Megan Anderson

Professor Jaeggli

Bioengineering Measurement, Engineering, and Statistics - Section 03

15 November 2024

Project 2: Linear Regressions to Investigate Causes of Cardiovascular Disease

Part 1 - Simple Linear Regression

1a. Exploratory Analysis into the Relationship between BMI and Systolic Blood Pressure

The assumptions of linear regression are linearity, independence of data, homoscedasticity, and normality of residuals. In order to determine whether a direct relationship between BMI and Systolic Blood Pressure (SBP) or a relationship between the log(BMI) and the log(SBP) met the assumptions of a linear regression better, these four assumptions were tested. Firstly, the independence of the data was ensured by sorting the data by period. The dataset provided included up to three observations of each measurement for each subject studied. However, these observations of the same individual do not meet the criteria of independence, as the data points are influenced by each other. In order to ensure independence, the data was sorted by observation number (referred to as period in this data set), and only data points from period 1 were studied in this analysis. Independence can not be used to determine which relationship (raw data or natural log of the data) better fits a linear regression, because after sorting the data both meet this assumption to the same degree.

Next, linearity and homoscedasticity were analyzed with a plot of the residuals of a linear fit vs. the fitted values. A simple linear model was created using the ‘fitlm’ function in MATLAB of both the raw data and the natural log of the data. This linear model predicted SBP values based on BMI x-values. A set of values that fit these linear models were created and the residuals were calculated by the equation

$$residuals\ =\ observed\ data - fitted\ data$$

Where the observed data is the raw data points or the natural log of the raw data points respectively. In order to analyze the linearity and homoscedasticity of the two relationships, the residuals were plotted against the fitted values, resulting in Figures 1 and 2 below. Based on these plots (Figures 1 and 2 below), the raw data better fits the assumptions of linear regression. This is because linearity and homoscedasticity are met by data that is symmetrically around the residual line y = 0. As seen in the figures, the red horizontal line represents the residual line y = 0 and the raw data in Figure 1 is relatively evenly distributed around this line. However, the log(data) in Figure 2 is not evenly distributed around the residual = 0 line and rather is distributed entirely above it.

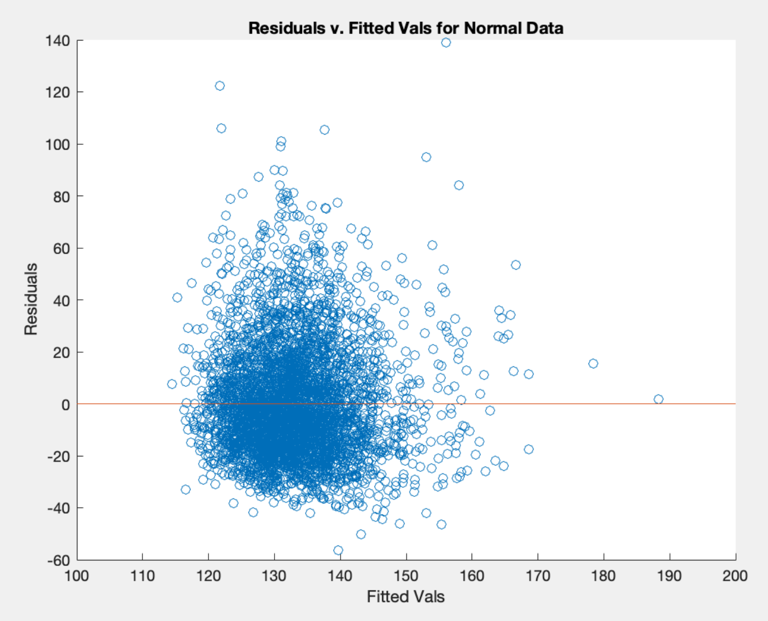

Figure 1: *Plot of Residuals vs. Fitted Values for a linear model of the Raw BMI and SYB data.*

