Performance Assessment: Linear Regression Modeling

Gabriela Howell

Master of Science Data Analytics, Western Governors University

D208 – Linear Regression Modeling

Professor Choudhury

March 30, 2024

Part I: Research Question

A. Describe the purpose of this data analysis by doing the following:

1. Summarize one research question that is relevant to a real-world organizational situation captured in the data set you have selected and that you will answer using multiple linear regression in the initial model.

The purpose of this paper is to investigate the potential relationship between income levels and health conditions using multiple linear regression analysis. By examining a medical dataset, the research question focuses on understanding how income influences the likelihood of readmission for health conditions. This is relevant to the real world because readmission is an overall issue for companies as they are getting fined. Understanding the impact of income on readmission rates can provide valuable insights for healthcare policies and interventions aimed at reducing readmission rates and improving overall patient outcomes.

2. Define the goals of the data analysis.

The goal of the data analysis is to find patterns and relationships using Regression. With a specific focus on understanding the important variables of health-related issues associated with income. By identifying key factors that influence readmissions in hospitals, this analysis aims to assist healthcare establishments in prioritizing their resources more effectively.

Part II: Method Justification

B. Describe multiple linear regression methods by doing the following:

1. Summarize four assumptions of a multiple linear regression model.

Multiple regression relies on several key assumptions to ensure the authenticity and reliability of the analysis. Firstly, it assumes a linear relationship between the outcome we're predicting and the factors we're analyzing, meaning that changes in predictors result in proportional changes in the outcome. Secondly, the factors involved in the model shouldn't be strongly correlated with each other to avoid multicollinearity, which complicates coefficient interpretation. Additionally, the data used for regression should be randomly selected and independent, allowing for generalization beyond the sample. Third, the differences among predicted values and actual outcomes, recognized as residuals, should follow a normal distribution pattern with an average of zero, ensuring the normality of residuals assumption. Lastly, increasing the complexity of the model by adding more factors should ideally improve its ability to explain outcome variation, but excessive predictors without justification can lead to overfitting. These assumptions collectively provide for the authenticity and reliability of the regression analysis.

2. Describe two benefits of using Python or R in support of various phases of the analysis.

I choose to use Python as the language to do my data analysis. There are numerous benefits to using Python, but I will just name two. Firstly, Python has an allotted amount of libraries for incorporating linear regression. The second is Python presents scalability, as it is compatible with large datasets. These two attributes make Python a perfect place to do linear regression analyses. Python is great as it sets up regression as a machine-learning problem. Moreover, Python's capacity to frame regression as a machine-learning problem enhances its utility and robustness, as emphasized by Srinivasan.

3. Explain why multiple linear regression is an appropriate technique to use for analyzing the research question summarized in part I.

Multiple linear regression is the suitable method for analyzing the research question due to its ability to simultaneously consider multiple factors. It is employed to investigate the relationship involving the dependent variable, which in this instance is income, and multiple independent variables. This approach provides insights into how every independent variable influences to the variation in the dependent variable while steering for the influences of another variables.

Part III: Data Preparation

C. Summarize the data preparation process for multiple linear regression analysis by doing the following:

1. Describe your data cleaning goals and the steps used to clean the data to achieve the goals that align with your research question including your annotated code.

The first step in the cleaning process is to ensure there are no duplicates, outliers, and missing values. All of which were done and came back with nothing. As outliers seemed reasonable to keep. These steps align with the research question to allow data to follow the Income to the various data points, I included all these steps in my Python code which will be attached to this document.

