Multiple Linear Regression

Leticia Lugo

Western Governor University

Case

Hospital readmissions from Medicaid patients alone cost the government $26 billion annually according to Robert Wood Johnson Foundation. According to Tom Prose, CEO at General Medicine, P.C., in his article “Top Reasons for Hospital Readmissions” some reasons to get re-admitted to a hospital are:

* The patients are not well informed about their care, which means patients do not know the extent of their conditions and the expectations of future treatment processes.
* Patients leave their hospital care too soon because of the hospital charges causing a potential extension of their illness.
* Patients naturally require more medical care with age.

(Prose, n.d.)

Research Question

Do Doctor Visits, Age, Income, Gender, Area, Total Charge, Additional Charges have an influence on the days spent in the hospital?

Objective

The analyst should detect the reasons that influence patients to spend more days in a hospital. Based on the article by Prose, the following columns were selected for the analysis: doctor visits, age, gender, area, total charge, additional charges, and readmissions.

Target Variable

Initial\_days = the number of days the patient stayed in the hospital during the initial visit. The Initial\_days variable is essential because patients have more time to get treated, communicate with their doctor and have a plan for recovery. This variable is continuous.

This report will be performed in Python language using jupiter notebook environment due to several reasons compiled on table 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Comparisons | Python | R | SAS | Option/Reason |
| Uses | Used for software development, web development, and data science. | Used for data mining and machine learning. | Used for business intelligence, data management, and predictive analytics. | Python/Widely usage. |
| Cost | Open-source programming and free language | Open-source programming and free language | Proprietary software, customers pay to use it. | Python/No cost |
| Speed | A high-level programming language. It is faster for building large applications and web development. | A low-level programming language. It manages extended codes for simple procedures, which results in low speed. | A high-level programming. It uses SAS SQL and automatic code generation with reusable code snippets. | Python/High-level language. |
| Accessibility | It is easy to learn. It uses friendly libraries, and it has become a phenomenon in data analytics. | It has a steep learning curve because it needs a working knowledge of coding. | It is easy to learn, and it has a simple GUI. It does not required knowledge on a prior programming language. | Python/Analyst more familiar with it. |
| Data Handling | It is easy to handle data with popular libraries such as Numpy and Pandas. | It is easy to handle data with packages like plyr, dplyr, and tdyr. However, small tasks will take time to run because it runs on RAM memory. | It is efficient to perform data handling and manipulation using the DATA step, which compiles and runs faster. | Python/Analyst familiar with libraries. |
| Data Visualization | Have packages such as matplotlib, seaborn, and vispy among others that makes it robust in graphical analysis. | Makes visualizations and analysis with packages like ggplot, lattice, ggvis, rgis, among others. | Work still in progress to improve its visualization and graphical capabilities, but it does not yet match Python and R's standards. | Python/Analyst familiar with libraries. |
| Customer and Community Support | It does not have customer support. However, it provides online community support. | It does not have customer support. However, it provides online community support. | It has customer support, and it has an online community to help solve customer’s questions. | Python/Large community support. |
| Popularity | 41% | 30% | 29% | Python/More popular. |

(Comparison of Python, R, and SAS languages)

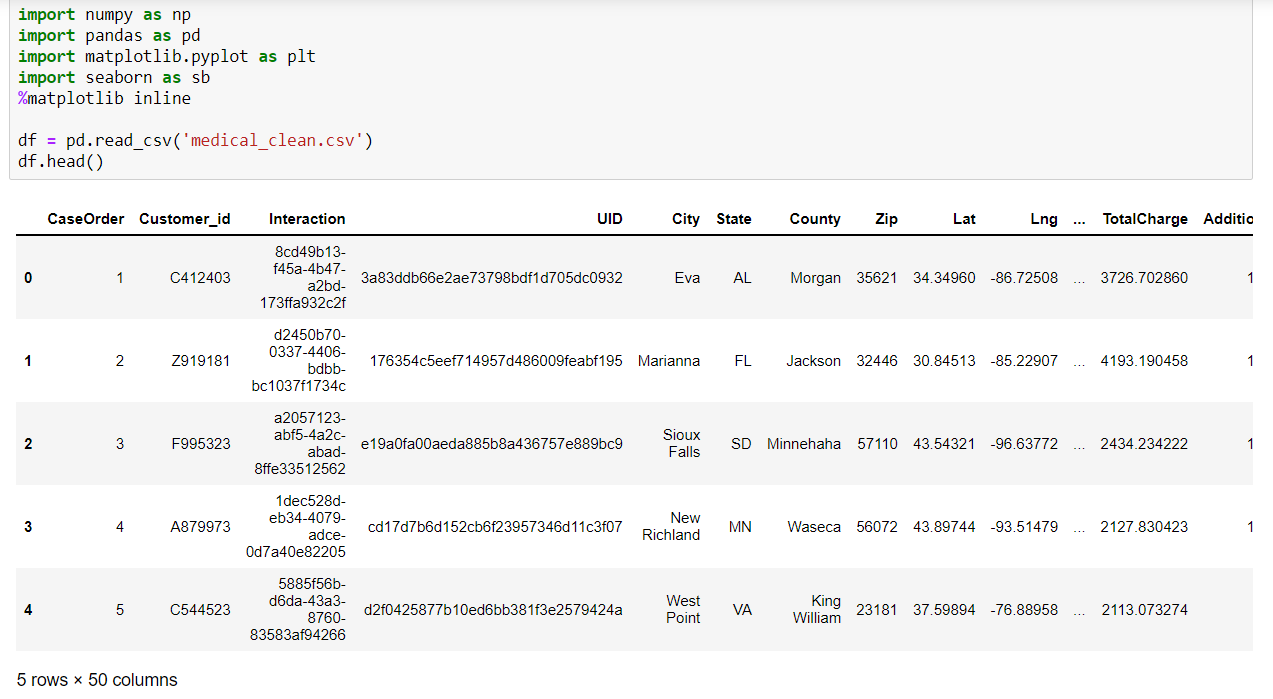
Table 1: Comparison of Python, R, and SAS languages.

