STAT 481 PROJECT REPORT

IDENTIFYING KEY PREDICTORS OF DIABETES PROGRESSION USING MULTIPLE LINEAR REGRESSION PROJECT

BADER REZEK

STAT 481 – APPLIED STATISTICAL METHODS II

PROFESSOR DR. WANG JING

DATE: 04/02/2025

INTRODUCTION

This project's main goal is to use Multiple Linear Regression to explain the progression of diabetes in patients one year after baseline. The dataset contains 442 observations and has ten predictor variables including age, sex, body mass index, blood pressure, and six blood serum measurements. The response variable is a measure of disease progression. By doing this analysis, our goal is to identify the significant predictors that lead to disease progression among patients. The analysis will be done using R, and results will be interpreted to understand the influence of each predictor on the progression of diabetes.

1. DATA INTRODUCTION AND DESCRIPTION

AGE:

To start, I created a histogram of age [Figure 1.1] so that I could see its distribution and get a visual of the five-number summary. The five-number summary output was Min: 19.00, 1st Qu :38.25, Median :50.00, Mean :48.52, 3rd Quarter: 59.00, and Max. :79.00. When I plotted this predictor, I noticed a relatively normal distribution as well.

SEX:

In the original dataset, SEX was coded as 1 or 2, where 1 represented Male and 2 represented Female. I transformed this variable into a factor using these labels, so that R could treat it as a categorical variable and give me easy to interpret results. A bar plot [Figure 1.2] shows that there are more Male patients than Female patients in the dataset. This transformation also ensures that "Male" is used as the baseline level in the regression model, making it easier to interpret the effect of being Female.

BMI:

I created a histogram of BMI [Figure 1.3] so that I could understand its distribution. In it, I saw a relatively right-skewed plot. This tells me that I may have to transform this predictor in the future to accommodate with our model selection properties. The five-number summary output was Min :18.00, 1st Quarter: 23.20, Median :25.70, Mean :26.38, 3rd Quarter: 29.27, and Max. :42.20.

BP:

In the histogram [Figure 1.4], I saw a relatively right-skewed plot. This tells me that I may have to transform this predictor in the future to accommodate with our model selection properties. The five-number summary output was Min.: 62.00, 1st Quarter: 84.00, Median: 93.00, Mean: 94.65, 3rd Quarter: 105.00 and Max. :133.00.

S1 - SERUM TOTAL CHOLESTEROL LEVEL:

In the histogram [Figure 1.5], I saw a relatively normal distribution. The five-number summary output was Min: 97.0, 1st Quarter: 164.2, Median: 186.0, Mean: 189.1, 3rd Quarter: 209.8, and Max: 301.0. When I plotted the standard deviation [Figure 1.11], this predictor was the highest in terms of variability. This tells us that this predictor can be influential because it varies so much across patients.

S2 - LOW-DENSITY LIPOPROTEINS (LDL):

In this histogram [Figure 1.6], I saw a relatively right-skewed plot. This tells me that I may have to transform this predictor in the future to accommodate with our model selection properties. The five-number summary output was Min: 41.60, 1st Qu.: 96.05, Median :113.00, Mean :115.44, 3rd Qu.:134.50, and Max. :242.40. When I plotted the standard deviation [Figure 1.11], this predictor was the second highest in terms of variability. This tells us that this predictor can be influential because it varies so much across patients.

S3 - High-density lipoproteins (HDL):

In this histogram [Figure 1.7], I saw a relatively right-skewed plot. This tells me that I may have to transform this predictor in the future to accommodate with our model selection properties. The five-number summary output was Min.   :22.00, 1st Qu.:40.25, Median :48.00, Mean   :49.79, 3rd Qu.:57.75, Max.   :99.00.

S4 - Total cholesterol / HDL ratio:

In this histogram [Figure 1.8], I saw a extremely right-skewed plot with the peak of this histogram being on the Y axis. This tells me that I may have to transform this predictor in the future to accommodate with our model selection properties. The five-number summary output was Min.   :2.00, 1st Qu.:3.00, Median :4.00, Mean   :4.07, 3rd Qu.:5.00, and Max.   :9.09. When I plotted the standard deviation [Figure 1.11], this predictor was the second lowest in terms of variability. This tells us that this predictor may not be influential because it doesn't seem to be contributing much information.

S5 - Serum triglyceride level:

In this histogram [Figure 1.9], I saw a relatively normal distribution. The five-number summary output was Min.   :3.258, 1st Qu.:4.277, Median :4.620, Mean   :4.641, 3rd Qu.:4.997, and Max.   :6.107. When I plotted the standard deviation [Figure Below], this predictor was the lowest in terms of variability. This tells us that this predictor may not be influential because it doesn't seem to be contributing much information.

S6 - Blood sugar level:

In this histogram [Figure 1.10], I saw a relatively normal distribution. The five-number summary output was Min.   : 58.00, 1st Qu.: 83.25 , Median : 91.00 , Mean   : 91.26 , 3rd Qu.: 98.00 , and Max.   :124.00.

FIGURE 1.1

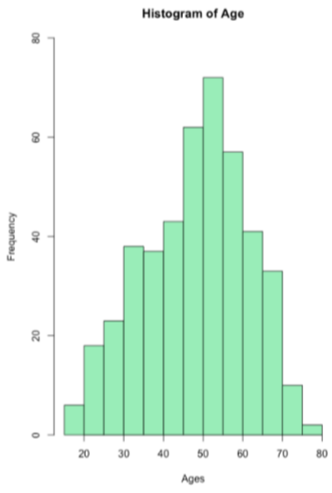


FIGURE 1.2

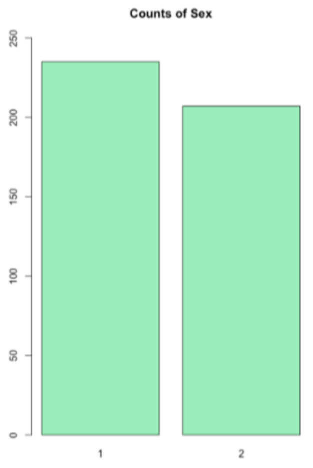


FIGURE 1.8

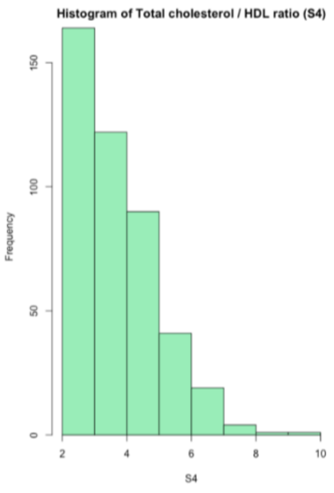
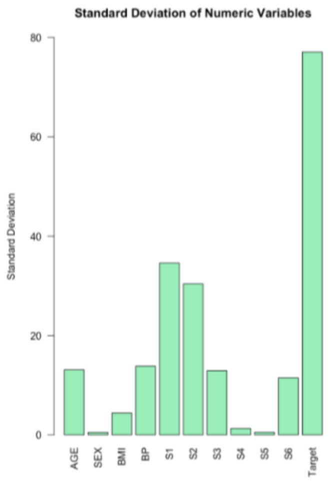


FIGURE 1.11



Additional plots are available in the Appendix.

