Principles of Data Science

Task 2 – Individual Portfolio

NDUBUAKU EKWUEME

QUESTION 1

```
#INDIVIDUAL TASK

library(broom)
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   3.4.3     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ 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

library(stats)
#install.packages('olsrr')
library(olsrr)

##
## Attaching package: 'olsrr'
##
## The following object is masked from 'package:datasets':
##
##     rivers

library(ggfortify)
library(psych)

##
## Attaching package: 'psych'
##
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha

library(factoextra)

## Welcome! Want to learn more? See two factoextra-related books at
https://goo.gl/ve3WBa

library(nFactors)

## Loading required package: lattice
##
## Attaching package: 'nFactors'
```

```
##
## The following object is masked from 'package:lattice':
##
##     parallel

library(factoextra)
library(cluster)
library(ISLR)
print('Individual Task Part 1')

## [1] "Individual Task Part 1"
```

QUESTION 1

```
set.seed(341)

diabetesdata <- read.csv('Pima_Indians_Diabetes_Dataset1.csv')

# To take stratified random samples
strata_sample = diabetesdata %>%
  group_by(Target) %>%
  slice_sample(n=200) %>%
  ungroup()

write_csv(strata_sample, 'portfoliodata.csv')
read_csv('portfoliodata.csv')

## Rows: 400 Columns: 9
## ── Column specification
────────────────────────────────────────────────────────

## Delimiter: ","
## chr (1): Target
## dbl (8): 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.

## # A tibble: 400 × 9
##    Number.of.times.pregnant Plasma.glucose.concentration
Diastolic.blood.press…¹
##                       <dbl>                        <dbl>
<dbl>
##  1                        0                          123
```

```
72
##  2                        5                          112
66
##  3                        6                          124
72
##  4                        9                          164
84
##  5                        0                          131
88
##  6                        2                          146
70
##  7                        4                          115
72
##  8                        0                          162
76
##  9                        1                          168
88
## 10                        8                          100
74
## # ℹ 390 more rows
## # ℹ abbreviated name: ¹Diastolic.blood.pressure
## # ℹ 6 more variables: Triceps.skinfold.thickness <dbl>,
## #   X2.Hour.serum.insulin <dbl>, Body.mass.index <dbl>,
## #   Diabetes.pedigree.function <dbl>, Age <dbl>, Target <chr>
```

```
diabetesdata <- read_csv('portfoliodata.csv')
```

```
## Rows: 400 Columns: 9
## ── Column specification
────────────────────────────────────────────────────────
## Delimiter: ","
## chr (1): Target
## dbl (8): 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.
```

```
diabetes_data_bi = select(diabetesdata,-Target)

#install.packages("GGally")
library(GGally)
```

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

```
ggpairs(diabetesdata, aes(color=diabetesdata$Target))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

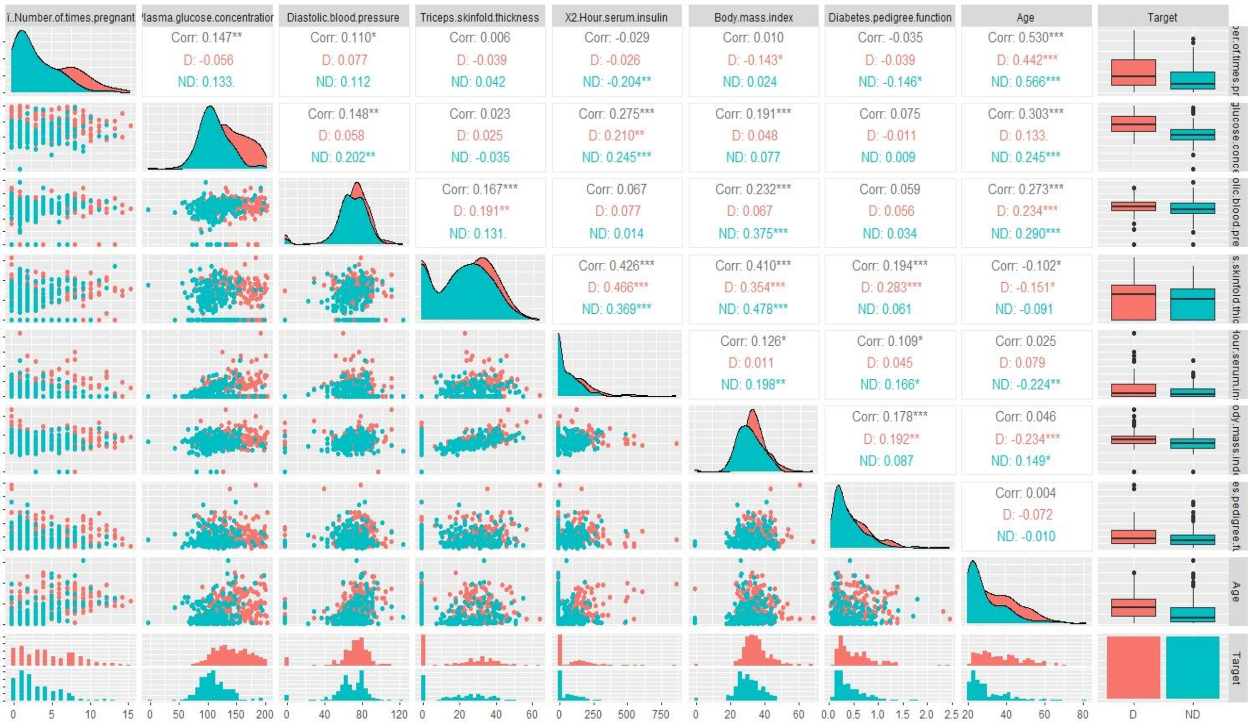

Figure 1: Scatterplot

Key Observations:

- **Age** and **Number of Times Pregnant** have a moderate correlation **(0.53)**. This relationship indicates older individuals tend to have more pregnancies.
- **Triceps Skinfold Thickness** shows the strongest relationship (0.194) with **Diabetes Pedigree Function**. Although weak, it is the strongest predictor in this dataset.

Why Triceps Skinfold Thickness?

- This variable was selected for the regression model due to its stronger correlation compared to others. Despite the low value, it captures a notable trend.