Data

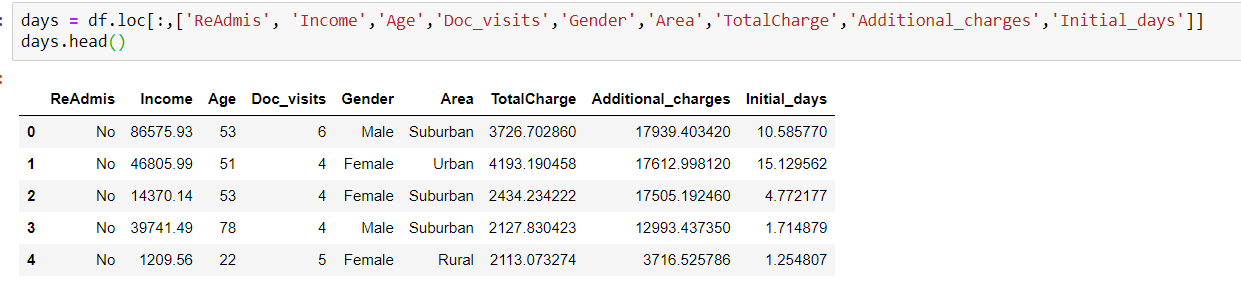
The data consist of 10,000 entries of patients admitted to the hospital with 50 columns of patients’ medical conditions, demographic information, and patient service in the hospital.

Import the data

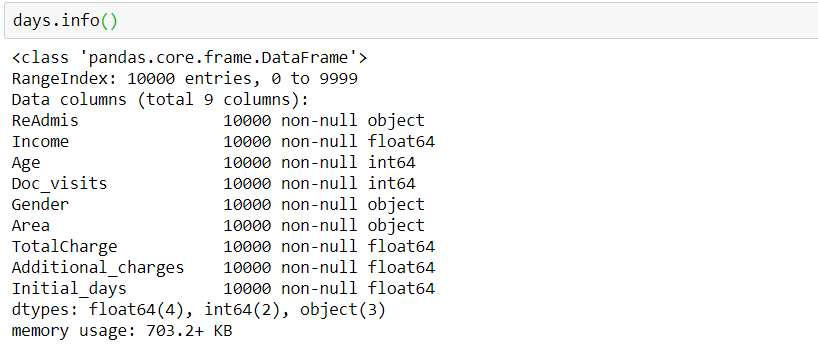
The dataset is imported using pandas library to the Python environment.

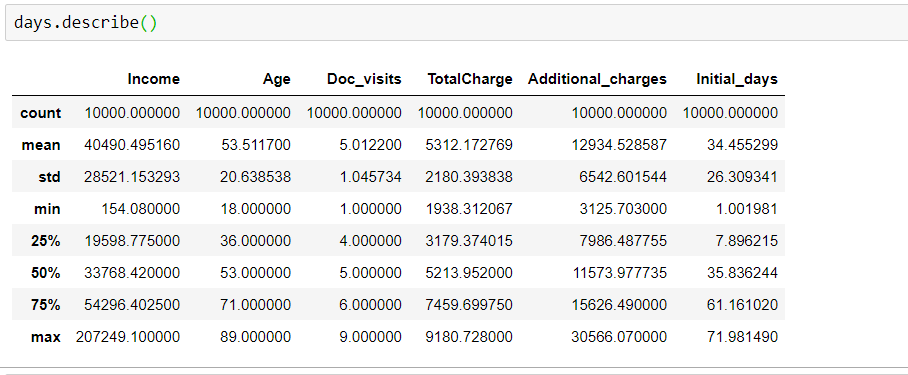


Select the variables and clean data

Based on the research question, the analyst's variables were taken from the dataset to a new dataset named ‘days’. 

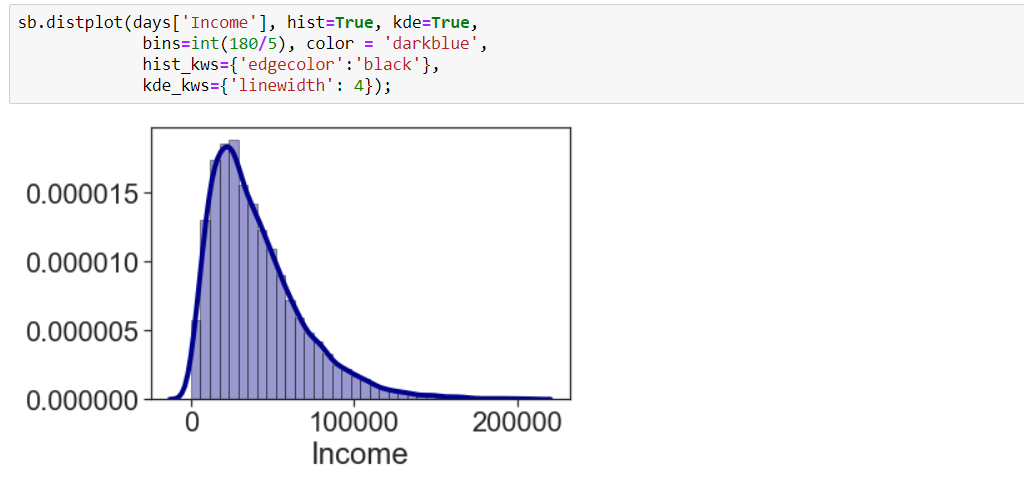
Starting the data cleaning process using info() method of pandas to detect missing values, based on the result table, there are not detected missing values.



