Using Linear Regression to Predict Hospital Traffic

Shweta Shambhavi & Stephanie C. Cyrill

Abstract- Predictive modeling is a useful technique that can aid in anticipating patient stays in the hospital. Knowing this information in advance can aid in hospital management, increase bed availability, improve medical supply ordering to treat patients, and medication ordering as they would know what people typically come in for. This paper describes our employment of linear regression modeling techniques to analyze and attempt to predict the number of admittances to the hospital using MIMIC-III data.

Introduction

Hospital admissions can fluctuate in a given year, and this can be due to a variety of reasons, such as insurance, diagnosis, or whether the patient needs urgent medical attention. Because of reasons like this and many more, it can be difficult for hospitals to predict when they will see the highest number of admissions, thus impacting the amount of medical resources they have on-hand. As a result, hospitals can be undersupplied, and unfortunately in times when it matters the most. When the COVID-19 pandemic arose, it exposed how hospitals are ill-equipped for major health crises. Some of the challenges that hospitals faced with the pandemic were decreased bed availability, difficulty maintaining and expanding hospital capacity to treat patients, and severe shortages of testing supplies (Grimm, 2020).

Based on this, hospitals could not have accommodated for serious illness outbreaks like COVID-19, but knowing beforehand when hospital admissions typically pick up during the year might have been helpful. Knowing beforehand an estimate of the number of people admitted to the hospital would allow for the hospital to be prepared. For example, knowing that flu season typically peaks between December and February is helpful because then hospitals know to order more doses of the flu shot, advocate for preventative measures their patients can take (i.e. getting the flu shot early), and accommodating for the elderly or patients with weakened immune systems by making sure they have enough beds for immediate care. This is just a snapshot of the possibilities that accompany forecasting hospital admissions. Therefore, in this study, we employed the use of linear regression models in an attempt to forecast hospital admissions based on MIMIC-III data.

Methodology

Using MIMIC-III data, we exported the admissions data for all the patients in the hospital and imported it into Google Colab. In Google Colab, we imported numpy (np), pandas (pd), matplotlib.pyplot (plt), and seaborn (sns). Throughout the course of the data preparation, however, we mainly used pd as it can be used to manipulate numerical tables and time series. In Google Colab, we prepared the dataset by dropping the columns that were not pertinent to what we were looking for. This included Row_ID, Subject_ID, DEATHTIME, Admission_Location, Discharge_Location, Insurance, Language, Religion, Marital_Status, Ethnicity, EDREGTIME, EDOUTTIME, Diagnosis, Hospital_Expire_Flag, Has_ChartEvents_Data, and Admission_Type. Thus leaving only the HADM_ID, ADMITTIME, and DISCHTIME columns. Following this, we began what we called ‘Data Preprocessing’. In this step, we converted the admission and discharge times to datetime type using the following code:

```
dataset['ADMITTIME'] =
pd.to_datetime(dataset['ADMITTIME'])
```

Following this, we used pd to create a pivot table that displayed ADMITTIME in YYYY-MM-DD format using the following code:

Dataset For Daily Admission

```
dataset_new = pd.pivot_table(dataset,values =
'HADM_ID', columns=['ADMITTIME'],
aggfunc=np.count_nonzero)
```

Dataset for Monthly Admission

```
df_month = pd.pivot_table(df_month,
values='Admissions',  columns=['Period'],
aggfunc=np.sum)
```

With the new dataset, associating admission IDs with the dates that patients entered the hospital was made easier and laid the foundation for the next steps. After reviewing the index of the entire dataset, we separated each of the time components, making admissions, weekday, month, date (i.e. the numerical day), and year.

This was done so that based on past information, we could use our linear regression model to predict what exact day, week, and month we could expect an increase in hospital admissions (Figure 1 & 2).

ADMITTIME	Admissions	DAY_OF_WEEK	DATE	MONTH	YEAR
2100-06-07	1	0	7	6	2100
2100-06-09	1	2	9	6	2100
2100-06-14	2	0	14	6	2100
2100-06-22	1	1	22	6	2100
2100-06-24	1	3	24	6	2100
...	...	...	...	...	...
2208-08-19	1	4	19	8	2208
2209-02-09	1	3	9	2	2209
2209-07-14	1	4	14	7	2209
2209-07-31	1	0	31	7	2209
2210-08-17	1	4	17	8	2210