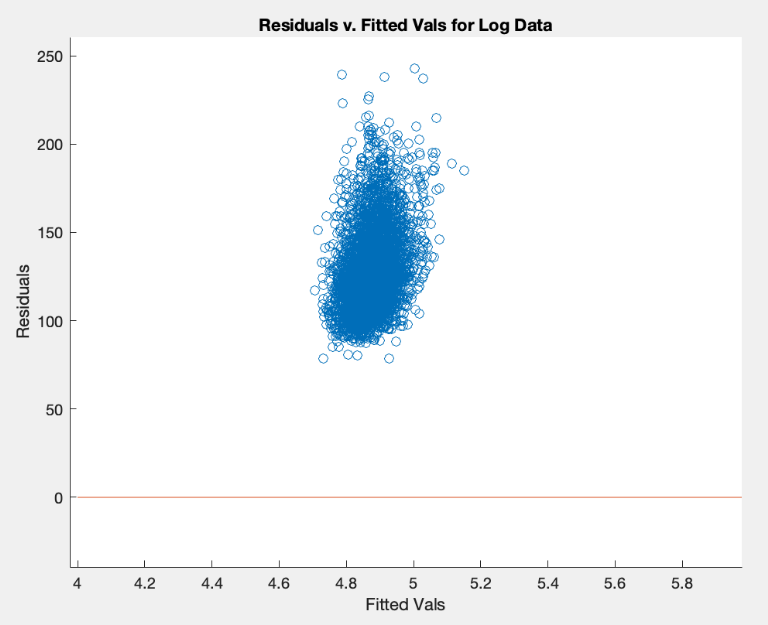

Figure 2: *Plot of Residuals vs. Fitted Values for a linear model of the log(BMI) and log(SYB) data.*

The last assumption of a linear regression to analyze is the normality of residuals. This was analyzed by creating a normal probability plot for each of the previously calculated residuals (raw data and log(data)), resulting in Figure 3 below.

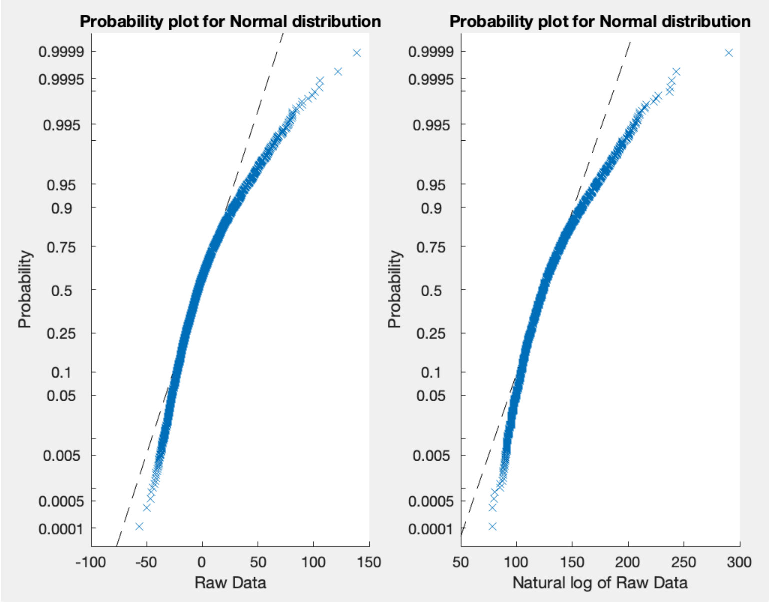

Figure 3: *Figure demonstrating the normal probability plots of both the raw and log data. The left subplot shows the probability plot of the raw data. The right subplot shows the probability plot of the log(data).*

These normal probability plots indicate that the residuals of both the raw data and log(data) linear models generally follow a normal distribution. Furthermore, neither follows the normal distribution better than the other.

Based on the analysis of each linear regression assumption, the raw data demonstrates the assumptions of a linear regression better than the natural log of the raw data does. The analysis of independence and normality of residuals showed equal results for both sets of data. However, the raw data set better demonstrated the assumptions of linearity and homoscedasticity. Therefore, the raw data will be used to create future linear models.

1b. Creation of Linear Regressions

Two simple linear regression models were created, one for men and one for women. These models were both created using the 'fitlm' function in MATLAB and a set of values that fit the resulting model was fabricated. The estimated regression coefficients of the resulting models are displayed in Table 1 below. x1 represents the slope of the linear regression, or how the SBP values are predicted to change in response to changing BMI. Meanwhile, the intercept represents the predicted SBP value at a BMI of 0. A BMI of 0 is illogical in reality, and therefore the intercepts don't hold much meaning in this circumstance of linear regression. However, the slopes (x1 values) of these regressions indicate that the SBP of women has greater increases than that of men for the same increase in BMI. This is concluded since the x1 of the women model is greater than that of the men model.

