**D208**

**Performance Assessment**

**NBM2 Task 1**

**Multiple Regression for Predictive Modeling**

**Western Governors University**

**By: Michael Rupert**

**May 2022**

# Table of Contents

[Table of Contents 2](#_Toc103945590)

[A1: Research Question 3](#_Toc103945591)

[A2: Objectives and Goals 3](#_Toc103945592)

[B1: Summary of Assumptions 3](#_Toc103945593)

[B2: Tool Benefits 3](#_Toc103945594)

[B3: Appropriate Technique 3](#_Toc103945595)

[C1: Data Goals 4](#_Toc103945596)

## C1a: Data Preparation.………………………………………………………………………………………….……………….…..4

[C2: Summary Statistics 6](#_Toc103945597)

[C3: Steps to Prepare 9](#_Toc103945598)

[C4: Visualizations 9](#_Toc103945599)

[C5: Prepared Data Set 17](#_Toc103945600)

[D1: Initial Model 17](#_Toc103945601)

[D2: Justification of Model Reduction 20](#_Toc103945602)

[D3: Reduced Multiple Regression Model 21](#_Toc103945603)

[E1: Model Comparison 23](#_Toc103945604)

[E2: Output and Calculations 25](#_Toc103945605)

[E3: Code 25](#_Toc103945606)

[F1: Results 28](#_Toc103945607)

[F2: Recommendations 29](#_Toc103945608)

[G: Panopto 29](#_Toc103945609)

[H: Sources of Third-Party Code 29](#_Toc103945610)

[I: Sources 29](#_Toc103945611)

# 

# A1: Research Question

1. Can we build a model from variables that could predict which patient factors contribute to vitamin D levels?

# A2: Objectives and Goals

# Stakeholder, i.e., doctors, nurses and care staff, can benefit from the knowledge and better provide understanding and knowledge what action to take or what patients might require to increase Vitamin D levels to required levels. This would result in better care for the patients.

# B1: Summary of Assumptions

# With multiple regression models we can assume:

# There must be a linear relationship between the dependent and independent variables.

# Residuals are normally distributed, Multivariate Normality.

# The independent variables are not highly correlated to each other, No multicollinearity.

# There is No autocorrelation, the relationship between the current value and the past values of the variable in question

# Homoscedasticity, or the error term is relatively the same across all variables

# (Zach, 2021)

# B2: Tool Benefits

# Python is my tool of choice, running in the Anaconda framework. Python is a very robust language with many libraries that allow analysis of data, visualizations. Python is a multiplatform language which allows users to perform analysis on any style computer. Python is suited to handling large datasets quickly. Lastly, I am much more familiar with it even though I have used R in the past.

# B3: Appropriate Technique

# When building a predictive model, we can use Multiple Regression Analysis to assess the relationship between the outcome, dependent variables, and several predictor variables. We can also examine the importance of each predictor to the relationship, often, with the effect of other predictors eliminated statistically. There could be several. Instead of just one, explanatory variables (Doc\_Visits, Soft\_drink, HighBlood, Stroke, Complication\_risk, Overweight, Arthritis, Diabetes, Hyperlipidemia, BackPain, Anxiety, Allergic\_rhinitis, Reflux\_esophagitis, Asthma, Services, and Initial\_days, Timely\_addmission, Timely\_treatment.) which can increase our understanding as we develop a predictive model that can help use lower readmission to various facilities. As we add or remove variables from the regression model or equation, it will reflect a negative or positive relationship to the variable we are focusing on and ho best to adapt a facilities strategy on dealing with this.

# C1: Data Goals

## **C1a: Data Preparation**

The data was prepared in the previous Class D206 Data Cleaning. I will be using the cleaned data set.

After looking at the data for duplicate records none were detected. The following variables were found to have missing values or values of zero would prove to be incorrect values. For example, the Population variable having a value of zero for any record will prove incorrect simply due to the fact if a person lives in an area then the population is at least 1. Other variables that had missing values were, 'Children', 'Age', 'Household\_Income', 'Soft\_drink', 'Overweight', 'Anxiety', 'Initial\_days'. All these values were replaces using the mean of all other records of that variable.

Also re-expressing qualitative data and convert to quantitative data by changing yes or no values to 1 or 0 as well as other categories being re-represented using integers.

