Practice SG MLR

Ahmad Firdaus Mohamed

2024-01-21

# 1. Data were collected to study factors that were related with HbA1c. The dataset is named as “metab\_syndrome” and in .xlsx formats. (FIRDAUS) (? Dataset [https://github.com/drkamarul/multivar\_data\_analysis/blob/main/data/metabolic\_syndrome.dta)](https://github.com/drkamarul/multivar_data_analysis/blob/main/data/metabolic_syndrome.dta))

A model was proposed to examine the relationship between HBA1C with these covariates: AGE, DMDX, MSBPR, FBS and HDL.

a. Write the regression model and equation for MODEL\_A. (15 marks)

# Prepare Environment

## Load Library

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ 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

library(broom)  
library(ggplot2)  
library(corrplot)

## corrplot 0.92 loaded

library(gtsummary)  
library(haven)

## Read Data

Data1<-read\_dta("metabolic\_syndrome.dta")  
summary(Data1)

## codesub age hpt smoking   
## Length:4340 Min. :18.00 Length:4340 Length:4340   
## Class :character 1st Qu.:38.00 Class :character Class :character   
## Mode :character Median :48.00 Mode :character Mode :character   
## Mean :47.84   
## 3rd Qu.:58.00   
## Max. :89.00   
##   
## dmdx height weight waist   
## Length:4340 Min. :1.270 Min. : 30.00 Min. : 50.80   
## Class :character 1st Qu.:1.510 1st Qu.: 53.80 1st Qu.: 77.00   
## Mode :character Median :1.560 Median : 62.00 Median : 86.00   
## Mean :1.568 Mean : 63.75 Mean : 86.32   
## 3rd Qu.:1.630 3rd Qu.: 71.97 3rd Qu.: 95.00   
## Max. :1.960 Max. :187.80 Max. :154.50   
## NA's :1 NA's :2 NA's :2   
## hip msbpr mdbpr hba1c   
## Min. : 61.00 Min. : 68.5 Min. : 41.50 Min. : 0.200   
## 1st Qu.: 91.00 1st Qu.:117.0 1st Qu.: 70.00 1st Qu.: 5.100   
## Median : 97.00 Median :130.0 Median : 77.50 Median : 5.400   
## Mean : 97.88 Mean :133.5 Mean : 78.47 Mean : 5.805   
## 3rd Qu.:104.00 3rd Qu.:146.5 3rd Qu.: 86.00 3rd Qu.: 5.800   
## Max. :160.00 Max. :237.0 Max. :128.50 Max. :15.000   
## NA's :2 NA's :70   
## fbs mogtt1h mogtt2h totchol   
## Min. : 2.500 Min. : 0.160 Min. : 0.160 Min. : 0.180   
## 1st Qu.: 4.480 1st Qu.: 6.540 1st Qu.: 5.150 1st Qu.: 4.970   
## Median : 5.180 Median : 8.590 Median : 6.600 Median : 5.700   
## Mean : 5.734 Mean : 9.106 Mean : 7.343 Mean : 5.792   
## 3rd Qu.: 6.020 3rd Qu.:10.840 3rd Qu.: 8.410 3rd Qu.: 6.530   
## Max. :28.010 Max. :41.500 Max. :37.370 Max. :23.140   
## NA's :248 NA's :604 NA's :608 NA's :54   
## ftrigliz hdl ldl gender   
## Min. : 0.110 Min. :0.080 Min. : 0.140 Length:4340   
## 1st Qu.: 0.930 1st Qu.:1.110 1st Qu.: 2.790 Class :character   
## Median : 1.260 Median :1.320 Median : 3.460 Mode :character   
## Mean : 1.534 Mean :1.345 Mean : 3.549   
## 3rd Qu.: 1.770 3rd Qu.:1.540 3rd Qu.: 4.245   
## Max. :12.660 Max. :4.430 Max. :10.560   
## NA's :53 NA's :52 NA's :53   
## crural   
## Length:4340   
## Class :character   
## Mode :character   
##   
##   
##   
##