```
ggplot(diabetesdata,aes(x=Triceps.skinfold.thickness,y=Diabetes.pedigree.func
tion))                                                                        +
  geom_point(aes(colour=Target))
```

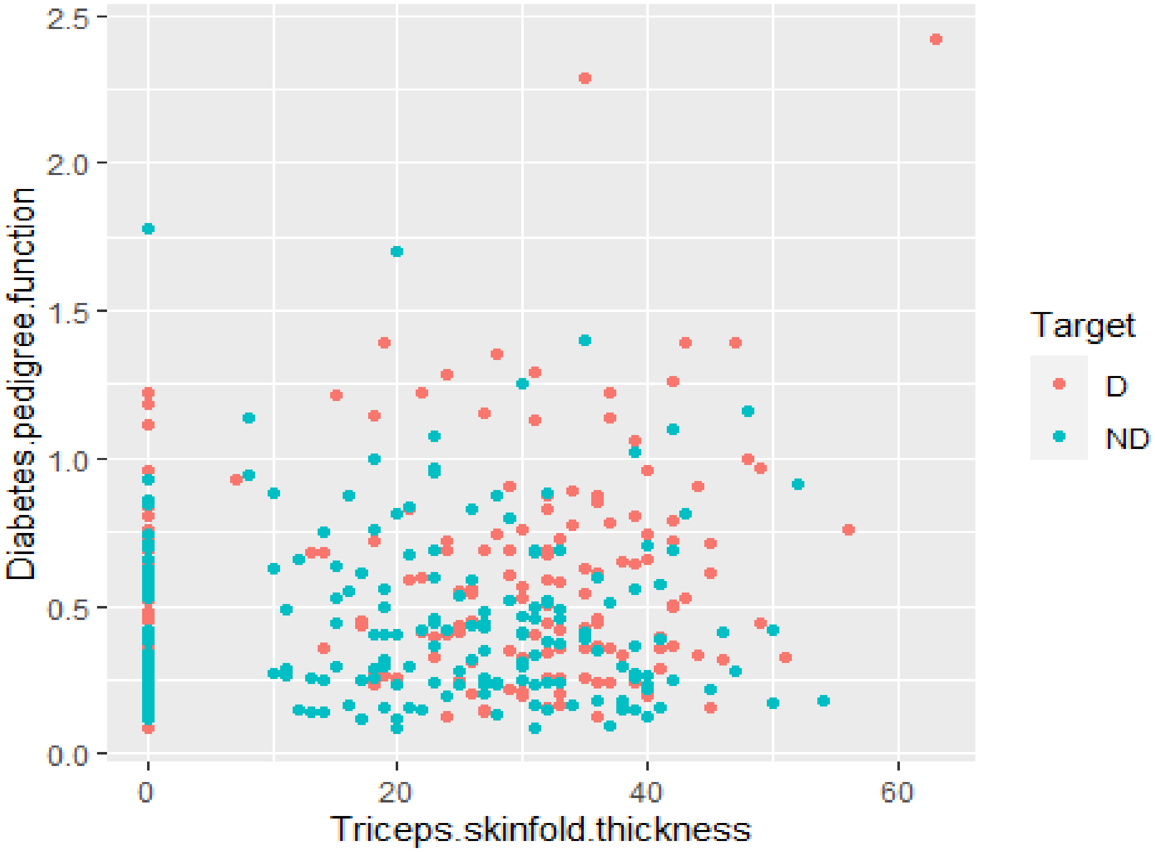

Figure 2

```
ekwuemedata = diabetes_data_bi
names(ekwuemedata) = c("Times_pregnant", "Plasma_GC", "Diastolic_BP",
"Tricep_ST",
                       "X2HSI", "BMI", "Diabetes_PF", "Age")

colnames(ekwuemedata)

