D208 Performance Assessment Task 1

September 24, 2021

**Part A**

1. One research question an organization using this data could ask how different personal factors affect the additional costs at the hospital.
2. My goal with this data set is to perform a multivariate linear regression analysis on different columns of hospital data to help try and predict how much an additional cost would be added depending on those variables and if there is a bias towards any one group of data.

**Part B**

1. With a multiple regression model, we must use a few assumptions. We need to make sure our dependent and independent variables are continuous, that they have a linear relationship and are stripped of as many outliers as possible otherwise it could mess with our model and give us a worse fit. We also need to assume that the variables are normally distributed to make sure we don’t distort our data by skewing towards outliers. We also need to make sure variables are measured with as little error as possible because this could skew our results. This just means we need to make sure our data is homogeneous and we can assume that all the variance is approximately the same. The final thing we need to make sure of is to assume that our data is homogenous and that all the data will have roughly the same number of errors.
2. I am using Python since I’m the most comfortable working with it, plus I’m able to use Replit and Jupyter Notebook at work so I can do all my analysis in web-based apps instead of installing special software onto any machine I use and allows me to save progress and move to a different machine and work on the same analysis.
3. Multiple regression is appropriate since it would allow us to use as many independent variables as we want to create a linear relationship which would allow us to accurately predict what an additional cost might be at the hospital. For this analysis, I’m choosing to use Children, Age, Income, Gender, Readmission, Vitamin D levels, Doctor visits, Complication risks, and days in the hospital to check the effect on Additional costs to keep the analysis simple yet demonstrative.

**Part C**

1. The data cleaning for this was actually pretty easy. First step was to make a data file that contained the 'Children', 'Age', 'Income', 'Gender', 'ReAdmis', 'VitD\_levels', 'Doc\_visits', 'Complication\_risk', 'Initial\_days', and 'Additional\_charges'. Then I had to convert the ‘Gender’ column into 1/0 but because of the “Nonbinary” gender, I converted Nonbinary and female to 0’s and male to 1’s. I also changed the ‘ReAdmis’ section from Yes/No to 1/0. The other column that I had to make a judgement call on was ‘Complication\_risk’. I chose to make High a 1 and Medium/Low a 0 since I wanted to see if higher complication risks were associated more with additional costs.
2. The summary statistics were obtained using the describe function in pandas. The results are in the notebook that’s been attached. Some of the main observations are just basic checks, like averages and maxes. There appears to be a fairly low mean of children of only 2 children per patient. Another thing that seems odd is how the average number of days in the hospital is 34 which seems really high. Also having an additional charge of 12934.53 seems rather high. A full display of the summary can be seen in the attached code.
3. First thing that I needed to do was to isolate the columns I wanted to analyze from the medical data csv file. I then had to alter the gender, readmission, and complication risk columns since they are categorical and multiple regression analysis doesn’t work with categories very well, so I changed certain values to 1/0 based on what I believed to be an adequate take on the analysis. The columns that were altered are listed above. I also checked for duplicates and null values which luckily showed that there were none in my dataset. I also checked for outliers in my data by running a z test on the standard deviations to eliminate anything not within the 99.7% confidence interval. I ended up having to remove 370 records from the dataset which was only 3.7% which still was >95% of the data so I felt confident removing them and analyzing the data with those few outliers deleted.
4. See attached code
5. See attached data set

