D208 Performance Assessment Task 2

October 12, 2021

**Part A**

1. One research question an organization using this data could ask how soft drink consumption is effected by various factors including pre-existing conditions, income, doctor visits, and a wide range of other factors being analyzed in this analysis.
2. My goal with this data set is to perform a multivariate logarithmic regression analysis on different columns of hospital data to help try and predict how much various factors are affected by soft drink consumption and if any predictions could be made based solely on if soft drinks are consumed by someone.

**Part B**

1. For a logistic regression model, there are 6 core assumptions that must be taken for our data set. The first thing is that the dependent variable is binary. It has to be something that can be yes/no, pass/fail, have or have not. It can only have 2 possible outcomes. A second thing we must assume is that all observations are independent and don’t rely on another variable to influence the results. The third thing we must assume is that there is no multicollinearity between variables, which simply means that the variables being used in the analysis can’t be highly correlated as this could mess up the calculation. The fourth assumption is that there are no extreme outliers. Usually this means anything that is beyond 3 standard deviations which can be removed to prevent the distortion of the data. The fifth assumption is that there is a linear relationship between the predictor variable and the logit of the dependent variable. The sixth and final assumption is that the data set is very large because a too small of a dataset would cause problems in the calculations which could interfere with our data analysis.
2. For this analysis I used Replit and Jupyter notebook. I used both of these tools since they are both available online through a web portal and since I do most of my homework at work, it allows me to work on this from anywhere without having to download various IDE’s and tool packages. I’m using python because I know how to use python best and I have access to libraries such as statsmodels, pandas, and matplotlib for all the analysis tools I could need for this analysis.
3. The reason logistic regression is good for this analysis, is because we are trying to simply predict whether a patient does or doesn’t drink soft drinks based on a wide variety of factors that we are analyzing. This could help us figure out if soft drinks are causing problems within patients or if there is a relationship between soft drink consumption and various factors of patients.

**Part C**

1. To prep my data for logistic regression analysis, I first needed to isolate my columns into a datafile that I could use for my analysis. Next, I checked for duplicates and nulls which turned out none so that was an easy check for my data. I also got rid of all outliers by deleting all data that wasn’t within a z-score of 3 since that is within 99.7% of all data. The final step of cleaning the data is to create a dictionary to replace all yes/no values with 1/0 values in order to perform logistic regression.
2. For this data analysis I decided to check out a wide multitude of factors to see if soft drink consumption had any relationship to various pre-existing conditions or other factors such as doctor visits, or income. My target variable is going to be Soft\_drink, while my predictor variables are Income, VitD\_levels, Doc\_visits, Full\_meals\_eaten, vitD\_supp, HighBlood, Stroke, Overweight, Arthritis, Diabetes, Hyperlipidemia, BackPain, Anxiety, Allergic\_rhinitis, Reflux\_esophagitis, and Asthma. For a summary of the statistics, I performed a logit function in python which printed out a summary of the data from this initial logical regression. Some examples from this is the pseudo r-squared values 0.002071, and LLR p-value of 0.1140 along with p-values of all variables ranging between 0.038 to 0.986 which is a drastic range and shows that some variables are fairly confident, while others are not confident at all.
3. As mentioned above, I isolated columns for analysis, eliminated outliers, and changed yes/no variables to 1/0. Code is attached.
4. See attached code.
5. See attached data set.

**Part D**

1. For my initial logistic regression model, I used a logit function from the statsmodels library in python, which gave me a regression model of all the predictor variables and the target variable. The initial regression equation is: -1.4062 + 0.0000007586\*Income + 0.0074\*VitD\_levels + 0.0266\*Doc\_visits + 0.0488\*Full\_meals\_eaten - 0.0504\*vitD\_supp - 0.0233\*HighBlood + 0.0231\*Stroke - 0.0280\*Overweight - 0.0393\*Arthritis + 0.0710\*Diabetes + 0.0943\*Hyperlipidemia + 0.0685\*BackPain + 0.0752\*Anxiety - 0.0812\*Allergic\_rhinitis - 0.0008\*Reflux\_esophagitis + 0.0481\*Asthma.
2. For my regression analysis, I used the logit function and found p values that were under 0.10 which is still a bit of a swing, but within 10 accuracy of the real result. I also did a double check with the python conf\_int function in order to see confidence intervals, which some factors like Income and Reflux\_esophagitis shows numbers incredibly close to 1 which indicate that they are not very significant.
3. After running my checks, I’ve decided to isolate Full\_meals\_eaten, Hyperlipidemia, and Allergic\_rhinitis as predictor variables for soft drink consumption since they had p values <0.10 and those were the most significant findings that were reaffirmed by my check with the confidence interval test. My reduced model comes out as: -1.1074 + 0.0498\*Full\_meals\_eaten + 0.0914\*Hyperlipidemia - 0.0818\*Allergic\_rhinitis.

