7144CEM – PRINCIPLES OF DATA SCIENCE

TASK 2- INDIVIDUAL TASK

Submitted by:

RAMANA Kulanthaivelu

Student ID: 14231913

Contents:

1. Introduction………………………….. 3
2. Scatter matrix and single predictor…... 3

Linear model

1. a. AIC…………………………………. 8

b. comparisons of models…………….. 9

4. Conclusion……………………………. 15

5. Reference……………………………… 16

Introduction:

This assessment aims to recognize the variables and their interaction with the 'Diabetes pedigree function'. In this assessment, we are understanding them through linear modeling and exploratory data analysis. By analyzing the data set and comparing the various linear models, the best predictor for diabetes pedigree function is identified. The data correlations are displayed first, followed by the analysis of single-predictor models. Several linear models are then constructed and evaluated using AIC.

The link for the entire code for the exploratory analysis and linear models,

<https://github.com/Ramana-Kulanthaivelu/Exploratory-data-analysis-and-linear-models>

1.Scatter matrix and single predictor linear model

The data sample file used for analysing the data in the group portfolio has been taken.

#Task 2 (Individual). Exploratory Data Analysis and Linear Models (MLO2)  
  
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

#data = read.csv("C:/Users/Admin/Pictures/Saved Pictures/Pima\_Indians\_Diabetes\_Dataset1.csv")  
# To take stratified random samples  
#set.seed(733)  
#strata\_sample = data %>%   
# group\_by(Target) %>%   
# slice\_sample(n=200) %>%   
# ungroup()  
# The data sample has been taken and stored in the folder for group and individual tasks.  
strata\_sample<-read\_csv("C:/Users/Admin/Downloads/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.

View(strata\_sample)  
  
#install.packages("tidyverse")  
#install.packages("factoextra")  
#install.packages("psych")  
#install.packages("GGally")  
#install.packages("olsrr")  
#Task 2 Exploratory Data Analysis and Linear Models (MLO2)  
# Task 2.1  
library(GGally)

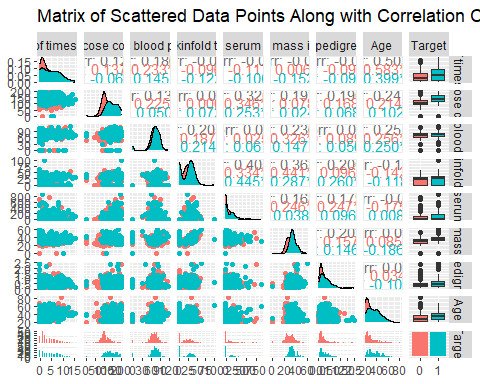
## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

# plotting a scatter matrix using ggpairs  
str(strata\_sample)

## spc\_tbl\_ [400 × 9] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ Number of times pregnant : num [1:400] 2 2 1 1 3 3 1 0 2 1 ...  
## $ Plasma glucose concentration: num [1:400] 111 106 119 79 96 128 100 105 108 0 ...  
## $ Diastolic blood pressure : num [1:400] 60 56 88 60 78 78 66 68 62 48 ...  
## $ Triceps skinfold thickness : num [1:400] 0 27 41 42 39 0 15 22 32 20 ...  
## $ 2-Hour serum insulin : num [1:400] 0 165 170 48 0 0 56 0 56 0 ...  
## $ Body mass index : num [1:400] 26.2 29 45.3 43.5 37.3 21.1 23.6 20 25.2 24.7 ...  
## $ Diabetes pedigree function : num [1:400] 0.343 0.426 0.507 0.678 0.238 0.268 0.666 0.236 0.128 0.14 ...  
## $ Age : num [1:400] 23 22 26 23 40 55 26 22 21 22 ...  
## $ Target : num [1:400] 0 0 0 0 0 0 0 0 0 0 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. `Number of times pregnant` = col\_double(),  
## .. `Plasma glucose concentration` = col\_double(),  
## .. `Diastolic blood pressure` = col\_double(),  
## .. `Triceps skinfold thickness` = col\_double(),  
## .. `2-Hour serum insulin` = col\_double(),  
## .. `Body mass index` = col\_double(),  
## .. `Diabetes pedigree function` = col\_double(),  
## .. Age = col\_double(),  
## .. Target = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

strata\_sample$Target<-as\_factor(strata\_sample$Target)  
ggpairs(strata\_sample,columns = 1:9,aes(color=Target)) +  
 ggtitle ("Matrix of Scattered Data Points Along with Correlation Coefficients")

## `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 01**: Plotting the scattered data points along with correlation coefficient

In the strata sample, the target variable has been converted into a categorical variable rather than being treated as continuous values. The Scatter matrix is plotted using the ‘GGally’ package, In this scatter matrixIn the strata sample, the target variable has been converted into a categorical variable rather than being treated as continuous values.

