7144CEM PRINCIPLES OF DATA SCIENCE

Individual Assignment (Task 2)

Student Name: Sujan Tumbaraguddi
Student ID: 14194733

Contents

Exploratory Data Analysis and Linear Models (MLO2)

```
# R Studio

# 7144CEM Principles of Data Science

# The original dataset has 768 rows (500 'no diabetes' and 268 'diabetes').
# We have taken a subset of 200 'no diabetes' cases and 200 'diabetes' cases.
# 400 observations and 8 variables.

#########################################################################
# INDIVIDUAL TASK -- Exploratory Data Analysis and  Linear Models (MLO2)
#########################################################################


# Install the tidyverse package if not already installed
#install.packages("tidyverse")
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.3     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.4     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
to become errors
```

```
#########################################################
# Uncomment below code when extracting the data sample
#########################################################

# Setting the seed to '733'
#set.seed(733)

# Read Pima Indians Diabetes Dataset as csv file
#data = read_csv('Pima_Indians_Diabetes_Dataset1.csv')
#View(data)

# To take stratified random sampling based on the 'Target' variable
#strata_sample = data %>%
#  group_by(Target) %>%
```

```r
#  slice_sample(n=200) %>%
#  ungroup()
#View(strata_sample)

# Save the obtained stratified sample
#write_csv(strata_sample, 'PI_Diabetes_Datasample.csv')

# Read PI_Diabetes_Datasample Dataset as csv file and convert to Tibble
strata_sample <- as_tibble(read_csv('PI_Diabetes_Datasample.csv'))
```

```
## Rows: 400 Columns: 9
## ── Column specification ────────────────────────────────────────────────────
## Delimiter: ","
## dbl (9): Number of times pregnant, Plasma glucose concentration, Diastolic b...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# Install the GGally package if not already installed
#install.packages("GGally")
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```r
# Renaming the 'Target' variable to more understandable categories
strata_sample$Target <- ifelse(strata_sample$Target == 0, "No Diabetes", "Diabetes")
strata_sample$Target <- as_factor(strata_sample$Target)
```

1.1 Single-predictor linear model

```
############Q No. 01##############

# Creating a scatterplot matrix to visualize relationships among variables
strataSample <- select(strata_sample, -Target)
ggpairs(strataSample)
```