**Part E**

1. For isolating my reduced regression equation, I picked variables that had a p value of <0.10. This is a bit larger than a normal <0.05 value, but none of my variables were precise enough for that. However the factors that were <0.10 were a few important factors that seem to be affected by the consumption of soft drinks since they also had a fairly close relationship. I used the LLR p-value to judge the accuracy of the fit and how confident we are that the fit works for our analysis. Our initial regression model had an LLR p-value of 0.1140 which isn’t very good confidence, but our reduced model has an LLR p-value of 0.01271 which is <0.05 which means we have some very high confidence that this function demonstrates a strong relationship between soft drink consumption along with the selected predictor variables. In the code I used a logit function in order compare the initial and reduced functions and compare the relationships between each model. Attached in the code there is a logit description that compares each model. A more solid way of comparing models is to break them down into residual plots with bar graphs comparing all factors with comparisons to consumption of Soft\_drinks vs all factors which are attached. I also ran a plot comparing the initial and reduced model together to see what kind of relationship we could draw from our results. However, the residual plot showed that both models tended to go along the same value range, but the rate of change between them was different, as the initial model showed that our models’ logistic regression is incredibly faint and so slow that it almost looks linear. I have attached a diagram showing the initial and reduced regression models. I also ran a confusion matrix on the sample which gave us a result of [1787, 0] [645, 0] which means that out of our predicted variables, 1787 were correctly predicted and 645 were not which gives us an accuracy of approximately 73%. This is not fantastic, but it does show that our model can approximate if a patient consumes soft drinks roughly 73% of the time.
2. See attached for output and calculations.
3. See attached for logistic regression model calculation.

**Part F**

1. See above for my reduced regression model. Because I’ve set soft drinks as the target variable, it appears from my analysis that the biggest factors that are affected by soft drink consumption are Hyperlipidemia and Allergic Rhinitis. Full meals eaten is also related to soft drink consumption, but it kind of makes sense that if you’re eating more meals, you’re more likely to have soft drinks with those meals. I think there are quite a few limitations of this analysis. One that I have to be called out for is that I’m mostly basing my analysis on the confidence intervals of the data. While the analysis is confident that these factors are related to soft drink consumption, it doesn’t necessarily tell us if soft drinks are the actual cause. This analysis mostly just says “Yes, soft drinks factor into these variables. How? We are not fully sure.” I think the best course of action would dive deep into the predictor variables that have been identified and check how affecting those affects soft drink consumption and vice versa. If we relook at the assumptions made at the beginning of this paper, we can find that no assumptions are being violated. Everything is in a binary state, multicollinearity has been established to not be a problem because of our heatmaps, and all variables are independent of each other. It appears that nothing has any direct relationship with each other, but some factors act as better predictors of soft drink consumption than others. The great thing about the confusion matrix that was made shows that our variables have very little relationship with each other making our logistic regression model work since the factors can’t be directly related to each other.
2. As stated above, I believe the best course of action would be to isolate each individual predictor variable and test it by seeing what happens if soft drinks are included/excluded. It’d also be interesting to do this full analysis on each predictor variable in the reduced regression equation could turn up on perhaps other factors affecting each predictor variable and thus still having a relationship to the target variable.

**Panopto Video**: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=c2db3e42-18b4-4794-b5e9-adca01729d2b>

**References:**

<https://www.statology.org/assumptions-of-logistic-regression/> Statology, The 6 assumptions of Logistic Regression (with Examples)

<https://stackoverflow.com/questions/67551380/eliminate-outliers-in-a-dataframe-with-different-dtypes-pandas> Eliminate outliers in a datafram with different dtypes – Pandas, Henry Ecker

<https://digitaschools.com/binary-logistic-regression-in-python/> Binary LKogistic Regression in Python – a tutorial Part 1 ,Digita Schools

<https://investigate.ai/regression/evaluating-logistic-regressions/> Evaluating a logistic regression, investigate.ai

<https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8> Building A Logistic Regression in Python, Step by Step, towards data science

<https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826> Confusion Matrix for Your Multi-Class Machine Learning Model, towards data science

<https://www.pythonfordatascience.org/logistic-regression-python/> Logistic Regression, Python for Data Science