## Select relevant variables

Data2 <- Data1 %>%  
 select(hba1c,age,dmdx,msbpr,fbs,hdl,mdbpr)

Clear NA data

Data2 <- Data2 %>%  
 na.omit()

summary(Data2)

## hba1c age dmdx msbpr   
## Min. : 0.200 Min. :18.00 Length:4068 Min. : 68.5   
## 1st Qu.: 5.100 1st Qu.:38.00 Class :character 1st Qu.:117.0   
## Median : 5.400 Median :48.00 Mode :character Median :130.0   
## Mean : 5.827 Mean :48.04 Mean :133.9   
## 3rd Qu.: 5.800 3rd Qu.:58.00 3rd Qu.:147.0   
## Max. :15.000 Max. :89.00 Max. :237.0   
## fbs hdl mdbpr   
## Min. : 2.500 Min. :0.08 Min. : 41.50   
## 1st Qu.: 4.480 1st Qu.:1.11 1st Qu.: 70.00   
## Median : 5.190 Median :1.32 Median : 78.00   
## Mean : 5.734 Mean :1.35 Mean : 78.56   
## 3rd Qu.: 6.020 3rd Qu.:1.55 3rd Qu.: 86.00   
## Max. :28.010 Max. :4.43 Max. :128.50

Data2 <- Data2 %>%  
 mutate(dmdx = as\_factor(dmdx))

summary(Data2)

## hba1c age dmdx msbpr fbs   
## Min. : 0.200 Min. :18.00 no :3607 Min. : 68.5 Min. : 2.500   
## 1st Qu.: 5.100 1st Qu.:38.00 yes: 461 1st Qu.:117.0 1st Qu.: 4.480   
## Median : 5.400 Median :48.00 Median :130.0 Median : 5.190   
## Mean : 5.827 Mean :48.04 Mean :133.9 Mean : 5.734   
## 3rd Qu.: 5.800 3rd Qu.:58.00 3rd Qu.:147.0 3rd Qu.: 6.020   
## Max. :15.000 Max. :89.00 Max. :237.0 Max. :28.010   
## hdl mdbpr   
## Min. :0.08 Min. : 41.50   
## 1st Qu.:1.11 1st Qu.: 70.00   
## Median :1.32 Median : 78.00   
## Mean :1.35 Mean : 78.56   
## 3rd Qu.:1.55 3rd Qu.: 86.00   
## Max. :4.43 Max. :128.50

## Model A

MODEL\_A <- lm (hba1c~age + dmdx + msbpr + fbs + hdl, data = Data2 )  
summary(MODEL\_A)

##   
## Call:  
## lm(formula = hba1c ~ age + dmdx + msbpr + fbs + hdl, data = Data2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.6758 -0.4002 -0.0591 0.3044 8.6793   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.7738286 0.1080849 34.915 < 2e-16 \*\*\*  
## age 0.0033929 0.0012295 2.759 0.00582 \*\*   
## dmdxyes 1.3374982 0.0554958 24.101 < 2e-16 \*\*\*  
## msbpr -0.0005795 0.0007397 -0.784 0.43337   
## fbs 0.3157897 0.0068808 45.894 < 2e-16 \*\*\*  
## hdl 0.0043690 0.0429465 0.102 0.91898   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9827 on 4062 degrees of freedom  
## Multiple R-squared: 0.546, Adjusted R-squared: 0.5455   
## F-statistic: 977.2 on 5 and 4062 DF, p-value: < 2.2e-16

tidy(MODEL\_A, conf.int = TRUE)

## # A tibble: 6 × 7  
## term estimate std.error statistic p.value conf.low conf.high  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) 3.77 0.108 34.9 8.23e-234 3.56 3.99   
## 2 age 0.00339 0.00123 2.76 5.82e- 3 0.000982 0.00580   
## 3 dmdxyes 1.34 0.0555 24.1 4.56e-120 1.23 1.45   
## 4 msbpr -0.000580 0.000740 -0.784 4.33e- 1 -0.00203 0.000871  
## 5 fbs 0.316 0.00688 45.9 0 0.302 0.329   
## 6 hdl 0.00437 0.0429 0.102 9.19e- 1 -0.0798 0.0886