After detecting no missing values, the described() method is run to observe the statistical values of the data. This method will only apply to numerical variables. For example, with the table of statistical analysis below, we can see how the mean of days in the hospital is 34, which is about a month. Another insight from the statistics table gives us the average age of 53. The average income is $40,490. The addition of total charges and additional charges is about $19,000, about half of the average income. 

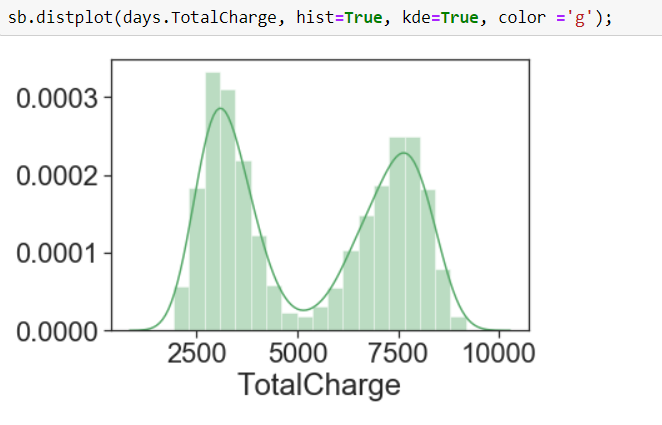
Univariate and Bivariate plots

The first univariate plot describes how the income average is $40,490. The tail to the right indicates that there are more people with lower income admitted to the hospital.



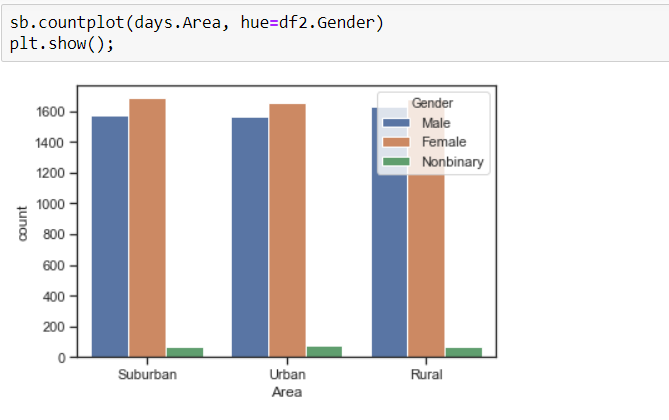
(Seaborn, n.d.)

The second univariate plot indicates a bimodal histogram on total charges. There are two modes, one approximately $3,000 and a second one approximately $8,000.



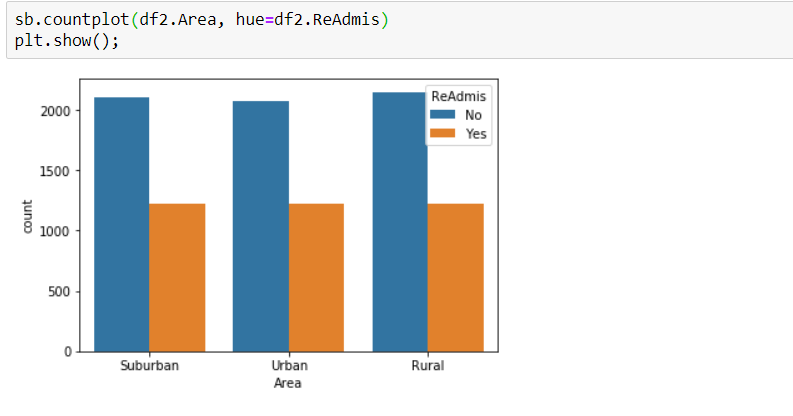
(Seaborn, n.d.)

The first bivariate plot shows how females are more likely to be in the hospital despite the area in which they live.



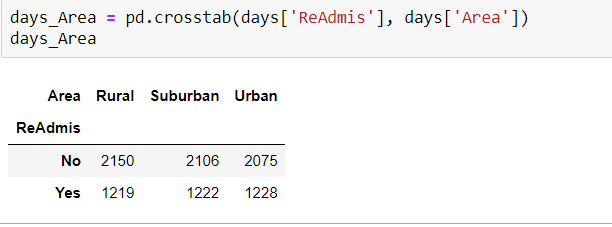
(Seaborn, n.d.)

The second bivariate plot shows that the admission to the hospitals is about the same in each type of area in which the patient lives.



(Seaborn, n.d.)

A cross tabulation table is created with the variables area and readmission where the results are very close for readmissions in each area.