## [1] "Times_pregnant" "Plasma_GC"      "Diastolic_BP"    "Tricep_ST"
## [5] "X2HSI"          "BMI"            "Diabetes_PF"     "Age"

#To get the best predictor linear models of diabetes pedigree function
model0 = lm(Diabetes_PF~Times_pregnant+Plasma_GC+Diastolic_BP+Tricep_ST +
X2HSI + BMI+ Age,data = ekwuemedata)
summary(model0)
```

```
##
## Call:
## lm(formula = Diabetes_PF ~ Times_pregnant + Plasma_GC + Diastolic_BP +
##     Tricep_ST + X2HSI + BMI + Age, data = ekwuemedata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49204 -0.23123 -0.09349  0.15961  1.67996
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.947e-01  1.005e-01   1.937   0.0535 .
## Times_pregnant -6.010e-03  5.861e-03  -1.025   0.3058
## Plasma_GC       4.760e-04  5.763e-04   0.826   0.4094
## Diastolic_BP   -9.338e-06  9.183e-04  -0.010   0.9919
## Tricep_ST       3.049e-03  1.290e-03   2.363   0.0186 *
## X2HSI           5.498e-05  1.683e-04   0.327   0.7441
## BMI             4.680e-03  2.459e-03   1.903   0.0577 .
## Age             9.346e-04  1.834e-03   0.510   0.6106
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3292 on 392 degrees of freedom
## Multiple R-squared:  0.05494,    Adjusted R-squared:  0.03807
## F-statistic: 3.256 on 7 and 392 DF,  p-value: 0.002261
```

```
ols_step_best_subset(model)
```

```
##                           Best Subsets Regression
## ----------------------------------------------------------------------------
--
## Model Index    Predictors
## ----------------------------------------------------------------------------
--
##      1         Tricep_ST
##      2         Tricep_ST BMI
##      3         Plasma_GC Tricep_ST BMI
##      4         Times_pregnant Plasma_GC Tricep_ST BMI
##      5         Times_pregnant Plasma_GC Tricep_ST BMI Age
##      6         Times_pregnant Plasma_GC Tricep_ST X2HSI BMI Age
##      7         Times_pregnant Plasma_GC Diastolic_BP Tricep_ST X2HSI BMI
Age
## ----------------------------------------------------------------------------
--
##
##                                                      Subsets Regression
Summary
```

```
## --------------------------------------------------------------------------
------------------------------------------------------
##                        Adj.        Pred
## Model    R-Square    R-Square    R-Square     C(p)       AIC         SBIC
SBC        MSEP       FPE       HSP      APC
## --------------------------------------------------------------------------
------------------------------------------------------
##   1        0.0378      0.0354      0.0261    3.1138    251.4073    -
883.7347    263.3817    43.4721    0.1092    3e-04    0.9719
##   2        0.0494      0.0447      0.0317    0.2825    248.5357    -
886.5286    264.5015    43.0543    0.1084    3e-04    0.9649
##   3        0.0520      0.0448        0.03    1.2360    249.4726    -
885.5412    269.4299    43.0487    0.1087    3e-04    0.9672
##   4        0.0540      0.0444      0.0265    2.3966    250.6179    -
884.3400    274.5666    43.0658    0.1090    3e-04    0.9700
##   5        0.0547      0.0427      0.0226    4.1072    252.3227    -
882.5874    280.2630    43.1436    0.1095    3e-04    0.9741
##   6        0.0549      0.0405      0.0166    6.0001    254.2135    -
880.6525    286.1452    43.2418    0.1100    3e-04    0.9787
##   7        0.0549      0.0381      0.0126    8.0000    256.2134    -
878.6118    292.1365    43.3524    0.1105    3e-04    0.9836
## --------------------------------------------------------------------------
------------------------------------------------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria



# best single predictor model of diabetes pedigree function
model = lm(Diabetes_PF~Tricep_ST, data=ekwuemedata)
summary(model) # R-squared = 0.03779



##
## Call:
## lm(formula = Diabetes_PF ~ Tricep_ST, data = ekwuemedata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45391 -0.22027 -0.09442  0.16310  1.76407
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.399031   0.027129  14.709  < 2e-16 ***
## Tricep_ST   0.004078   0.001031   3.954  9.1e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3297 on 398 degrees of freedom
## Multiple R-squared:  0.03779,    Adjusted R-squared:  0.03538
## F-statistic: 15.63 on 1 and 398 DF,  p-value: 9.103e-05

AIC(model)     # AIC value = 251.4073

## [1] 251.4073

a= model$coefficients[1] #(Intercept) 0.3990308
b = model$coefficients[2]# Tricep_ST 0.004077741
print(a)

## (Intercept)
##   0.3990308

print(b)

##   Tricep_ST
## 0.004077741
```

```
# fitted regression model
# Diabetes_PF = 0.399 + 0.004078*Tricep_ST
# the r-square value of 0.03779 only explains 3.8% of the variance of
Diabetes pedigree function
```

QUESTION 2A

AIC also known as Akaike information criterion (AIC) is a mathematical method for evaluating how well a model fits the data it was generated from. In statistics, AIC is used to compare different possible models and determine which one is the best fit for the data. AIC is calculated from:

- the number of independent variables used to build the model.
- the maximum likelihood estimate of the model (how well the model reproduces the data). (Bevans, 2023).

QUESTION 2B

MODEL 1

```
# Model 1: best predictor model between Diastolic blood pressure and body
mass index of diabetes pedigree function
model_1a = lm(Diabetes_PF~Diastolic_BP+BMI, data=ekwuemedata)
summary(model_1a) # R-squared = 0.03203

##
## Call:
## lm(formula = Diabetes_PF ~ Diastolic_BP + BMI, data = ekwuemedata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49240 -0.24328 -0.08911  0.15573  1.73532
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.2115292  0.0854291   2.476 0.013699 *
## Diastolic_BP 0.0003180  0.0008804   0.361 0.718109
## BMI          0.0076201  0.0022258   3.424 0.000682 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3311 on 397 degrees of freedom
## Multiple R-squared:  0.03203,    Adjusted R-squared:  0.02715
## F-statistic: 6.568 on 2 and 397 DF,  p-value: 0.001562

AIC(model_1a) # AIC value = 255.7964

## [1] 255.7964

a= model_1a$coefficients[1] #(Intercept) 0.2115292
b = model_1a$coefficients[2] #Diastolic_BP 0.0003180339
c = model_1a$coefficients[3] #BMI 0.007620086
print(a)

## (Intercept)
##   0.2115292

print(b)

## Diastolic_BP
## 0.0003180339

print(c)

##         BMI
## 0.007620086

#Diabetes_PF = 0.2115 + 0.00032*Diastolic_BP + 0.007620086*BMI
# the r-square value of 0.03203 only explains 3.2% of the variance of
```

```
Diabetes pedigree function
autoplot(model_1a)
```

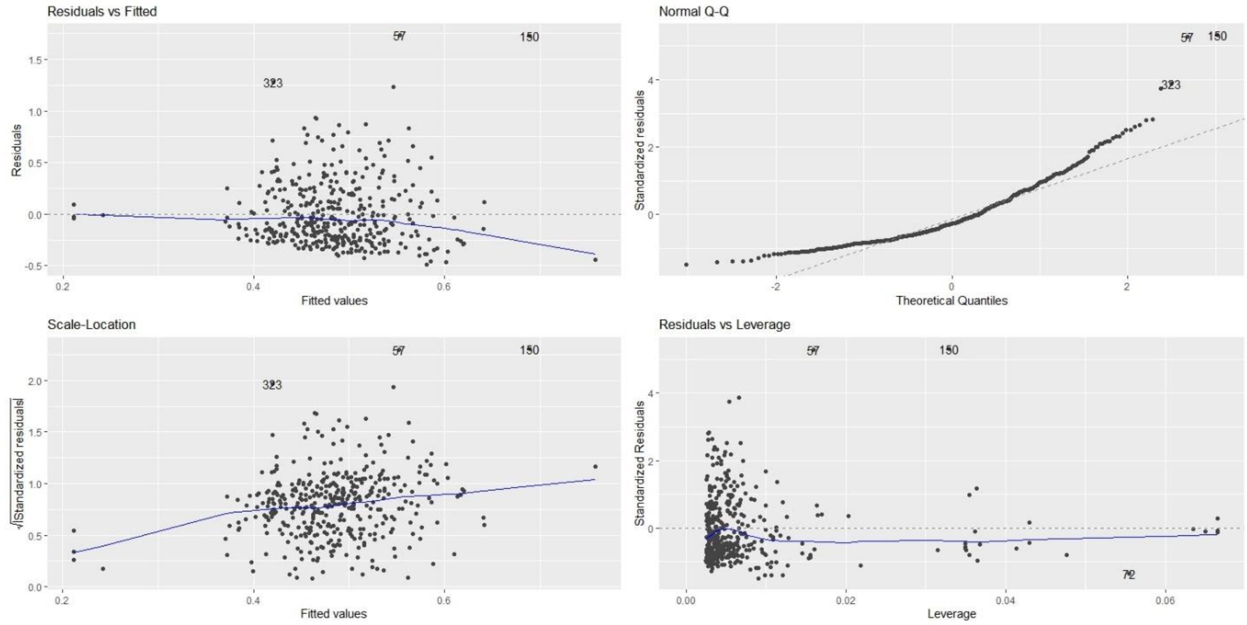

Figure 3

From figure 3 (Model 1):

Residual vs fitted: The "Residuals vs Fitted" plot (top left) checks the linear relationship assumption between **Triceps Skinfold Thickness** and **Diabetes Pedigree Function**. Ideally, the residuals should show no clear pattern around the horizontal blue line. In this case, the residuals exhibit a mild curvature, suggesting that the linear model does not fully capture the relationship. This curvature indicates that a non-linear relationship may exist between the predictor and the response variable.

Normal Q-Q: The "Normal Q-Q" plot (top right) is used to assess whether the residuals follow a normal distribution. Ideally, the residuals should align along the dashed line. However, in this plot, there are significant deviations at the tails, which indicate the residuals are not normally distributed. This lack of normality suggests that the model struggles to explain extreme values in the data.

Scale-Location: The "Scale-Location" plot (bottom left) is used to check for homoscedasticity, or equal variance of residuals across fitted values. Here, the points appear slightly more spread out as the fitted values increase, and the blue line shows a slight upward trend. This indicates mild heteroscedasticity, where the variance of residuals increases with larger predictions.

Residuals vs Leverage: The “Residuals vs Leverage” plot (bottom right) is used to identify influential observations. Observations with high leverage can disproportionately influence the regression model. In this plot, point **323** has a higher leverage value, indicating it could have a notable impact on the model’s results. However, there are fewer extreme points overall compared to more complex models.

In summary, Model 1 explains only **3.8%** of the variability in Diabetes Pedigree Function (R² = 0.0378), which indicates a weak relationship between **Triceps Skinfold Thickness** and the response variable. The diagnostic plots reveal issues with non-linearity, non-normality, and mild heteroscedasticity, suggesting that a single predictor is insufficient to explain the observed variability in the data

MODEL2

```
# Model 2: best two predictor model of diabetes pedigree function
model2 = lm(Diabetes_PF~Tricep_ST+BMI, data=ekwuemedata)
summary(model2) # R-squared = 0.0494
##
## Call:
## lm(formula = Diabetes_PF ~ Tricep_ST + BMI, data = ekwuemedata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49005 -0.22521 -0.09342  0.16257  1.70759
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.249720   0.072882   3.426 0.000676 ***
## Tricep_ST   0.003061   0.001125   2.721 0.006791 **
## BMI         0.005186   0.002351   2.206 0.027985 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3281 on 397 degrees of freedom
## Multiple R-squared:  0.04944,    Adjusted R-squared:  0.04465
## F-statistic: 10.32 on 2 and 397 DF,  p-value: 4.254e-05

AIC(model2)     #AIC value = 248.5357
```

```
## [1] 248.5357

a= model2$coefficients[1] #(Intercept) 0.2497198
b = model2$coefficients[2] # Tricep_ST 0.003061499
c = model2$coefficients[3] #BMI 0.005186392
print(a)

## (Intercept)
##   0.2497198

print(b)

##   Tricep_ST
## 0.003061499

print(c)

##         BMI
## 0.005186392

# y = a + b*x1 + c*x2...+ e

#Diabetes_PF = 0.2497 + 0.0031*Tricep_ST + 0.0052*BMI + 1.7076
# the r-square value of 0.03203 only explains 3.2% of the variance of
Diabetes pedigree function
autoplot(model2)
```

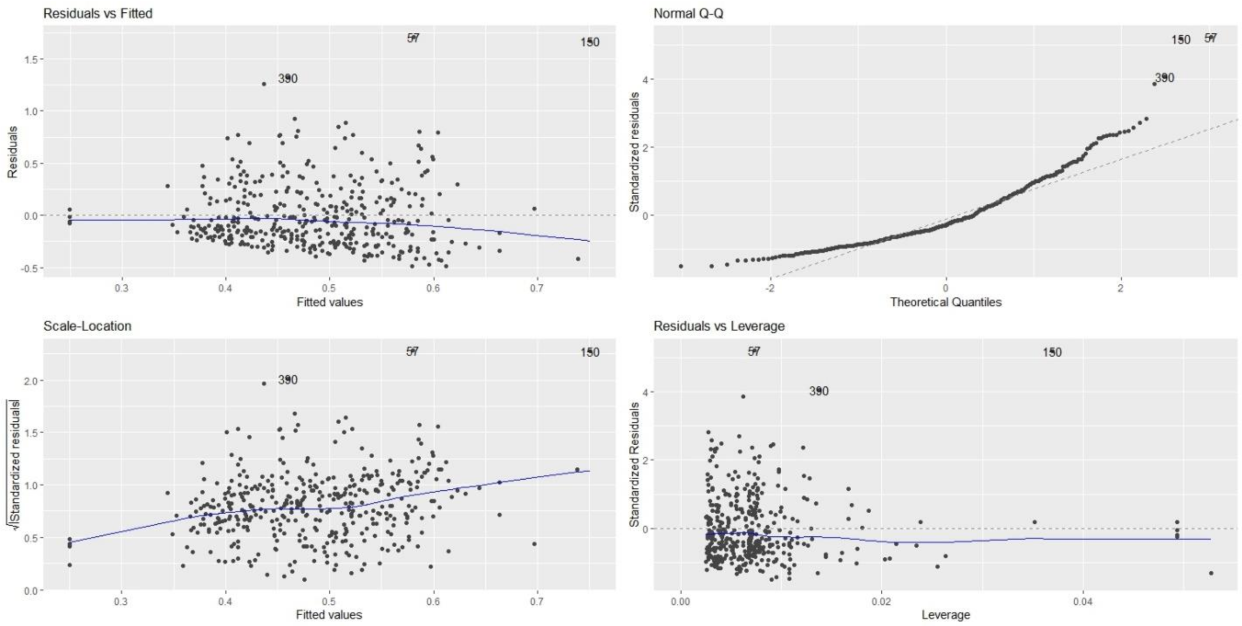

Figure 4

From figure 4:

Residual vs fitted: The "Residuals vs Fitted" plot (top left) is used to check the linear relationship assumption. Ideally, the plot will show no fitted pattern, this means that the blue line should be approximately horizontal at zero.

In this plot, there is no pattern in the residual plot, which suggests that we can assume a linear relationship between the predictor and the response variables.

Normal Q-Q: The "Normal Q-Q" plot (top right) is used to visually check whether the residuals are approximately normally distributed. The points should approximately follow a straight line.

In this plot, the points do not fall along the dashed line so we can assume abnormality of residuals.

Scale-location: The "Scale-Location" plot (bottom left) is used to check the homogeneity of variance of the residuals (homoscedasticity).

In this plot, the line doesn't always follow a straight path across the plot, but it never veers off course. Also, the points are equally spread out therefore, that the assumption of equal variance is upheld.

Residual vs leverage: The "Residuals vs Leverage" plot (bottom right) is used to identify influential cases, that is extreme values that might influence the regression results when included or excluded from the analysis. These would indicate lack of independence of residuals. Observations whose standardized residuals are greater than 3 or less than −3 are possible outliers.

Points 323, 57, 150 are potential influential points. Leverage values indicate they could have a strong impact on the model.

MODEL 3