**Part D**

1. After identifying my predictor variables, I was able to come out with a regression formula of y = 30.0\*Children + 228\*Age - 0.000126\*Income + 190\*Gender + 390\*ReAdmis + 5.47\*VitD\_levels + 40.7\*Doc\_visits - 8.42\*Initial\_days + 500\*Complication\_risk + 288. The 288 is the intercept for the Additional\_charges column. I chose these predictor variables since intuitively I figured they would have some effect on the additional costs of a patient at the hospital which is what this whole analysis is trying to understand. The R2 value was 0.901 which isn’t spectacular, but still linear and encompasses a vast majority of the data be analyzed.
2. By running an OLS analysis on the regression model, I was able to determine that since the R2 value is >90, we have a significant linearity with the model and data. We can also look at the heatmap generated earlier and see that there seems to be a direct relationship between additional charges and age of a patient. We can also see that there seems to be a direct relationship between readmission and initial days in the hospital. It’s not greatly significant because it’s still <1, but the fact that it stands out from the other factors quite obviously shows that we should take these variables and analyze them further. I justified these variables because the correlation coefficients are >0.75 which seems to show there is a direct relationship.
3. Since we are going to be analyzing the 2 continuous and 2 categorical variables, we can isolate them for further examination. I pulled out Initial\_days and Age as a continuous variable and ReAdmis as a categorical variable. These 3 were used as predictor variables for Additional\_charges. A reduced multiple regression model looks like y = 227.65\*Age + 400.56\*ReAdmis – 8.5782\*Initial\_days + 900.31 with 905.96 being the Additional costs coefficient.

**Part E**

1. The reason I ended up choosing age, readmission, and initial days to test for Additional costs, is because it seems like those factors might have an effect. When we look at the bivariate analysis, we can see that these variables have relationships with each other, so it makes sense to dig into the relationships between all the variables and additional costs to tabulate the relationships. The research question wants to see if they have some effect on additional costs and by looking at if those factors were related at all and from our heatmap. To get better understandings of this, a residual plot of the relationship between additional charges and age, readmission, and initial days was created.
2. After creating the scatterplot, I did another OLS analysis on the relationship between the additional charges and age. The results are attached.
3. See code in attached documentation.

**Part F**

1. From this analysis we can see that there is a small connection between age and additional charges and readmission with days in the hospital. A quick regression equation derived from this analysis is above in D1. From our OLS regression analysis we can find that we have an F-statistic which is 29070 which means we can reject our null hypothesis which also means that there is connection between age, readmission, days in the hospital, and additional costs. Our P-value is so small that it doesn’t fully show up on our calculation which means that there is high confidence that our predictive variables influence additional costs. However, the downside of this analysis is that our relationship is still kind of weak. All this analysis has been able to show is that there is some relationship between age and additional costs and that there is some high confidence in that. One thing that I had to do was to remove some outliers in my data set that I got to using a z statistic which required me to delete around 400 records from the datafile. When I found the outlier data, I decided to delete the data instead of replacing with the median or average because I figured substituting the variables would cause a shift in the dataset which could screw up the analysis just as much as deleting them, and when deleting them they can no longer really affect the results of the analysis. I am now missing 370 records from the dataset, but that’s a minuscule amount compared to the whole dataset of 10000 records. The major limitation of this is that because I’m deleting records instead of substituting average/median values is that I’m literally getting rid of data that could be useful for the analysis. I figured this would be fine since again, I’d only be getting rid of 370 records so out of 10000 which is rather minuscule and shouldn’t affect the overall analysis in any significant way.
2. The best course of action after this would be to analyze the relationship between age and additional costs by running t-tests and possibly a blind ANOVA test to make sure that there really is some underlying bias going on and to double check the results that this multiple regression analysis provides.

**Part G.**

1. Panopto video: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=5902fefd-959e-484f-a7e4-adb5017ddb40>

**Resources:**

<https://statistics.laerd.com/spss-tutorials/multiple-regression-using-spss-statistics.php> Laerd Statistics

<https://online.stat.psu.edu/stat462/node/135/> The General Linear F-Test, PennState

<https://www.geeksforgeeks.org/ordinary-least-squares-ols-using-statsmodels/> Ordinary Least Squares (OLS) using statsmodels, GeeksforGeeks

[https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1111&context=pare#:~:text=Specifically%2C%20we%20will%20discuss%20the,measurement%2C%20homoscedasticity%2C%20and%20normality](https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1111&context=pare#:~:text=Specifically%2C we will discuss the,measurement%2C homoscedasticity%2C and normality). Four assumptions of multiple regression that researchers should always test, University of Massachusetts.

<https://stackoverflow.com/questions/23199796/detect-and-exclude-outliers-in-pandas-data-frame> Detect and exclude outliers in Pandas data frame, tanemaki