Based on the scatter matrix observation, we can identify variables that are highly correlated. The 'Body mass index' has a strong correlation with the 'Diabetes pedigree function' , with a correlational coefficient of 0.202. This is followed by the 'Triceps skin thickness', which has a correlational coefficient of 0.200.

# Create a linear model using the appropriate variable name.  
x<-strata\_sample$`Body mass index`  
y<-strata\_sample$`Diabetes pedigree function`  
dpf\_lm<- lm(y~x)   
summary(dpf\_lm)

##   
## 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

# predicting the best predictor  
correlation\_btw\_Diabetes\_pedigree\_function <- cor(strata\_sample[, 1:6], strata\_sample$`Diabetes pedigree function`)  
best\_predictor <- rownames(correlation\_btw\_Diabetes\_pedigree\_function)[which.max(abs(correlation\_btw\_Diabetes\_pedigree\_function))]  
cat(" The Best Predictor for the Diabetes Pedigree Function is", best\_predictor, "\n")

## The Best Predictor for the Diabetes Pedigree Function is Body mass index

The summary output of the linear model provides valuable information about the model's accuracy. It includes data on coefficients, standard errors, t-values, p-values, residual statistics, and R-squared values.

**The residual standard error : 0.3517**

It indicates a more precise estimate in linear regression, assuming it's small relative to the range of the dependent variable and the nature of the dataset.

**The adjusted R-squared value**

The adjusted R\_squared value **0.03856** is low, indicating that the model only explains one-third of the variance in the "Diabetes pedigree function".

**The F-statistic:**

The F-statistic test the overall significance of the model. A low p-value for the F-statistic **4.546e-05** suggests that at least one of the independent variables is significant in predicting the dependent variable.

**The body mass index** coefficient is the best predictor in linear models. It represents the change in the response variable for a one-unit change in the predictor variable, holding all other variables constant.

2.a) AIC

“The **Akaike information criterion (AIC)** is a mathematical method for evaluating how well a model fits the data it was generated from. In [statistics](https://www.scribbr.com/category/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](https://www.scribbr.com/methodology/independent-and-dependent-variables/#independent) used to build the model.
* the maximum likelihood estimate of the model (how well the model reproduces the data).

The best-fit model according to AIC is the one that explains the greatest amount of variation using the fewest possible independent variables. In statistics, AIC is most often used for model selection. By calculating and comparing the AIC scores of several possible models, you can choose the one that is the best fit for the data”. **(Rebecca Bevans. Revised on June 22, 2020)**

A lower AIC indicates that the model fits better, making it a more efficient and effective option for explaining the data. AIC is an essential statistical modeling and analysis tool. When a set of models are evaluated, the one with the lowest AIC is usually considered as the best model, given that it sufficiently represents the data.

b) Comparing the Models

Model#1

#Task 2.2(b) compare and accessing the given models  
# Model #1  
model1 <- lm(formula = `Diabetes pedigree function` ~ `Diastolic blood pressure` + `Body mass index`,strata\_sample)  
summary(model1)

##   
## Call:  
## lm(formula = `Diabetes pedigree function` ~ `Diastolic blood pressure` +   
## `Body mass index`, data = strata\_sample)  
##   
## 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   
## `Diastolic blood pressure` 0.0006413 0.0009734 0.659 0.510398   
## `Body mass index` 0.0092102 0.0023926 3.849 0.000138 \*\*\*  
## ---  
## 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

This code utilizes the predictor variables "Diastolic blood pressure" and "Body mass index" to predict the "Diabetes pedigree function" using a linear model (Model 1). The lm() function in R is used to build the linear model. When comparing the coefficients of 'Diastolic blood pressure' and Body Mass Index (BMI), it was found that ‘Diastolic blood pressure’ has a lesser effect. Therefore, the coefficient of the BMI has a higher value and possesses a greater effect than Diastolic blood pressure. The Adjusted r squared value is very small indicates that the model only explains the one-third of the sample

