Exploring Relationships between a Person's BMI and Factors Related to Diabetes

Stat 306 Project Final Report

Group A7: Maggie Ruan, Shaoyun Tong, Yimin You, Kevin Yu

1 Introduction

1.1 Background and Motivation

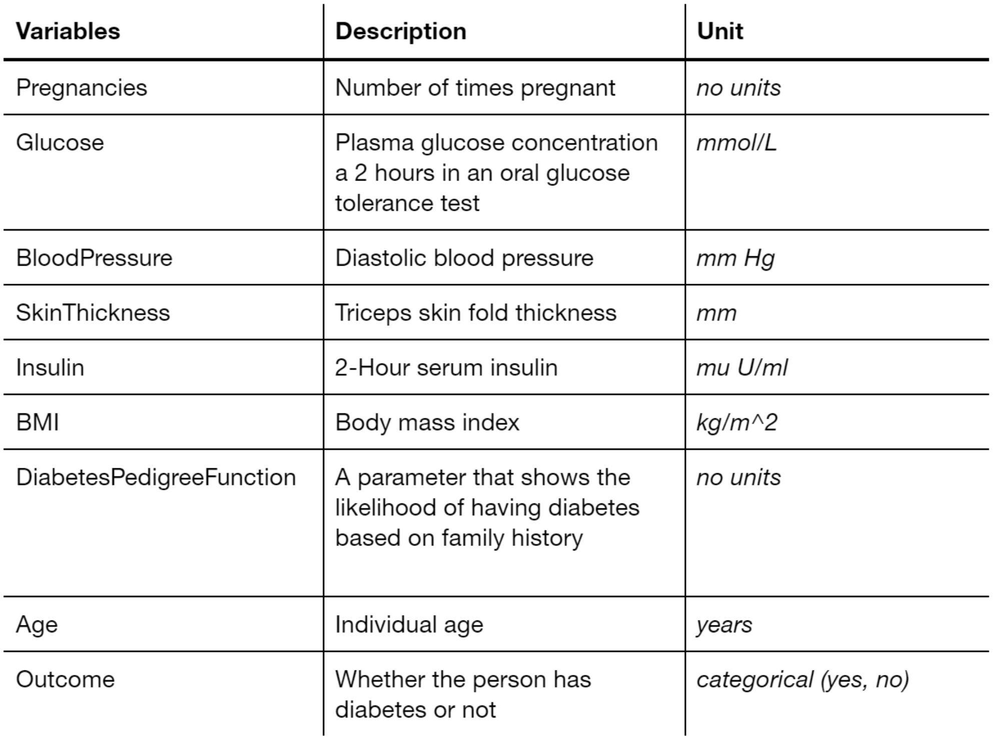
The Body Mass Index (BMI) is a measurement of body fat based on a person's height and weight. People are paying more attention to their BMI index (Chatswoodmedical, 2020) as a higher BMI may result in a higher risk for getting certain diseases such as heart disease, high blood pressure, type 2 diabetes, etc. In the world right now, there are 537 million adults (20-79 years) living with diabetes - meaning approximately 1 in 15 (International Diabetes Federation, 2022). Moreover, diabetes will increase the risk of heart disease by about four times in women and may have worse outcomes towards diabetes-related complications such as blindness, kidney disease, and depression (WHO).

Research has shown (Murphy, 2019) that there is a correlation between a person's BMI and a person's chance of having diabetes, and further research has shown that an individual having a higher BMI does increase the chance of having diabetes.

This project hopes to view this correlation in the opposite direction. While we do not conclude any implications between diabetes and BMI, our goal is to build the most accurate linear model that we can. The dataset used in this project was originally used as a classification project for factors related to diabetes. We hope to exploit major multicollinearities found within the dataset, such that future research could find possible implications within how diabetes factors could affect the BMI of a person, as well as find other potential multicollinearities within the dataset.

1.2 Dataset source

According to Statista.com, India comes in second in the rank of the highest number of diabetics in 2021 (Elflein, 2022). The National Institute of Diabetes and Digestive and Kidney Diseases collected the data on females over 21 years old of Pima Indian heritage on May 9th, 1990, and was donated to Kaggle.com two years ago as of the time of publication. A link to the source of the dataset can be found in the references section.



Variables	**Description**	**Unit**
Pregnancies	Number of times pregnant	*no units*
Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	*mmol/L*
BloodPressure	Diastolic blood pressure	*mm Hg*
SkinThickness	Triceps skin fold thickness	*mm*
Insulin	2-Hour serum insulin	*mu U/ml*
BMI	Body mass index	*kg/m^2*
DiabetesPedigreeFunction	A parameter that shows the likelihood of having diabetes based on family history	*no units*
Age	Individual age	*years*
Outcome	Whether the person has diabetes or not	*categorical (yes, no)*

Figure 1: A brief overview of the variables in the dataset

2 Analysis

We treated BMI as the response variable, while Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, DiabetesPedigreeFunction, Age, and Outcome as the explanatory variables.

2.1 Data Wrangling

Before we begin, we created a ggpairs plot for our dataset (Figure 1). From this figure, we can see that the correlation values with BMI are both significantly higher than other variables, which shows that there could be a relationship between BMI and other variables.