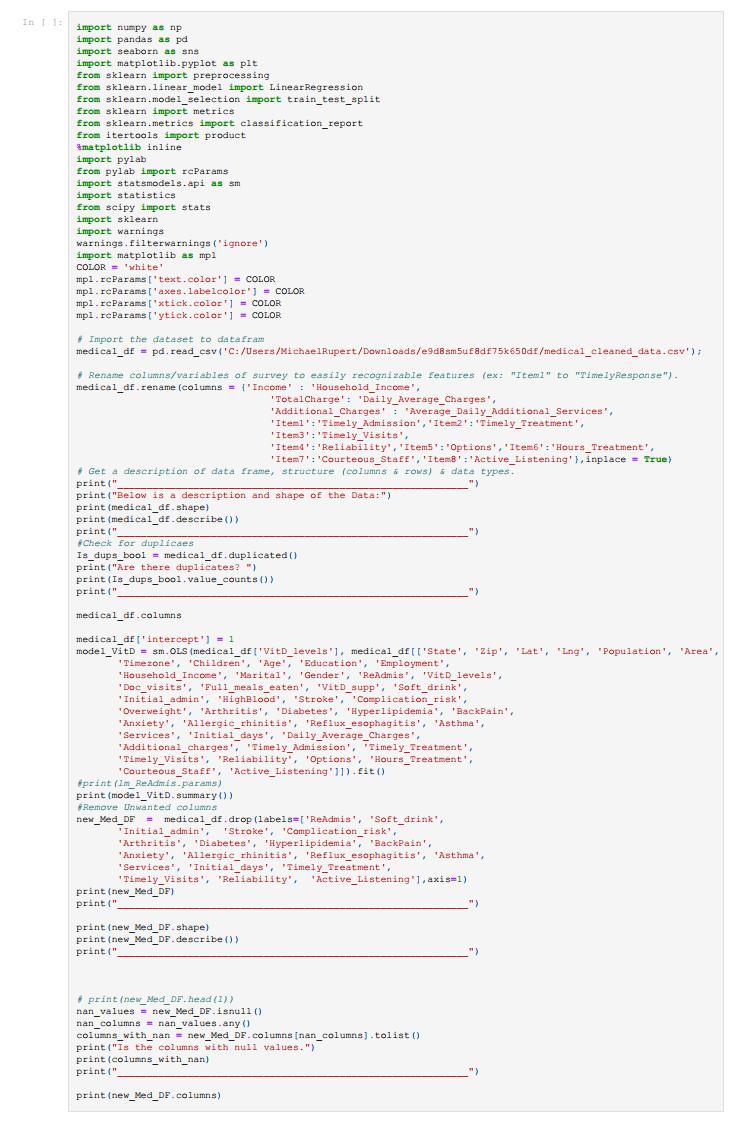
Several the variables present in the data had outliers, some of these outliers were kept, Population, Household\_Income, Total\_Average\_Daily\_Charge, were left unchanged. Other outliers were normalized.

As discussed in the previous Data Cleaning section, variables were replaced using the mean of the values of that variable that did exist. This is a common way to handle missing data points but must be used with caution so as not to reduce the variability of the data. ([https://seleritysas.com/blog/2020/03/03/the-best-way-to-handle-missing-data//](https://seleritysas.com/blog/2020/03/03/the-best-way-to-handle-missing-data/))

The normalizing of data with outliers was used to deal with many of the variables that had outliers. This causes the data to be utilized across all records in the database. (<https://blog.insycle.com/normalize-data>)

The most relevant variable to our decision-making process is VitD\_level, this will be our focus variable to determine what variables effect Vitamin D levels. The model mut be trained and tested on the cleaned data set to develop a model that will give us an idea of what effects Vitamin D levels in patients given their respective data points for selected predictor variables.

**Cleaning code shown below:**

****

# C2: Summary Statistics

# Originally the data set consisted of 53 fields and 10000 records. Not all this is needed for this analysis and hence some for the fields were removed i.e.; ’CaseOrder’, ’Customer\_id’, ’Interaction’, ’UID’, ’City’, ’State’, ’County’, ’Zip’, ’Area’, ’TimeZone’, ’Job’, ’Marital’. All nulls were removed and replaced using the mean, the data has been normalized. Any erroneous data was repaired.

