuracy exist some problems when we plot Q-Q plot. Because head and tail data are not close to the line.

Discussion:

During the period when we analyze and build the model, we find the variables that we used are limited. We need to find more variables to realize the relationship with patients' average length of stay. Besides, the low accuracy of the model verified again that we need to find more correlated variables with patients' average length of stay. Even though there are many things we can do in the future, we know the relationship between APR.Medical.Surgical.Description, Type.of.Insurance, Patient.Gender, Discharged.Dead.or.Alive and Patient.Age.Group and patients' average length of stay and give us a mind so that we know how to study patients' average length of stay.

Reference:

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59(4), 390-412. doi:10.1016/j.jml.2007.12.005