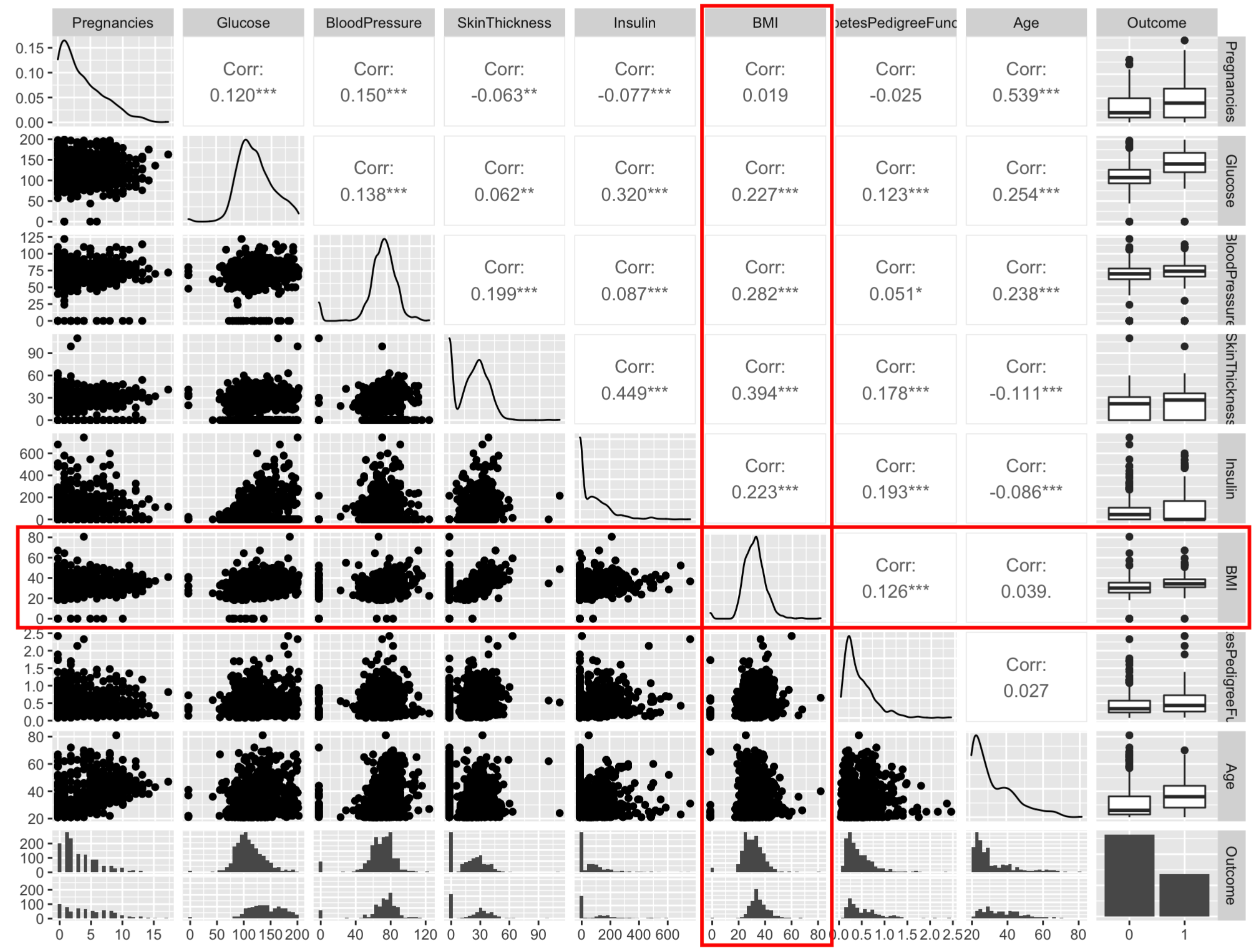


Figure 2: ggpairs plot on the original dataset, the row/column with BMI is highlighted

Taking a closer look at Figure 2, we noticed that there are lots of 0 values in pregnancies and other variables, which will substantially affect our findings. To counteract this issue, we decided on two procedures; one, to remove the Pregnancies column from our dataset, and two, to filter out all the zeros from the other columns.

Removing the Pregnancies column was a difficult choice, as we initially hypothesized that it could have an impact; but we have discovered that having so many zeros has created many problems in our tests later on. It is known that BMI has a negative influence on pregnancies, but it is not known if the reverse is true (HealthLink BC, 2022).

There were also many zero values within the data, especially in columns such as BloodPressure, SkinThickness, Insulin and BMI. Normally, those variables should not contain zero values, but it may happen because of missing data in collecting this dataset. (The source of this dataset made no comment on this matter.) To make our analysis more accurate, we assumed that these values were empty, and filtered them out.

After this filter, our dataset was reduced by approximately half, ending with n = 1035, compared to the original 2000 we started with.

2.2 Splitting our Data

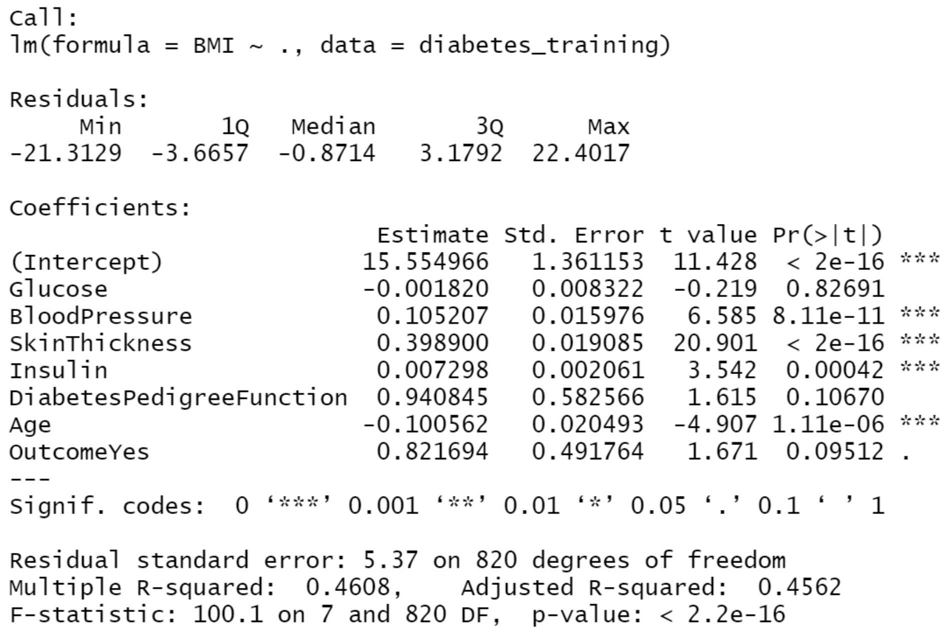
The (now wrangled) data is now randomly split into two groups. 80% of the data is set into a “training” dataset, and 20% of the data is set into a “testing” dataset. The majority of this project will utilize the training data; one analysis later on will require using the testing data. So to be consistent, we will utilize the training data for most of the testing done.
The training data had n = 828 observations, and the testing data had n = 207.