Figure 1. Display of daily admission dataset in separate components

Period	Admissions	Month_Apr	Month_Aug	Month_Dec	Month_Feb	Month_Jan	Month_July	Month_Jun	Month_Mar	Month_May	Month_Nov	Month_Oct	Month_Sep
2100-06	7	0	0	0	0	0	0	1	0	0	0	0	0
2100-07	30	0	0	0	0	0	1	0	0	0	0	0	0
2100-08	43	0	1	0	0	0	0	0	0	0	0	0	0
2100-09	51	0	0	0	0	0	0	0	0	0	0	0	1
2100-10	37	0	0	0	0	0	0	0	0	0	0	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
2208-05	1	0	0	0	0	0	0	0	0	1	0	0	0
2208-08	1	0	1	0	0	0	0	0	0	0	0	0	0
2209-02	1	0	0	0	1	0	0	0	0	0	0	0	0
2209-07	2	0	0	0	0	0	1	0	0	0	0	0	0
2210-08	1	0	1	0	0	0	0	0	0	0	0	0	0

1291 rows × 13 columns

Figure 2. Display of monthly admission dataset in separate components

With this dataset being complete, it was possible to explore the admissions data by plotting the points for the year, day, month, and weekday (Figures 2-6).

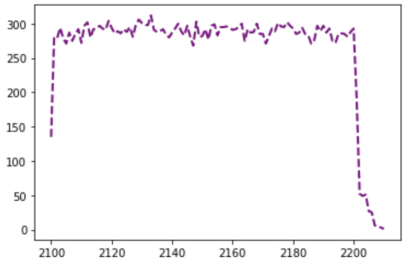


Figure 3. Analysis of admissions to the hospitals over the years

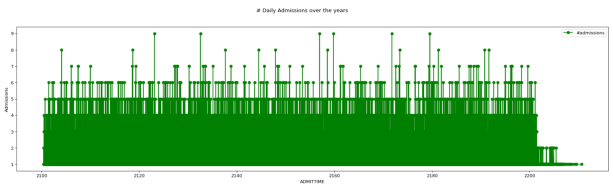


Figure 4. Analysis of daily admissions to the hospitals over the years (enlargement of figure on page 5)

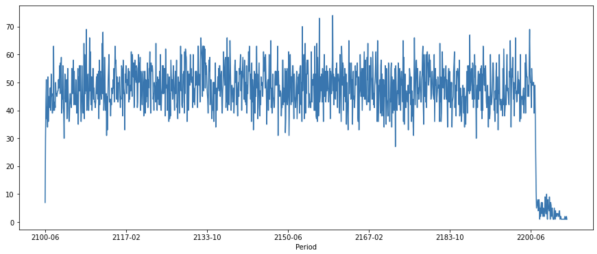


Figure 5. Analysis of monthly admissions to the hospitals over the years

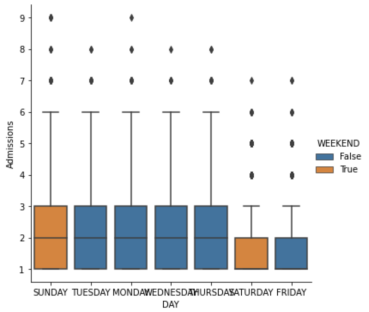


Figure 6. Analysis of which day of the week received the most admissions

Since we wanted to predict the number of admissions based on timeframes and have prepared the dataset as such, it could fit the simple regression model well. We prepared two datasets and two models, for daily (Figure 1) and monthly admission predictions (Figure 2). For both of these models, we chose Ordinary Least Square (OLS) because we wanted the model to be easily interpreted. Linear regressions use OLS models to "assume that the analysis is fitting a model of a relationship between one or more explanatory variables and a continuous or at least interval outcome variable that minimizes the sum of square errors, where an error is the difference between the actual and the predicted value of the outcome variable" (Zdaniuk, 2014). The OLS Model was imported from the stats model library. In addition to this, admissions were based on days, weeks, months, and the year. We expressed our target variable 'Admissions' dependent on the feature set as follows for Daily and Monthly model:

<u>Expressions for Daily Model</u>

expr = 'ADMISSION ~ DAY_OF_WEEK + DATE + MONTH + YEAR'

Expressions for Monthly Model

expr_month = 'Admissions ~ Month_Apr + Month_Aug + Month_Dec + Month_Feb + Month_Jan + Month_July + Month_Jun + Month_Mar + Month_May + Month_Nov + Month_Oct + Month_Sep'