```
# Model 3: best four predictor model of diabetes pedigree function
model3 = lm(Diabetes_PF~Times_pregnant + Plasma_GC + Tricep_ST + BMI,
data=ekwuemedata)
summary(model3) # R-squared = 0.05399
```

```
##
## Call:
## lm(formula = Diabetes_PF ~ Times_pregnant + Plasma_GC + Tricep_ST +
##     BMI, data = ekwuemedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4967 -0.2318 -0.1023  0.1596  1.6873
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.2068606  0.0899617   2.299  0.02200 *
## Times_pregnant -0.0045585  0.0049592  -0.919  0.35855
## Plasma_GC       0.0006088  0.0005286   1.152  0.25016
## Tricep_ST       0.0031443  0.0011274   2.789  0.00554 **
## BMI             0.0046498  0.0024001   1.937  0.05342 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3281 on 395 degrees of freedom
## Multiple R-squared:  0.05399,    Adjusted R-squared:  0.04441
## F-statistic: 5.636 on 4 and 395 DF,  p-value: 0.0002025
```

```
AIC(model3)     #AIC value =  250.6179
```

```
## [1] 250.6179
```

```
a = model3$coefficients[1] #(Intercept) 0.2068606
b = model3$coefficients[2] # Times_pregnant -0.004558482
c = model3$coefficients[3] #Plasma_GC 0.0006087658
d = model3$coefficients[4] #Tricep_ST 0.003144331
e = model3$coefficients[5] #BMI 0.004649767
print(a)
```

```
## (Intercept)
##   0.2068606
```

```
print(b)
```

```
## Times_pregnant
##   -0.004558482
```

```
print(c)
```

```
##    Plasma_GC
## 0.0006087658
```

```
print(d)
```

```
##   Tricep_ST
## 0.003144331
```

```
print(e)

##         BMI
## 0.004649767

#fitted regression model

#Diabetes_PF = 0.2069 + (-0.00456)*Times_pregnant +0.0006*Plasma_GC +
0.0031*Tricep_ST + 0.0046*BMI + 1.6873