# Below is a description and summary of statistics of variables including the target variable.

# Description

# Graphical user interface Description automatically generated with medium confidence

# OLS Summary

# A picture containing graphical user interface Description automatically generated

## **Summary Discussion**

# After reviewing the variables using OLS Regression, we can see that if we used all the variables, we may end up with an overfitting problem. A number of these can be removed due to a lack of relevancy to the target variable, VitD\_levels. The most likely candidates to keep are indicated by their p-Value indicated in the 5th column. An of the variables below the value of 0.05 are likely candidates for selection for further investigation. The variables, 'State', 'Zip', 'Lat', 'Lng', 'Population', 'Area', 'Timezone', 'Children', 'Age', 'Education', 'Employment', 'Household\_Income', 'Marital', 'Gender', 'Doc\_visits', 'Full\_meals\_eaten', 'VitD\_supp', 'HighBlood', 'Overweight’, 'Options', 'Hours\_Treatment', 'Courteous\_Staff', ‘Daily\_Average\_Charges', 'Additional\_charges', 'Timely\_Admission', all could have some impact on the level of Vitamin D in a patient according to the output. We can explore some of these variables and see if further removal of unneeded variables will make the data even more manageable and help reduce the possibility of overfitting.

# If we look both at the description of the data and the OLS summary from this initial model which contains all pertinent data remaining after removing the identification data and other needless columns. We can see the mean of each, variable, the values represented in each quartile, the standard deviation, and the minimum and maximum values.

# The summary gives us a more granular idea of the information about each variable. As can be seen in the table above. This header region shows us, the dependent variable. It explains that it is an OLS model is Least squares model. This method minimizes the sum of the squares of the residuals between the observed targets of the dataset, and the targets predicted by linear approximation. The summary also gives us the R-squared value which indicates how good of a fit this model is. The closer the value is to 1 the better the fit. 1 is a perfect fit, the fit of this initial model is 0.984. It also provided the adjusted R-squared which is adjusted for the number of variables in the model. The 0.984 explains the percentage of variation in dependent is explained by the independent variables. In this model 98% variation in VitD\_levels are explained by the variables listed in the output above.

# The Prob(F-statistic) indicates the overall significance of the over all regression. This assesses the significance of all the variables together. The closer this is to zero indicates the regression is meaningful. This is zero.

# In the middle body of the summary, the first column is the variable name. In the second column is the Coefficient of each variable or its weight in the general initial model. This can also be gotten by using <model>.params. The third column shows the standard deviation error for each variable. The next column shows the t value, measures the size of the difference relative to the variation in your sample data. The greater the magnitude of T, the greater the evidence against the null hypothesis. In the next column of fifth column, we find the p value. The p value is the value we will use to determine what variables to remove from the initial value. A p value above 0.05 will be grounds to eliminate the variable. Below 0.05 means the variable is significant and will be kept.

# There are currently 32 variables within this model that are significant.

# The final value of not is the Durbin-Watson score This shows the level of Homoscedasticity. A good score is between 1 and 2. (Yadav, 2019) The score for this model is 1.925.

# C3: Steps to Prepare

# Import dataset to Python data frame.

# Rename columns/variables of survey to easily recognizable features (ex: "Item1" to "TimelyResponse").

# Get a description of data frame, structure (columns & rows) & data types.

# View summary of statistics of the dataset.

# Remove data that will unlikely impact the model (ex: "Customer\_id") & demographic columns (ex: zip code) from the dataset.

# Check for missing data fields & impute missing data with measures of central tendency (mean, median or mode) or remove outliers that are several standard deviations above the mean or normalize.

# Encode categorical, yes/no data points into 1/0 numerical values.

# This was done in a previous D206 Assessment.

# Encode other categorical data with integer data that makes sense.

# This was done in a previous D206 Assessment.

# View univariate & bivariate visualizations.

# Place VitD\_levels at end of Dataset.

# Rename the prepared dataset to " medical\_cleaned\_data.csv"

# Run initial regression model

# Eliminate variables

# 

# C4: Visualizations

# It can be seen in the boxplots below there are several outliers. There is no valid reason to remove these after checking the validity of these outliers it is decided to keep them to give a more accurate picture of the data and properly fit a regression model.

# Univariate Visualizations

# The code used to produce the histograms below can be seen below. Two arrays were created to handle continuous and categorical variables. These will be used in further along in this analysis.

# Text Description automatically generated

# Below are the output histograms of continuous variables selected. Here we can see the probability distribution for the continuous values we selected.

# A screenshot of a computer Description automatically generated with low confidence

# Boxplots

# Here we can see the measures of central tendency as well as the variability of the data in this data set. Below is the code that produces the visualizations.

