Question 3 (Syahid)

Ahmad Firdaus Mohamed

2024-02-10

1. Dataset named as Q3.csv has six (6) variables. (SYAHID) (? Dataset [https://github.com/drkamarul/multivar\_data\_analysis/blob/main/data/metabolic\_syndrome.csv)](https://github.com/drkamarul/multivar_data_analysis/blob/main/data/metabolic_syndrome.csv))

1. HBA1C: in percentage (the outcome variable)

ii.FBS: Fasting blood sugar in mmol/L

1. AGE: Age of participants
2. DMDX: History of diabetes (0 = No, 1 = Yes)
3. MOGTT2H: 2-hours post prandial sugar in mmol/L
4. TOTCHOL: Total cholesterol in mmol/L

## (a) The analyst has proposed two (2) candidate models:

Model\_A: covariates are DMDX, FBS, and MOGTT2H

Model\_A <- lm (hba1c~ dmdx + fbs + mogtt2h, data = Data1)

summary(Model\_A)

##   
## Call:  
## lm(formula = hba1c ~ dmdx + fbs + mogtt2h, data = Data1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.0771 -0.3296 -0.0299 0.2715 9.0948   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.883382 0.036739 105.701 <2e-16 \*\*\*  
## dmdxyes 0.319341 0.119560 2.671 0.0076 \*\*   
## fbs 0.149135 0.009134 16.328 <2e-16 \*\*\*  
## mogtt2h 0.114737 0.004567 25.126 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7207 on 3542 degrees of freedom  
## Multiple R-squared: 0.4502, Adjusted R-squared: 0.4498   
## F-statistic: 967 on 3 and 3542 DF, p-value: < 2.2e-16

Model\_B: covariates are AGE and TOTCHOL

Model\_B <- lm (hba1c~ age + totchol, data = Data1)

summary(Model\_B)

##   
## Call:  
## lm(formula = hba1c ~ age + totchol, data = Data1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.1613 -0.3952 -0.1482 0.1202 9.2199   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.557589 0.080656 56.506 < 2e-16 \*\*\*  
## age 0.006790 0.001163 5.837 5.8e-09 \*\*\*  
## totchol 0.113329 0.013246 8.556 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9516 on 3543 degrees of freedom  
## Multiple R-squared: 0.04127, Adjusted R-squared: 0.04072   
## F-statistic: 76.25 on 2 and 3543 DF, p-value: < 2.2e-16

Compare Model\_A and Model\_B. Show your results. (4 marks)

anova(Model\_A, Model\_B)

## Analysis of Variance Table  
##   
## Model 1: hba1c ~ dmdx + fbs + mogtt2h  
## Model 2: hba1c ~ age + totchol  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 3542 1839.7   
## 2 3543 3208.3 -1 -1368.6 2635 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

When significant level is set at 0.05, there is significant difference between the 2 models.

Model A adj R-sq is 0.4498

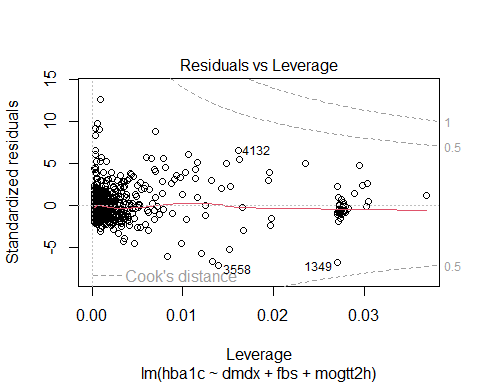
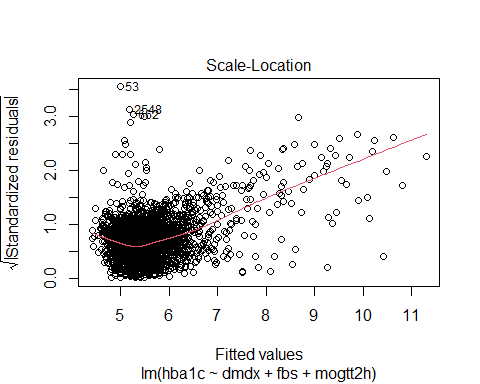
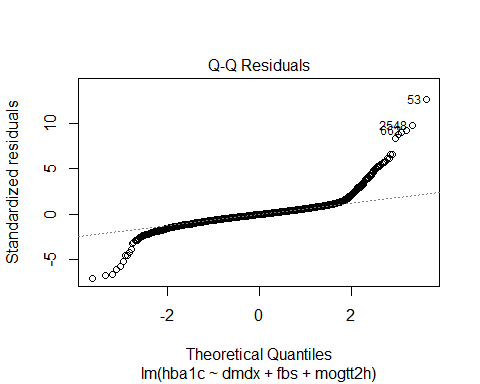
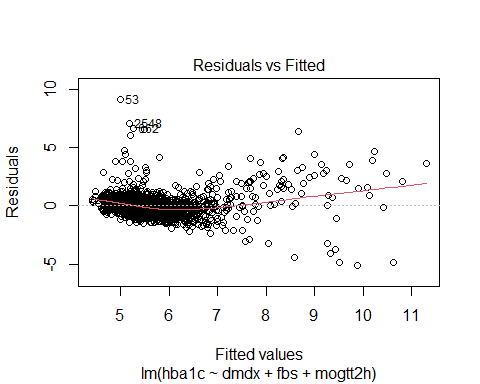
Model B adj R-sq is 0.0407