This allowed us to have the distribution of monthly admission count and calculated the rolling mean and standard deviation of the admission column (Figure 7). This makes the entire date a series and windows over the file. To get the mean and standard deviation, we used the pd rolling method with a window of five:

rolmean = pd.Series(ts).rolling(window=5).mean()

rolstd = pd.Series(ts).rolling(window=5).std()

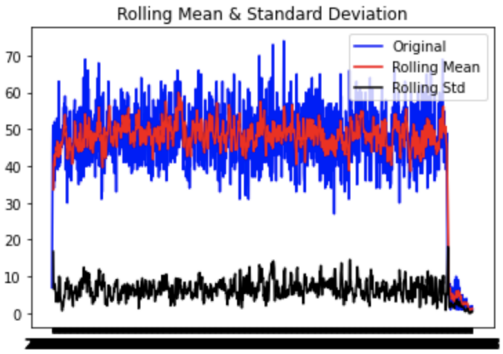


Figure 7. Rolling Mean and Standard Deviation of the Model

Having this figure allows us to analyze the distribution of monthly admission better and will help us in understanding the results of our model.

To access our Google Colab file:
https://colab.research.google.com/drive/1Rd3KrNxLCUqf3Vit2ZQ6Nke9fuf6YJZL?authuser=2#scrollTo=n7k0AvMnaYX0

Results

We created different OLS Linear models for daily and monthly predictions. As mentioned above, In the Daily Model, Admission (target variable) is dependent on the following features: Weekday, Date, Month, and Year. With these features, we observed that the model lacked precision and predicted close to the mean of the admission value of the actual admissions (Figure 8). The daily model had an R2 score of 0.02. The low score of R2 signifies close to no learning by the model in predicting the admission count. In Figure 9, we see the ACF plot of the residuals in the daily model; this plot indicates the correlation between the residuals.

Similarly, we have an admission count of each month as our target variable for the monthly prediction model. In this model, the target variable, admission, is dependent on each month of the year. The months are dummy encoded to fit the regression model. We had similar findings (Figure 10) that the model couldn't learn to predict the admission count and had the R2 value of 0.017 (Figure 10). In Figure 11, we see the linear fit of the monthly admission; the linear fit doesn't justify the actual admission count in the dataset. Lastly, in figure 12, we have the ACF plot of monthly admission residuals.

Daily Modeling Results

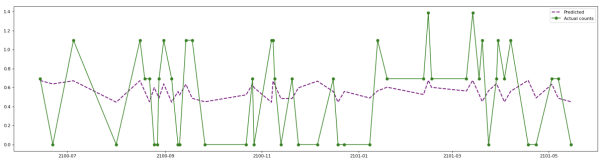


Figure 8. Predicted and actual results for daily admissions (enlargement of figure on page 6)

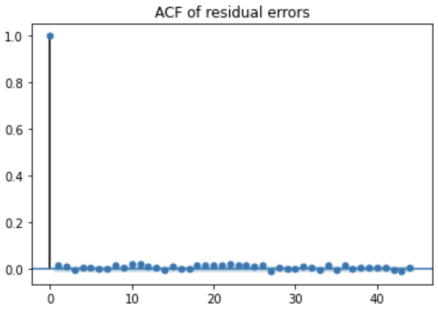


Figure 9. Visualizing errors of the daily model

Monthly Modeling Results

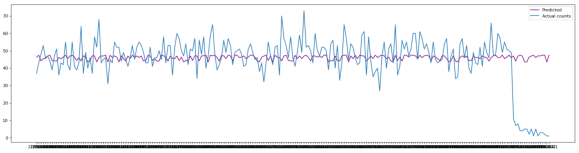


Figure 10. Predicted and actual results for monthly admissions (enlargement of figure on page 6)

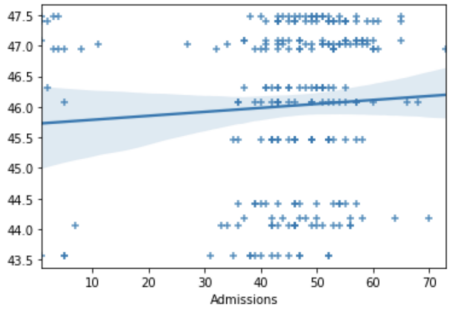


Figure 11. OLS Linear Regression Fit for Monthly Admissions (Intercept = 42.3078)

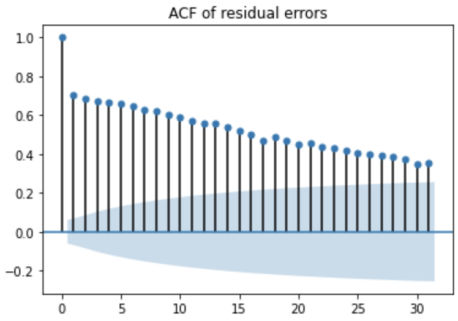


Figure 12. Visualising error terms of the monthly model

Conclusion

In this project we used the MIMIC-III admissions dataset to predict the expected number of admissions to the hospital in daily and monthly periods which we hoped would help the hospital prepare equipment and prepare as needed.