Linear Regression Model for Men		Linear Regression Model for Women	
Intercept	x1	Intercept	x1
99.176	1.2428	80.956	2.0659

Table 1*: Estimated regression coefficients calculated by MATLAB's 'fitlm' function for the two linear regression models (men and women).*

In order to analyze the goodness of fit, the R^2 values were compared. These values are displayed in Table 2 below. An R^2 of 1 represents a perfect linear fit, and thus the closer an R^2 value is to one, the better it fits a linear regression. The R^2 values calculated from these linear regressions are both not close to 1, indicating an insufficient goodness of fit. The R^2 value of the women's linear regression is higher than that of the men, indicating it has a better fit to the linear model. However, both are relatively low, indicating that neither data set fits a simple linear regression well.

R^2 of Linear Regression Model for Men	R^2 of Linear Regression Model for Women
0.0482	0.148

Table 2*: Estimated R^2 values for the two linear regression models (men and women) used to determine goodness of fit.*

<u>*1c. Vizualization of Linear Regressions*</u>

Both linear regressions were visualized using a scatter plot with the linear regression fit line superimposed on the graph. These plots are demonstrated in Figure 4 below.

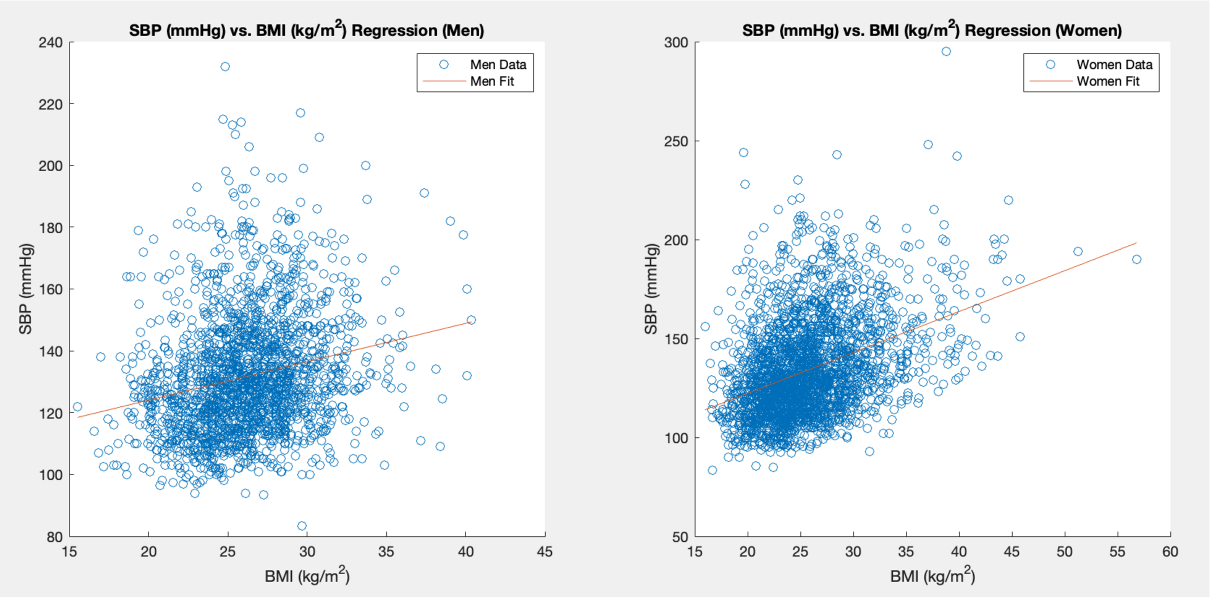

Figure 4: *Figure demonstrating the linear regression of both the men's and women's data sets.*

Based on these visualizations, it appears the women's data set better fits the linear regression as the data points are more closely distributed to the line representing the fit. This visualization and conclusion reinforce the previous analysis that the women's data better fits the linear regression since the R^2 value is higher. However, as concluded from the R^2 values these graphical visualizations also reinforce the conclusion that neither regression fits linearly particularly well. This is demonstrated by the fact that in both plots, the data points stray from the line of regression significantly.

1d. Analysis of Assumptions

These linear regression models follow the same assumptions as described in part 1a, linearity, independence of data, homoscedasticity, and normality of residuals. Since the men and women data sets are a subset of the original data set analyzed in part 1a, the independence of these data sets can be inferred. To analyze the linearity and homoscedasticity, a plot of the residuals vs. the fitted values was constructed as described before, resulting in Figures 5 and 6 below. In the same figures, (Figures 5 and 6 below), the normality of the residuals was also analyzed using a normal probability plot.