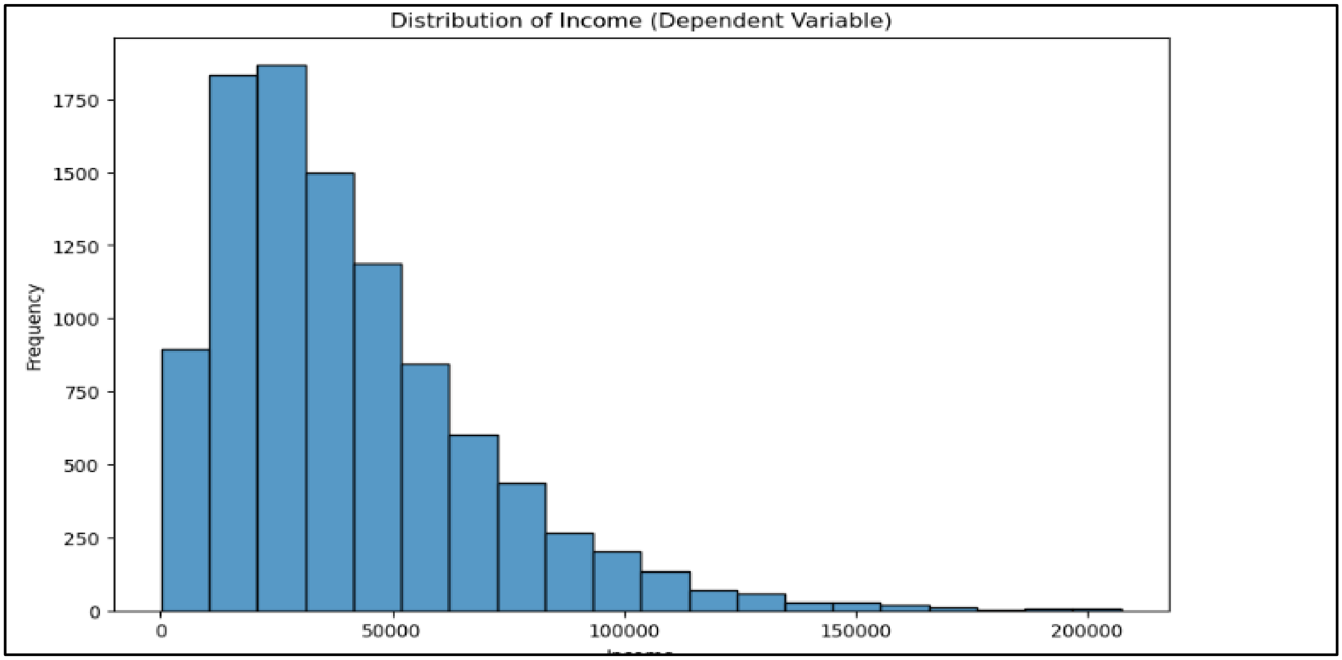
2. Describe the dependent variable and all independent variables using summary statistics that are required to answer the research question, including a screenshot of the summary statistics output for each of these variables.