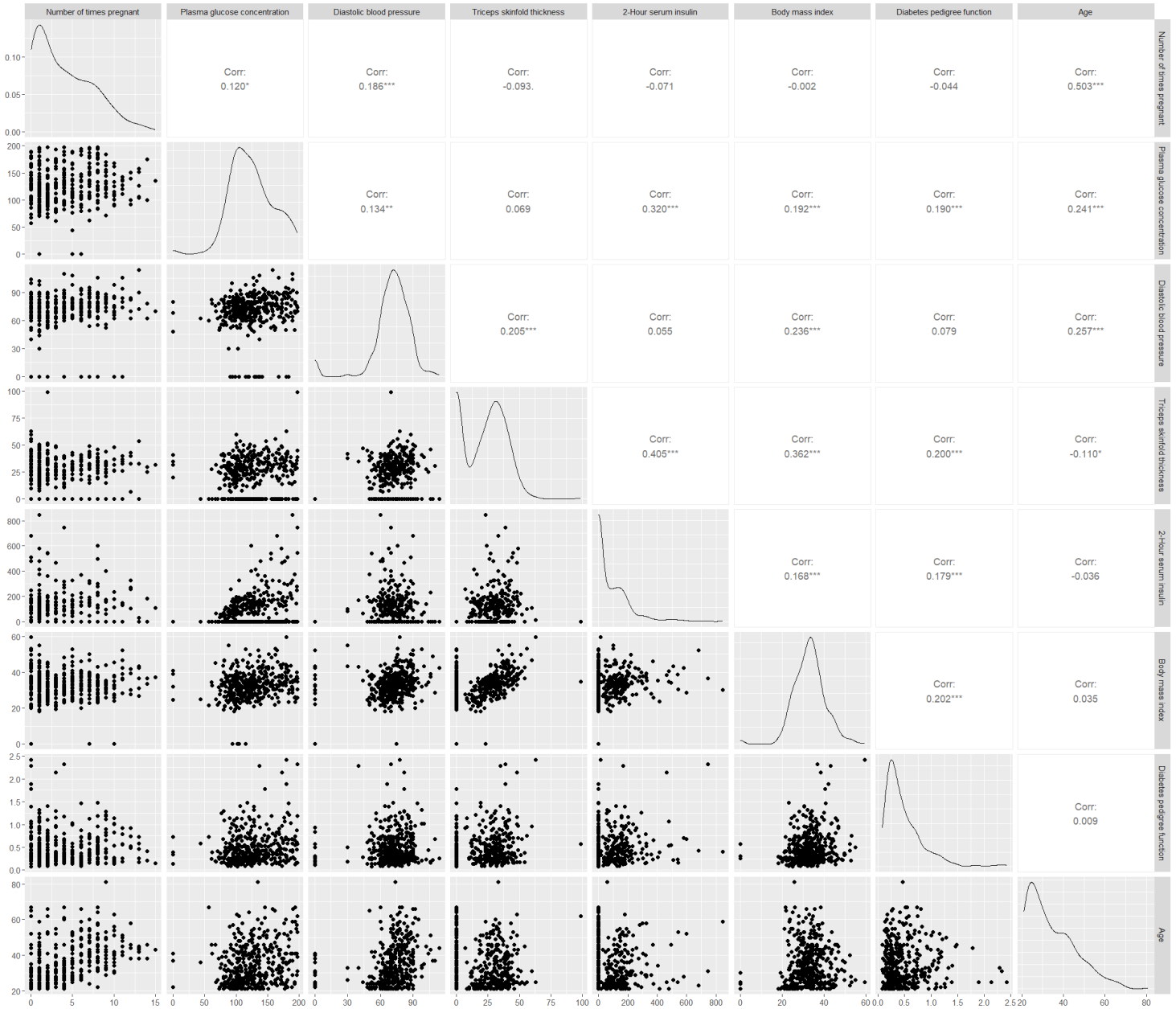

Image 01 – Scatter Plot Matrix for all 8 variables from the sample acquired.

When comparing the correlation values for the response variable, “Diabetes pedigree function” to the other seven variables in the Scatter Plot Matrix, the one with the highest value is “Body mass index” or 0.202, which is a predictor variable. We can conclude that the “Diabetes pedigree function” variable is best predicted by the

“Body mass index” variable. Further, we could consider using a linear regression model to fit these two variables.

```
# Fitting a linear regression model and summarizing its results
x <- strataSample$`Body mass index`
y <- strataSample$`Diabetes pedigree function`
lm_01 = lm(y~x)
summary(lm_01)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47491 -0.24755 -0.09607  0.13469  1.81140
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.165958   0.078629   2.111   0.0354 *
## x           0.009581   0.002324   4.123 4.55e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3517 on 398 degrees of freedom
## Multiple R-squared:  0.04097,    Adjusted R-squared:  0.03856
## F-statistic:    17 on 1 and 398 DF,  p-value: 4.546e-05

AIC(lm_01)

## [1] 303.064
```

Following the fitting of the “Diabetes pedigree function” as response variable and the “Body mass index” as predictor variable into a linear regression model, the following summary was obtained:

1. Residuals: The discrepancies between the observed and anticipated values are represented by residuals. The residuals in this instance vary from -0.47491 to 1.81140.

2. Coefficients: The intercept’s “estimate” is 0.165958, while the body mass index’s is 0.009581. For the specified predictor variable, these estimates show the regression line’s slope and intercept.

3. Significance: The ‘t value’ for the intercept is 2.111, and for “Body mass index” is 4.123. Both coefficients’ “Pr(>|t|)” values are less than 0.05, and the coefficient for the “Body mass index” variable is much lower (4.55e-05), indicating that it has a statistically significant impact on the dependent variable.