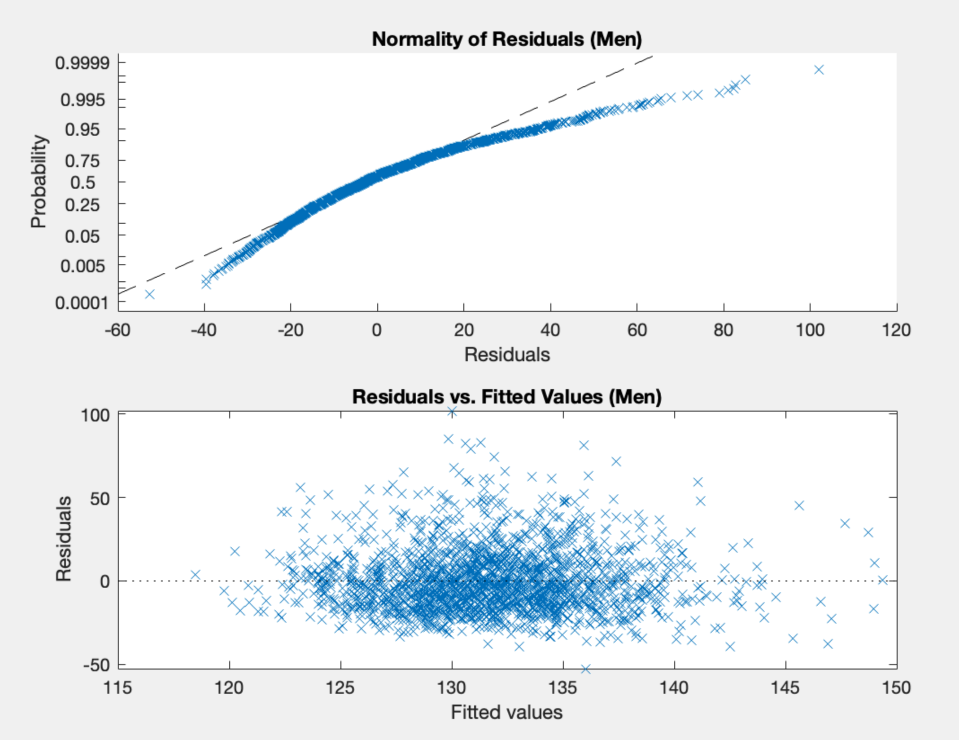

Figure 5: *Figure demonstrating the linearity, homoscedasticity, and normality of residuals of men's data set. The top subplot shows a normal probability plot of the residuals of the men's data set. The bottom subplot shows Residuals vs. Fitted Values for a linear model of this data.*