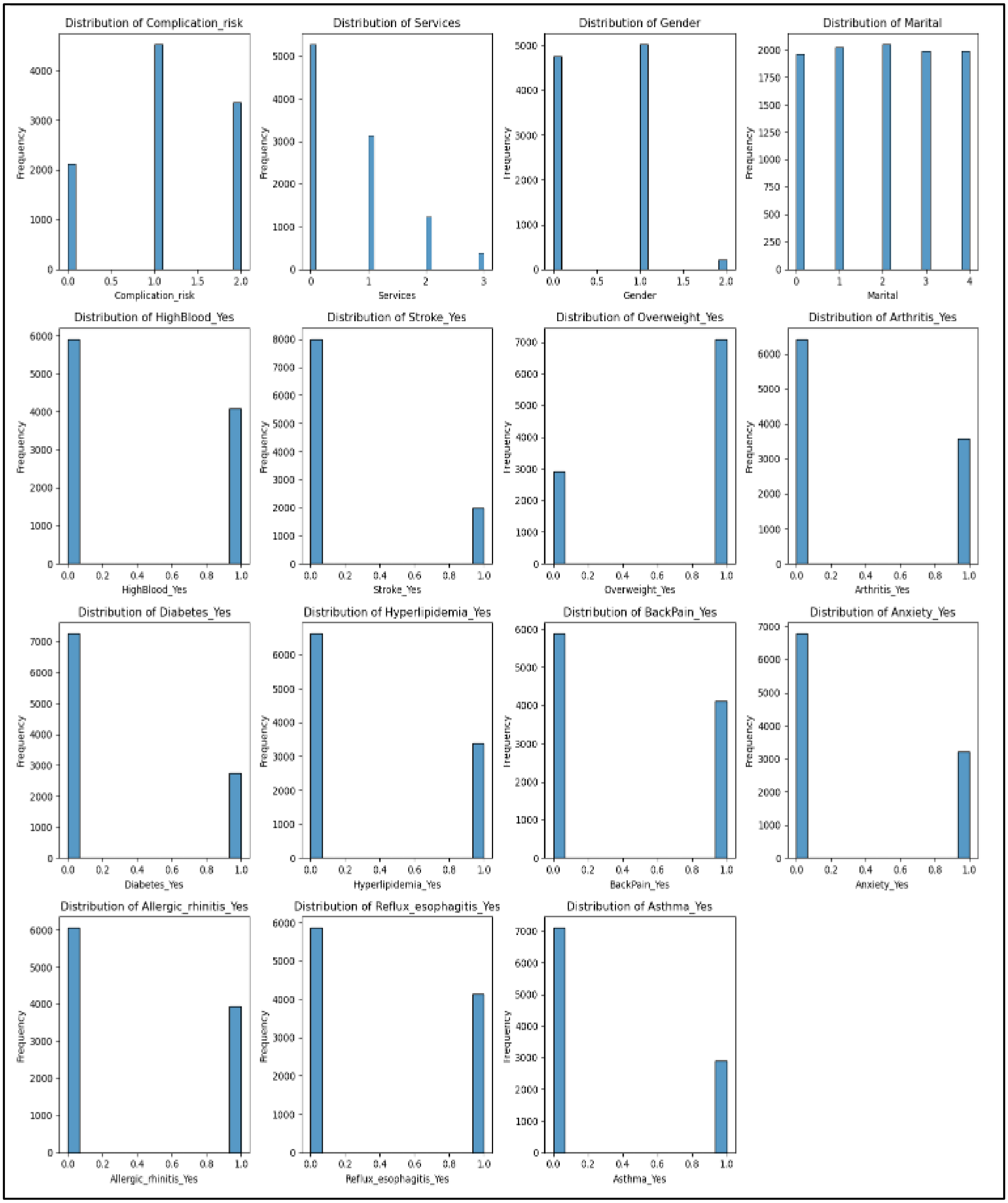
For this research the dependent variable is Income and the independent variables are BackPain, HighBlood, Complication_risk, Asthma, Services, Overweight, Arthritis, Diabetes, Hyperlipidemia, Anxiety, Stroke, Allergic_rhinitis, Reflux_esophagitis, Gender and Marital. Enclosed below are summary statistics for each independent variable alongside the dependent variable.

```
Summary statistics for X:
       Complication_risk      Services        Gender       Marital  \
count       10000.000000  10000.000000  10000.000000  10000.000000
mean            1.123300      0.672000      0.544600      2.001300
std             0.730172      0.832758      0.539296      1.407159
min             0.000000      0.000000      0.000000      0.000000
25%             1.000000      0.000000      0.000000      1.000000
50%             1.000000      0.000000      1.000000      2.000000
75%             2.000000      1.000000      1.000000      3.000000
max             2.000000      3.000000      2.000000      4.000000

       HighBlood_Yes    Stroke_Yes  Overweight_Yes  Arthritis_Yes  \
count   10000.000000  10000.000000    10000.000000   10000.000000
mean        0.409000      0.199300        0.709400       0.357400
std         0.491674      0.399494        0.454062       0.479258
min         0.000000      0.000000        0.000000       0.000000
25%         0.000000      0.000000        0.000000       0.000000
50%         0.000000      0.000000        1.000000       0.000000
75%         1.000000      0.000000        1.000000       1.000000
max         1.000000      1.000000        1.000000       1.000000

       Diabetes_Yes  Hyperlipidemia_Yes  BackPain_Yes   Anxiety_Yes  \
count   10000.00000        10000.000000  10000.000000  10000.000000
mean        0.27380            0.337200      0.411400      0.321500
std         0.44593            0.472777      0.492112      0.467076
min         0.00000            0.000000      0.000000      0.000000
25%         0.00000            0.000000      0.000000      0.000000
50%         0.00000            0.000000      0.000000      0.000000
75%         1.00000            1.000000      1.000000      1.000000
max         1.00000            1.000000      1.000000      1.000000

       Allergic_rhinitis_Yes  Reflux_esophagitis_Yes  Asthma_Yes
count           10000.000000            10000.000000  10000.00000
mean                0.394100                0.413500      0.28930
std                 0.488681                0.492486      0.45346
min                 0.000000                0.000000      0.00000
25%                 0.000000                0.000000      0.00000
50%                 0.000000                0.000000      0.00000
75%                 1.000000                1.000000      1.00000
max                 1.000000                1.000000      1.00000

Summary statistics for Y:
count    10000.000000
mean        53.511700
std         20.638538
min         18.000000
25%         36.000000
50%         53.000000
75%         71.000000
max         89.000000
Name: Age, dtype: float64
```