2.3 Full Model

We first defined the variables used in this dataset.

Response Variable	Explanatory Variable
y = BMI (kg/m^2)	x_1 = Glucose (mg/dL)
	x_2 = BloodPressure ($mmHg$)
	x_3 = SkinThickness (mm)
	x_4 = Insulin ($mu\ U/ml$)
	x_5 = DiabetesPedigreeFunction
	x_6 = Age ($years$)
	x_7 = Dummy Variable for when Outcome = Yes

Table 1: Quick Summarization of the Variables



```
Call:
lm(formula = BMI ~ ., data = diabetes_training)

Residuals:
     Min       1Q   Median       3Q      Max 
-21.3129  -3.6657  -0.8714   3.1792  22.4017 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)              15.554966   1.361153  11.428  < 2e-16 ***
Glucose                  -0.001820   0.008322  -0.219  0.82691    
BloodPressure             0.105207   0.015976   6.585 8.11e-11 ***
SkinThickness             0.398900   0.019085  20.901  < 2e-16 ***
Insulin                   0.007298   0.002061   3.542  0.00042 ***
DiabetesPedigreeFunction  0.940845   0.582566   1.615  0.10670    
Age                      -0.100562   0.020493  -4.907 1.11e-06 ***
OutcomeYes                0.821694   0.491764   1.671  0.09512 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.37 on 820 degrees of freedom
Multiple R-squared:  0.4608,	Adjusted R-squared:  0.4562 
F-statistic: 100.1 on 7 and 820 DF,  p-value: < 2.2e-16
```

Figure 3: Model summary for the full model (model 7)

Our full model (model 7) is: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7$

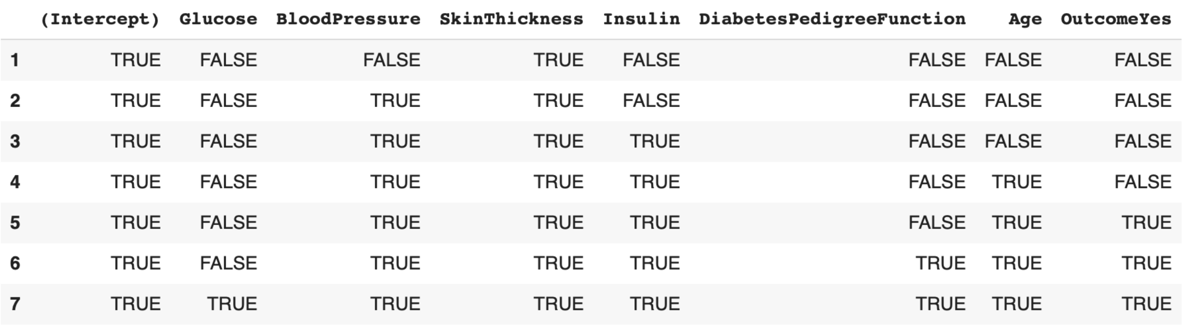
We first do a full model (use all variables from the training dataset), the results summarized (Figure 3). According to the summary table, variables that weren't found to be significant within the 5% significance level include Glucose, DPF, and the Outcome dummy variable. We will not remove any variables now; we will be relying on techniques conducted later on to determine which variables may be removed in our final model.

2.4 Model Selection Using Various Techniques

We now begin the model selection techniques. We opted to use four techniques/statistics to justify our final model. Exhaustive Search Methods, Alkaline Information Criterion, Backwards Elimination, and Root Mean Squared Error through 4-fold Cross Validation.

2.4.1 Exhaustive Search - Mallows C_p, $AdjR^2$

The first test we will conduct is the Exhaustive Search Method. The table below shows the result of searching through all possible models by selecting different numbers of explanatory variables. The R^2, $AdjR^2$ and Mallows's C_p statistics are all calculated to be analyzed. The results are shown in Figure 4 and Table 2.



	(Intercept)	Glucose	BloodPressure	SkinThickness	Insulin	DiabetesPedigreeFunction	Age	OutcomeYes
1	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
2	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE
3	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE
4	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE
5	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE
6	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
7	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

Figure 4: Outcome of Exhaustive Search; FALSE means the variable has been taken out of the best model

Model	Number of Predictors	R^2	$AdjR^2$	C_p
1	1	0.40866	0.40794	75.3518
2	2	0.43117	0.42980	43.1039
3	3	0.44294	0.44091	27.2065
4	4	0.45661	0.45397	8.4254
5	5	0.45908	0.45579	6.6698
6	6	0.46080	0.45686	6.0478
7	7	0.46083	0.45623	8.0000

Table 2: Summary of Exhaustive Search result

R^2 is an increasing function of the number of independent variables, meaning by adding an independent variable, R^2 will likely increase or maintain the same. As a result, we do not interpret R^2, rather we look at $AdjR^2$. It makes a more accurate view of the correlation between one variable and another by taking into account how many independent variables are added which a higher $AdjR^2$ implies a better fit to the model. According to Table 2, Model 6 has the highest $AdjR^2$.