autoplot(model3)
```

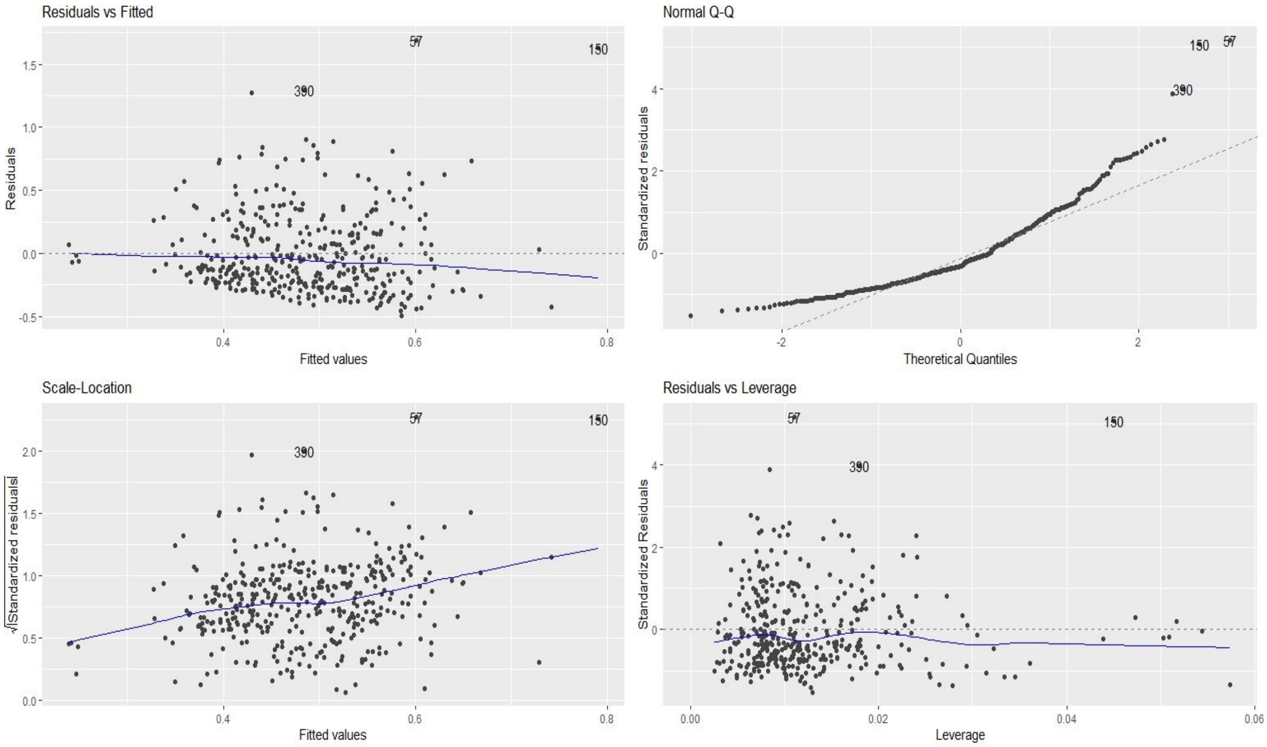

Figure 5.

From figure 5 (Model 3):

Residual vs fitted: The “Residuals vs Fitted” plot (top left) for Model 3 examines the linear relationship between the four predictors (**Times Pregnant, Plasma Glucose, Triceps Skinfold Thickness, BMI**) and the response variable. In this plot, the residuals exhibit a scattered pattern but with a more noticeable curvature, indicating the model still does not fully meet the linearity assumption. This suggests that although adding predictors improves the model slightly, there are still non-linear trends that the model cannot capture

Normal Q-Q: The "Normal Q-Q" plot (top right) is used to assess the normality of residuals. Ideally, the residuals should align along the dashed line. In this plot, the residuals deviate significantly at the upper and lower ends, particularly at the tails. This deviation suggests that the residuals are not normally distributed, which may affect the accuracy of predictions for extreme values.

Scale-location: The "Scale-Location" plot (bottom left) evaluates the homoscedasticity assumption. Here, the blue line shows an upward trend, and the residuals become increasingly dispersed as fitted values increase. This indicates the presence of mild heteroscedasticity, where the variance of residuals is not constant across the range of predictions. The increasing spread suggests that the model struggles to generalize for larger fitted values.

Residual vs leverage: The "Residuals vs Leverage" plot (bottom right) identifies influential data points that may disproportionately impact the regression results. In this plot, points **390**, **57**, and **150** have high leverage values and stand out as potential influential observations. These points suggest that the model is sensitive to specific extreme values in the dataset, which may introduce bias into the results.

In summary, Model 3 explains **5.4%** of the variability in Diabetes Pedigree Function (R^2 = 0.0540), which is slightly higher than Model 1 but still quite low. The diagnostic plots indicate that adding more predictors slightly improves the model fit but introduces additional challenges, such as increased sensitivity to influential points and persistent heteroscedasticity. The model still struggles with normality and linearity assumptions, highlighting the complexity of the relationships in the dataset.

Model	R_Squared	AIC	Comment
Model 1	0.0378	251.41	Single predictor (Tricep_ST)
Model 2	0.0494	248.54	Two predictors (Tricep + BMI)
Model 3	0.0540	250.62	Four predictors, higher AIC

Table 1: Comparison of Model Results

Critical Comparison of Model Results

From table 1, Model 3 achieves the highest **R^2** value (0.0540), indicating that it explains the largest proportion of the variability in the observed data compared to Model 1 (0.0378) and Model 2 (0.0494). However, the increase in R^2 is marginal and comes at the cost of added complexity, as reflected in the **AIC** (250.62), which is higher than that of Model 2 (248.54).

R^2 measures how well the model explains the observed data, while the **Akaike Information Criterion (AIC)** balances model fit with model complexity. A lower AIC indicates better predictive performance on new, unseen data by penalizing excessive model complexity.

In this case, **Model 2** strikes a better balance between goodness of fit and simplicity. Although Model 3 has a slightly higher R^2, the corresponding increase in AIC suggests that the additional predictors do not significantly improve the model's predictive accuracy. In fact, adding more predictors often leads to **overfitting**, where the model performs well on the existing data but poorly on new data.

Conclusion

While Model 3 explains a slightly greater amount of variance, Model 2 provides a more optimal trade-off between model simplicity and predictive performance. With a lower AIC (248.54), **Model 2** is likely the superior option for generalization to new data while maintaining interpretability.

.

QUESTION 3

VARIABLES	PCA	FACTOR ANALYSIS	COEFFICIENT OF LINEAR MODEL
Number of times pregnant	0.2567	0.354	-0.00601
Plasma Glucose Concentration	0.3922	0.405	0.000476
Diastolic Blood Pressure	0.3707	0.240	-0.00000933
Triceps Skinfold Thickness	0.4025	0.225	0.003049
Body Mass Index	0.4199	0.122	0.004680
X2 hour serum insulin	0.3726	0.709	0.00005498
Age	0.3353	0.719	0.0009346

Table 2

Key Findings from Linear Models

1. **Linear Models**:
   - Three regression models were developed to predict **Diabetes Pedigree Function (DPF)**:
     - **Model 1**: Single predictor (Triceps Skinfold Thickness).