2. Model Selection and Fitting

2.1 – Preliminary Hypothesis:

To start I tested if the key linear regression assumptions were met. Specifically, I tested normality of residuals, constant variance (homoscedasticity), and independence of residuals. Although independence is primarily a concern for time-series data, it was included here for completeness.

Normality Check	Equal Variance Check	Independence Check
$H_0: \varepsilon_i \sim N(0, \sigma^2)$ $H_1: \varepsilon_i \notin N(0, \sigma^2)$	$H_0: Var(\varepsilon_i) = \sigma^2$ $H_1: Var(\varepsilon_i) \neq \sigma^2$	$H_0: Cov(\varepsilon_i, \varepsilon_{i-1}) = 0$ $H_1: Cov(\varepsilon_i, \varepsilon_{i-1}) \neq 0$
To check normality, I created a Q-Q plot of the residuals [Figure 2.1]. The points follow the reference line, which indicates approximate normality. I also wanted to be sure, so I conducted a Shapiro test which gave me a P-Value of 0.612 which indicates normality. Thus, we do not reject the null hypothesis that our data is not normally distributed.	Constant variance was tested using a residuals vs. fitted values plot [Figure 2.2]. There was no clear pattern in this plot. Thus we do not reject the null hypothesis. The homoscedasticity assumption is met.	Finally, I plotted a residuals vs. lagged residuals plot (e_i vs e_{i-1}) [Figure 2.3] was created to assess independence. There is no pattern or trend which tells us that the independence assumption is met. Because of this, we do not reject the null hypothesis that claims we have independence of Errors.

Linearity

To determine if the relationship between each predictor and the response is linear, I plotted the residuals against each predictor individually. In a linear model, these plots should show no curves or patterns. I did not see any curvature within any of these plots, and they were for the most part random. This tells us that these predictors meet the linearity constraint. These plots are included in Appendix

Together, these diagnostics support the validity of the linear regression model assumptions.

2.2 – Primary Hypothesis:

This Hypothesis tests if none of the predictors are significant.

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_{10} = 0$$
$$H_1: \text{At least one predictor } \neq 0$$
$$(\text{is significant})$$

To test this hypothesis, I used the F-test with the full model. Based off of the model summary, the P-Value was < 2.2e-16. Since the p-value was less than 0.05, we reject the null hypothesis and conclude that the model explains a significant amount of variation in disease progression.

2.3 – Post Test Hypothesis':

To test each predictors significance on the model, I used individual t-tests. Predictors with p-values below 0.05 were considered statistically significant contributors to the model.

AGE	$H_0: \beta_1 = 0$ $H_1: \beta_1 \neq 0$	P-value: 0.867031 Decision/ Conclusion: Since the p-value is greater than 0.05, we do not reject the null hypothesis and conclude that the predictor Age does not explain a significant amount of variation in disease progression.
SEX	$H_0: \beta_2 = 0$ $H_1: \beta_2 \neq 0$	P-value: 0.000104 Decision/ Conclusion: Since the p-value is less than 0.05, we reject the null hypothesis and conclude that the predictor Sex explains a significant amount of variation in disease progression.
BMI	$H_0: \beta_3 = 0$ $H_1: \beta_3 \neq 0$	P-value: 4.30e-14 Decision/ Conclusion: Since the p-value is less than 0.05, we reject the null hypothesis and conclude that the predictor BMI explains a significant amount of variation in disease progression.
BP	$H_0: \beta_4 = 0$ $H_1: \beta_4 \neq 0$	P-value: 1.02e-06 Decision/ Conclusion: Since the p-value is less than 0.05, we reject the null hypothesis and conclude that the predictor for BP explains a significant amount of variation in disease progression.
S1	$H_0: \beta_5 = 0$ $H_1: \beta_5 \neq 0$	P-value: 0.057948 Decision/ Conclusion: Since the p-value is greater than 0.05, we do not reject the null hypothesis and conclude that the predictor for Serum total cholesterol level does not explain a significant amount of variation in disease progression.
S2	$H_0: \beta_6 = 0$ $H_1: \beta_6 \neq 0$	P-value: 0.160390 Decision/ Conclusion: Since the p-value is greater than 0.05, we do not reject the null hypothesis and conclude that the predictor for Low-density lipoproteins (LDL) does not explain a significant amount of variation in disease progression.
S3	$H_0: \beta_7 = 0$ $H_1: \beta_7 \neq 0$	P-value: 0.634723 Decision/ Conclusion: Since the p-value is greater than 0.05, we do not reject the null hypothesis and conclude that the predictor for High-density lipoproteins (HDL) does not explain a significant amount of variation in disease progression
S4	$H_0: \beta_8 = 0$ $H_1: \beta_8 \neq 0$	P-value: 0.273459 Decision/ Conclusion: Since the p-value is greater than 0.05, we do not reject the null hypothesis and conclude that the predictor for Total cholesterol/ HDL ratio does not explain a significant amount of variation in disease progression
S5	$H_0: \beta_9 = 0$ $H_1: \beta_9 \neq 0$	P-value: 1.56e-05 Decision/ Conclusion: Since the p-value is less than 0.05, we reject the null hypothesis and conclude that the predictor for Serum triglyceride level explains a significant amount of variation in disease progression.
S6	$H_0: \beta_{10} = 0$ $H_1: \beta_{10} \neq 0$	P-value: 0.305990 Decision/ Conclusion: Since the p-value is greater than 0.05, we do not reject the null hypothesis and conclude that the predictor for Blood sugar levels does not explain a significant amount of variation in disease progression.