## Regression Model & Equation

hba1c = 3.77 + 1.34 (dmdxyes) + 0.32 (fbs)

Adjusted R-squared: 0.5455

# b. Add variable MDBPR to MODEL\_A and name the new model as MODEL\_B. Describe the effect of this addition to the model. (5 marks)

MODEL\_B <- lm (hba1c ~ age + dmdx + fbs + hdl + msbpr + mdbpr, data = Data2)  
summary (MODEL\_B)

##   
## Call:  
## lm(formula = hba1c ~ age + dmdx + fbs + hdl + msbpr + mdbpr,   
## data = Data2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.5179 -0.4013 -0.0567 0.2993 8.6937   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.494560 0.123378 28.324 < 2e-16 \*\*\*  
## age 0.004371 0.001244 3.513 0.000448 \*\*\*  
## dmdxyes 1.332653 0.055365 24.070 < 2e-16 \*\*\*  
## fbs 0.314633 0.006868 45.812 < 2e-16 \*\*\*  
## hdl 0.016450 0.042916 0.383 0.701521   
## msbpr -0.004019 0.001044 -3.849 0.000120 \*\*\*  
## mdbpr 0.008700 0.001869 4.655 3.34e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9802 on 4061 degrees of freedom  
## Multiple R-squared: 0.5485, Adjusted R-squared: 0.5478   
## F-statistic: 822.1 on 6 and 4061 DF, p-value: < 2.2e-16

Adjusted R-squared: 0.5478

By adding mdbpr into the model, the model is able to predict the hba1c better from 54.55 % to 54.78%

# c. Perform comparison between MODEL\_A and MODEL\_B. Give comment. (5 marks)

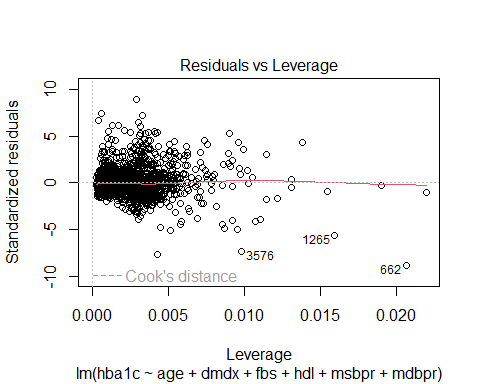
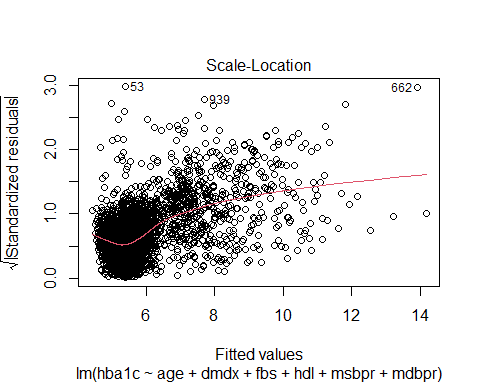
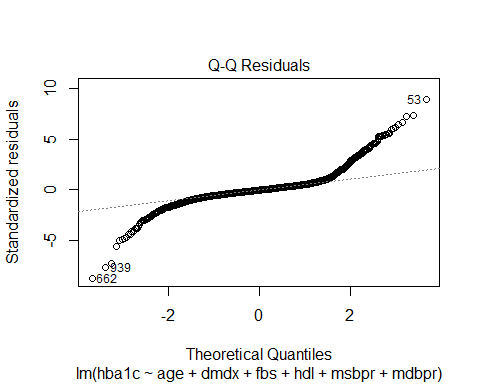
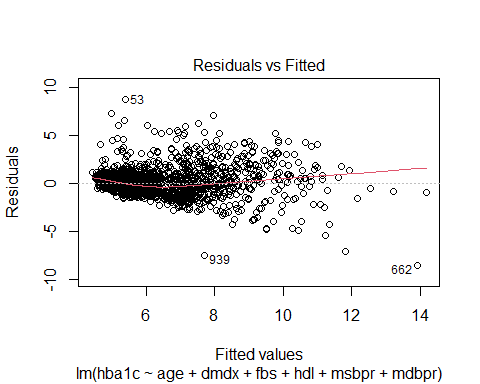
anova (MODEL\_A, MODEL\_B)