Model#2

# Model #2 (best two predictor linear model)  
model2 <- lm(formula = `Diabetes pedigree function` ~ `Body mass index`+`Triceps skinfold thickness` ,strata\_sample)   
summary(model2)

##   
## Call:  
## lm(formula = `Diabetes pedigree function` ~ `Body mass index` +   
## `Triceps skinfold thickness`, data = strata\_sample)  
##   
## 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

In this model Diabetes pedigree function is predicted by Body mass index and Triceps skinfold thickness. This section provides information about the residuals of the model, which are the differences between the observed values and the values predicted by the model. The intercept value estimated in this study represents the predicted value of "Diabetes pedigree function" when both "Body mass index" and "Triceps skinfold thickness" are at zero level. The "Body mass index" has an estimated coefficient of 0.007078, indicating the estimated change in the "Diabetes pedigree function" that will occur due to a one-unit change in "Body mass index." Similarly, the estimated coefficient for "Triceps skinfold thickness" is 0.003086, which represents the estimated change in the "Diabetes pedigree function" due to a one-unit change in "Triceps skinfold thickness

Model#3

# Model #3 (best four-predictor liner model)  
model3 <- lm(formula = `Diabetes pedigree function` ~+`Body mass index`+`Triceps skinfold thickness`+`Plasma glucose concentration`+`2-Hour serum insulin` ,strata\_sample)   
summary(model3)

##   
## Call:  
## lm(formula = `Diabetes pedigree function` ~ +`Body mass index` +   
## `Triceps skinfold thickness` + `Plasma glucose concentration` +   
## `2-Hour serum insulin`, data = strata\_sample)  
##   
## 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 \*   
## `Plasma glucose concentration` 0.0014284 0.0005365 2.662 0.00807 \*\*  
## `2-Hour serum insulin` 0.0001858 0.0001567 1.186 0.23637   
## ---  
## 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

The formula includes the ‘Body mass index’, Triceps skinfold thickness’,’plasma glucose concentration’ and ‘2-Hour serum inslin’. . The "Body mass index" has an estimated coefficient of 0.0057518, indicating the estimated change in the "Diabetes pedigree function" that will occur due to a one-unit change in "Body mass index." The coefficient of ‘2-Hour serum insulin’ does not indicate any effect on the variable. The F-statistic (9.363) and the associated p-value (3.074e-07) indicate that the model as a whole is statistically significant, suggesting that at least one of the predictor variables is significant in predicting the response variable. The third model (model3) uses four predictor variables to predict the "Diabetes pedigree function." Although the model is statistically significant (as indicated by the low p-value for the F-statistic), the R-squared values are still relatively low, indicating that the model only explains a moderate portion of the variability in the response variable. Further analysis and consideration of the practical significance of these predictor variables may be necessary to gain more insights.

Comparing the models using AIC

# By using AIC comparing the models  
AIC\_1<-AIC(model1)  
AIC\_2<-AIC(model2)  
AIC\_3<-AIC(model3)  
AIC\_vector<- c(AIC\_1, AIC\_2, AIC\_3)  
model\_names <- c("Model#1", "Model#2", "Model#3")  
  
# Creating a summary table  
summarized <- data.frame(Model = model\_names, AIC = AIC\_vector)  
print(summarized)

## Model AIC  
## 1 Model#1 304.6269  
## 2 Model#2 297.2691  
## 3 Model#3 289.5617

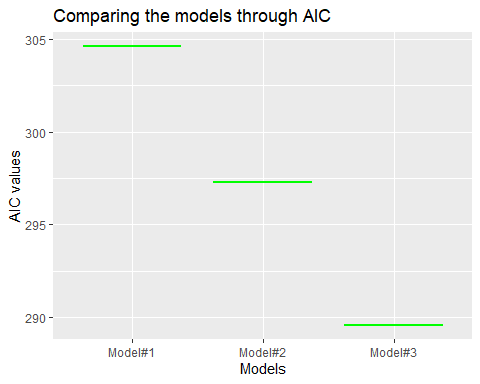
#predicting the best model  
print(best\_model<- summarized[which.min(summarized$AIC), ])

## Model AIC  
## 3 Model#3 289.5617

By using a AIC comparing method we are comparing the three models AIC contrasts the three models The code sample contrasts three linear models (Models 1, 2, and 3) based on the quality of fit as well as complexity of each model as measured by the AIC (Akaike Information Criterion) value. The "AIC\_values" vector contains the computed and saved AIC values for each model. The summary table below shows the results for Model 3, which has the lowest AIC value of 289.5617 and is the best-performing model among the three.