2.4 - MULTICOLLINEARITY CHECK:

I checked for multicollinearity by calculating the Variance Inflation Factor (VIF) for each predictor. VIF values greater than 10 indicate multicollinearity.

	Age	Sex	BMI	BP	S1	S2	S3	S4	S5	S6
VIF	1.217	1.278	1.509	1.459	59.203	39.193	15.402	8.891	10.076	1.485

From this, test we learn that the predictors S1, S2, S3, and S5 contain multicollinearity.

3. Variable Selection

3.1 – Variable Selection:

Using AIC:

Backward Selection	**Procedure**: I started with the full model including all ten predictors. I then removed variables one at a time based on which removal led to the greatest improvement in AIC. The process continued until no further improvement could be made by removing predictors. **Variables Chosen**: BMI, S5, BP, S1, SEX, S2 **Residual Standard Error**: 54.06 **Adjusted $\mathbf{R^2}$**: 0.5082 **AIC**: 4790.603
Forward Selection	**Procedure**: I started with a model which only contained the intercept (β_0). I then added predictors one by one that resulted in the largest reduction in AIC. At each step, the variable that most improved the model was added, continuing until I couldn't lower AIC anymore. **Variables Chosen**: BMI, S5, BP, S1, SEX, S2 **Residual Standard Error**: 54.06 **Adjusted $\mathbf{R^2}$**: 0.5082 **AIC**: 4790.603
Stepwise Selection	**Procedure**: In stepwise selection, both forward and backward strategies were applied step by step. I either added or removed predictors at each step based on which change most improved the model according to AIC. **Variables Chosen**: BMI, S5, BP, S1, SEX, S2. **Residual Standard Error**: 54.06 **Adjusted $\mathbf{R^2}$**: 0.5082 **AIC**: 4790.603

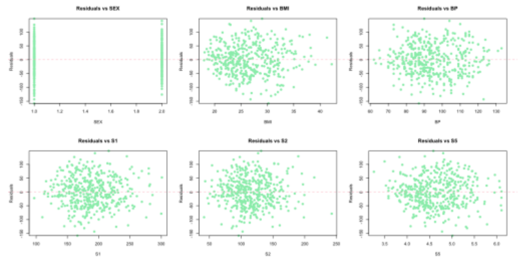
Using P-Value

Backward Selection	**Procedure**: I started with the full model including all ten predictors. I then removed the least significant variable (p>0.1) and looked at the summary of the new model to choose the next variable to remove. The process continued until no more variables could be removed. **Variables Chosen**: BMI, S5, BP, S1, SEX, S2 **Residual Standard Error**: 54.06 **Adjusted $\mathbf{R^2}$**: 0.5082
Forward Selection	**Procedure**: I started with a model which only contained the intercept (β_0). I then added predictors one by one that gave the smallest p-value in the model. I only added it if the p-value was below 0.10. **Variables Chosen**: BMI, S5, BP, S1, SEX, S2 **Residual Standard Error**: 54.06 **Adjusted $\mathbf{R^2}$**: 0.5082

After running these procedures, I interestingly got the same output each time. This tells us that the data contains predictors that are significant enough to be chosen by each method, and that the data used is very good as well. The chosen variables were BMI, S5, BP, S1, SEX, S2. We also got a common Residual Standard error of 54.06, and a AIC of 4790.603. The Adjusted R^2 for each of these procedures was 0.5082 which tells us that approximately 50% of the variability in disease progression is explained by the model.

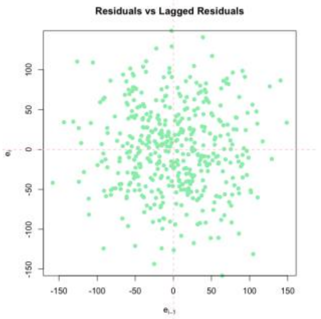
3.2 – New Model Assumptions:

Linearity: Residuals vs Each Predictor



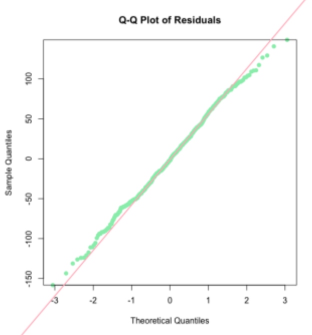
Residuals were plotted against each predictor in the final model (SEX, BMI, BP, S1, S2, and S5). Each plot showed a generally random scatter around zero, with no strong curvature or any patterns/ trends.

Independence: Residuals vs Lagged Residuals



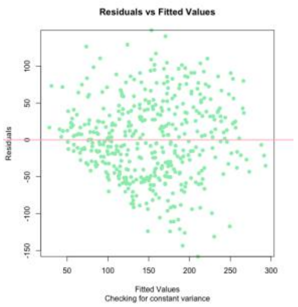
I plotted residuals and the lagged residuals (e_{i-1}) to test independence. The points are randomly scattered with no obvious trend or pattern, which supports the independent errors assumption.

Normality: Q-Q Plot and Shapiro-Wilk Test



The Q-Q plot of the residuals shows the points following the reference line, with only slight deviations at the tails. This suggests the residuals are approximately normally distributed. Additionally, the Shapiro-Wilk test gave us a p-value of 0.6719, so there's no significant evidence against normality.

Equal Variance (Homoscedasticity): Residuals vs Fitted



I plotted residuals and fitted values to test the equal variance assumption. In the plot we see no pattern, funneling, or increasing/decreasing variance. This plot supports the assumption of constant variance (homoscedasticity).

After testing these model assumptions on the new model containing SEX, BMI, BP, S1, S2, and S5, we can safely say that we meet all of the L.I.N.E standards.

4. Conclusions

Our goal was to create a better model to predict disease progression one year after baseline. Initially we started with 10 predictors in our full model. Those predictors include age, sex, body mass index, blood pressure, and six blood serum measurements (Serum total cholesterol level, Low-density lipoproteins (LDL), High-density lipoproteins (HDL), Total cholesterol/ HDL ratio, Serum triglyceride level, and Blood Sugar Level). After using backwards selection with AIC and with the P-value, we were left with 6 final predictors. They include sex, body mass index, blood pressure, Serum total cholesterol level, Low-density lipoproteins (LDL), and Serum triglyceride level.