# Text Description automatically generated

# Graphical user interface Description automatically generated with low confidence A screenshot of a computer Description automatically generated with low confidence

# Graphical user interface Description automatically generated A picture containing graphical user interface Description automatically generated

# Bivariate

# For the non-continuous variables use of categorical scatter plot.

# The code used to produce the categorical plots is below.

# Text Description automatically generated

# Graphical user interface, application Description automatically generated Graphical user interface Description automatically generated with medium confidence

# Graphical user interface Description automatically generated with low confidence Graphical user interface, application Description automatically generated

# Timeline Description automatically generated Timeline Description automatically generated with medium confidence

# Below are scatterplots to show direct or inverse relationships between target & independent.

# The code below produces the scatterplots

# A screenshot of a computer Description automatically generated

# Graphical user interface Description automatically generatedScatter chart, qr code Description automatically generated

# Chart, scatter chart Description automatically generated Chart, scatter chart Description automatically generated

# A model of the selected and cleaned variables will now be produced. Below is the code and new model output.

# C5: Prepared Data Set

# *A copy of the prepared data set is included in the submission.*

# D1: Initial Model

# We created the model below, which is step one. I wish to mention one thing of note, if we look at the metric R-squared, shown here to be 0.984, this indicated this model is a near perfect fit. This makes me wonder if it is also an overfit, even the R-squared adjusted is 0.984 indicating this might not accurately represent the general population and include too many variables. I will continue the analysis and see what might be removed further to make this a good fit for general population. There are 45 column or variables in this data.

# The summary gives us a more granular idea of the information about each variable. As can be seen in the table above. This header region shows us, the dependent variable. It explains that it is an OLS model is Least squares model. This method minimizes the sum of the squares of the residuals between the observed targets of the dataset, and the targets predicted by linear approximation. The summary also gives us the R-squared value which indicates how good of a fit this model is. The closer the value is to 1 the better the fit. 1 is a perfect fit, the fit of this initial model is 0.984. It also provided the adjusted R-squared which is adjusted for the number of variables in the model. The 0.984 explains the percentage of variation in dependent is explained by the independent variables. In this model 98% variation in VitD\_levels are explained by the variables listed in the output above.

# The Prob(F-statistic) indicates the overall significance of the overall regression. This assesses the significance of all the variables together. The closer this is to zero indicates the regression is meaningful. This is zero.

# In the middle body of the summary, the first column is the variable name. In the second column is the Coefficient of each variable or its weight in the general initial model. This can also be gotten by using <model>.params. The third column shows the standard deviation error for each variable. The next column shows the t value, measures the size of the difference relative to the variation in your sample data. The greater the magnitude of T, the greater the evidence against the null hypothesis. In the next column of fifth column, we find the p value. The p value is the value we will use to determine what variables to remove from the initial value. A p value above 0.05 will be grounds to eliminate the variable. Below 0.05 means the variable is significant and will be kept.

# There are currently 32 variables within this model that are significant.

# Text Description automatically generated

# A picture containing graphical user interface Description automatically generated

# 

# D2: Justification of Model Reduction

# Reason for Variable Removal able

|  |  |  |
| --- | --- | --- |
| Variable | Reason | p-value |
| State | p-value | 0.808 |
| Timezone | p-value | 0.279 |
| VitD\_supp | p-value | 0.161 |
| Soft\_Drink | p-value | 0.297 |
| HighBlood | p-value | 0.467 |
| Stroke | p-value | 0.083 |
| Allergic\_rhinitis | p-value | 0.104 |
| Asthema | p-value | 0.481 |
| Additional\_charges | p-value | 0.069 |
| Timely\_Treatment | p-value | 0.334 |
| Timely\_Visits | p-value | 0.601 |
| Courteous\_Staff | p-value | 0.145 |
| Active\_Listening | p-value | 0.310 |
| Zip | redundant | 0.000 |
| Area | redundant | 0.000 |
| Doctor\_Visits | unneeded | 0.000 |
| Hours\_Treatment | ambiguous | 0.003 |
| Options | ambiguous | 0.000 |
|  |  |  |

# There are 45 column or variables in this data. This is a large number of variables. Some of these can be removed. Justification for removing these is as follows:

# We can examine and remove first any of the variables with a P value above 0.05, as mentioned above the significant variables are below 0.05. the others are not considered significant. The following variables fit these criteria: State, Timezone, VitD\_supp, Soft\_drink, HighBlood, Stroke, Allergic\_rhinitis , Asthma, Additional\_charges, Timely\_Treatment, Timely\_Visits, Courteous\_Staff, Active\_Listening.