     - **Model 2**: Two predictors (Triceps Skinfold Thickness + BMI).
     - **Model 3**: Four predictors (Times Pregnant, Plasma Glucose, Triceps Skinfold Thickness, BMI).
   - **Results**:
     - All models had relatively low R^2 values (3.7% to 5.4%), indicating that only a small portion of the variance in DPF is explained by these predictors.

        - **Triceps Skinfold Thickness** and **BMI** emerged as the most important predictors of DPF.
        - Adding more variables improved the fit slightly but did not significantly increase the explained variance.

2. **Critical Assessment**:
    - Linear models showed weak predictive power, likely due to:
        - High variability in the dataset.
        - Potential multicollinearity among predictors.
        - Non-linear relationships that linear models cannot capture effectively.
    - This suggests that **Triceps Skinfold Thickness** and **BMI**, while important, are not the sole determinants of Diabetes Pedigree Function.

Comparison with FA, PCA, and Cluster Analysis

1. PCA (Principal Component Analysis):

- **Results**:
    - The first three principal components (PC1, PC2, PC3) explained ~58.9% of the total variance.
    - Variables like **BMI**, **Triceps Skinfold Thickness**, and **Plasma Glucose Concentration** contributed most to PC1 and PC2.
- **Comparison**:
    - PCA highlighted the same key variables as the linear models (**BMI** and **Triceps**), confirming their importance.
    - However, PCA accounts for more variance by transforming correlated predictors into independent components, while regression models focus on fitting a straight-line relationship.

2. FA (Factor Analysis):

- **Results**:
    - FA identified **three latent factors**:
        - **Factor 1**: Obesity (BMI, Triceps Skinfold Thickness).
        - **Factor 2**: Insulin Resistance (Serum Insulin, Plasma Glucose).

        - **Factor 3**: Age and Number of Pregnancies.
- **Comparison**:
    - FA grouped variables into meaningful categories that provide deeper medical insights.
    - While linear models highlight individual predictors (BMI and Triceps), FA captures broader trends such as **obesity** and **insulin resistance** as key factors influencing diabetes risk.

3. Cluster Analysis:

- **Results**:
    - Hierarchical clustering revealed groups of participants based on **Euclidean distance** and **Manhattan distance**:
        - Clusters reflected variable similarities, such as groups with higher BMI or higher Plasma Glucose values.
- **Comparison**:
    - Cluster analysis aligns with PCA and FA by identifying patterns among participants (e.g., groups with higher BMI and Triceps Skinfold Thickness).
    - Linear models, however, focus only on the relationship between predictors and DPF without grouping participants.

<u>How I could communicate my conclusions to the participants represented in my dataset</u>

What I Found:

1. The study looked at factors like **body mass index (BMI)**, **skin thickness**, and **age** to see how they relate to diabetes risk.

2. **BMI** (weight relative to height) and **Triceps Skinfold Thickness** (a measure of body fat) were the most important factors linked to diabetes risk.

3. While these factors were helpful, they do not tell the full story. Other things like **insulin levels** and **age** also play a role.

Group Patterns:

1. People with higher **BMI** and **Triceps Skinfold Thickness** tended to have a higher risk of diabetes. This was consistent across all methods we used (PCA, Factor Analysis, and Clustering).
2. We also saw that participants could be grouped into patterns:
   - Some had higher weight and blood sugar levels.
   - Others showed different risk factors like age or insulin levels.

What This Means for You:

- Maintaining a healthy weight and body fat level (as measured by BMI and Triceps Skinfold Thickness) can lower diabetes risk.
- Other factors, like age and insulin levels, also need attention. It's important to get regular check-ups to monitor these.

Why This Matters:

- Even though our models couldn't perfectly predict diabetes risk, they give us clues about which factors are important.
- These results can help people focus on lifestyle changes (like diet and exercise) and understand their personal health risks.

REFERENCES

Bevans, R. 2023. *Akaike Information Criterion | When & How to Use It (Example).* Scribbr. https://www.scribbr.com/statistics/akaike-information-criterion/

Understanding diagnostic plots for linear regression Analysis | UVA Library. (n.d.). https://library.virginia.edu/data/articles/diagnostic-plots