All things considered; the model indicates that the “Diabetes pedigree function” variable is statistically significantly impacted by the “Body mass index” variable. It’s crucial to remember that the model’s R-squared value is somewhat low, indicating that the independent variable is not the only factor that can adequately explain the variance in the dependent variable.

1.2 AIC (Akaike Information Criterion)

```
############Q No. 02(a)#############
```

AIC (Akaike Information Criterion) is a statistical tool for evaluating the relative merits of various models and is used for model selection. Models with more parameters are penalized by AIC, which strikes a balance between the trade-off between complexity and goodness of fit. It is calculated using the formula: AIC = -2 * log-likelihood + 2 * number of parameters. The model that fits the data the best is the one with the lowest AIC value, since a lower value denotes a better model. (Akaike, 1974)

1.3 Several-predictor linear models

```
############Q No. 02(b)#############
# MODEL #1:
# Fitting a linear model with 'Diabetes pedigree function' as the response and 'Body mass
index' and 'Diastolic blood pressure' as predictors
lm_02B_1 = lm(data = strataSample, `Diabetes pedigree function` ~ `Body mass index` +
                `Diastolic blood pressure`)
summary(lm_02B_1)
```

```
##
## Call:
## lm(formula = `Diabetes pedigree function` ~ `Body mass index` +
##     `Diastolic blood pressure`, data = strataSample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48025 -0.24675 -0.09943  0.13860  1.81306
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.1330383  0.0932105   1.427 0.154283
## `Body mass index`          0.0092102  0.0023926   3.849 0.000138 ***
## `Diastolic blood pressure` 0.0006413  0.0009734   0.659 0.510398
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3519 on 397 degrees of freedom
## Multiple R-squared:  0.04202,    Adjusted R-squared:  0.03719
## F-statistic: 8.706 on 2 and 397 DF,  p-value: 0.0001993
```

```
AIC(lm_02B_1)
```

```
## [1] 304.6269
```

```
# Install the ggfortify package if not already installed
#install.packages("ggfortify")
library(ggfortify)
```

```r
# Visualizing the model using ggfortify package
autoplot(lm_02B_1)
```