In the course of our work, we created daily and monthly datasets and models, which included pivot tables, plotted charts, box-and-whiskers plots, rolling mean and standard deviation tables. By creating these, we could visualize the data better and draw conclusions about hospital admissions. For example from the box-and-whiskers plot, we could conclude that Sunday saw the highest admissions compared to other days of the week (Figure 6). From this point, we began working on models to predict hospital admissions, hoping that by using machine learning methods, specifically linear regression, we could predict hospital admissions in the future. However, as we worked, we found that the models that we generated with our data had trouble predicting hospital admissions as the predicted sequence was consistently maligned with the actual number of admissions. Therefore, we hope to use other linear regression tools in the future to predict hospital admissions.

Lastly, we would also create an additional feature set, by grouping on existing features in the MIMIC-III dataset such as LOS, patients and gender, etc. Adding additional features could help in the learning of the model and thus result in better admissions predictions.

Related Works

Similar studies have attempted to use linear regression to predict hospital admissions, however they employed the use of Auto Regressive Integrated Moving Average (ARIMA) modeling to reach this end. Seeing how the ARIMA model outperformed other forecasting models in predicting patient admissions according to a similar study done by Kim et al (2013). In this study, the ARIMA model was capable of predicting hospital admissions one week in advance within 7.8% of the mean absolute percentage error (MAPE) and a month ahead with less than 8.8% MAPE (Kim et al, 2014). These findings were thoroughly impressive and gave us hope that when attempting to predict hospital admissions in the future using the ARIMA model would help us reach our goal of predicting hospital admissions using linear regression. In another study conducted by Boyle et al (2008), they also used ARIMA modeling to forecast patient admissions to the emergency department in addition to others. In their study, they tested various linear regression models and based the results on where their model fell in MAPE, which is something that could be useful in future studies to predict hospital admissions.

References

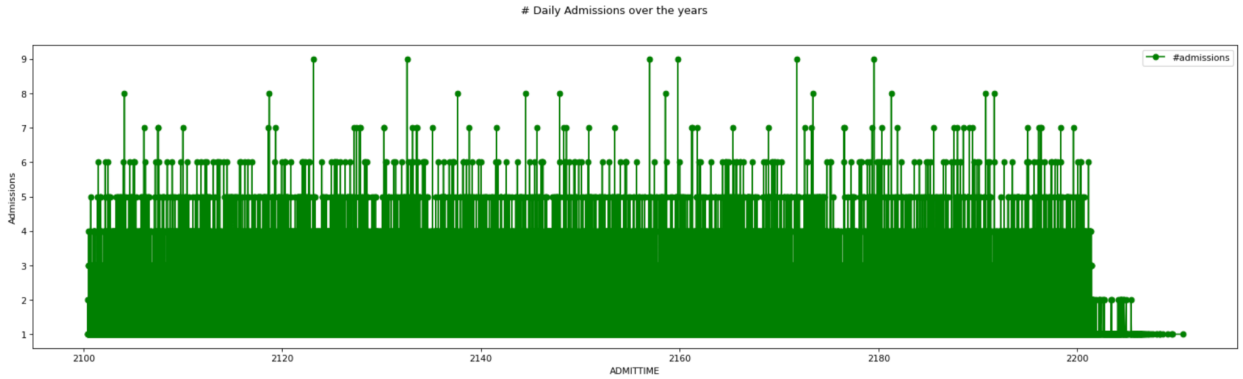
Boyle, J., Wallis, M., Jessup, M., Crilly, J., Lind, J., Miller, P., & Fitzgerald, G. (2008, August). Regression forecasting of patient admission data. In 2008 30th

Annual International Conference of the IEEE Engineering in Medicine and Biology Society (pp. 3819-3822). IEEE.