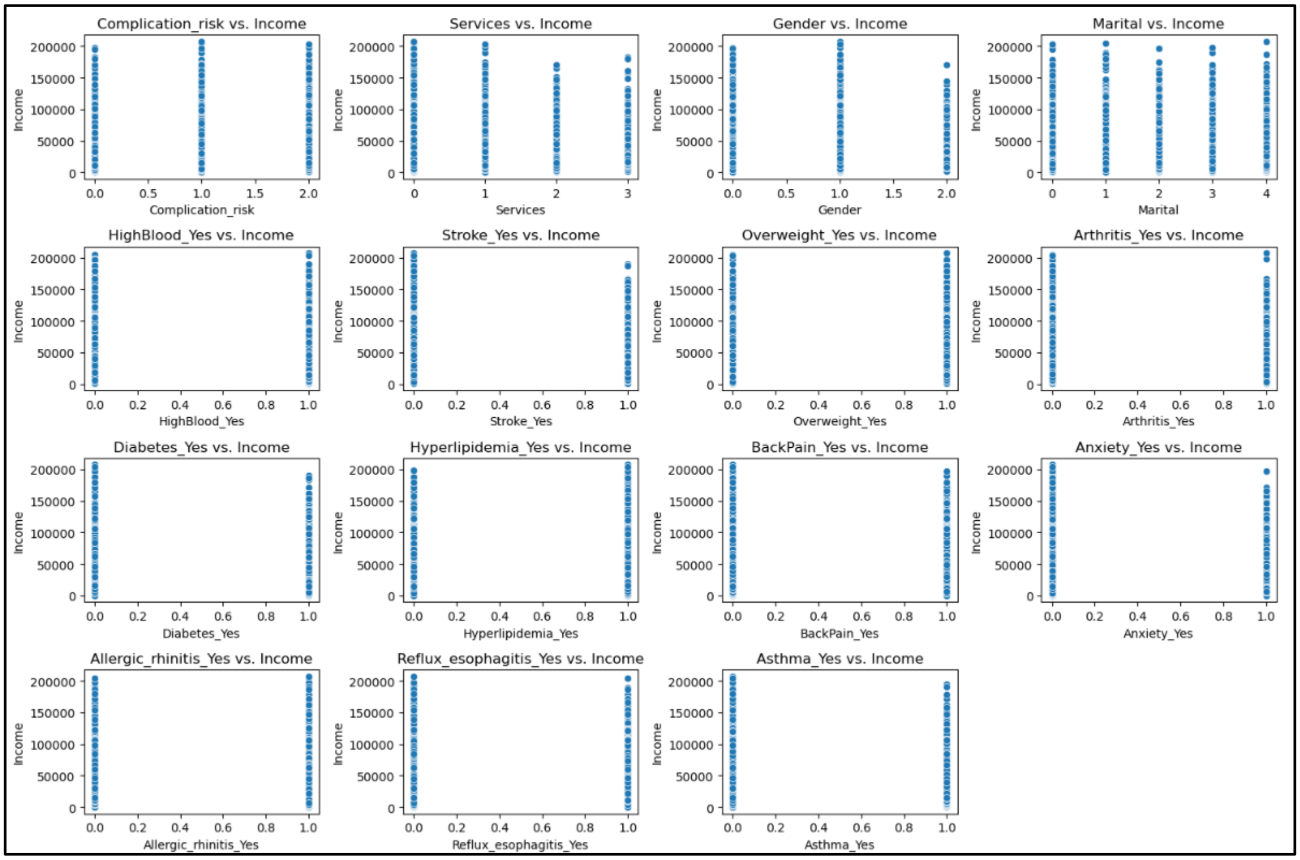
3. Generate univariate and bivariate visualizations of the distributions of the dependent and independent variables, including the dependent variable in your bivariate visualizations.

Below is my univariate :




Distribution of Complication_risk
Complication_risk
Distribution of Services
Services
Distribution of Gender
Gender
Distribution of Marital
Marital
Distribution of HighBlood_Yes
HighBlood_Yes
Distribution of Stroke_Yes
Stroke_Yes
Distribution of Overweight_Yes
Overweight_Yes
Distribution of Arthritis_Yes
Arthritis_Yes
Distribution of Diabetes_Yes
Diabetes_Yes
Distribution of Hyperlipidemia_Yes
Hyperlipidemia_Yes
Distribution of BackPain_Yes
BackPain_Yes
Distribution of Anxiety_Yes
Anxiety_Yes
Distribution of Allergic_rhinitis_Yes
Allergic_rhinitis_Yes
Distribution of Reflux_esophagitis_Yes
Reflux_esophagitis_Yes
Distribution of Asthma_Yes
Asthma_Yes
Frequency

Here is the Bivariant variables:



4. Describe your data transformation goals that align with your research question and the steps used to transform the data to achieve the goals, including the annotated code.

Since some of the variables I picked were categorical variables, they required transformation into numerical data to be included in the analysis and model training. Categorical variables, such as 'Services' indicating different medical services or 'Marital' denoting marital status, contain non-numeric values that cannot be directly used in mathematical computations. So, transforming these categorical variables into numerical, enables statistical analyses and machine learning models. This process provides multiple purposes. Firstly, it ensures that all variables in the dataset are in a consistent format, facilitating comparison and interpretation across different variables. Secondly, it aligns with the assumptions of linear regression, which normally involve numerical inputs. Linear regression always assumes a linear relationship amongst the independent and dependent variables, and using numerical representations allows for the computation of coefficients and predictions within this framework. Additionally, transforming categorical variables into numerical form through techniques like one-hot encoding (creating dummy variables) or ordinal encoding (mapping categories to numerical values) enables the model to capture the inherent relationships between categories. For example, if a categorical variable represents different levels of severity (e.g., ‘Low’, ‘Medium’, ‘High’), encoding it numerically preserves the ordinal relationship between these levels, which can be important for certain analyses.

In the process of transforming categorical variables into numerical data, the purpose is to ensure comparability and meet the assumptions of linear regression. These transformations are implemented using Python libraries like scikit-learn and numpy, enhancing the reliability and interpretability of the analysis for multiple linear regression.

5. Provide the prepared data set as a CSV file.

The provided CSV file is called ‘cleaned_data.csv’