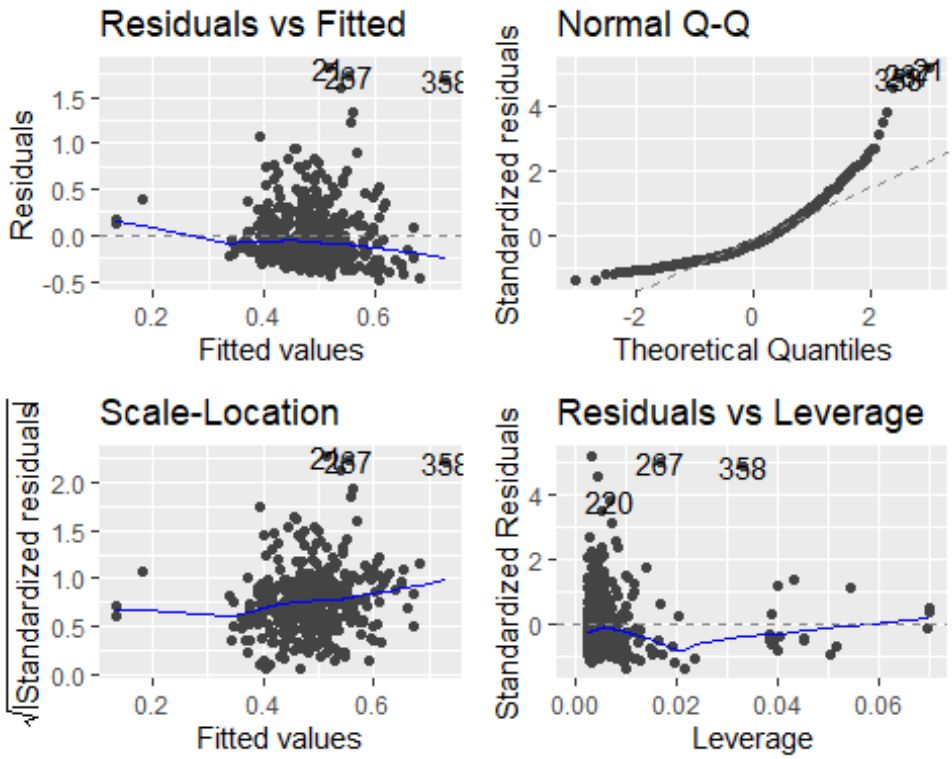

Image 02 – Diagnostic plot of Model #1

```r
# MODEL #2:

# Fitting a linear model with 'Diabetes pedigree function' as the response and 'Body mass index' and 'Triceps skinfold thickness' as predictors
lm_02B_2 = lm(data = strataSample, `Diabetes pedigree function` ~ `Body mass index` +
                `Triceps skinfold thickness`)
summary(lm_02B_2)

##
## Call:
## lm(formula = `Diabetes pedigree function` ~ `Body mass index` +
##     `Triceps skinfold thickness`, data = strataSample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48828 -0.23650 -0.09635  0.13647  1.76675
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   0.182105   0.078179   2.329  0.02034 *
## `Body mass index`             0.007078   0.002472   2.863  0.00441 **
## `Triceps skinfold thickness` 0.003086   0.001104   2.795  0.00544 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3487 on 397 degrees of freedom
## Multiple R-squared:  0.05948,    Adjusted R-squared:  0.05474
## F-statistic: 12.55 on 2 and 397 DF,  p-value: 5.173e-06
```

```
AIC(lm_02B_2)
```

```
## [1] 297.2691
```

```
autoplot(lm_02B_2)
```