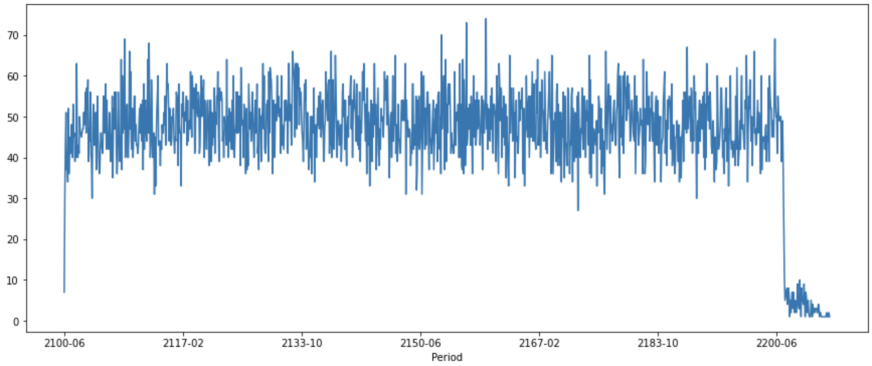
Grimm, C. A. (2020). Hospital experiences responding to the COVID-19 pandemic: results of a national pulse survey march 23–27, 2020. US Department of Health and Human Services. Office of Inspector General. Available online at: https://www. oig. hhs. gov/oei/reports/oei-06-20-00300. Pdf.

Kim, K., Lee, C., O'Leary, K., Rosenauer, S., & Mehrotra, S. (2014). Predicting patient volumes in hospital medicine: A comparative study of different time series forecasting methods. Northwestern University, Illinois, USA, Scientific Report.

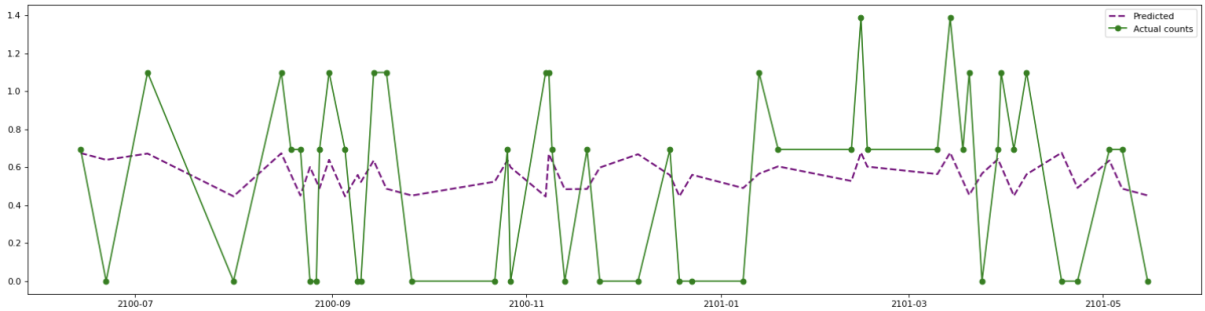
Zdaniuk B. (2014) Ordinary Least-Squares (OLS) Model. In: Michalos A.C. (eds) Encyclopedia of Quality of Life and Well-Being Research. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-0753-5_2008



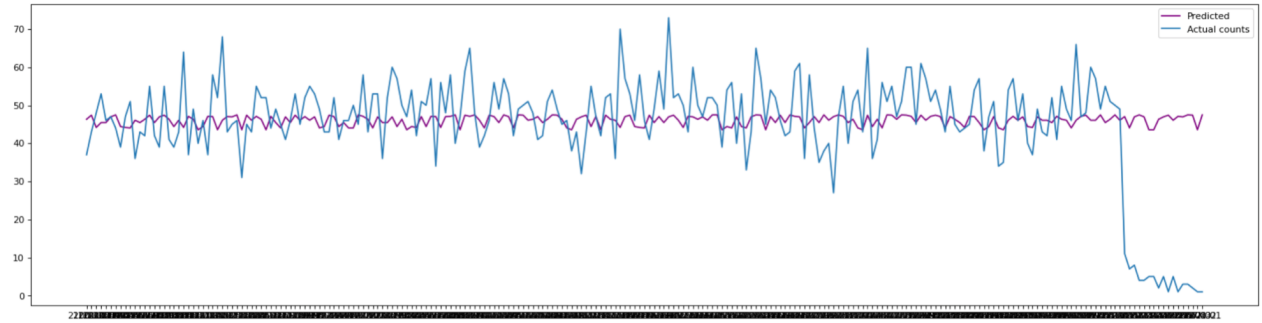
Enlargement of Figure 4



Enlargement of Figure 5



Enlargement of Figure 8



Enlargement of Figure 10