# Final Paper Code

Safiya Geelani

5/10/2022

```r
knitr::opts_chunk$set(echo = F, message = F, warning = F)

set.seed(1579)

#libraries used
library(faraway)
library(tidyverse)
library(GGally)
library(MASS)
library(car)
library(lmtest)
library(caret)

#cleaning data and descriptive statistics
data(diabetes)
dim(diabetes)
min(diabetes$glyhb, na.rm = T)
max(diabetes$glyhb, na.rm = T)
diabetes2 <- na.omit(diabetes)
dim(diabetes2)
summary(lm(glyhb ~ ., data = diabetes2))

#scatterplot/ correlation/ density matrix
ggpairs(diabetes2[,c(6,1:5)])
ggpairs(diabetes2[,c(6,7:13)])
ggpairs(diabetes2[,c(6,14:19)])
#or ggpairs(diabetes2)

#variable selection using BIC
lm_full <- lm(glyhb ~ ., data = diabetes2)
n <- nrow(diabetes2)
lm_reduced <- step(lm_full, k = log(n))
summary(lm_reduced)

#checking assumptions
plot(lm_reduced, which = 2)
plot(rstandard(lm_reduced) ~ predict(lm_reduced), xlab = "Fitted values", ylab = "Standard residuals")
shapiro.test(resid(lm_reduced))

#box-cox transformation
boxcox(lm_reduced, lambda=seq(-2.5, 1.3, by=0.05))
summary(powerTransform(lm_reduced))
```

```r
diabetes2$glyhb_new <- 1/(sqrt(diabetes2$glyhb))
lm_final <- lm(glyhb_new ~ stab.glu + ratio + age, data = diabetes2)
summary(lm_final)

#rechecking assumptions
performance::check_model(lm_final)
bptest(lm_final)

#outliers and high leverage
plot(hatvalues(lm_final), rstandard(lm_final),
     xlab = "Leverage", ylab = "Standardized Residuals")
n <- nrow(diabetes2)
p <- 3
abline(v = 2*(p+1)/n, lty = 2) # threshold for high leverage
abline(h = c(-3,3), lty = 2) # threshold for outliers
ind <- which(abs(rstandard(lm_final)) > 3 | hatvalues(lm_final) > 8/n)
diabetes2[ind, ]

#predictions using cross validation
ctrl <- trainControl(method = "cv", number = 7)
model <- train(glyhb_new ~ stab.glu + ratio + age, data = diabetes2, method = "lm", trControl = ctrl)
print(model)

#1-1 ggplot
preds <- predict(lm_final, newdata = diabetes2)
pred_df <- data.frame(Actual = diabetes2$glyhb_new, Pred = preds)
ggplot(pred_df, aes(x = Actual, y = Pred)) +
  geom_point(size = 0.5) +
  geom_abline(intercept = 0, slope = 1) +
  xlab("Actual values of glyhb_new") +
  ylab("Predicted values of glyhb_new") +
  ggtitle("")
```