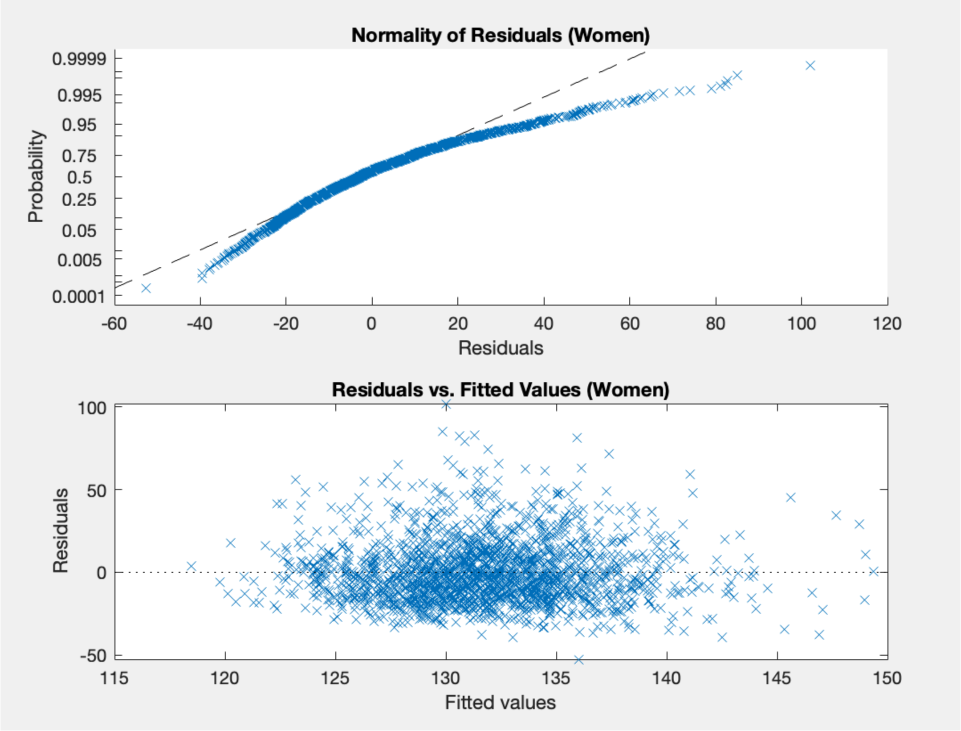

Figure 6: *Figure demonstrating the linearity, homoscedasticity, and normality of residuals of women's data set. The top subplot shows a normal probability plot of the residuals of the women's data set. The bottom subplot shows Residuals vs. Fitted Values for a linear model of this data.*

Based on Figures 5 and 6, both the men's and women's data sets meet the assumptions of linear regression. Both normal probability plots demonstrate a generally good fit along the line of normal distribution, demonstrating the normality of residuals. Additionally, both plots of Residuals vs. Fitted Values are generally symmetrically distributed around the residual line of 0, demonstrating linearity and homoscedasticity. Therefore, it can be concluded that the assumptions of linear regression are met for both the men's and women's data sets.

1e. Predictions Based on Linear Regression

The linear regressions created in part 1b were used to predict the SBP levels of a man and woman with a BMI of 33 kg/m^2. The resulting values are displayed in Table 3 below.

Estimated SBP for a Man with a BMI of 33 kg/m^2	Estimated SBP for a Woman with a BMI 33 of kg/m^2
140.1897 mmHg	0. 149.1291 mmHg

Table 3*: Predicted SBP values (in mmHg) for a man and woman of BMI 33 kg/m^2 based on the linear regression models created.*

Part 2 - Multiple Regression

2a. Development of Multiple Regression to Predict SBP

In order to develop a multiple regression to predict SBP, four predictors were chosen that likely have some sort of impact on SBP. The four predictors chosen were age, BMI, glucose levels, and total cholesterol levels. These predictors were then used to create two multiple regressions using the 'fitlm' function in MATLAB, one for men's data and one for women's. The resulting regression coefficients are displayed in Table 4 below. These regressions indicate how the SBP values will change in response to a unit change in predictor respectively.

Multiple Regression Model for Men					Multiple Regression Model for Women				
Intercept	Age	BMI	Glucose	Total Cholesterol	Intercept	Age	BMI	Glucose	Total Cholesterol
61.945	0.55832	1.2284	0.025271	0.03271	25.662	1.125	1.5188	0.074565	0.029972

Table 4*: Estimated regression coefficients calculated by MATLAB's 'fitlm' function for the two multiple regression models (men and women).*

In order to analyze the goodness of fit, the R^2 values were compared. These values are displayed in Table 5 below. As previously described, an R^2 of 1 represents a perfect linear fit, and thus the closer a R^2 value is to one, the better it fits a linear regression. The R^2 values calculated from these linear regressions are both not close to 1, indicating an insufficient goodness of fit. The R^2 value of the women's linear regression is higher than that of the men, indicating it has a better fit to the linear model. Additionally, each R^2 value in these multiple regressions is higher than the corresponding R^2 value from the simple linear regressions in part 1b, indicating that the multiple regression is a better predictor of SBP than the simple linear regression. However, both R^2 are still relatively low, indicating that neither data set fits a simple linear regression well.

R^2 of Multiple Regression Model for Men	R^2 of Multiple Regression Model for Women
0.121	0.329

Table 5*: Estimated R^2 values for the two multiple regression models (men and women) used to determine goodness of fit.*

2b. Vizualization of Multiple Regressions

Both multiple regressions were visualized using the ‘plotmatrix’ function in MATLAB which outputs a matrix of sub-axes containing scatter plots of the columns of the predictor values against the columns of SBP values. These plots are demonstrated in Figures 7 and 8 below.