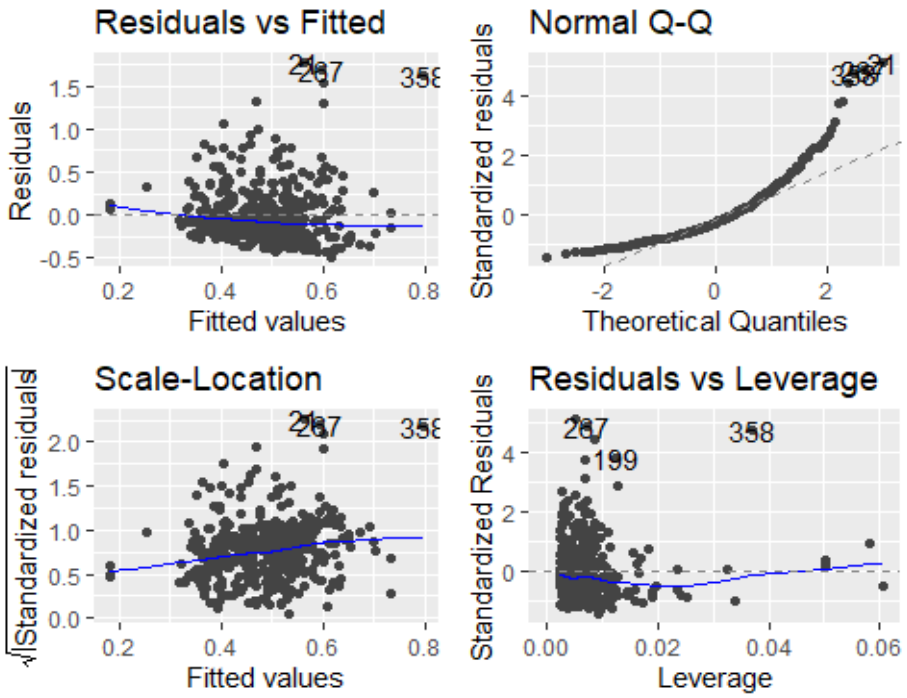

Image 03 – Diagnostic plot of Model #2

```
# MODEL #3:

# Fitting a linear model with 'Diabetes pedigree function' as the response and multiple
predictors including 'Body mass index', 'Diastolic blood pressure', '2-Hour serum
insulin', and 'Plasma glucose concentration'
lm_02B_3 = lm(data = strataSample, `Diabetes pedigree function` ~ `Body mass index` +
                `Triceps skinfold thickness` +
                `2-Hour serum insulin` +
                `Plasma glucose concentration`)
summary(lm_02B_3)
```

```
##
## Call:
## lm(formula = `Diabetes pedigree function` ~ `Body mass index` +
##     `Triceps skinfold thickness` + `2-Hour serum insulin` + `Plasma glucose
```

```
concentration`,
##     data = strataSample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.55721 -0.23578 -0.08914  0.11557  1.68106
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    0.0433680  0.0926211   0.468  0.63988
## `Body mass index`              0.0057518  0.0024839   2.316  0.02109 *
## `Triceps skinfold thickness`   0.0025362  0.0011860   2.138  0.03309 *
## `2-Hour serum insulin`         0.0001858  0.0001567   1.186  0.23637
## `Plasma glucose concentration` 0.0014284  0.0005365   2.662  0.00807 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3445 on 395 degrees of freedom
## Multiple R-squared:  0.08661,    Adjusted R-squared:  0.07736
## F-statistic: 9.363 on 4 and 395 DF,  p-value: 3.074e-07

AIC(lm_02B_3)

## [1] 289.5617

autoplot(lm_02B_3)
```

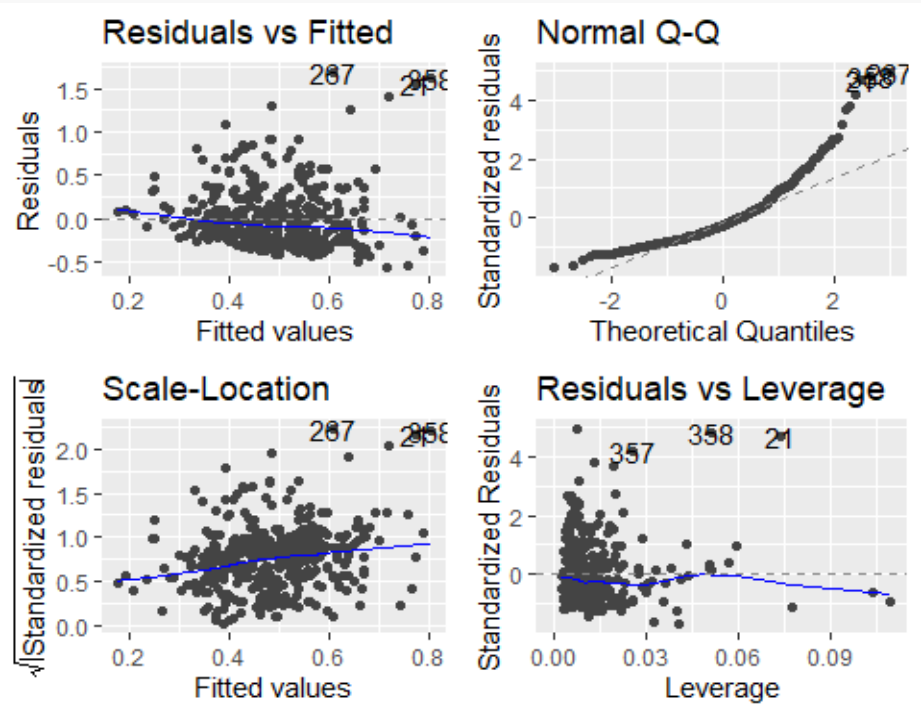

Image 04 – Diagnostic plot of Model #3

Attributes for Comparison	Model #1	Model #2	Model #3

Predictors	`Body mass index`
	`Diastolic blood pressure`
		`Body mass index`
		`Triceps skinfold thickness`
			`Body mass index`
			`Triceps skinfold thickness`
			`2-Hour serum insulin`
			`Plasma glucose concentration`

AIC	304.6269	297.2691	289.5617
Multiple R-squared	0.04202	0.05948	0.08661
Adjusted R-squared	0.03719	0.05474	0.07736
F-statistic	8.706	12.55	9.363
p-value	0.0001993	5.173e-06	3.074e-07

Table 01 – Comparison of Model's

Criteria to pick a "best" linear regression model is as follows:

- AIC must be lower.
- Multiple R-squared must be higher.
- Adjusted R-squared must be higher.
- F-statistic must be higher.
- p-value must be lower.