Establish the model of your choice. Give reasons. (4 marks)

The selected model is Model A because of higher R-sq and there is significant difference between the 2 models

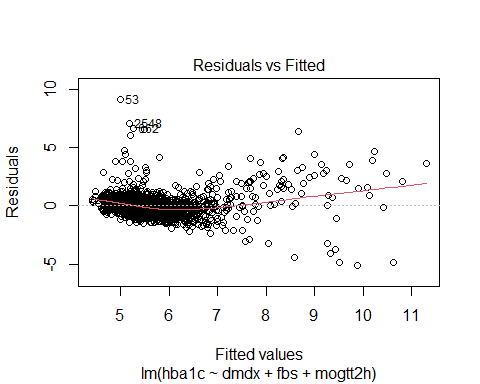
## (b) Using your model of choice from (a), perform a model assessment. Show the results and comment on the findings. (8 marks)

plot(Model\_A)



### Line

plot(Model\_A, which = 1)



The red line through the scatter plot straight, horizontal and not curved indicating linearity assumption is satisfied. The homoscedasticity assumption is met as the residuals are equally spread around the y = 0 line

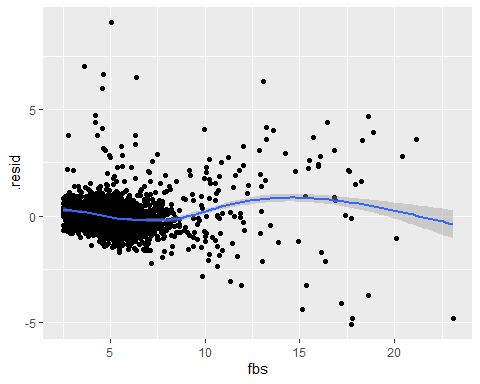
summary(Model\_A)

##   
## Call:  
## lm(formula = hba1c ~ dmdx + fbs + mogtt2h, data = Data1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.0771 -0.3296 -0.0299 0.2715 9.0948   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.883382 0.036739 105.701 <2e-16 \*\*\*  
## dmdxyes 0.319341 0.119560 2.671 0.0076 \*\*   
## fbs 0.149135 0.009134 16.328 <2e-16 \*\*\*  
## mogtt2h 0.114737 0.004567 25.126 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7207 on 3542 degrees of freedom  
## Multiple R-squared: 0.4502, Adjusted R-squared: 0.4498   
## F-statistic: 967 on 3 and 3542 DF, p-value: < 2.2e-16

Furthermore for each CV

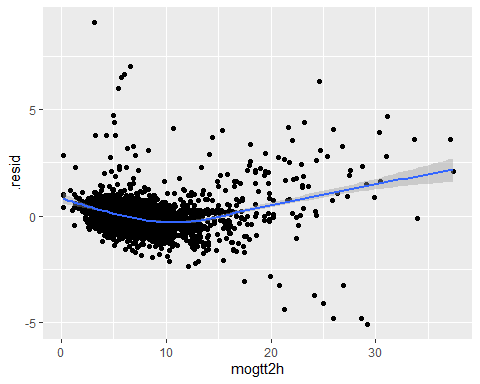
augment(Model\_A)%>%  
 ggplot(aes(x=fbs, y=.resid))+  
 geom\_point()+  
 geom\_smooth()

## `geom\_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'



augment(Model\_A)%>%  
 ggplot(aes(x=mogtt2h, y=.resid))+  
 geom\_point()+  
 geom\_smooth()

## `geom\_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'



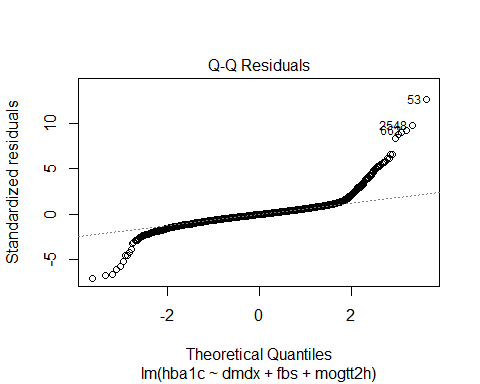
Residuals shows individual linearity against mogtt2h and fbs