# There are a few variables that are redundant State, Zip, both can be removed since Lat and Lng are more precise, all of these put the patient in the same general geographical area. This is also the case for Area and population a high population is urban a low population is rural. I propose removing these simply because they are redundant.

# Other variables I propose removing are, Options will be removed simply because it is too vague. There seems no reason to keep Doctor\_Visits since much of the care is provided by the nurses. Hours\_Treatment is vague, is it surgery or how much attention the patient gets from the staff.

# Inspecting the p-values in the above model we can see which values are most likely to affect this model and which are not. We can remove several variables and bring this model in to an easier more tractable size but reducing the number of variables again. After reducing the variables further, we are left with: 'Lat', 'Lng', 'Population', 'Children', 'Age', 'Education', 'Employment', 'Household\_Income', 'Marital', 'Gender', 'ReAdmis', 'Full\_meals\_eaten', 'Initial\_admin', 'Complication\_risk', 'Overweight', 'Arthritis', 'Diabetes', 'Hyperlipidemia', 'BackPain', 'Anxiety', 'Reflux\_esophagitis', 'Services', 'Initial\_days', 'Daily\_Average\_Charges', 'Timely\_Admission', 'Reliability. By reducing the model, we can avoid the potential for overfitting. Overfitting could misrepresent the data and not give us a model that is suitable to the general population. Some of this data is redundant for example Zip code and longitude/latitude tend to put the patient in the same area as does state and city information.

# D3: Reduced Multiple Regression Model

# If I adjust the model further and remove a few more variables I finally end up with only these variables, 'Lat', 'Lng', 'Population', 'Children', 'Age', 'Education', 'Employment', 'Household\_Income', 'Marital', 'Gender', 'ReAdmis', 'Full\_meals\_eaten', 'Initial\_admin', 'Complication\_risk', 'Overweight', 'Arthritis', 'Diabetes', 'Hyperlipidemia', 'BackPain', 'Anxiety', 'Reflux\_esophagitis', 'Services', 'Initial\_days', 'Daily\_Average\_Charges', 'Timely\_Admission', 'Reliability ' The other values were removed due to p-values or the type of information, (i.e., Zip codes are numeric but are categorical and latitude and longitude are better representations of location). Running a model with these variables yields a very minor change in R-squared, 0.983, a very slight change and a very good model according to the R-Squared value.

# Graphical user interface, text Description automatically generated with medium confidence

# Next, we get the Coefficients.

# Text Description automatically generated

# The coefficients yield the following equation:

# VitD\_Level = 0.086826(Lat) + -0.022287(Lng) +0.000011(Population) + 0.054354(Children) + 0.011618(Age) + 0.106996(Education) + 0.081692(Employment) + (0.000005)Household\_Income +0.079221(Marital) + 0.283736(Gender) + -2.767702(ReAdmis) + 0.102105(Full\_meals\_eaten) + 0.745789(Initial\_admin ) + -0.234681(Complication\_risk) + 0.257335(Overweight) + -0.124040(Arthritis) + -0.245011(Diabetes) + -0.125870( Hyperlipidemia) -0.157625(BackPain) + -0.167394(Anxiety) + -0.109512(Reflux\_esophagitis) + 0.178840(Services) + -0.148032(Initial\_days) + 0.002374(Daily\_Average\_Charges) + 0.242683(Timely\_Admission) + 0.244920 (Reliability) (The Trustees of Princeton University, 2007)

# E1: Model Comparison

|  |  |  |
| --- | --- | --- |
| Statistic | Initial | Reduced |
| R-Squared | 0.984 | 0.983 |
| Prob (F-statistic) | 0.00 | 0.00 |
| Durbin-Watson | 1.925 | 1.915 |
| Number of Vars | 45 | 26 |
| MSE | 6.894 | 7.262 |
| RSE | 2.626 | 2.695 |

# If we compare the models, we can see the is a very negligible change in the R-squared values, the same is with r-squared adjusted. This indicates the model is very good since a perfect model is 1. The variation in Vitamin D levels can be explained 94.7% of the time. The RSE value for the initial model to reduced model is 2.626 vs. 2.695 , this is calculated by taking the square root of the MSE, which indicates either of these is a good fit. If we look at the Initial model, we have a R-squared of 0.984 and RSE above, we see that indeed, the initial model was a better fit with the Residual Standard Error slightly less than the reduced model and the R-Squared closer to 1. The reduced model runs the analysis and predictions much faster on the computer which can be beneficial, and the difference is negligible between the two.