From the above table and criteria, **Model #3** has the lowest AIC value compared. It has significantly higher 'Multiple R-squared' & 'Adjusted R-squared' values which indicates larger variance in 'Diabetes pedigree function' is explained by the 4 Predictor Variables. Furthermore, the F-statistic value and significantly lower p-value agrees that the **Model #3 is the "best" fit** for the above data sample among the 3 models.

Accessing **Model #3:**

1. Residuals: The residuals have a good distribution because most of them fall between the first and third quantiles. Additionally, the range seems fair, suggesting that there are no major problems with the model.

2. Coefficients: The "Body mass index" predictor variable has a higher estimate than the other three, indicating that it has a greater impact on the response variable. "Triceps skinfold thickness" and "Plasma glucose concentration" come next. "2-Hour serum insulin" has the least impact on the response variable.

3. Significance: 'Plasma glucose concentration' has lowest 'Pr(>|t|)' values which means it has high significance level. Followed by 'Triceps skinfold thickness' & 'Body mass index'. '2-Hour serum insulin' is relatively has the least significant level.

4. Degree of fit: The Multiple R-squared values suggests that 8.661% of variance in 'Diabetes pedigree function' is explained by the four predictor variables. But Adjusted R-squared is lower compared, which imply that the four predictor variables don't fully capture the variance in 'Diabetes pedigree function'.

5. F-statistic & p-value: The higher F-statistic and lower p-value simply suggests that the model is statistically significant.

6. Diagnostic Plot (Image 04):

    6.1. Residuals vs Fitted plot - There is no pattern in the residual plot & the blue line is approximately horizontal at zero, which suggests that we can assume a linear relationship between the predictor and the response variables.

    6.2. Normal Q-Q plot – Majorly, many points are on the Normal distribution line (dashed line), so we can assume normality of residuals. But note that there are few outliers and further analysis is required as to what factors are affecting them.

    6.3. Scale-Location plot - The variability of the residual points is not constant with the value of the fitted response variable, suggesting nonconstant variances in the residuals.

    6.4. Residuals vs Leverage plot – There are around 7 outliers which appear to be extreme cases which are affecting the interpretation of the model. Further, we might want to take a closer look at them individually to check if there is anything special for the subject or if it could be simply data entry errors.

1.4 Summary and Conclusions

Several advantages of Model #3 are highlighted by the assessment, such as the identification of relevant predictors, variable effect prioritization based on coefficients, and a comparatively even distribution of residuals. The model has statistical significance, but it may not be able to adequately capture the variance in the response variable given the lower adjusted R-squared value. Furthermore, the diagnostic plots highlight possible problems like outliers and nonconstant variances in residuals, highlighting the necessity of additional research into these anomalies to guarantee the robustness and dependability of the model.

The following tactics should be taken into consideration in order to properly inform the participants included in the dataset of the conclusions:

- Simplify the English: Avoid using jargon like "adjusted R-squared" or "residuals" and instead use common English to explain the findings.
- Make Use of Visual Aids: To help people understand complex relationships, use charts and graphs to highlight important aspects.
- Give Context: Explain how body mass index and blood sugar levels impact a person's chance of developing diabetes by connecting the research to the participants' everyday lives.
- Highlight Important Factors: To help with understanding, highlight important factors that affect the onset of diabetes using straightforward examples.
- Address Implications and Recommendations: Discuss the significance of the findings and offer doable solutions, such as modifying one's lifestyle or scheduling routine checkups.
- Encourage Questions and Feedback: To improve understanding of the results and their implications, invite attendees to pose questions.
- Utilize Plain-language Summaries: Make Use of Simple Language Brief summaries should be written, with an emphasis on the applications and suggested courses of action.

References

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6), 716-723.