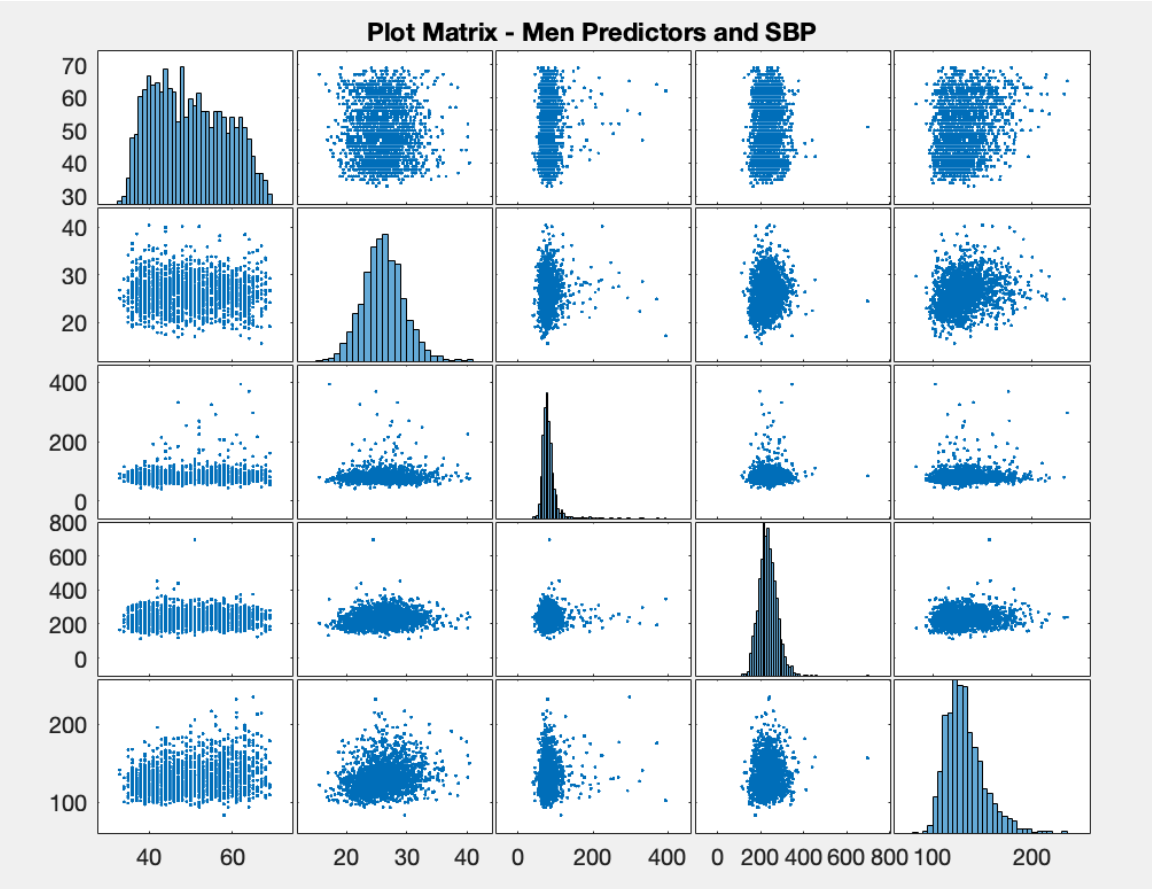

Figure 7: *Plot matrix of men’s multiple regression where the predictor (x) values are age, BMI, glucose levels, and cholesterol levels.*