# The logic behind variable selection

# There are a few techniques when selecting variables in a multiple regression model. The first is the full model approach. In this approach we remove terms one at a time to see how the model responds. We focus only on removing the variables with a p value above a certain threshold, in this case 0.05, as mentioned previously. This is the method used in this analysis. Mentioned previously, some of the variables were removed due to redundancy or ambiguousness of the information they represented.

# Other methods of variable selection include , forward selection, starting with only the intercept and slowly adding more variables in one at a time starting with the predictor with the lowest pvalue and continuing until all remaining values are above the specific pvalue threshold.

# Another way is the mixed approach this allows nonsignificant terms to be removed.

# And lastly is the best subsets regression and all possible models get fit from the combination of predictors. This could be very cumbersome with large models of many variables. (Variable Selection in Multiple Regression, n.d.)

# Below is the residuals plot with trend line and code to produce it.

# A screenshot of a computer Description automatically generated with medium confidence

# A picture containing text, monitor, screenshot, screen Description automatically generated

# E2: Output and Calculations

# A screenshot of a computer Description automatically generated with medium confidence

# Below is a prediction performed using the reduced model. The code and output are shown in the frame.

# Text Description automatically generated

# E3: Code

# *\*\*code will also be included in a file\*\**

# Text Description automatically generated

# Text Description automatically generated

# Text Description automatically generated

# Graphical user interface, text, application Description automatically generated

# F1: Results

# After exploring this data to determine what factors seem to affect the Vitamin D levels in patients. It would seem from the model that people who have a higher income bracket, have a higher daily average, more meals during their visit all benefit from higher Levels of Vitamin D. This can be seen in the distribution of the scatterplots and other analysis. We can, indeed, build a model that reflects and fits the data provided to predict with in a 5% CI that predicts what factors influence Vitamin D levels. Latitude and Longitude play a part since Vitamin D is naturally produced by the sun we get on our skin. If we look at the data and see that people at a 60+ latitude, (most likely living in Alaska or higher northern latitudes), receive less sun light than people at lower latitudes.

# As discussed above the models according to the R-squared values, RSE, and MSE values seem to be a good fit for predicting Vitamin D levels.

# Some surprising results are that supplementing vitamin D levels while at the facility didn’t seem to have much affect on the target variable, VitD\_levels.

# F2: Recommendations

# From these results it is recommended that people who live in facilities at more northern latitudes be educated to perhaps supplement their Vitamin D intake. It is also a recommendation that people various lower income patients might benefit from increased Vitamin D intake. Within the facility, it is recommended that patients eat more full meals.

# The stakeholders most interested in this information would be primary care physicians who could prescribe diet changes, supplements and other recommendations that could educate their patients and help decrease Vitamin D deficiency.

# G: Panopto

# H: Sources of Third-Party Code

# <https://app.datacamp.com/learn/courses/intermediate-regression-with-statsmodels-in-python>

# <https://seaborn.pydata.org/generated/seaborn.catplot.html>

# I: Sources

# Definitions of Statistical terms

# <https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/>

# Bibliography

team, C. c. (n.d.). *Histograms and Box Plots*. Retrieved from CIToolkit: https://citoolkit.com/articles/histograms-and-boxplots/

The Trustees of Princeton University. (2007). *Interpreting Regression Output*. Retrieved from Princeton University Library: https://dss.princeton.edu/online\_help/analysis/interpreting\_regression.htm#:~:text=a%20miniscule%20effect.-,Coefficients,the%20direction%20of%20the%20effect.

*Variable Selection in Multiple Regression*. (n.d.). Retrieved from JMP Statistical Discovery: https://www.jmp.com/en\_us/statistics-knowledge-portal/what-is-multiple-regression/variable-selection.html

Yadav, J. (2019, August 15). *Statistics: How Should I interpret results of OLS?* Retrieved from https://jyotiyadav99111.medium.com/: https://jyotiyadav99111.medium.com/statistics-how-should-i-interpret-results-of-ols-3bde1ebeec01#:~:text=Prob(F%2DStatistic)%3A%20This,measures%20it%20for%20individual%20variables.

Zach. (2021, November 16). *Assumptions of Multiple Linear Regression*. Retrieved from statology: ps://www.statology.org/multiple-linear-regression-assumptions/

: