Question 1 (Firdaus)

Ahmad Firdaus Mohamed

2024-02-10

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)

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

summary(Model\_A)

##   
## Call:  
## lm(formula = hba1c ~ age + dmdx + msbpr + fbs + hdl, data = Data1)  
##   
## 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  
## (272 observations deleted due to missingness)  
## Multiple R-squared: 0.546, Adjusted R-squared: 0.5455   
## F-statistic: 977.2 on 5 and 4062 DF, p-value: < 2.2e-16

Regression model

Y= B0 + B1X1 + B2X2 + B3X3 + B4X4 + B5X5

hba1c = 3.7738 + 0.0033(age) + 1.3374 (dmdxyes) - 0.0005 (msbpr) + 0.0043 (hdl)

## 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 + msbpr + fbs + hdl + mdbpr, data = Data1)   
summary(Model\_B)

##   
## Call:  
## lm(formula = hba1c ~ age + dmdx + msbpr + fbs + hdl + mdbpr,   
## data = Data1)  
##   
## 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 \*\*\*  
## msbpr -0.004019 0.001044 -3.849 0.000120 \*\*\*  
## fbs 0.314633 0.006868 45.812 < 2e-16 \*\*\*  
## hdl 0.016450 0.042916 0.383 0.701521   
## 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  
## (272 observations deleted due to missingness)  
## Multiple R-squared: 0.5485, Adjusted R-squared: 0.5478   
## F-statistic: 822.1 on 6 and 4061 DF, p-value: < 2.2e-16

The effect of adding mdbr, the adjusted R-Sq changes from 0.5455 to 0.5478

## 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 + msbpr + fbs + hdl + 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

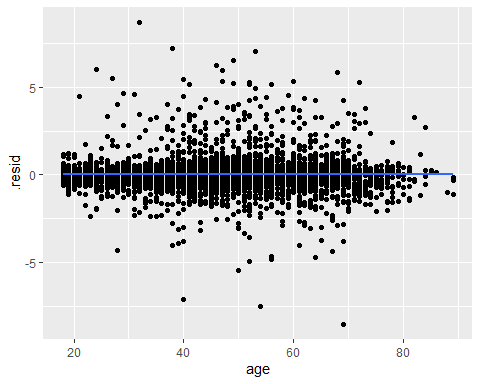
When the significant level is set at 0.005, there is significant difference between the 2 models

## 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).

Linearity for each numerical Covariate

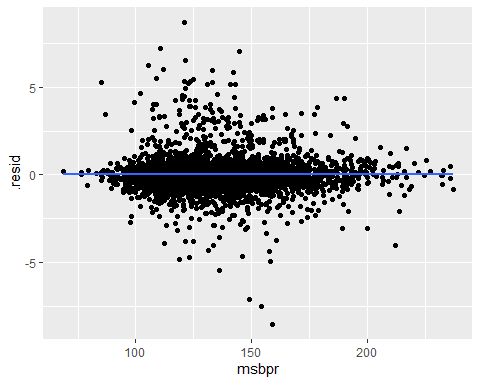
augment(Model\_B) %>%  
 ggplot(aes(x=age, y=.resid)) +  
 geom\_point()+  
 geom\_smooth()

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



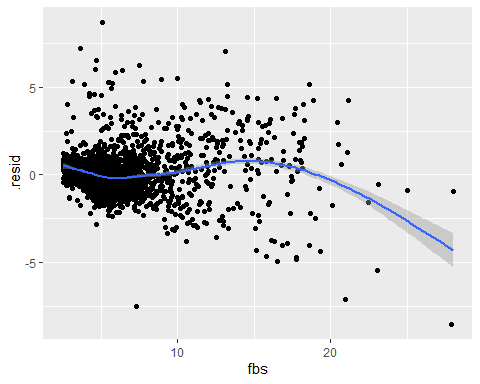
augment(Model\_B) %>%  
 ggplot(aes(x=msbpr, y=.resid)) +  
 geom\_point()+  
 geom\_smooth()

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



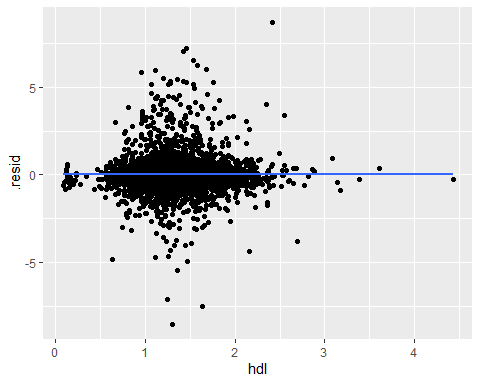
augment(Model\_B) %>%  
 ggplot(aes(x=fbs, y=.resid)) +  
 geom\_point()+  
 geom\_smooth()