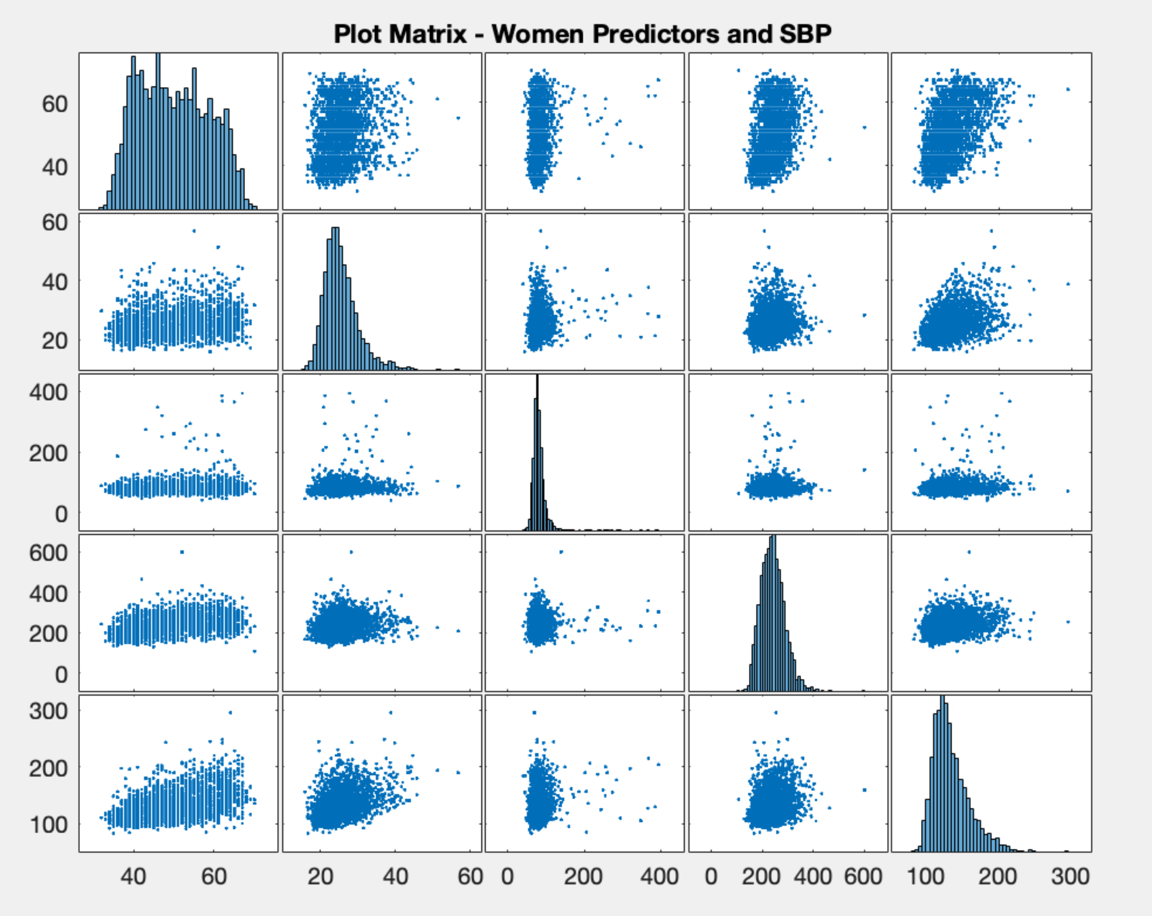

Figure 8: *Plot matrix of women's multiple regression where the predictor (x) values are age, BMI, glucose levels, and cholesterol levels.*

The histogram in the top row of plots represents the distribution of age in the data, and the scatter plots in this row represent age compared to the other predictor and SBP values. The histogram in the second row of plots represents the distribution of BMI in the data, and the scatter plots in this row represent BMI compared to the other predictor and SBP values. The histogram in the third row of plots represents the distribution of glucose levels in the data, and the scatter plots in this row represent glucose levels compared to the other predictor and SBP values. The histogram in the fourth row of plots represents the distribution of total cholesterol levels in the data, and the scatter plots in this row represent total cholesterol compared to the other predictor and SBP values. Lastly, the histogram in the last row of plots represents the distribution of SBP in the data, and the scatter plots in this row represent SBP compared to the predictor values.

2c. Analysis of Assumptions

These multiple regression models follow the same assumptions as described in part 1a, linearity, independence of data, homoscedasticity, and normality of residuals. Since the men and women data sets are a subset of the original data set analyzed in part 1a, the independence of these data sets can be inferred. To analyze the linearity and homoscedasticity, a plot of the residuals vs. the fitted values was constructed as described before, resulting in Figures 8 and 9 below. In the same figures, (Figures 8 and 9 below), the normality of the residuals was also analyzed using a normal probability plot.