## Analysis of Variance Table  
##   
## Model 1: hba1c ~ age + dmdx + msbpr + fbs + hdl  
## Model 2: hba1c ~ age + dmdx + fbs + hdl + msbpr + mdbpr  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 4062 3922.5   
## 2 4061 3901.7 1 20.82 21.67 3.342e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The p-value is <0.001, hence the re is significant difference between the models

# d. Perform model assessment for MODEL\_B. Save the relevant plots in the thumb-drive and write the names of the plots in the answer sheet. (10 marks).

plot(MODEL\_B)



library(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

bptest(MODEL\_B)

##   
## studentized Breusch-Pagan test  
##   
## data: MODEL\_B  
## BP = 639.89, df = 6, p-value < 2.2e-16

p-value is <0.005, hence assumption of equal residual variance is violated

shapiro.test(MODEL\_B$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: MODEL\_B$residuals  
## W = 0.8074, p-value < 2.2e-16

p-value is <0.005, hence assumption of normality of residual is violated

# e. Predict the average value of HBA1C for a population with:

AGE = 48 years

DMDX = has history of diabetes (1)

MSBPR = 133 mmHg

MDBPR = 79 mmHg

FBS = 5.6 mmol/l

HDL = 1.3 mmol/l

Show your calculation. (5 marks)

summary(MODEL\_B)

##   
## Call:  
## lm(formula = hba1c ~ age + dmdx + fbs + hdl + msbpr + mdbpr,   
## data = Data2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.5179 -0.4013 -0.0567 0.2993 8.6937   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.494560 0.123378 28.324 < 2e-16 \*\*\*  
## age 0.004371 0.001244 3.513 0.000448 \*\*\*  
## dmdxyes 1.332653 0.055365 24.070 < 2e-16 \*\*\*  
## fbs 0.314633 0.006868 45.812 < 2e-16 \*\*\*  
## hdl 0.016450 0.042916 0.383 0.701521   
## msbpr -0.004019 0.001044 -3.849 0.000120 \*\*\*  
## mdbpr 0.008700 0.001869 4.655 3.34e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9802 on 4061 degrees of freedom  
## Multiple R-squared: 0.5485, Adjusted R-squared: 0.5478   
## F-statistic: 822.1 on 6 and 4061 DF, p-value: < 2.2e-16

equation

hba1c = 3.49 + 1.33 (dmdxyes) + 0.31 (fbs)

summary(model.matrix(MODEL\_B))

## (Intercept) age dmdxyes fbs hdl   
## Min. :1 Min. :18.00 Min. :0.0000 Min. : 2.500 Min. :0.08   
## 1st Qu.:1 1st Qu.:38.00 1st Qu.:0.0000 1st Qu.: 4.480 1st Qu.:1.11   
## Median :1 Median :48.00 Median :0.0000 Median : 5.190 Median :1.32   
## Mean :1 Mean :48.04 Mean :0.1133 Mean : 5.734 Mean :1.35   
## 3rd Qu.:1 3rd Qu.:58.00 3rd Qu.:0.0000 3rd Qu.: 6.020 3rd Qu.:1.55   
## Max. :1 Max. :89.00 Max. :1.0000 Max. :28.010 Max. :4.43   
## msbpr mdbpr   
## Min. : 68.5 Min. : 41.50   
## 1st Qu.:117.0 1st Qu.: 70.00   
## Median :130.0 Median : 78.00   
## Mean :133.9 Mean : 78.56   
## 3rd Qu.:147.0 3rd Qu.: 86.00   
## Max. :237.0 Max. :128.50

newData1 <- expand.grid(age = 48, dmdx = 2, msbpr =133, mdbpr = 79, fbs = 5.9, hdl = 1.3)  
head(newData1)

## age dmdx msbpr mdbpr fbs hdl  
## 1 48 2 133 79 5.9 1.3