### Independence

Based on study design

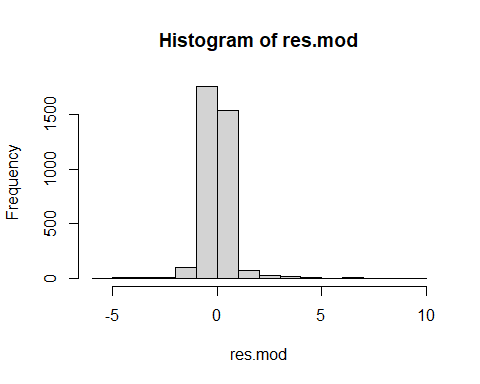
### Normality

plot(Model\_A, which = 2)



In Q-Q plot, there is curvature at head and tail indicating normality assumption is violated

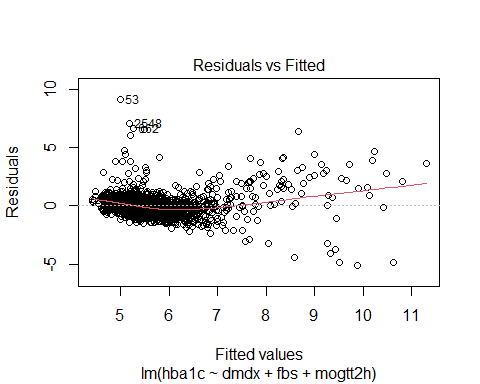
res.mod <- residuals(Model\_A)  
hist(res.mod)



However, based on histogram the residuals appear normally distributed

### Equal Variance

plot(Model\_A, which = 1)



horizontal line with randomly scattered data points around it observed, suggesting that the homoscedasticity assumption is satisfied.

## (c) Using Model\_A, build a model with an interaction term between DMDX and FBS. Present the results in a table. (6 marks)

Model.ia <- lm(hba1c ~ dmdx + fbs + mogtt2h + dmdx:fbs, data = Data1)  
summary(Model.ia)

##   
## Call:  
## lm(formula = hba1c ~ dmdx + fbs + mogtt2h + dmdx:fbs, data = Data1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.0302 -0.3279 -0.0310 0.2662 9.0887   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.900742 0.037038 105.319 < 2e-16 \*\*\*  
## dmdxyes -0.688552 0.319315 -2.156 0.031124 \*   
## fbs 0.147790 0.009129 16.190 < 2e-16 \*\*\*  
## mogtt2h 0.113352 0.004578 24.761 < 2e-16 \*\*\*  
## dmdxyes:fbs 0.156291 0.045924 3.403 0.000673 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7196 on 3541 degrees of freedom  
## Multiple R-squared: 0.452, Adjusted R-squared: 0.4514   
## F-statistic: 730.3 on 4 and 3541 DF, p-value: < 2.2e-16

Tblmod <- tbl\_regression(Model.ia)  
Tblmod

## Table printed with {flextable}, not {gt}. Learn why at  
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html  
## To suppress this message, include `message = FALSE` in the code chunk header.

Table 1 Multiple Linear Regression with Interaction (n= 3546)

| **Characteristic** | **Beta** | **95% CI**1 | **p-value** |
| --- | --- | --- | --- |
| dmdx |  |  |  |
| no | — | — |  |
| yes | -0.69 | -1.3, -0.06 | 0.031 |
| fbs | 0.15 | 0.13, 0.17 | <0.001 |
| mogtt2h | 0.11 | 0.10, 0.12 | <0.001 |
| dmdx \* fbs |  |  |  |
| yes \* fbs | 0.16 | 0.07, 0.25 | <0.001 |
| 1CI = Confidence Interval  Intercept = 3.9007  Adj R-sq = 0.4514 | | | |

## (d) Write the regression equation. (4 marks)

summary(Model.ia)

##   
## Call:  
## lm(formula = hba1c ~ dmdx + fbs + mogtt2h + dmdx:fbs, data = Data1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.0302 -0.3279 -0.0310 0.2662 9.0887   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.900742 0.037038 105.319 < 2e-16 \*\*\*  
## dmdxyes -0.688552 0.319315 -2.156 0.031124 \*   
## fbs 0.147790 0.009129 16.190 < 2e-16 \*\*\*  
## mogtt2h 0.113352 0.004578 24.761 < 2e-16 \*\*\*  
## dmdxyes:fbs 0.156291 0.045924 3.403 0.000673 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7196 on 3541 degrees of freedom  
## Multiple R-squared: 0.452, Adjusted R-squared: 0.4514   
## F-statistic: 730.3 on 4 and 3541 DF, p-value: < 2.2e-16

Y= B0 + B1X1 + B2X2 + B3X3 + B4X4

hba1c = 3.9007 - 0.688552(dmdxyes) + 0.147790 (fbs) + 0.1133(mogtt2h) + 0.1562 (dmdxyes\*fbs)