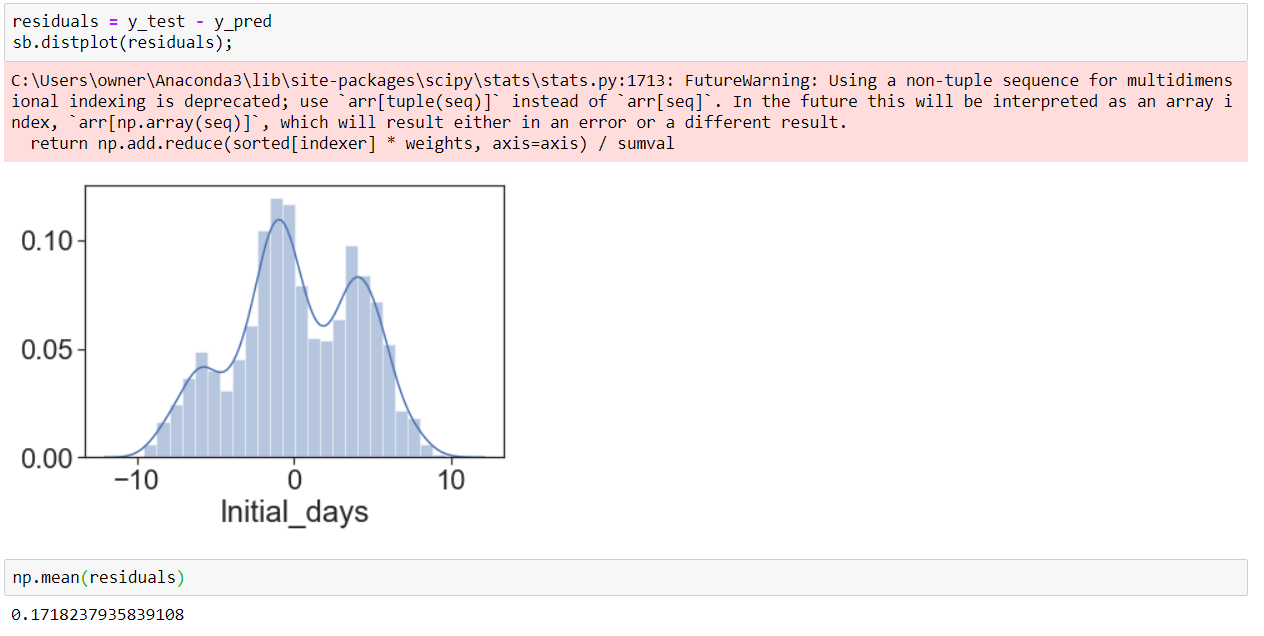
Statistical Method

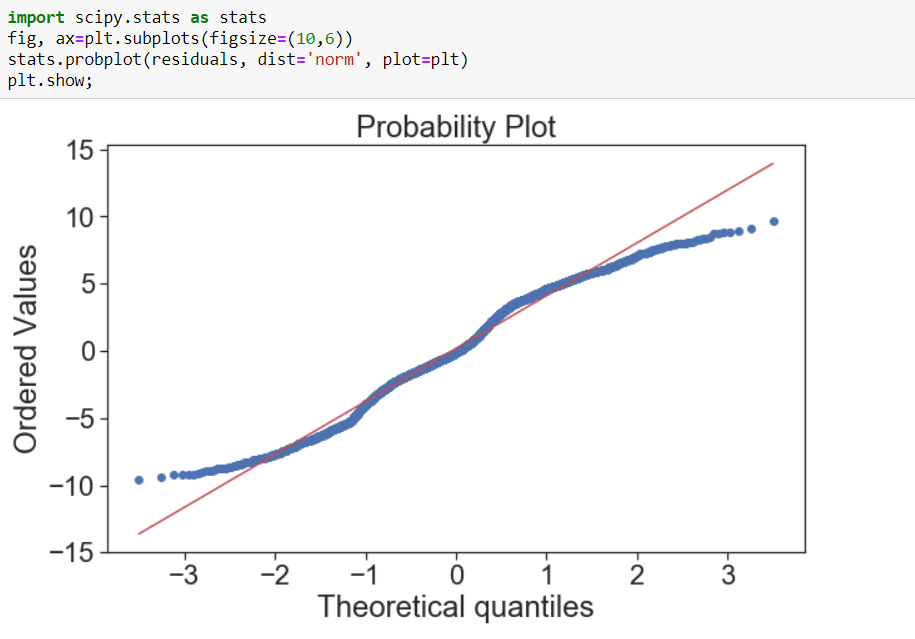
Multiple Linear Regression is a statistical prediction model based on relationships between the dependent and independent variables.

Multiple linear regression is the statistical prediction model for the analysis. The reasons for opting for multiple linear regression are that the data has a continuous target variable, and the predictor variables are a mix of continuous and categorical variables (Linear Regression, 2021).

Assumptions

1. Residuals should be normally distributed with a mean of 0. Residuals are the difference between the independent variable and the predicted variable. They explain the variation by the regression model. The next plot shows the residuals distribution which is close to normal since the mean is 0.17 which is close to 0. Another way to check normal distribution on residuals is by creating a Q-Q plot, which shows two distributions against each other. If both distributions are the same, the plot should show a roughly straight line. In the plot below the blue line is very close to the red line, meaning that the residuals are normally distributed. Even when the probability plot indicates outliers on each end of the blue line, the assumption is not violated.