Sex was a significant categorical predictor. Compared to Male patients, being Female was associated with an average decrease of approximately 21.59 units in disease progression, holding all other predictors constant. BMI and S5 (serum triglyceride level) had the largest positive coefficients (5.711107, and 73.306526 respectively) and smallest p-values (6.687185e-15, and 1.938635e-21 respectively), indicating they are the strongest positive contributors to diabetes progression. BP (blood pressure), S1 (serum cholesterol), and S2 (LDL) also showed statistically significant associations with the response, though their effect sizes were more moderate.

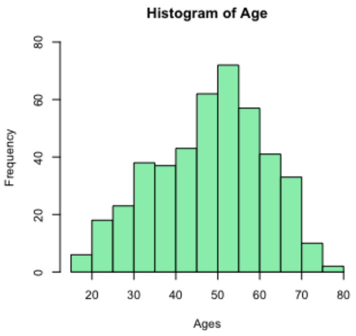
Our first model had an R^2 of 0.5177 and an Adjusted R^2 of 0.5066. After removing some of the predictors, we were left with an R^2 of 0.5149 and an Adjusted R^2 of 0.5082. This is a minor improvement in our model that allowed us to be able to go from approximately 50.66% of the variability in disease progression being explained to 50.82%.

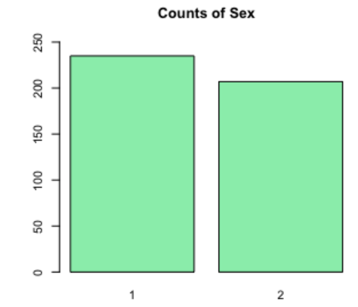
Appendix:

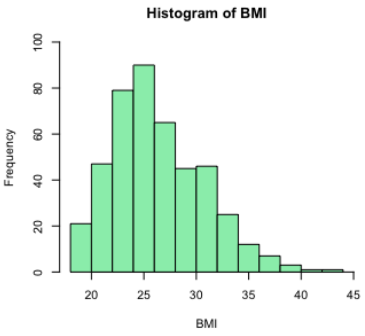
Additional Graphs:

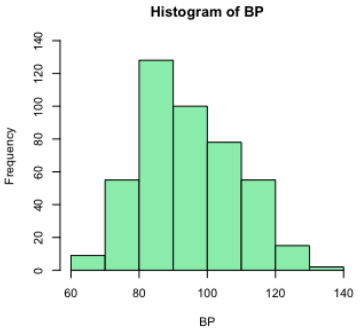
Below I will show all plots I created within each step of this project

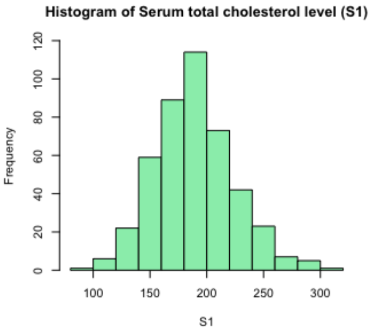
Exploring the Predictors:

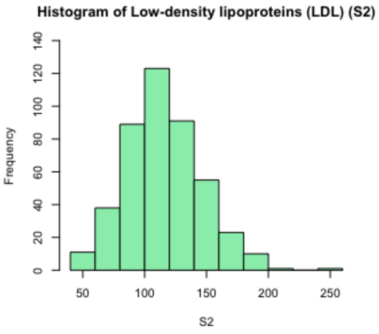


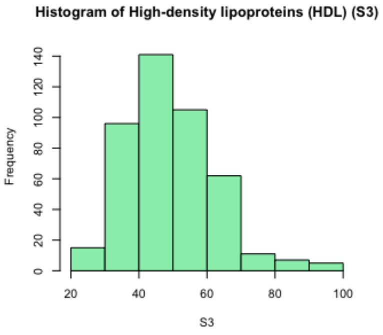


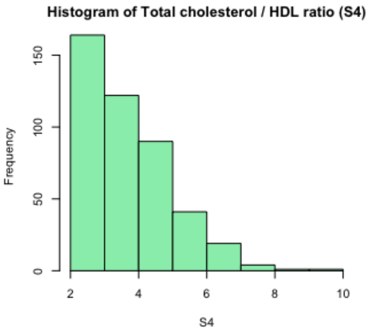


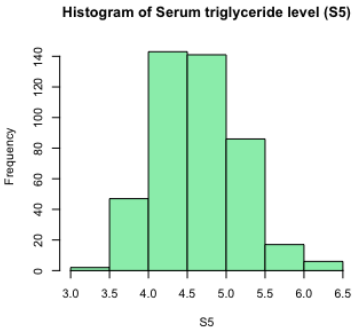


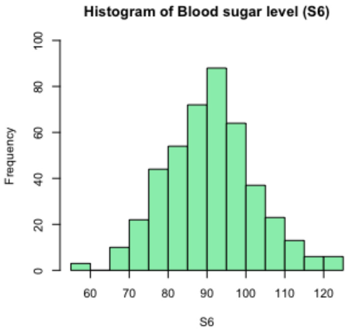




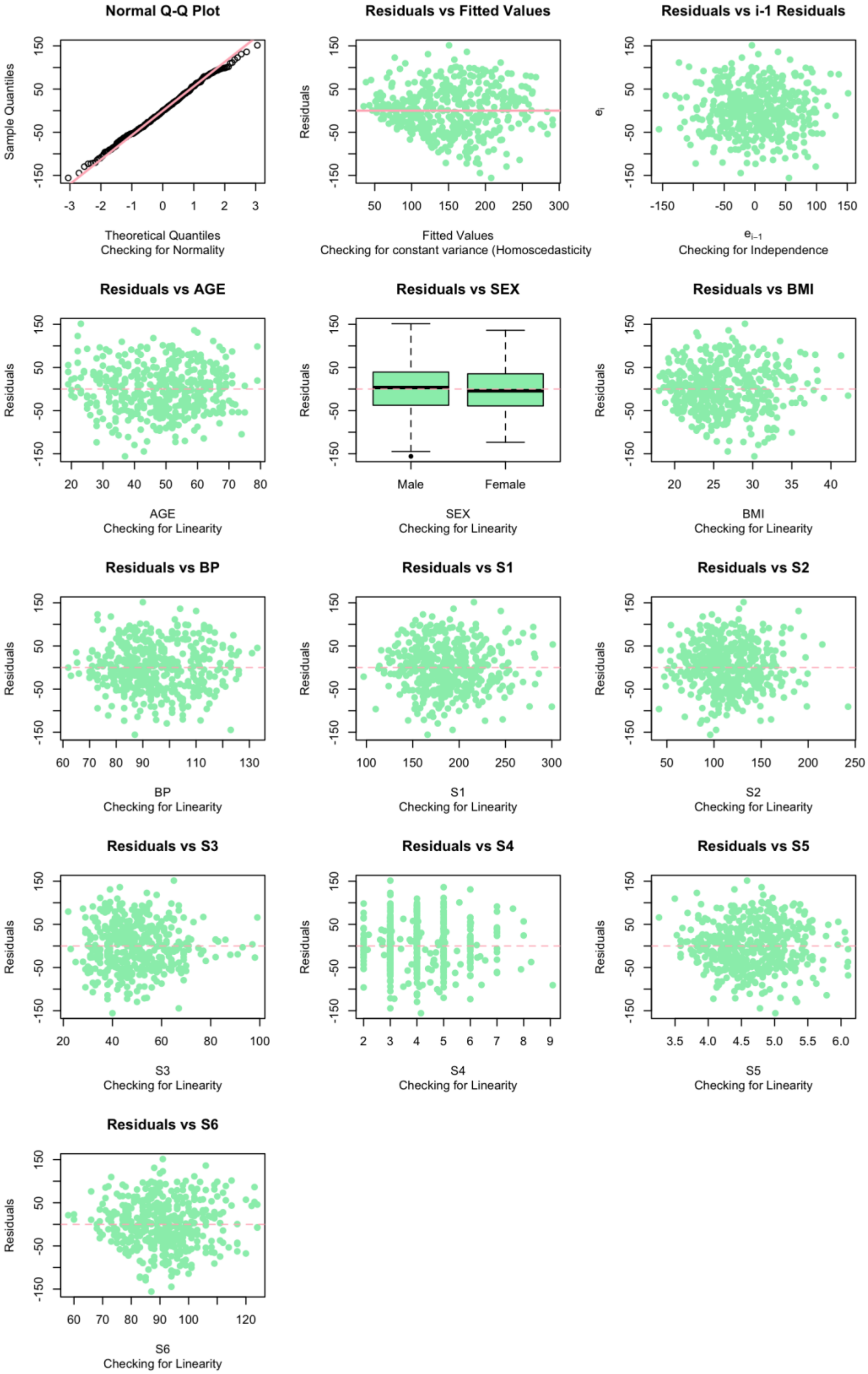




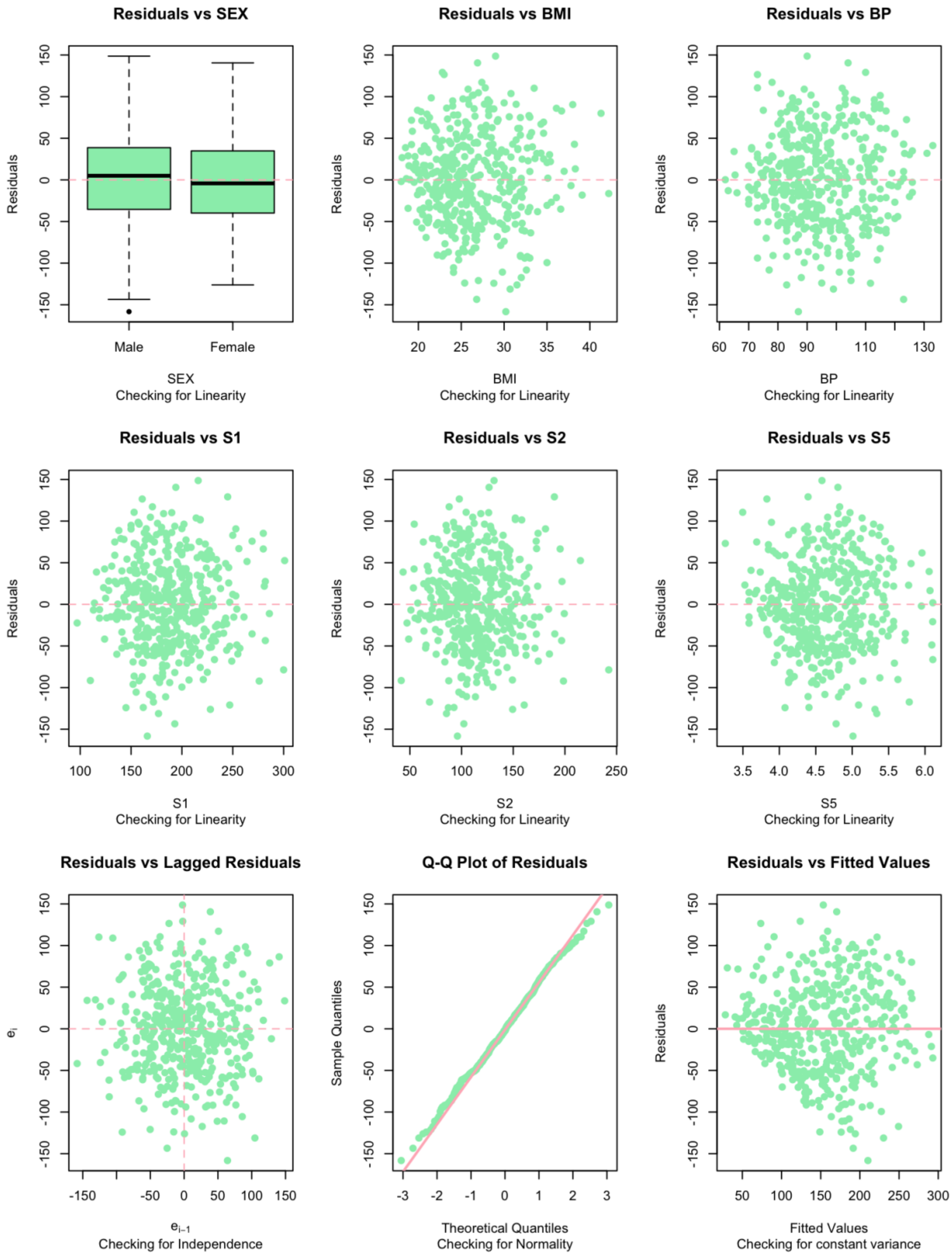




Model Selection and Fitting:



Model Assumptions on New Model:



R CODE:

```
# Data Exploration

    colSums(is.na(data))
    # AGE    SEX    BMI     BP     S1     S2     S3     S4     S5     S6 Target
    #   0      0      0      0      0      0      0      0      0      0      0
    # Our data is clean and ready to explore with

    summary(data)
    # AGE            || SEX            || BMI           ||  BP            ||   S1
    # Min.   :19.00  || Min.   :1.000  || Min.   :18.00 || Min.   : 62.00 ||  Min.   : 97.0
    # 1st Qu.:38.25  || 1st Qu.:1.000  || 1st Qu.:23.20 || 1st Qu.: 84.00 ||  1st Qu.:164.2
    # Median :50.00  || Median :1.000  || Median :25.70 || Median : 93.00 ||  Median :186.0
    # Mean   :48.52  || Mean   :1.468  || Mean   :26.38 || Mean   : 94.65 ||  Mean   :189.1
    # 3rd Qu.:59.00  || 3rd Qu.:2.000  || 3rd Qu.:29.27 || 3rd Qu.:105.00 ||  3rd Qu.:209.8
    # Max.   :79.00  || Max.   :2.000  || Max.   :42.20 || Max.   :133.00 ||  Max.   :301.0

    # S2             ||  S3            ||  S4           || S5             || S6
    # Min.   : 41.60 ||  Min.   :22.00 ||  Min.   :2.00 || Min.   :3.258  || Min.   : 58.00
    # 1st Qu.: 96.05 ||  1st Qu.:40.25 ||  1st Qu.:3.00 || 1st Qu.:4.277  || 1st Qu.: 83.25
    # Median :113.00 ||  Median :48.00 ||  Median :4.00 || Median :4.620  || Median : 91.00
    # Mean   :115.44 ||  Mean   :49.79 ||  Mean   :4.07 || Mean   :4.641  || Mean   : 91.26
    # 3rd Qu.:134.50 ||  3rd Qu.:57.75 ||  3rd Qu.:5.00 || 3rd Qu.:4.997  || 3rd Qu.: 98.00
    # Max.   :242.40 ||  Max.   :99.00 ||  Max.   :9.09 || Max.   :6.107  || Max.   :124.00

    # Target
    # Min.   : 25.0
    # 1st Qu.: 87.0
    # Median :140.5
    # Mean   :152.1
    # 3rd Qu.:211.5
    # Max.   :346.0

    # Standard Deviation
    Std_Dev <- sapply(data[, sapply(data, is.numeric)], sd)
    Std_Dev

    barplot(Std_Dev,
        main = "Standard Deviation of Numeric Variables",
        ylab = "Standard Deviation",
        col = "#99EDB8FF",
  ylim = c(0, 80),
  las = 2)
# BMI, S4, and S5 have the lowest variability
# S1 (cholesterol) and S2 (LDL) have the highest variablility

data$SEX <- factor(data$SEX, levels = c(1, 2), labels = c("Male", "Female"))

# Exploring Predictors

  # Age
  hist(
    data$AGE,
    main = "Histogram of Age",
    xlab = "Ages",
    ylim = c(0, 80),
    col = "#99EDB8FF")
  # Looks relatively normal

  # Sex
  barplot(
    table(SEX),
    main = "Counts of Sex",
    ylim = c(0, 250),
    col = "#99EDB8FF")
  # There seems to be more Males than Females in the dataset

  # BMI
  hist(
    data$BMI,
    main = "Histogram of BMI",
    xlab = "BMI",
    ylim = c(0, 100),
    col = "#99EDB8FF")
  # This variable seems to be right-skewed
  # I may have to transform this variable in the future because its not normally distributed

  # BP
  hist(
    data$BP,
    main = "Histogram of BP",
    xlab = "BP",
```

```r
    ylim = c(0, 140),
    col = "#99EDB8FF")
# This variable also seems to be right-skewed
# I may have to transform this variable in the future

# S1 - Serum total cholesterol level
hist(
    data$S1,
    main = "Histogram of Serum total cholesterol level (S1)",
    xlab = "S1",
    ylim = c(0, 120),
    col = "#99EDB8FF")
# This variable has a normal distribution

# S2 - Low-density lipoproteins (LDL)
hist(
    data$S2,
    main = "Histogram of Low-density lipoproteins (LDL) (S2)",
    xlab = "S2",
    ylim = c(0, 140),
    col = "#99EDB8FF")
# This variable also seems to be right-skewed
# I may have to transform this variable in the future

# S3 - High-density lipoproteins (HDL)
hist(
    data$S3,
    main = "Histogram of High-density lipoproteins (HDL) (S3)",
    xlab = "S3",
    ylim = c(0, 150),
    col = "#99EDB8FF")
# This variable also seems to be right-skewed
# I may have to transform this variable in the future

# S4 - Total cholesterol / HDL ratio
hist(
    data$S4,
    main = "Histogram of Total cholesterol / HDL ratio (S4)",
    xlab = "S4",
    ylim = c(0, 160),
    col = "#99EDB8FF")
```

```
        # This variable looks extremely right-skewed, the peak of this histogram is on the Y axis
        # I may have to transform this variable in the future

        # S5 - Serum triglyceride level
        hist(
            data$S5,
            main = "Histogram of Serum triglyceride level (S5)",
            xlab = "S5",
            ylim = c(0, 150),
            col = "#99EDB8FF")
        # This variable seems relatively normal

        # S6 - Blood sugar level
        hist(
            data$S6,
            main = "Histogram of Blood sugar level (S6)",
            xlab = "S6",
            ylim = c(0, 100),
            col = "#99EDB8FF")
        # This variable seems relatively normal


# Model Selection and Fitting

    # Preliminary Hypothesis Testing
    # Before interpreting the model, I will test key assumptions: normality of residuals, equal variance (homoscedasticity), and independence.

        # Normality
        main_model <- lm(Target ~ ., data = data)
        residuals <- main_model$residuals
        qqnorm(residuals)
        qqline(residuals, col = "lightpink", lwd = 2)

            # Shapiro-Wilk Test
            shapiro.test(residuals)
            # W = 0.99706, p-value = 0.6162



        # Equal Variances (Homoscedasticity)
        plot(main_model$fitted.values, residuals,
        main = "Residuals vs Fitted Values",
        xlab = "Fitted Values",
        ylab = "Residuals",
        pch = 19, col = "#99EDB8FF")
    abline(h = 0, col = "lightpink", lwd = 2)
    title(sub = "Checking for constant variance (Homoscedasticity)")

    # Independence
    plot(residuals[-length(residuals)], residuals[-1],