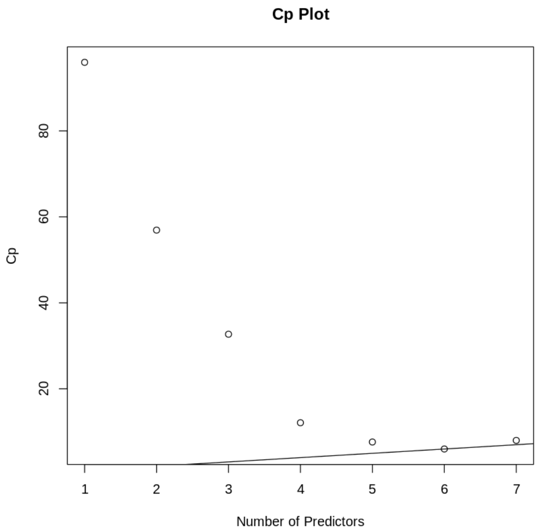


Figure 5: C_p plot

Furthermore, the C_p values assess the fit of a model where it is better when C_p is small and close to the number of predictors p.

Model	Difference between P and C_p
5	1.6698
6	0.0478
7	1.0000

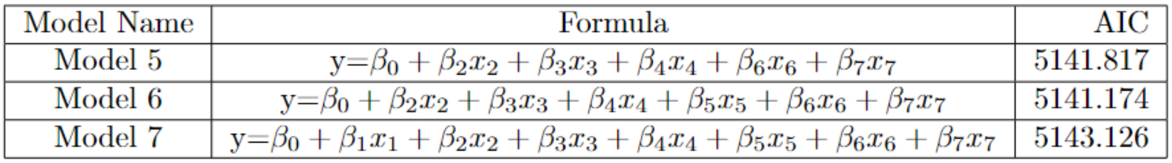
Table 3: Partial Table of P - C_p

By taking into account the C_p values, the differences between P and C_p are relatively small for Model 5, 6 and 7 (Table 3). The group decides to use these three models for further investigation.

2.4.2 Alkaline Information Criterion (AIC)

We can check the Alkaline Information Criterion (AIC) for each of models 5, 6, and 7. A lower AIC statistic means a smaller prediction error.

The results are shown in Table 4.



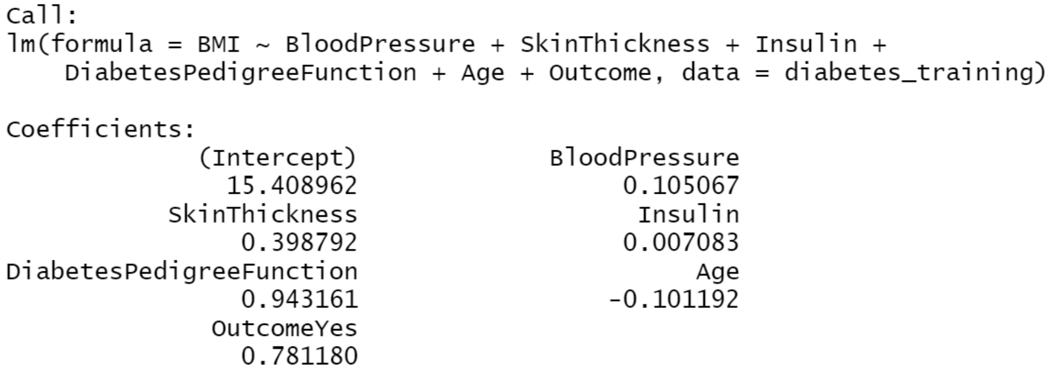
Model Name	Formula	AIC
Model 5	y=$\beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_6 x_6 + \beta_7 x_7$	5141.817
Model 6	y=$\beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7$	5141.174
Model 7	y=$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7$	5143.126

Table 4: Partial table of AIC statistics

To check the models using AIC, we determine which model results in the lowest AIC statistic. Looking at this table, while the statistics are very similar, we can see that Model 6 has the lowest AIC statistic. Hence, this test concludes with Model 6 being the best; however sampling variation could result in different AIC statistics, so this test is not completely conclusive, and further tests have to be made.

2.4.3 Backwards Elimination

The next technique we will be using is Backwards Elimination. We start with the full model (model 7), and compute the highest p-value for the coefficients; if it is statistically insignificant (greater than 0.05), remove the factor, and refit the model. When each predictor becomes statistically significant, we output the result.
The result of our findings is displayed in Figure 6.



```
Call:
lm(formula = BMI ~ BloodPressure + SkinThickness + Insulin + 
    DiabetesPedigreeFunction + Age + Outcome, data = diabetes_training)

Coefficients:
             (Intercept)             BloodPressure  
               15.408962                  0.105067  
           SkinThickness                   Insulin  
                0.398792                  0.007083  
DiabetesPedigreeFunction                       Age  
                0.943161                 -0.101192  
              OutcomeYes  
                0.781180  
```

Figure 6: Backward Elimination Results

According to the output on R, we conclude the same model with 6 predictors.