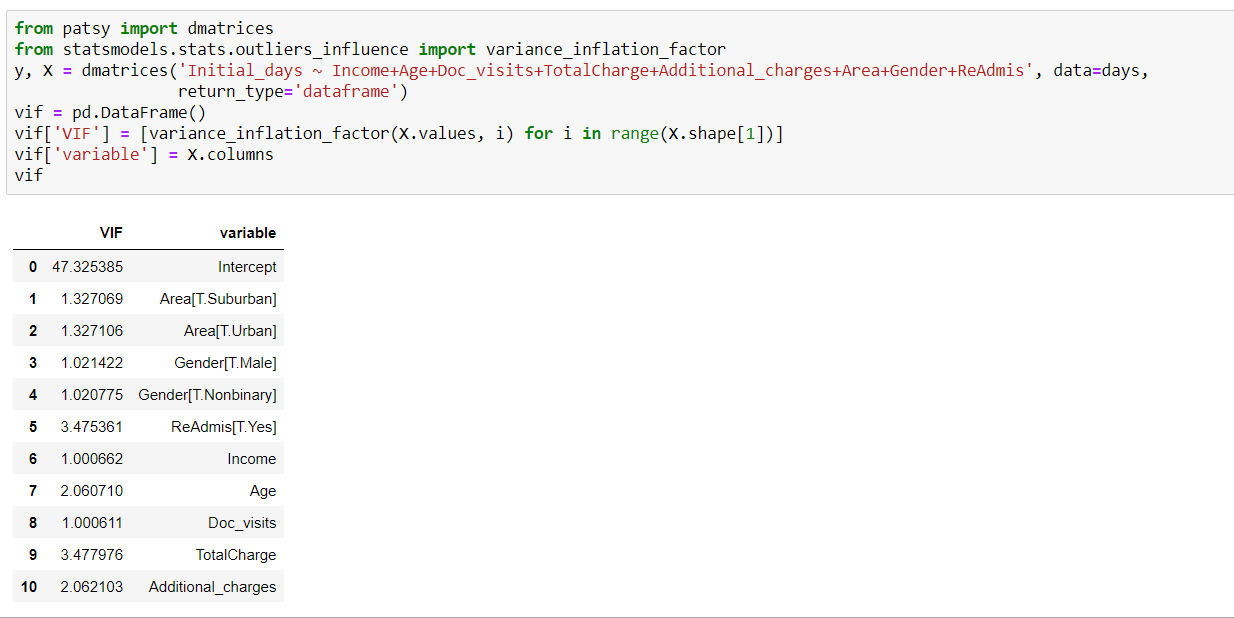


(Zach,2020)

1. Multicollinearity is the relationship among variables. However, the assumption is that there should not be a relationship among the independent variables. Multicollinearity affects the coefficients, causing them to be sensitive to any changes in the model.

By calculating VIF (variance inflation factor), we can define which variables have high multicollinearity. A value of one indicates no correlation between two independent variables. A value between 1 and 5 indicates moderate correlation. Moreover, a value greater than 5 is a potentially high correlation.

Looking at the VIF results, we can conclude that there are some moderate correlations, but they will not be significant enough to violate the assumption.



(Zach,2020)

1. Homoscedasticity measures if the variances are equal for all samples. If the samples have unequal variances (heteroscedasticity), that can affect the Type 1 error rate and lead to false positives.

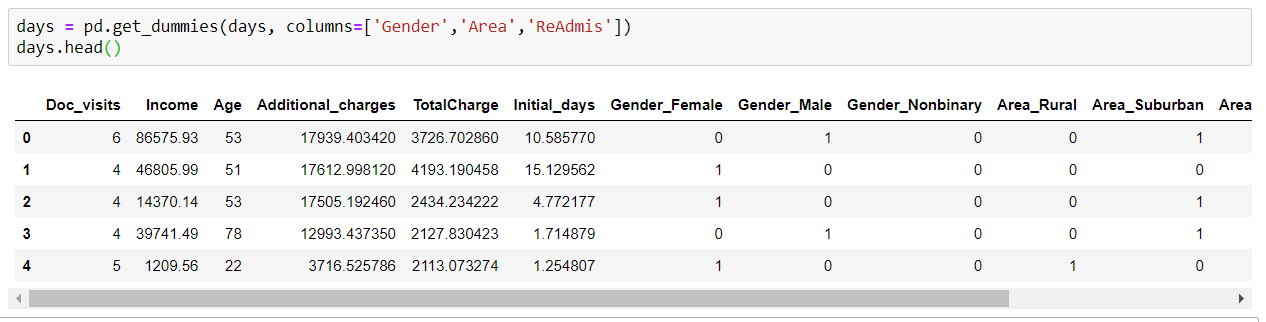
(Zach,2020)

1. There should be a linear relationship between the dependent and the independent variables. There should not be a relationship among the independent variables.

(Zach,2020)

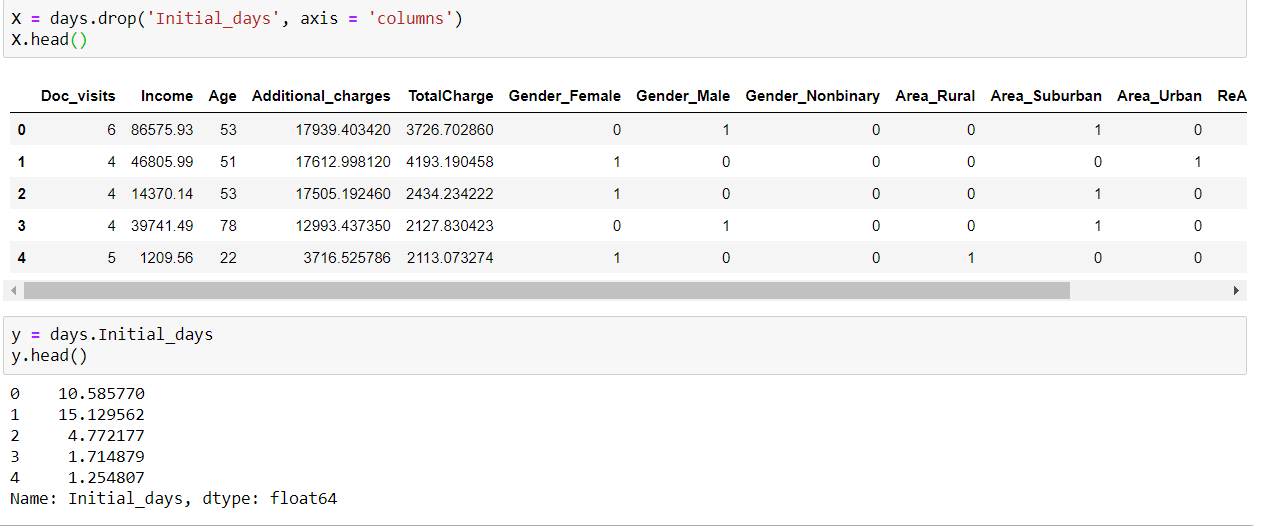
Preprocessing the data

There is a mix of categorical and numerical variables in the dataset. However, since multiple linear regression requires only numerical variables, the categorical variables are encoded, transforming them into numerical variables.