        main = "Residuals vs i-1 Residuals",
        xlab = expression(e[i-1]),
        ylab = expression(e[i]),
        pch = 19, col = "#99EDB8FF")
    title(sub = "Checking for Independence")

    # Linearity
    par(mfrow = c(3, 4))  # 3 rows, 4 columns (adjust as needed)

    # Loop through all predictors except the response
    for (var in names(data)[names(data) != "Target"]) {
    plot(data[[var]], residuals(main_model),
        main = paste("Residuals vs", var),
        xlab = var,
        ylab = "Residuals",
        pch = 19,
        col = "#99EDB8FF")
    abline(h = 0, col = "lightpink", lty = 2)
    }

    # Reset plotting layout
    par(mfrow = c(1, 1))



# Primary Hypothesis Testing & Post test Hypothesis Testing
# test to see if all of the predictors are insignificant
# test to see if each of the predictors are significant

main_model <- lm(Target ~ ., data = data)
summary(main_model)
    # Residuals:
    #     Min      1Q  Median      3Q     Max
```

```
        # -155.827  -38.536   -0.228   37.806  151.353
        #
        # Coefficients:
        #              Estimate Std. Error t value Pr(>|t|)
        # (Intercept) -334.56714   67.45462  -4.960 1.02e-06 ***
        # AGE           -0.03636    0.21704  -0.168 0.867031
        # SEX          -22.85965    5.83582  -3.917 0.000104 ***
        # BMI            5.60296    0.71711   7.813 4.30e-14 ***
        # BP             1.11681    0.22524   4.958 1.02e-06 ***
        # S1            -1.09000    0.57333  -1.901 0.057948 .
        # S2             0.74645    0.53083   1.406 0.160390
        # S3             0.37200    0.78246   0.475 0.634723
        # S4             6.53383    5.95864   1.097 0.273459
        # S5            68.48312   15.66972   4.370 1.56e-05 ***
        # S6             0.28012    0.27331   1.025 0.305990
        # ---
        # Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
        #
        # Residual standard error: 54.15 on 431 degrees of freedom
        # Multiple R-squared:  0.5177,  Adjusted R-squared:  0.5066
        # F-statistic: 46.27 on 10 and 431 DF,  p-value: < 2.2e-16

    # Multicollinearity
    # I will check for multicollinearity by calculating the Variance Inflation Factor (VIF) for each predictor.
    # VIF values greater than 10 indicate multicollinearity.

    library(car)
    vif(main_model)
        # AGE       SEX       BMI        BP        S1        S2        S3        S4        S5
        # 1.217307  1.278071  1.509437  1.459428 59.202510 39.193370 15.402156  8.890986 10.075967
        # S6
        # 1.484623

# Variable Selection

    # Backward selection using AIC

    full_model <- lm(Target ~ ., data = data)

    backward_model <- step(full_model, direction = "backward", trace = TRUE)
```

```
summary(backward_model)

  # Residuals:
  #   Min       1Q   Median       3Q      Max
  # -158.275  -39.476   -2.065   37.219  148.690
  #
  # Coefficients:
  #   Estimate Std. Error t value Pr(>|t|)
  # (Intercept) -313.7666    25.3848 -12.360  < 2e-16 ***
  #   SEX          -21.5910     5.7056  -3.784 0.000176 ***
  #   BMI            5.7111     0.7073   8.075 6.69e-15 ***
  #   BP             1.1266     0.2158   5.219 2.79e-07 ***
  #   S1            -1.0429     0.2208  -4.724 3.12e-06 ***
  #   S2             0.8433     0.2298   3.670 0.000272 ***
  #   S5            73.3065     7.3083  10.031  < 2e-16 ***
  #   ---
  #   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
  #
  # Residual standard error: 54.06 on 435 degrees of freedom
  # Multiple R-squared:  0.5149,    Adjusted R-squared:  0.5082
  # F-statistic: 76.95 on 6 and 435 DF,  p-value: < 2.2e-16

##-------------------------------##
# Backward Selection with P-value
pval_backward_model <- lm(Target ~ ., data = data)

repeat {
model_summary <- summary(pval_backward_model)
pvals <- coef(model_summary)[, "Pr(>|t|)"][-1]  # Exclude intercept
max_pval <- max(pvals, na.rm = TRUE)

if (max_pval > 0.10) {
    var_to_remove <- names(which.max(pvals))
    current_vars <- all.vars(formula(pval_backward_model))[-1]
    updated_vars <- setdiff(current_vars, var_to_remove)
    updated_formula <- as.formula(paste("Target ~", paste(updated_vars, collapse = " + ")))
    pval_backward_model <- lm(updated_formula, data = data)
} else {
    break
}
}
```

```
  # summary(pval_backward_model)
  #
  # Residuals:
  #   Min       1Q   Median       3Q      Max
  # -158.275  -39.476   -2.065   37.219  148.690
  #
  # Coefficients:
  #   Estimate Std. Error t value Pr(>|t|)
  # (Intercept) -313.7666    25.3848 -12.360  < 2e-16 ***
  #   SEX          -21.5910     5.7056  -3.784 0.000176 ***
  #   BMI            5.7111     0.7073   8.075 6.69e-15 ***
  #   BP             1.1266     0.2158   5.219 2.79e-07 ***
  #   S1            -1.0429     0.2208  -4.724 3.12e-06 ***
  #   S2             0.8433     0.2298   3.670 0.000272 ***
  #   S5            73.3065     7.3083  10.031  < 2e-16 ***
  #   ---
  #   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
  #
  # Residual standard error: 54.06 on 435 degrees of freedom
  # Multiple R-squared:  0.5149,    Adjusted R-squared:  0.5082
  # F-statistic: 76.95 on 6 and 435 DF,  p-value: < 2.2e-16

##-------------------------------##

# Forward selection using AIC

null_model <- lm(Target ~ 1, data = data)

forward_model <- step(null_model,
                      scope = list(lower = null_model, upper = full_model),
                      direction = "forward",
                      trace = TRUE)

summary(forward_model)