Part IV: Model Comparison and Analysis

D. Compare an initial and a reduced linear regression model by doing the following:

1. Construct an initial multiple linear regression model from all independent variables that were identified in part C2.

The initial model, for the multiple linear regression includes 15 independent variables and reports an R-squared value of 0.001, indicating poor performance in explaining income variation. The adjusted R-squared, considering the number of predictors, is negative, suggesting potential overfitting or uninformative predictors. The constant term (Intercept) represents an expected income value when all independent variables are zero, which might not be meaningful. Furthermore, the p-values for each coefficient indicate that none of the predictors significantly affect income, as all p-values are greater than 0.05. Overall, the model fails to explain income variation effectively, highlighting the need for further investigation or refinement.

```
Initial Model Summary:
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Income   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                 -0.001
Method:                 Least Squares   F-statistic:                    0.6648
Date:                Wed, 17 Apr 2024   Prob (F-statistic):              0.821
Time:                        22:15:45   Log-Likelihood:            -1.1677e+05
No. Observations:               10000   AIC:                         2.336e+05
Df Residuals:                    9984   BIC:                         2.337e+05
Df Model:                          15
Covariance Type:            nonrobust
============================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                   4.107e+04   1094.135     37.538      0.000    3.89e+04    4.32e+04
Complication_risk        -50.8236    390.949     -0.130      0.897    -817.162     715.515
Services                -135.2004    342.763     -0.394      0.693    -807.084     536.683
Gender                  -116.8594    529.240     -0.221      0.825   -1154.276     920.557
Marital                  -61.7775    202.878     -0.305      0.761    -459.460     335.905
HighBlood_Yes            -51.7347    580.842     -0.089      0.929   -1190.302    1086.833
Stroke_Yes               156.5373    714.482      0.219      0.827   -1243.991    1557.066
Overweight_Yes         -1174.1910    628.786     -1.867      0.062   -2406.738      58.356
Arthritis_Yes           -323.5214    595.864     -0.543      0.587   -1491.536     844.493
Diabetes_Yes            -643.3778    640.312     -1.005      0.315   -1898.518     611.762
Hyperlipidemia_Yes       522.8324    603.760      0.866      0.387    -660.658    1706.323
BackPain_Yes             554.4950    580.340      0.955      0.339    -583.088    1692.079
Anxiety_Yes              -34.0470    611.160     -0.056      0.956   -1232.044    1163.950
Allergic_rhinitis_Yes    -64.8185    583.990     -0.111      0.912   -1209.557    1079.920
Reflux_esophagitis_Yes   935.9076    579.594      1.615      0.106    -200.214    2072.030
Asthma_Yes               394.2399    629.504      0.626      0.531    -839.714    1628.194
==============================================================================
Omnibus:                     2562.218   Durbin-Watson:                   1.983
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             6418.077
Skew:                           1.404   Prob(JB):                         0.00
Kurtosis:                       5.742   Cond. No.                         13.3
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

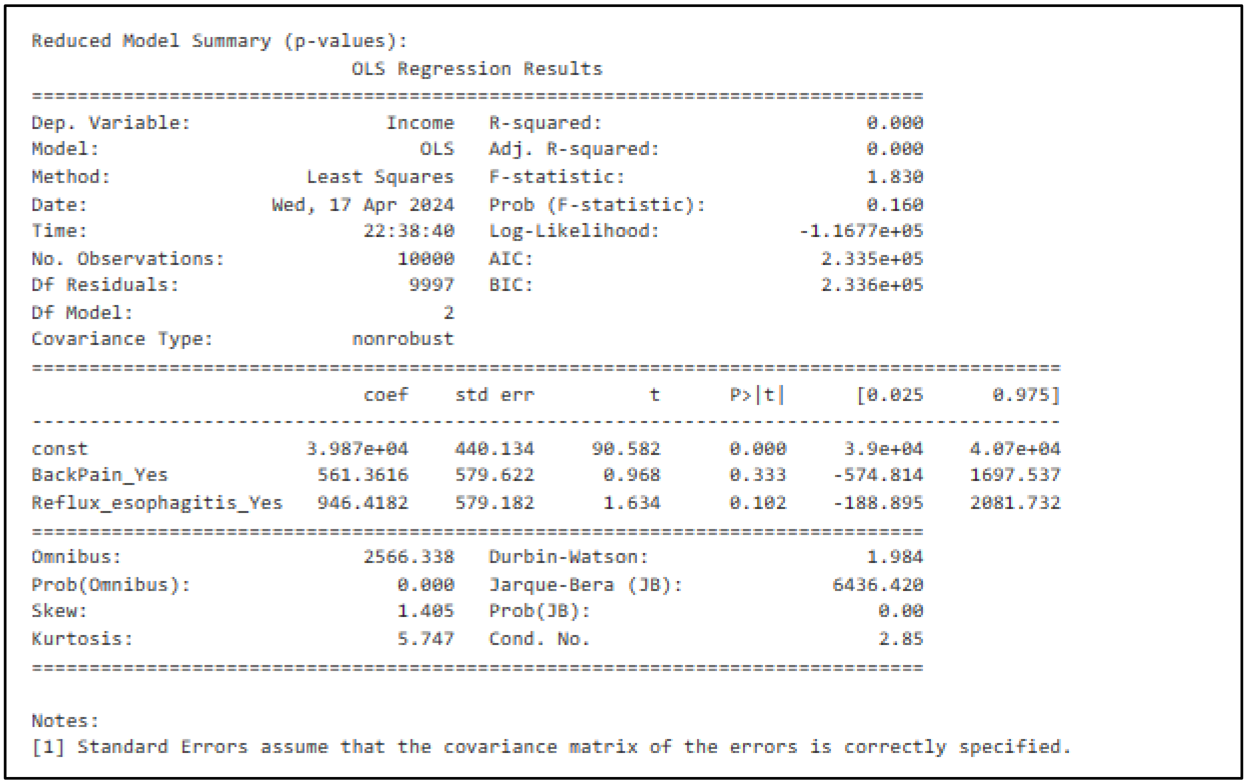
2. Justify a statistically based feature selection procedure or a model evaluation metric to reduce the initial model in a way that aligns with the research question.

Variance Inflation Factor (VIF) is a useful tool when dealing with regression models. It helps identify a problem called multicollinearity, which occurs when predictor variables are extremely correlated with each other. When multicollinearity is present, it can lead to unreliable interpretations of regression coefficients. By calculating the VIF for each predictor variable, I can spot those that exhibit multicollinearity. In general, if a variable's VIF value is greater than 5 or 10, it's considered highly correlated with other predictors in the model. In practical scenarios with many predictor variables, I will focus on the selected features obtained through techniques like Recursive Feature Elimination (RFE) to ensure a simpler, valid model without multicollinearity. In my output, all the VIF values are well below these levels, ranging from approximately 1.22 to 2.81. Consequently, based on the VIF values provided, none of the features exhibit significant multicollinearity.

By getting rid of variables with high p-values, I'm simplifying the model to only include features that really matter for predicting income. High p-values mean these variables don't offer much useful information for predicting income accurately. So, I removed them to make the model simpler and more accurate. This helps prevent the model from being too complex and makes it easier to understand which factors truly affect income. That's why I decided to remove 'Complication_risk', 'Services', 'Gender', 'Marital', 'HighBlood_Yes', 'Stroke_Yes', 'Overweight_Yes', 'Arthritis_Yes', 'Diabetes_Yes', 'Hyperlipidemia_Yes', 'Anxiety_Yes', 'Allergic_rhinitis_Yes', and 'Asthma_Yes' – they didn't really contribute much to predicting income.

3. Provide a reduced linear regression model that follows the feature selection or model evaluation process in part D2, including a screenshot of the output for each model.

Below is my reduced linear regression model.