Define X and Y

The data features ‘Initial\_days’ as a target variable or ‘Y’ dataset, and the rest of the variables from ‘days’ dataset are going to be the ‘X’ dataset.

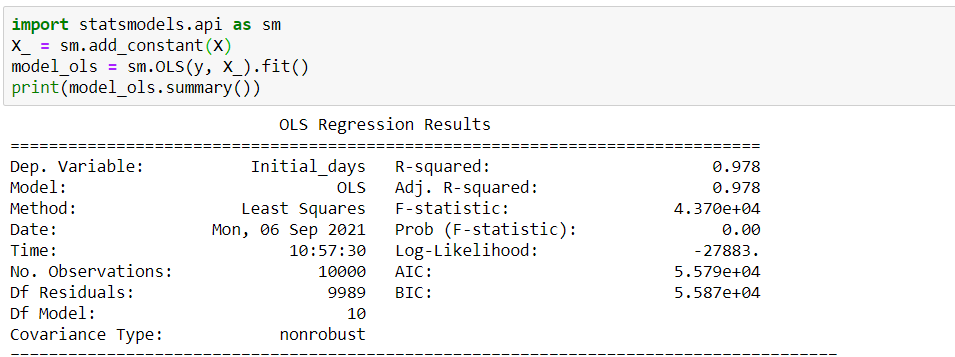


Model

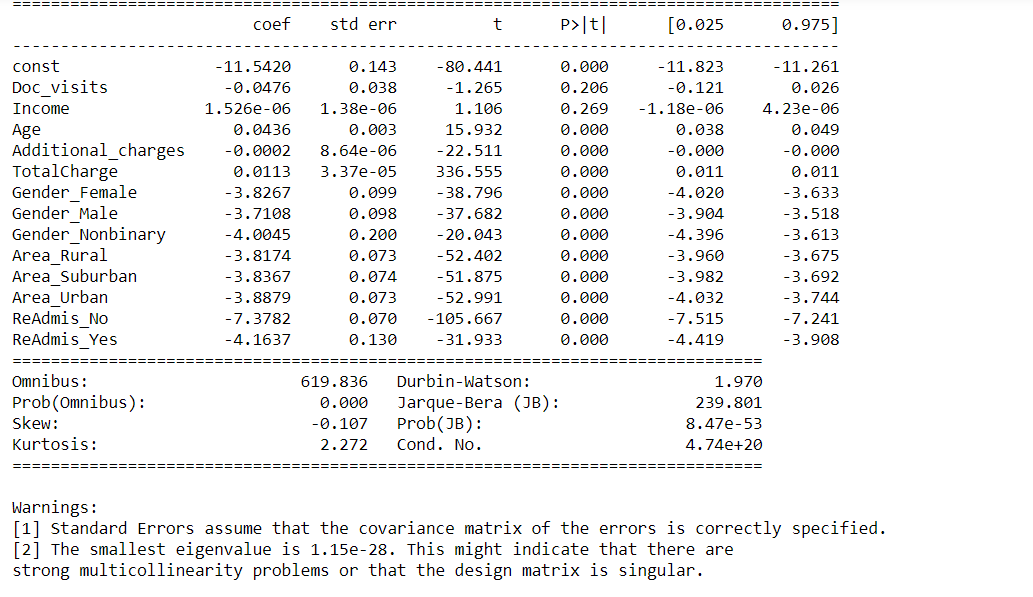
To model the data, it requires the ‘X’ and ‘y’ that were identified in the last step. The model gives the constant and coefficients for the multiple regression equation.

R-Square is the proportion of the variation in the dependent variable that is explained from the independent variables.

The model results identify the dependent variable as ‘Initial\_days’. The method used to fit the data is OLS (Ordinary least square/linear regression). The method used is Least Squares.

The date and time were captured when the model was created. The numbers of observations used in the model are stated in the summary table. The residuals and model degrees of freedom are calculated and added in the summary table. The R-squared is the coefficient of determination. The F-statistic shows if the model is different from a simple average by looking at the prob (F-statistic) where if p-value is equal or lower than 0.05 then the model is considered significant. The model prob (F-statistic) is 0.00, so the model is significant from the average model. The AIC number evaluates the model complexity. The AIC is 55,790 which indicates a weak model. The lower the result of the AIC, the stronger the model. BIC is similar to AIC except it punishes models with more parameters, and the smaller number is better for the model. The summary of results of the model is provided below: 