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



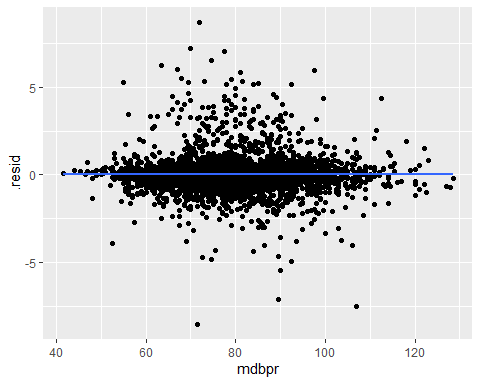
augment(Model\_B) %>%  
 ggplot(aes(x=hdl, y=.resid)) +  
 geom\_point()+  
 geom\_smooth()

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



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

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

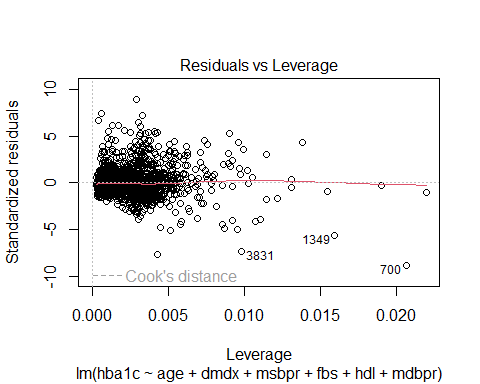
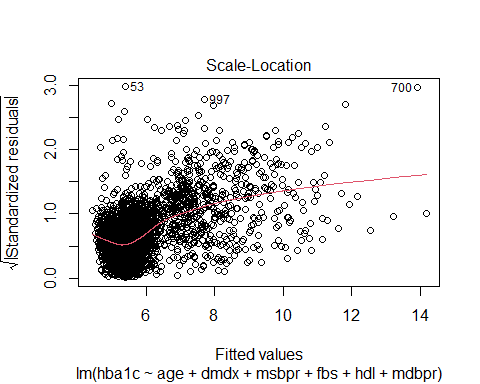
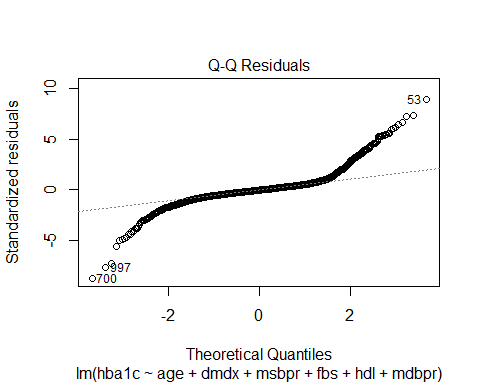
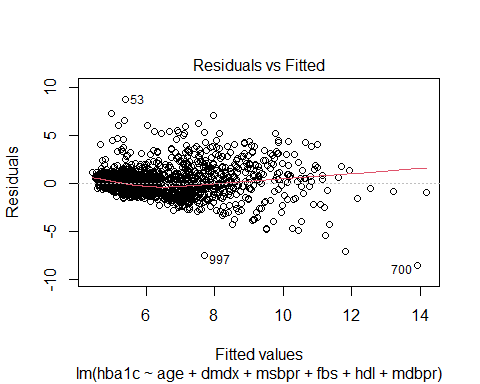


Independent

Based from study design

Normality of residual

plot(Model\_B)



## 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)

new\_data <- expand.grid(age=48, dmdx = "yes", msbpr = 133, fbs = 5.6, hdl = 1.3, mdbpr =79)  
new\_data

## age dmdx msbpr fbs hdl mdbpr  
## 1 48 yes 133 5.6 1.3 79

predict.Model\_B <- predict(Model\_B, newdata = new\_data)  
predict.Model\_B

## 1   
## 6.973205

summary(Model\_B)

##   
## Call:  
## lm(formula = hba1c ~ age + dmdx + msbpr + fbs + hdl + mdbpr,   
## data = Data1)  
##   
## 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 \*\*\*  
## msbpr -0.004019 0.001044 -3.849 0.000120 \*\*\*  
## fbs 0.314633 0.006868 45.812 < 2e-16 \*\*\*  
## hdl 0.016450 0.042916 0.383 0.701521   
## 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  
## (272 observations deleted due to missingness)  
## Multiple R-squared: 0.5485, Adjusted R-squared: 0.5478   
## F-statistic: 822.1 on 6 and 4061 DF, p-value: < 2.2e-16

hba1c = 3.4945 + 0.0043 (age) + 1.3326(dmdxyes) - 0.0040 (msbpr) + 0.3146 (fbs) + 0.0164 (hdl) + 0.0087 (mdbpr)

3.4945 + (0.0043\*48) + 1.3326 - (0.0040\*133) + (0.3146\*5.6) + (0.0164\*1.3) + (0.0087\*79)

## [1] 6.97188