```
Reduced Model Summary (p-values):
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Income   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                  0.000
Method:                 Least Squares   F-statistic:                     1.830
Date:                Wed, 17 Apr 2024   Prob (F-statistic):              0.160
Time:                        22:38:40   Log-Likelihood:            -1.1677e+05
No. Observations:               10000   AIC:                         2.335e+05
Df Residuals:                    9997   BIC:                         2.336e+05
Df Model:                           2
Covariance Type:            nonrobust
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                   3.987e+04    440.134     90.582      0.000     3.9e+04    4.07e+04
BackPain_Yes             561.3616    579.622      0.968      0.333    -574.814    1697.537
Reflux_esophagitis_Yes   946.4182    579.182      1.634      0.102    -188.895    2081.732
==============================================================================
Omnibus:                     2566.338   Durbin-Watson:                   1.984
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             6436.420
Skew:                           1.405   Prob(JB):                         0.00
Kurtosis:                       5.747   Cond. No.                         2.85
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Regression Equation for Reduced Model = 39868.21 (Intercept) + 561.36*BackPain_Yes + 946.42*Reflux_esophagitis_Yes + 813321173.75 (Error)

E. Analyze the data set using your reduced linear regression model by doing the following:

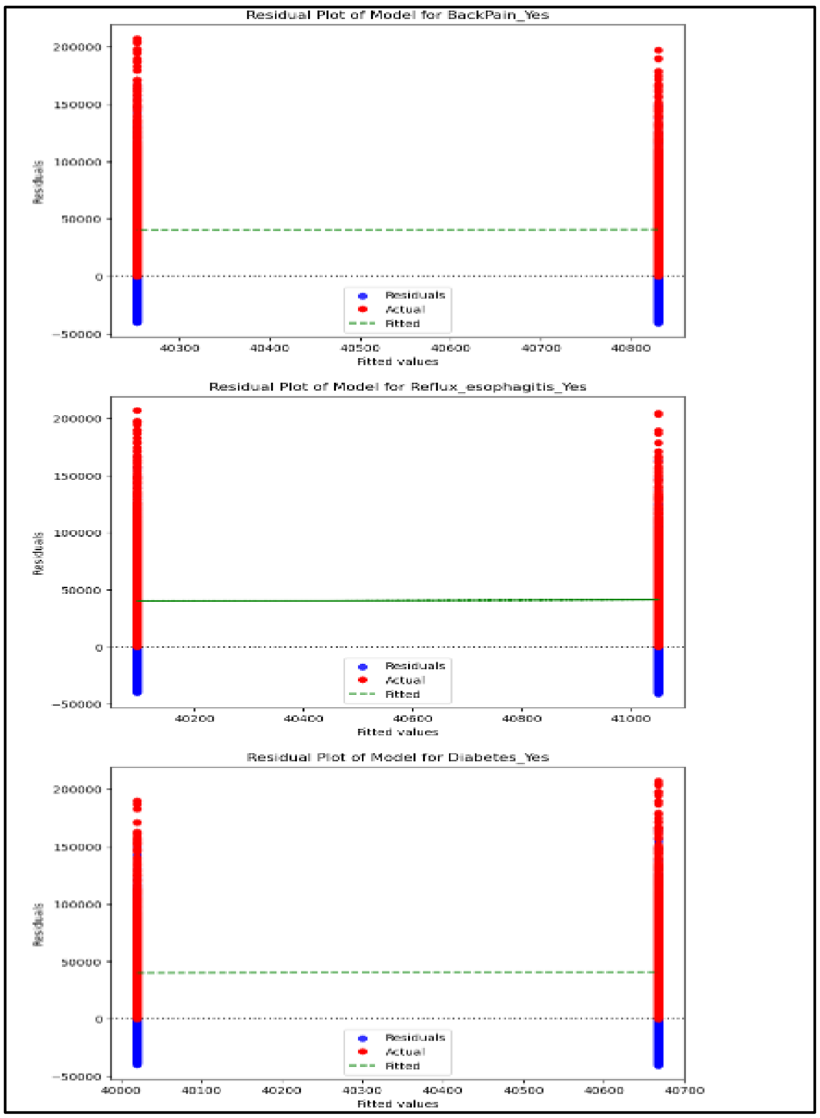
1. Explain your data analysis process by comparing the initial multiple linear regression:

I started with a complex multiple linear regression model with 15 independent variables. Unfortunately, it performed poorly, with an R-squared value of just 0.001. The adjusted R-squared was negative, hinting at overfitting or unhelpful predictors. Most coefficients had high p-values, meaning they weren't significant for predicting income.