  # Residuals:
  #   Min       1Q   Median       3Q      Max
  # -158.275  -39.476   -2.065   37.219  148.690
  #
  # Coefficients:
  #   Estimate Std. Error t value Pr(>|t|)
  # (Intercept) -313.7666    25.3848 -12.360  < 2e-16 ***
  #   BMI            5.7111     0.7073   8.075 6.69e-15 ***
  #   S5            73.3065     7.3083  10.031  < 2e-16 ***
  #   BP             1.1266     0.2158   5.219 2.79e-07 ***
  #   S1            -1.0429     0.2208  -4.724 3.12e-06 ***
  #   SEX          -21.5910     5.7056  -3.784 0.000176 ***
  #   S2             0.8433     0.2298   3.670 0.000272 ***
  #   ---
  #   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
  #
  # Residual standard error: 54.06 on 435 degrees of freedom
  # Multiple R-squared:  0.5149,    Adjusted R-squared:  0.5082
  # F-statistic: 76.95 on 6 and 435 DF,  p-value: < 2.2e-16

##-------------------------------##

# Forward Selection with P-value
null_model <- lm(Target ~ 1, data = data)
scope <- setdiff(names(data), "Target")
selected <- c()

repeat {
pvals <- c()
for (var in setdiff(scope, selected)) {
    temp_formula <- as.formula(paste("Target ~", paste(c(selected, var), collapse = " + ")))
    temp_model <- lm(temp_formula, data = data)
    p <- summary(temp_model)$coefficients[var, "Pr(>|t|)"]
    pvals[var] <- p
}

if (length(pvals) == 0 || min(pvals, na.rm = TRUE) > 0.10) {
    break
}

selected <- c(selected, names(which.min(pvals)))
}

final_formula <- as.formula(paste("Target ~", paste(selected, collapse = " + ")))
pval_forward_model <- lm(final_formula, data = data)
summary(pval_forward_model)
```

```
 # Call:
 #   lm(formula = final_formula, data = data)
 #
 # Residuals:
 #   Min       1Q   Median       3Q      Max
 # -158.275  -39.476   -2.065   37.219  148.690
 #
 # Coefficients:
 #   Estimate Std. Error t value Pr(>|t|)
 # (Intercept) -313.7666    25.3848 -12.360  < 2e-16 ***
 #   BMI            5.7111     0.7073   8.075 6.69e-15 ***
 #   S5            73.3065     7.3083  10.031  < 2e-16 ***
 #   BP             1.1266     0.2158   5.219 2.79e-07 ***
 #   S1            -1.0429     0.2208  -4.724 3.12e-06 ***
 #   SEX          -21.5910     5.7056  -3.784 0.000176 ***
 #   S2             0.8433     0.2298   3.670 0.000272 ***
 #   ---
 #   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
 #
 # Residual standard error: 54.06 on 435 degrees of freedom
 # Multiple R-squared:  0.5149,    Adjusted R-squared:  0.5082
 # F-statistic: 76.95 on 6 and 435 DF,  p-value: < 2.2e-16

##-------------------------------##

# Stepwise Selection
stepwise_model <- step(full_model, direction = "both", trace = TRUE)
summary(stepwise_model)

  # Coefficients:
  #              Estimate Std. Error t value Pr(>|t|)
  # (Intercept) -313.7666    25.3848 -12.360  < 2e-16 ***
  # SEX          -21.5910     5.7056  -3.784 0.000176 ***
  # BMI            5.7111     0.7073   8.075 6.69e-15 ***
  # BP             1.1266     0.2158   5.219 2.79e-07 ***
  # S1            -1.0429     0.2208  -4.724 3.12e-06 ***
  # S2             0.8433     0.2298   3.670 0.000272 ***
  # S5            73.3065     7.3083  10.031  < 2e-16 ***
  # ---
  # Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
  #
```

```
    # Residual standard error: 54.06 on 435 degrees of freedom
    # Multiple R-squared:  0.5149,  Adjusted R-squared:  0.5082
    # F-statistic: 76.95 on 6 and 435 DF,  p-value: < 2.2e-16

##-------------------------------##

# Compare with AIC
AIC(full_model, backward_model, forward_model, stepwise_model)

     # df      AIC
     # full_model     12 4795.986
     # backward_model  8 4790.603
     # forward_model   8 4790.603



##-------------------------------##
# Model Assumptions on new model

final_model <- lm(Target ~ SEX + BMI + BP + S1 + S2 + S5, data = data)
summary(final_model)$coefficients
#               Estimate Std. Error    t value     Pr(>|t|)
#   (Intercept) -313.766623 25.3847708 -12.360428 2.750430e-30
#   SEX          -21.591011  5.7056378  -3.784154 1.758474e-04
#   BMI            5.711107  0.7072624   8.074948 6.687185e-15
#   BP             1.126553  0.2158433   5.219308 2.786882e-07
#   S1            -1.042876  0.2207508  -4.724225 3.122573e-06
#   S2             0.843277  0.2297536   3.670354 2.723024e-04
#   S5            73.306526  7.3082565  10.030645 1.938635e-21

# 1. Linearity: Residuals vs Each Predictor
par(mfrow = c(2, 3))  # Layout for 6 predictors
for (var in c("SEX", "BMI", "BP", "S1", "S2", "S5")) {
plot(data[[var]], residuals(final_model),
    main = paste("Residuals vs", var),
    xlab = var, ylab = "Residuals",
    pch = 19, col = "#99EDB8FF")
abline(h = 0, col = "lightpink", lty = 2)
}
par(mfrow = c(1, 1))  # Reset layout

# 2. Independence: Residuals vs Lagged Residuals
```

```
plot(residuals(final_model)[-length(residuals(final_model))],
    residuals(final_model)[-1],
    main = "Residuals vs Lagged Residuals",
    xlab = expression(e[i-1]), ylab = expression(e[i]),
    pch = 19, col = "#99EDB8FF")
abline(h = 0, v = 0, col = "lightpink", lty = 2)

# 3. Normality: Q-Q Plot and Shapiro-Wilk Test
qqnorm(residuals(final_model),
    main = "Q-Q Plot of Residuals",
    col = "#99EDB8FF", pch = 19)
qqline(residuals(final_model), col = "lightpink", lwd = 2)

shapiro.test(residuals(final_model))
    # data:  residuals(final_model)
    # W = 0.99724, p-value = 0.6719

# 4. Equal Variance (Homoscedasticity): Residuals vs Fitted
plot(final_model$fitted.values, residuals(final_model),
    main = "Residuals vs Fitted Values",
    xlab = "Fitted Values", ylab = "Residuals",
    pch = 19, col = "#99EDB8FF")
abline(h = 0, col = "lightpink", lwd = 2)
title(sub = "Checking for constant variance")
```