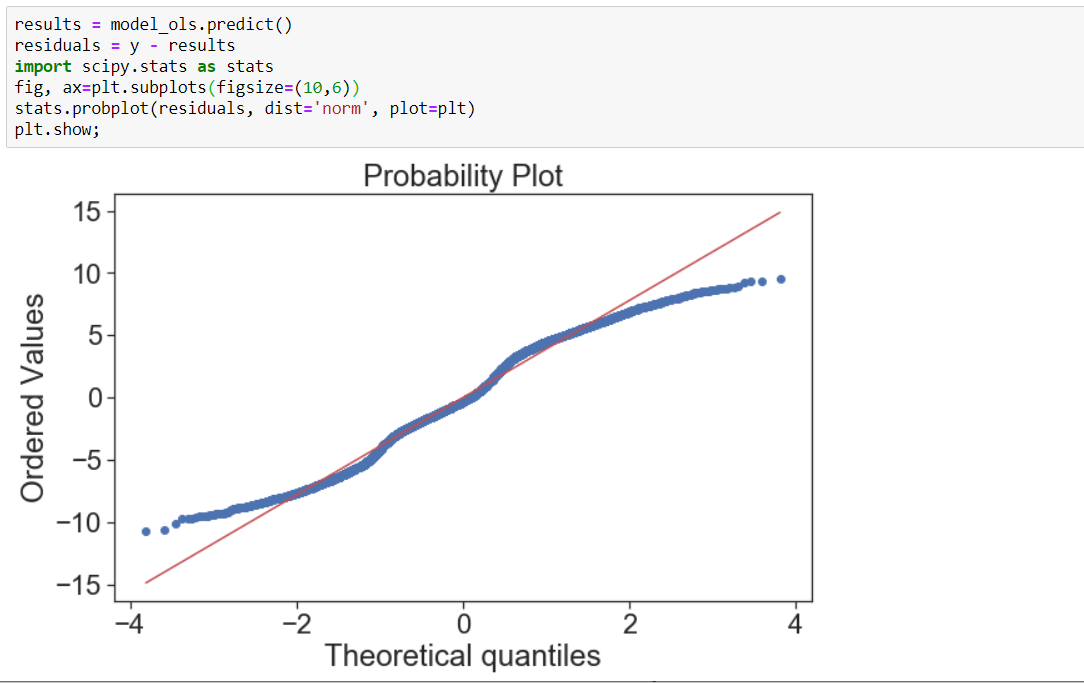
The p-value shows which of the variables are significant in the model at alpha level of 0.05. The no-significant values are: Income with p-value of 0.269, and Doc\_visits with p-value of 0.206 as shown in the results below. The last part of the table provides residuals tests of the model. The omnibus test is a combination of skew and kurtosis tests and its p-value is 0.00 making it significant. The skew is -0.107, which is the result of the measurement of the symmetry of the residuals. A negative result of the skew number is a long tail to the left. The kurtosis is 2.272, which is the measurement of the shape of the residuals. Normal distribution has kurtosis equal to 0. A positive kurtosis means a higher peak than normal distribution. The Durbin-Watson is a measurement of the correlation of the residuals. The Jarque-Bera is a combined test of skew and kurtosis with its p-value being very small. The Cond. No. is for testing multicollinearity. A result over 30 indicates that the residuals are unstable.



Multiple linear regression equation contributes us with a constant of -11.5420 which is the value of Y when all of the independent variables are 0. Then, the ‘Income’ variable has a positive effect to ‘Y’. Every time ‘Income’ increases, ‘Y’ increases 1.5e-0.06 times. ‘Additional\_charges’ has a negative effect on ‘Y’. Every time ‘Additional\_charges’ increase, ‘Y’ decreases 0.0002 times. The equation is shown next:

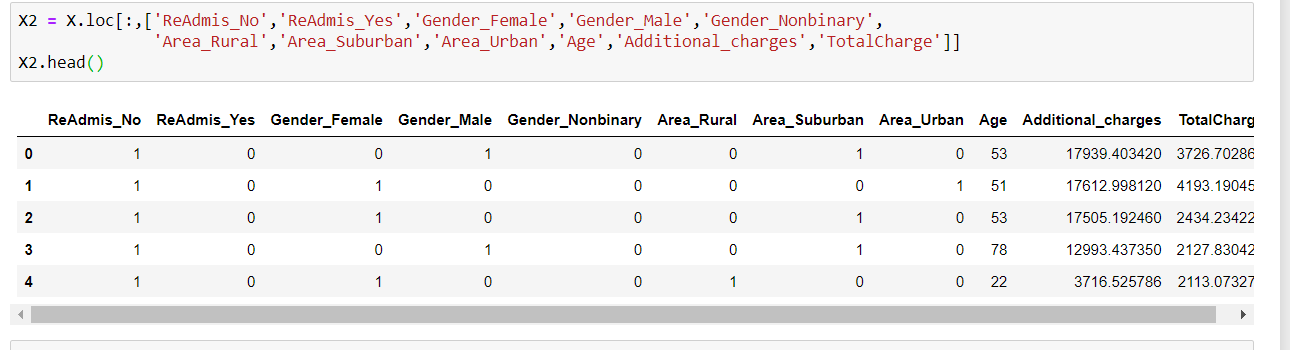
Y = -11.5420 + 1.5e-06(Income) + 0.0436(Age) – 0.0002(Addictional\_charges) -0.0476(Doctor\_visits) + 0.0113(TotarCharge) – 3.8267(Gender\_Female) -3.7108(Gender\_Male) – 4.0045(Gender\_Nonbinary) -3.8174(Area\_Rural) -3.8367(Area\_Suburban) – 3.8879(Area\_Urban) – 7.3782(ReAdmis\_No) – 4.1637(ReAdmis\_Yes)

The residuals plot is made to confirm the distribution of both the actual (theoretical) and the predicted (ordered) values. A plot of the residuals is generated to observe the variation between the actual and the predicted values. The probability plot below demonstrates two very close lines indicating that they are normally distributed.