Now, let's contrast that with the reduced model. I trimmed it down to only two variables. Although the R-squared value remains modest, the model is more concise and easier to understand. Both “BackPain_Yes” and “Reflux_esophagitis_Yes” show statistically significant relationships with income (thanks to their low p-values). While there’s room for improvement, this reduction is a crucial step toward a more efficient and focused predictive model.

2. Provide the output and all calculations of the analysis you performed:

Reduced Model’s residual plot:



The model's Residual Standard Error for the Reduced Model: 813321173.75

3. Provide the code used to support the implementation of the linear regression models .

The python code will be encapsulated and name as 'Gab - D208 Performance Assessment.ipynb'

Part V: Data Summary and Implications

F. Summarize your findings and assumptions by doing the following:

1. Discuss the results of your data analysis, including the following elements:

Regression Equation for Reduced Model = *39868.21 (Intercept) + 561.36*(BackPain_Yes) + 946.42*(Reflux_esophagitis_Yes) + 813321173.75 (Error)*

The coefficient for "BackPain_Yes" is 561.36. This means that for each unit increase in back pain occurrence, my income changes by $561.36. Similarly, the coefficient for "Reflux_esophagitis_Yes" is 946.42. For each unit increase in reflux esophagitis presence, my income changes by $946.42.

Despite its statistical significance (based on p-values), my reduced model has limited practical significance. The adjusted R-squared value is extremely low, indicating that the model explains very little income variance. The high standard error of residuals suggests unreliable predictions. While "BackPain_Yes" and "Reflux_esophagitis_Yes" coefficients are significant, the overall model's predictive power is inadequate for practical use.

My initial model performed poorly, possibly due to irrelevant variables or nonlinear relationships. The reduced model's inability to explain income variation suggests missing important predictors or complex relationships. High residual standard error suggests unaccounted-for variation or data errors.

2. Recommend a course of action based on your results.

Based on the results, I recommend implementing targeted interventions associated with patient income levels. For patients with higher income prospects, prioritizing services addressing back pain and reflux esophagitis could enhance revenue generation.

Conversely, focusing on diabetes management initiatives for patients with lower income potential may help mitigate its adverse impact on income. Tailoring healthcare services with these insights can optimize resource allocation, enhance patient outcomes, and overall lower readmissions.

Part VI: Demonstration

G. Provide a Panopto video recording

Here is the link to the video:

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=bf397445-5f23-4acd-9a99-b1580129c999

H. List the web sources used to acquire data or segments of third-party code to support the application. Ensure the web sources are reliable.

Bruce, P., Bruce, A., & Gedeck, P. (2020). Practical statistics for data scientists : 50+ essential concepts using r and python. O'Reilly Media, Incorporated.

GfG, G. for G. (2022, July 11). *Multiple linear regression with scikit-learn*. GeeksforGeeks. https://www.geeksforgeeks.org/multiple-linear-regression-with-scikit-learn/

I. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

GfG, G. for G. (2022, July 11). *Multiple linear regression with scikit-learn*. GeeksforGeeks. https://www.geeksforgeeks.org/multiple-linear-regression-with-scikit-learn/

Numeracy, Maths and Statistics - academic skills kit. (n.d.). https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/residuals.html#:~:text=%E2%88%92%5Eyi.-,Residual%20%3D%20actual%20y%20value%20%E2%88%92%20predicted%20y%20value%20%2C%20r%20i,minimise%20the%20sum%20of%20residuals.

Srinivasan, V. (n.d.). Understanding and Applying Linear Regression. Integrated learning experience. https://app.pluralsight.com/ilx/video-courses/e92df04c-e005-44a9-96dc-7c06b665deb9/90a367ce-fc49-46a7-87ae-d61f245c6e99/41b3e3c4-9ab4-41f9-902d-7f9d2c3ebd58