2.4.4 Normalized Root Mean Square Error (NRMSE)

We performed a 4-fold cross-validation on Models 5, 6, and 7 in which all training data are split into four groups. One group is set as the test data and the remaining three groups are set as the training and validation data. Results were taken for each of the four folds, and the average is taken to determine the results of this test. The results are shown below.

Model	NRMSE (Training)	NRMSE (Testing)
5	0.16153	0.13403
6	0.16154	0.13366
7	0.16193	0.13371

Table 5: Result of Cross-validation on NRMSE

The NRMSE calculates the absolute value between predicted and observed values using different types of normalization methods. A low NRMSE indicates that the observed and predicted data are close to each other meaning better accuracy. By looking at Table 5, Models 5 and 6 have fairly similar NRMSE results for the training set. However, Model 6's NRMSE is even smaller when we are predicting using the testing set. Ultimately, we prefer Model 6 when accessing NRMSE.

2.5 Model Diagnostics

We have looked at 4 model selection techniques, and it suggested that the best model is model 6; that is, the predictors being Blood Pressure, Skin Thickness, Insulin, DPF, Age, and the Outcome.

Before we jumped to conclusions, however, we had a look at some model diagnostics, just to be sure that our findings and our assumptions were valid.

2.5.1 Residual Plot for BMI

We had a look at the residual plot; that is, whether the model may have required a transformation. A residual plot is a diagnostic plot that helps us determine whether a transformation may have been required in our linear model and whether some assumptions (like constant variance) are valid.

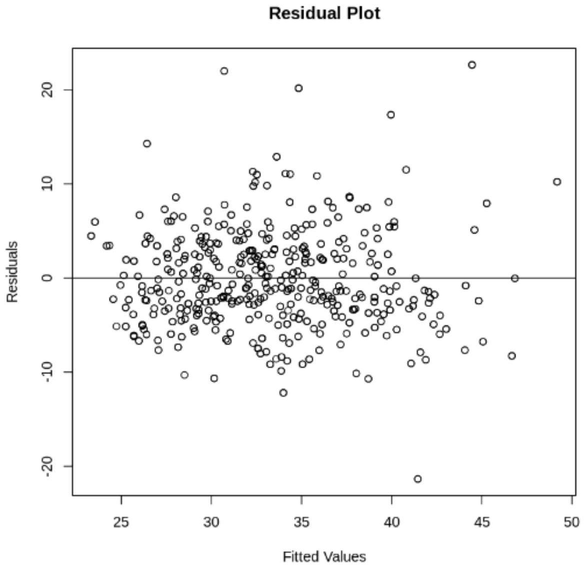


Figure 7: Residual plot for model 6 on our training set

Since the residual plot (Figure 7) does not seem to display an obvious pattern, (No quadratic trend, constant random spread… etc) we can conclude that no transformations were required for our model and that a constant variance can be assumed.

2.5.2 Q-Q Plot for Residuals

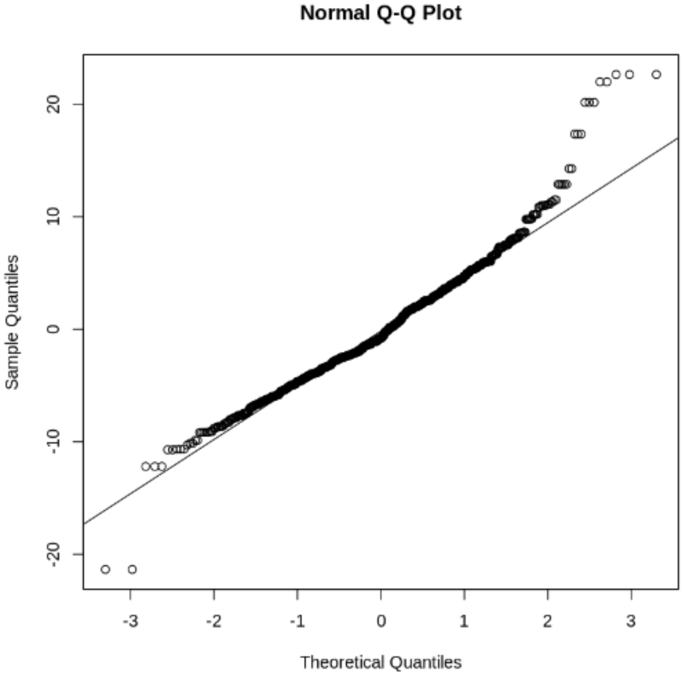


Figure 8: Residual Q-Q plot

The qqplot (Figure 8) for BMI demonstrates that the BMI variable is approximately normally distributed, which is a major assumption required for the majority of these tests used in this project. There does seem to be a major tail on the right, (which may suggest a right skew of the data,) and a few outliers on the left, but since the majority of the data is rested on the line, we believed that this is a good assumption to make.

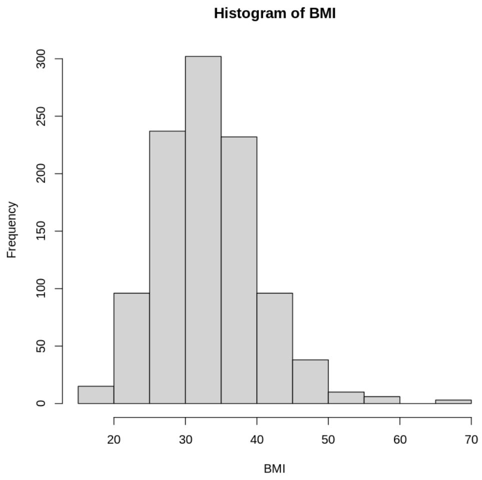
3 Conclusion

Using the four model selection techniques, we conclude that the best linear model that fits our guidelines is model 6, which can be written as follows:

$$BMI = 15.79 + 0.10*BloodPressure + 0.39*SkinThickness + 0.0072*Insulin + 0.95*DiabetesPedigreeFunction - 0.095*Age + 0.90*Outcome$$

With an $AdjR^2$ of 0.4648, a difference of 0.0029 between the model's number of parameters and the model's C_p statistic, a normalized RMSE of 0.13366, an AIC of 5141.174, and the selected result from a backwards elimination process, this is the best model that can be found within the guidelines set at the beginning. We made this finding not to find potential causation, but mainly to see how correlated we can build a linear model and see how accurate it can get, perhaps for future projects to continue with this analysis and dive deeper into our results and improve our model more, or to discover deeper possible causations and correlations between a person's BMI and factors used to help classify diabetes.

It is worth noting some assumptions made during this project, some of which have been mentioned in section 2.5 (Model Diagnostics), but will be recited here:



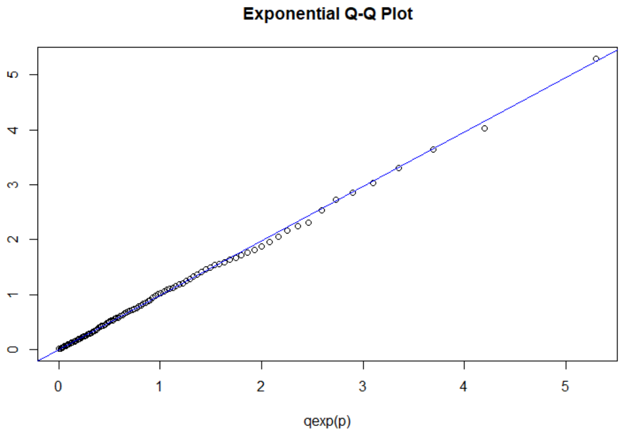


Figure 9: Histogram of BMI

Figure 10: Exponential Q-Q plot of BMI

1. To conduct our analysis, we assume the dataset follows the normal distribution, but the histogram plot (Figure 9)shows that our response value does not follow the normal distribution perfectly, and finds that it may follow an exponential distribution. We had a glance at an exponential Q-Q plot (Figure 10), and notice that the dataset fits it well. In a future report, this could be a factor that should be considered.
2. We assume that our dataset is a good sample of the population. From the dataset description, it says the data is collected from women of Pima Indian Heritage. For this project, we assumed that the population is the population of Earth; but we do accept the fact that there may be a huge sampling bias when it comes to this analysis.
3. We assume that the data collected is sampled independently from the population. As far as we know, the population of the dataset is Indian women, who are over 21 years old of Pima Indian heritage, with the population assumed to be taken sometime around 1990.

References

Akturk, Mehmet. “Diabetes Dataset.” *Kaggle*, 5 Aug. 2020, https://www.kaggle.com/datasets/mathchi/diabetes-data-set.

Chatswoodmedical. “The Importance of Knowing Your Body Mass Index (BMI): Chatswood Medical & Dental Centre.” *Chatswood Medical & Dental Centre* |, 23 Oct. 2020, https://www.chatswoodmedicalanddental.com.au/articles/body-mass-index/#:~:text=It's%20a%20good%20way%20to,Type%202%20Diabetes.

Elflein, John. “Diabetics Number Top Countries 2021.” *Statista*, 27 July 2022, https://www.statista.com/statistics/281082/countries-with-highest-number-of-diabetics/.

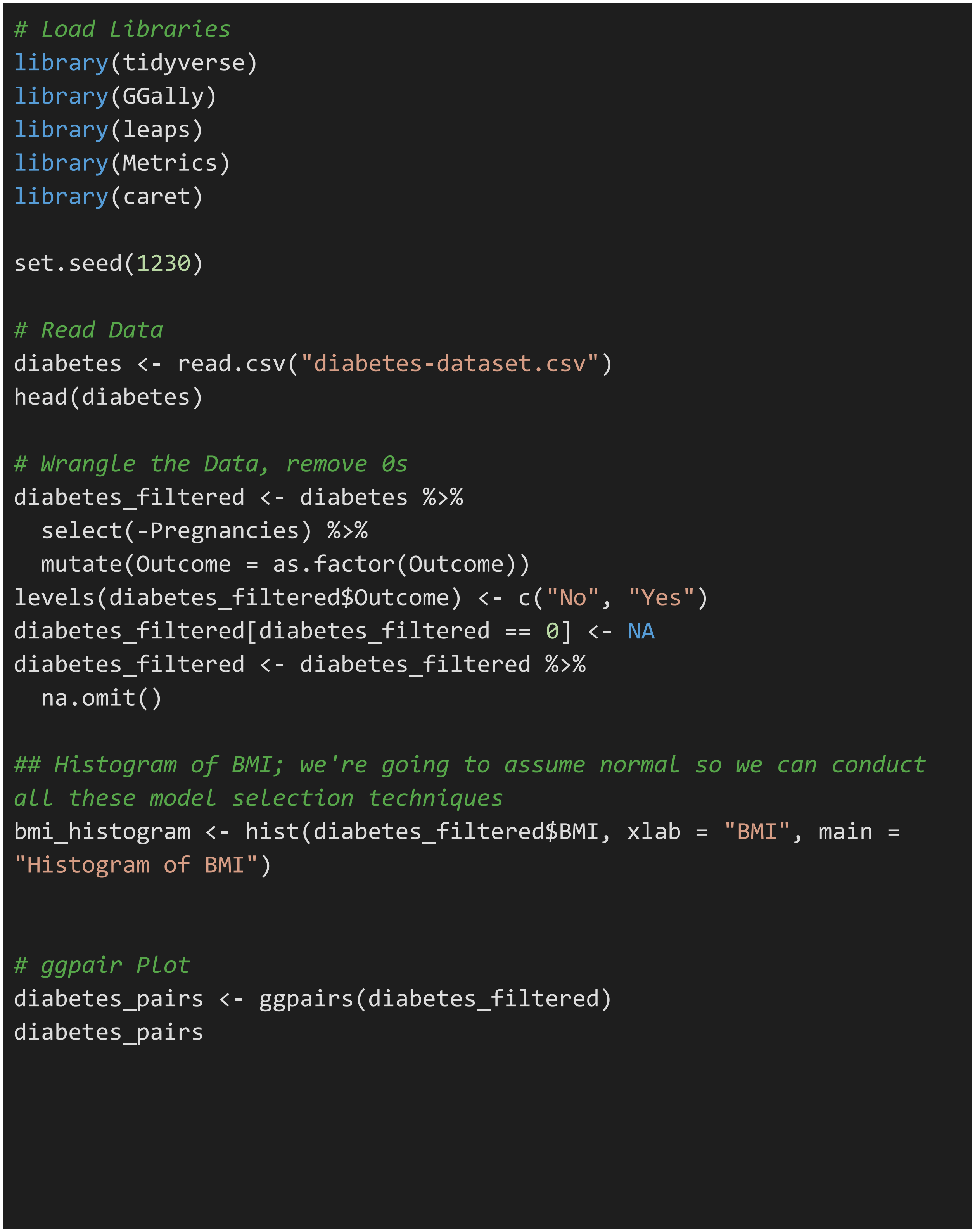
HealthLink BC. “Obesity and Pregnancy.” *Obesity and Pregnancy | HealthLink BC*, Apr. 2022, https://www.healthlinkbc.ca/pregnancy-parenting/pregnancy/health-conditions-and-pregnancy/obesity-and-pregnancy.

International Diabetes Federation. “Diabetes Is Spiralling out of Control .” *IDF Diabetes Atlas*, 2022, https://diabetesatlas.org/.

Murphy, Shane. “Diabetes and Cardiovascular Disease Risk in Women.” *Healthline*, Healthline Media, 3 May 2019, https://www.healthline.com/health/type-2-diabetes/diabetes-heart-disease-women.

WHO. “Diabetes.” *World Health Organization*, World Health Organization, https://www.who.int/health-topics/diabetes#tab=tab_1.

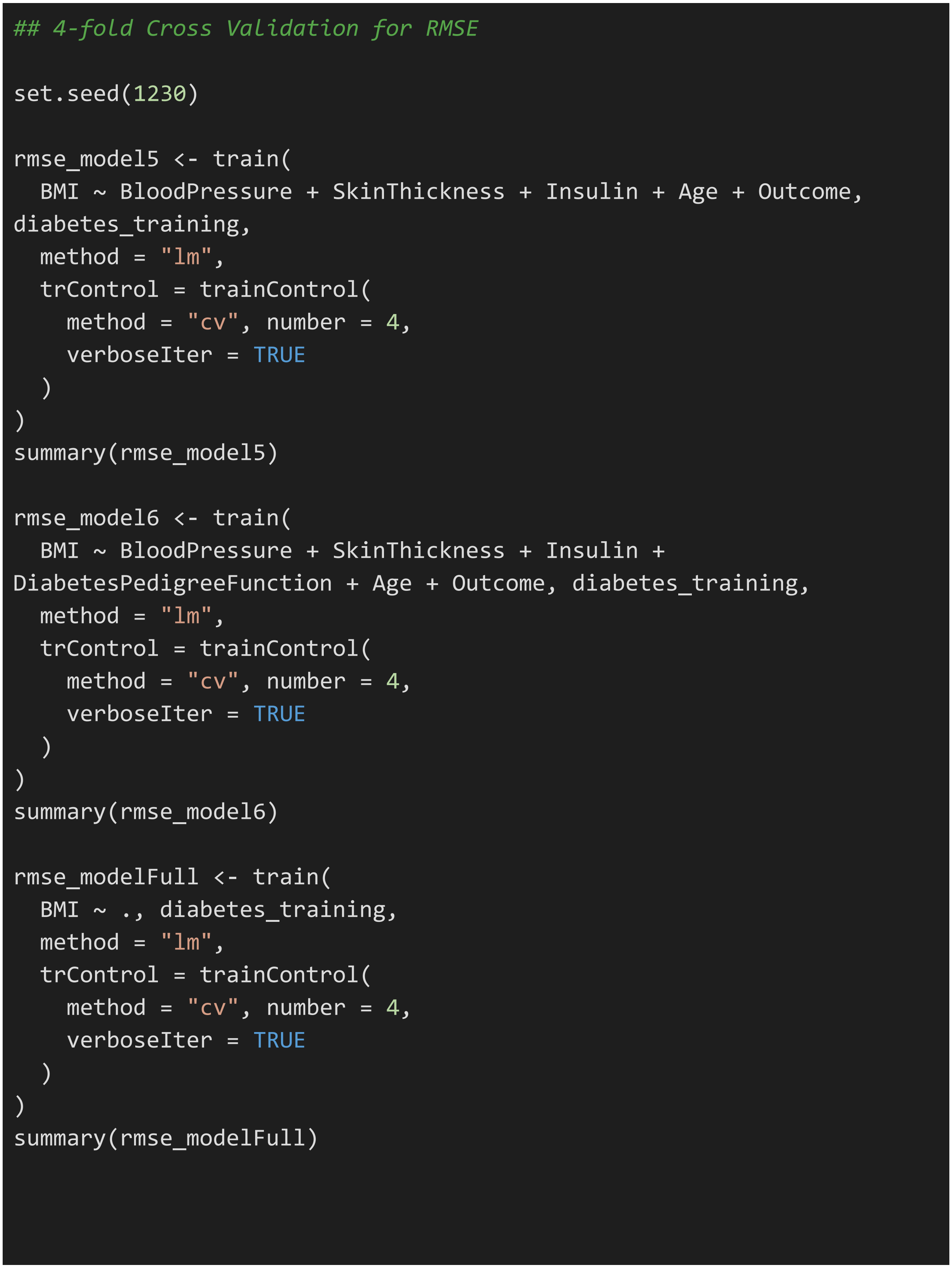
4 R Script



```r
# Load Libraries
library(tidyverse)
library(GGally)
library(leaps)
library(Metrics)
library(caret)

set.seed(1230)

# Read Data
diabetes <- read.csv("diabetes-dataset.csv")
head(diabetes)

# Wrangle the Data, remove 0s
diabetes_filtered <- diabetes %>%
  select(-Pregnancies) %>%
  mutate(Outcome = as.factor(Outcome))
levels(diabetes_filtered$Outcome) <- c("No", "Yes")
diabetes_filtered[diabetes_filtered == 0] <- NA
diabetes_filtered <- diabetes_filtered %>%
  na.omit()

## Histogram of BMI; we're going to assume normal so we can conduct
all these model selection techniques
bmi_histogram <- hist(diabetes_filtered$BMI, xlab = "BMI", main =
"Histogram of BMI")


# ggpair Plot
diabetes_pairs <- ggpairs(diabetes_filtered)
diabetes_pairs
```

```
## Split data into training and testing (80% training, 20% testing)
set.seed(1230)
diabetes_split <- sort(sample(nrow(diabetes_filtered),
nrow(diabetes_filtered)*.8))

diabetes_training <- diabetes_filtered[diabetes_split,]
diabetes_testing <- diabetes_filtered[-diabetes_split,]


# Linear Model
model <- lm(BMI ~ ., data = diabetes_training)
summary(model)


# Leaps
diabetes_models <- regsubsets(BMI~., data = diabetes_training, method
= "exhaustive")
summary_models <- summary(diabetes_models)
summary_models$which

summary_models$adjr2
summary_models$cp
summary_models$rsq

##R^2 increase, not useful, model get larger
##according to adjR^2 choose model 6 (max)

# Cp Plot:
plot(summary_models$cp, main = "Cp Plot", xlab =  "Number of
Predictors", ylab = "Cp")
abline(a = 0, b = 1)

cp_table <- data.frame(predictors = c(1:7),
                       cp = summary_models$cp) %>%
    mutate(diff = cp - predictors)
```

```
## According to Cp, we the model wht 6 predictors because its Cp is
the closest to the number of predictors, with a difference of
0.002864.
## We will continue to look at models with 5 and 7 predictors as
reference

#model 5
reg_model5 = lm(BMI ~ BloodPressure + SkinThickness + Insulin + Age +
Outcome, data = diabetes_training)
AIC(reg_model5)

#model 6
reg_model6 = lm(BMI ~ BloodPressure + SkinThickness + Insulin +
DiabetesPedigreeFunction + Age + Outcome, data = diabetes_training)
AIC(reg_model6)

#model 7 (aka. full model)
reg_modelFull = lm(BMI ~ ., data = diabetes_training)
AIC(reg_modelFull)

summary(reg_modelFull)

aic_table <- data.frame(model = c(5, 6, 7),
                        AIC = c(AIC(reg_model5), AIC(reg_model6),
AIC(reg_modelFull)))

aic_table




## Backwards Elimination

step(reg_modelFull, direction = "backward", trace = FALSE)

# Using backwards elimination, we conclude the same model with 6
predictors; so we will use this model.
```



```
## 4-fold Cross Validation for RMSE

set.seed(1230)

rmse_model5 <- train(
  BMI ~ BloodPressure + SkinThickness + Insulin + Age + Outcome,
diabetes_training,
  method = "lm",
  trControl = trainControl(
    method = "cv", number = 4,
    verboseIter = TRUE
  )
)
summary(rmse_model5)

rmse_model6 <- train(
  BMI ~ BloodPressure + SkinThickness + Insulin +
DiabetesPedigreeFunction + Age + Outcome, diabetes_training,
  method = "lm",
  trControl = trainControl(
    method = "cv", number = 4,
    verboseIter = TRUE
  )
)
summary(rmse_model6)

rmse_modelFull <- train(
  BMI ~ ., diabetes_training,
  method = "lm",
  trControl = trainControl(
    method = "cv", number = 4,
    verboseIter = TRUE
  )
)
summary(rmse_modelFull)
```

```
## Diagnostic plots

# Normal Q-Q Plot for Fitted Values
# Exponential Q-Q plot for Fitted Values
Z <- rexp(1000)
p <- ppoints(100)
q <- quantile(Z, p = p)

plot(qexp(p), q, main = "Exponential Q-Q Plot", xlab = "Theoretical
Quantiles", ylab = "Sample Quantiles")
qqline(q, distribution=qexp, col = "blue")

# Residual Plot
plot(reg_model6$fitted.values, reg_model6$residuals, xlab = "Fitted
Values", ylab = "Residuals", main = "Residual Plot")
abline(a = 0, b = 0)

# Normal Q-Q Plot for Residuals
qqnorm(reg_model6$residuals)
qqline(reg_model6$residuals)
```