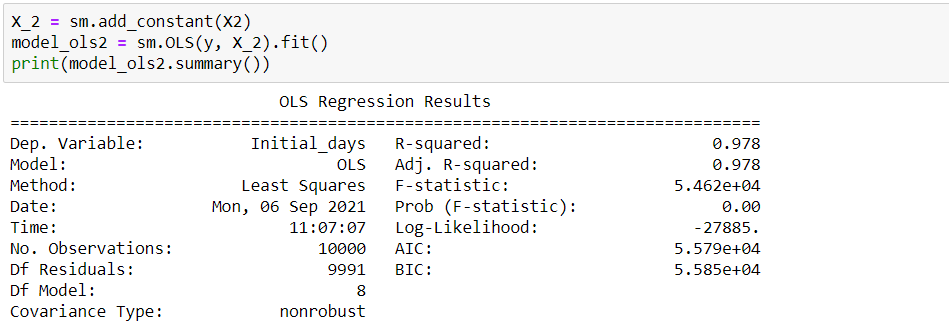
Reduce Model

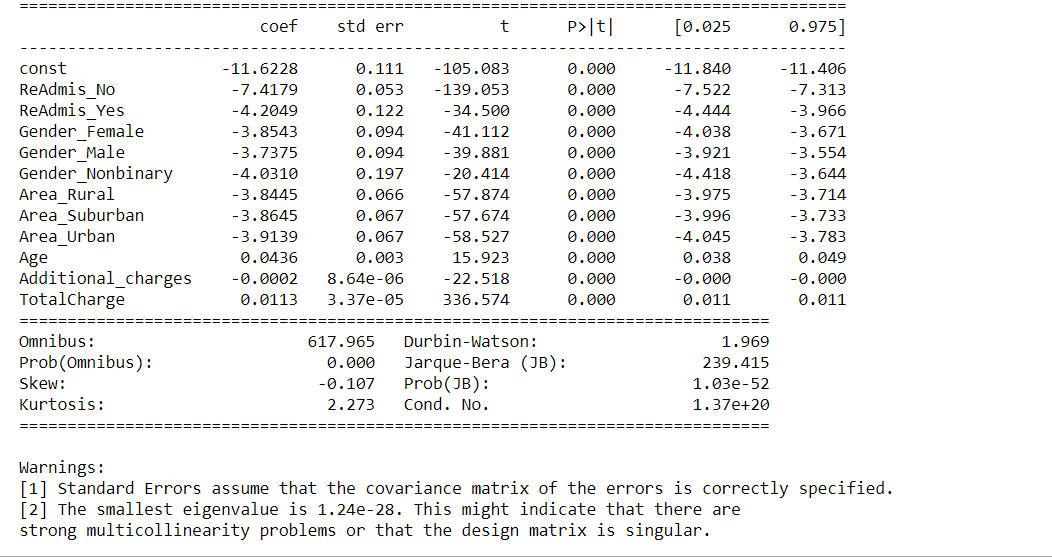
The significant variables were chosen based on their p-values to produce the reduce model. First, the no-significant variables were removed from the ‘X’ dataset.



Then, the model was made.

A statistical summary without the not-significant variables is created:





Multiple Linear regression reduced model equation:

Y = -11.6228 + 0.0436(Age) – 0.0002(Addictional\_charges) + 0.0113(TotarCharge) – 3.8543(Gender\_Female) – 3.7375(Gender\_Male) – 4.0310(Gender\_Nonbinary) – 3.8445(Area\_Rural) – 3.8645(Area\_Suburban) – 3.9139(Area\_Urban) – 7.4179(ReAdmis\_No) - 4.2049(ReAdmis\_Yes)

Conclusion

As a result of the analysis, we can conclude that the doctor visits and patients’ income have no influence on the days spent in the hospital. However, the age, gender, total charge, additional charges and the area in which the patient lives have an effect on the total days patients spent in the hospital.

The two residual plots are very similar for both models. The results of the two R-squares, adj. R-squares, F-statistics, and log-likelihood numbers are very close to provide the same evaluation of the model. Consequently, the reduced model is selected for its simplicity.

Recommendations for future research

* Run the model with a different set of variables. Where the correlation between the dependent variable and independent variables is stronger.
* Select a different target variable. Different target variable might its distribution closer to normal.
* Split the data in training and test datasets. This step can help the model accuracy.
* Use Sklearn library instead of statsmodels or other libreries available in python.
* Use different statistical method than multiple linear regression.
* Standardized the dataset to have all variables with the same scale of measurement.

References

Comparison of Python, R, and SAS languages:

<https://www.bing.com/videos/search?q=comparatives+between+Python%2c++and+SAS&&view=detail&mid=786FA178678969B7B5A0786FA178678969B7B5A0&&FORM=VRDGAR&ru=%2Fvideos%2Fsearch%3Fq%3Dcomparatives%2Bbetween%2BPython%252C%2B%2Band%2BSAS%26go%3DSearch%26qs%3Dds%26form%3DQBVDMH>

Medicine, Tom ProseCEO at General. “Top Reasons for Hospital Readmissions from Snfs " General Medicine.” General Medicine, [www.generalmedicine.com/top-reasons-for-hospital-readmissions-from-snfs](http://www.generalmedicine.com/top-reasons-for-hospital-readmissions-from-snfs)

Wikimedia Foundation. (2021, August 11). Linear regression. Wikipedia. [https://en.wikipedia.org/wiki/Linear\_regression#Simple\_and\_multiple\_linear\_regression](https://en.wikipedia.org/wiki/Linear_regression).

Zach. (2020, July 24). How to calculate vif in python. Statology. <https://www.statology.org/how-to-calculate-vif-in-python>

Wikimedia Foundation. (2021, August 11). Linear regression. Wikipedia. [https://en.wikipedia.org/wiki/Linear\_regression#Simple\_and\_multiple\_linear\_regression](https://en.wikipedia.org/wiki/Linear_regression)

Wikimedia Foundation. (2021, August 23). Coefficient of determination. Wikipedia. <https://en.wikipedia.org/wiki/Coefficient_of_determination>

Galarnyk, M. (2021, February 3). *PCA using Python (scikit-learn)*. Medium. <https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60>

*Visualizing distributions of data¶*. Visualizing distributions of data - seaborn 0.11.2 documentation. (n.d.). <https://seaborn.pydata.org/tutorial/distributions.html?highlight=univariate>