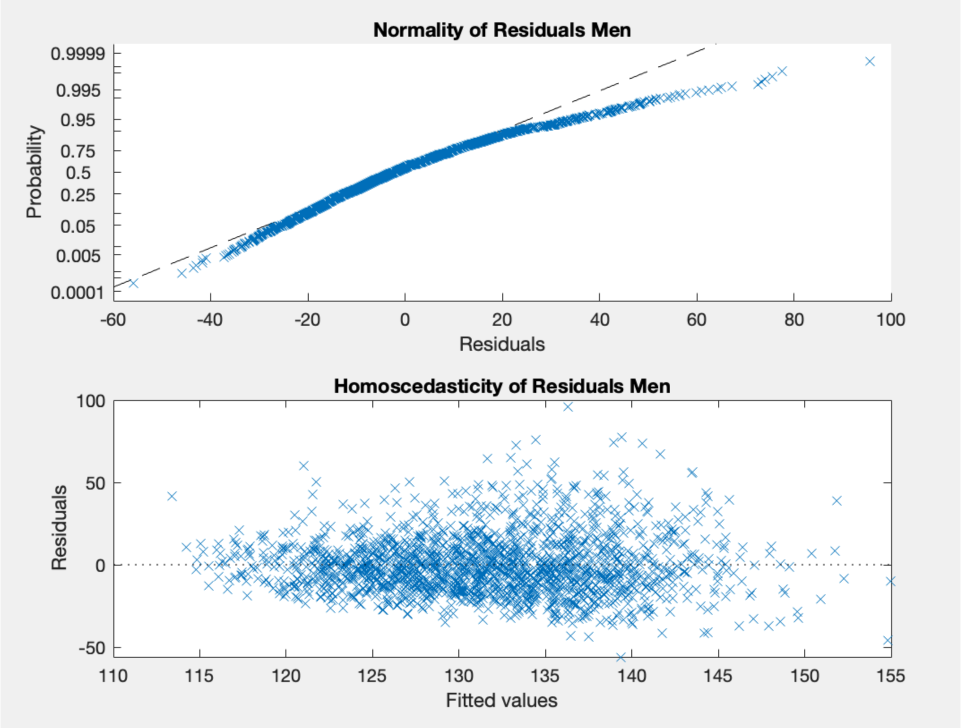

Figure 8: *Figure demonstrating the linearity, homoscedasticity, and normality of residuals of men's data set. The top subplot shows a normal probability plot of the residuals of the men's data set. The bottom subplot shows Residuals vs. Fitted Values for a linear model of this data.*

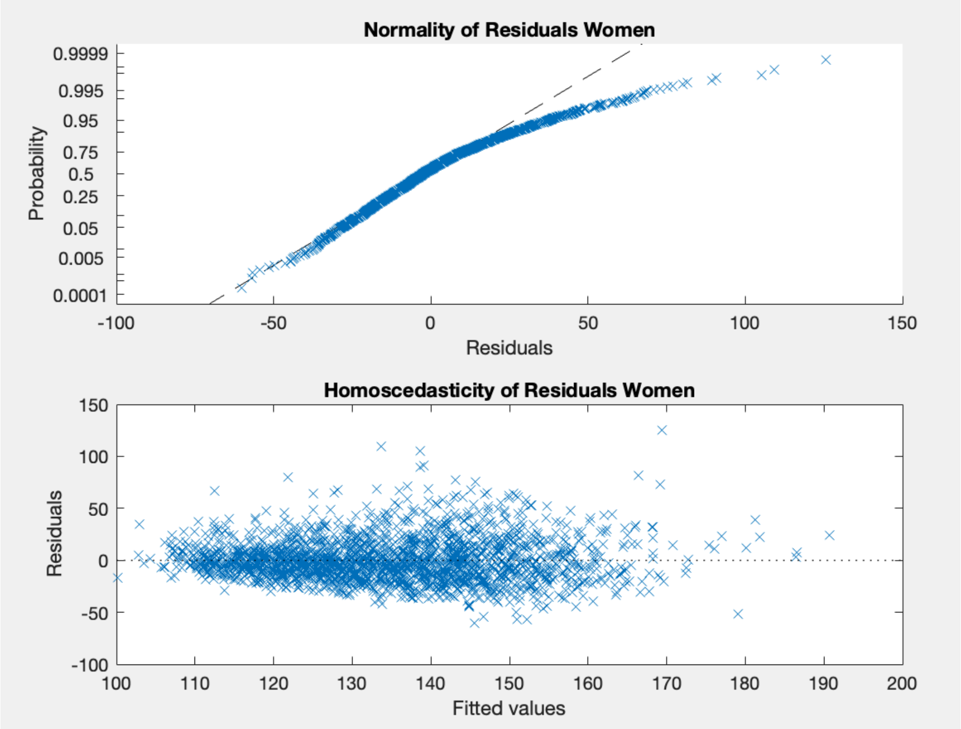

Figure 9: *Figure demonstrating the linearity, homoscedasticity, and normality of residuals of women's data set. The top subplot shows a normal probability plot of the residuals of the women's data set. The bottom subplot shows Residuals vs. Fitted Values for a linear model of this data.*

Based on Figures 8 and 9, both the men's and women's data sets meet the assumptions of multiple linear regression. Both normal probability plots demonstrate a generally good fit along the line of normal distribution, demonstrating the normality of residuals. Additionally, both plots of Residuals vs. Fitted Values are generally symmetrically distributed around the residual line of 0, demonstrating linearity and homoscedasticity. Therefore, it can be concluded that the assumptions of linear regression are met for both the men's and women's data sets.