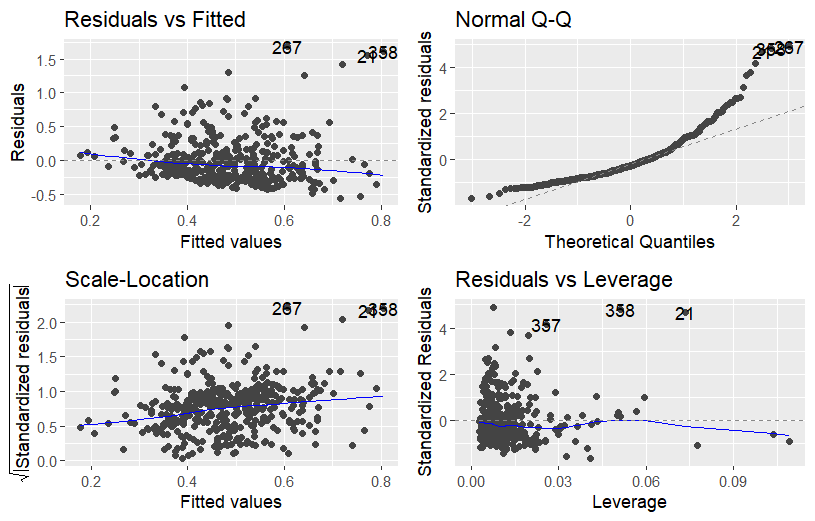
**AIC plot:**

# Load the ggplot2 library  
library(ggplot2)  
  
  
# Create a violin plot using ggplot2  
ggplot(data = summarized, aes(x = Model, y = AIC)) +  
 geom\_boxplot(fill = "Red", color = "green") +  
 labs(title = "Comparing the models through AIC", x = "Models", y = "AIC values")



**Figure 02:**  Comparisons of AIC

**Diagnostic plots of the model3:**

****

**Figure 03: Diagnostic plots**

The residual plots exhibits the good pattern shows the significant relation between the residual and the fitted values.

In the Q-Q plot the quantile of the residuals is normally distributed and it barely departs form the fitted values.

4. conclusion

The linear models used to predict the "Diabetes pedigree function" in the dataset have their limitations due to a strictly linear approach. Despite being the strongest model, Model 3 only accounts for a small fraction of the volatility in the target variable.Our group task included Factor Analysis (FA), Principal Component Analysis (PCA), and Cluster Analysis to gain further insight into the dataset's structures.

Participants should be informed that although linear models are helpful, they cannot accurately capture the complex nature of diabetes. To gain a more complete and precise understanding of the risk factors for diabetes, It is suggested to combine methods such as identifying groups of people with similar risk profiles and analyzing how different variables interact with each other. It is important to take into account the specific objectives of the analysis, the dataset context, and the intended audience's needs when conducting a critical assessment. In communicating the results and implications to non-experts, it is essential to prioritize clarity, relevance, and practicality to ensure effective communication.

It is true that visualizations can often provide a more detailed and comprehensive understanding than a purely theoretical or mathematical approach, especially for non-professionals. Luckily, R has a wide range of visualization properties that can be used to analyze data in a more visual way. By presenting data using visualizations and making it more accessible to non-professionals, we can effectively communicate important points and insights to a wider audience.

References:

**Published on March 26, 2020 by**[**Rebecca Bevans**](https://www.scribbr.com/author/beccabevans/)**. Revised on June 22, 2023.**

[**https://www.scribbr.com/statistics/akaike-information-criterion/**](https://www.scribbr.com/statistics/akaike-information-criterion/)

**For rectifying grammatical mistakes and to improve the quality of the sentences at some palces GRAMMARLY AI has been used.**