Project\_PSTAT126

Xihao Wu & Anhui Shi

9/10/2018

#Load the data set, delete the data that we don't want  
library(readr)  
auto\_mpg <- read\_table2("auto-mpg.txt", col\_names = FALSE)

auto\_mpg<-auto\_mpg[,-c(9,10,11)]  
names(auto\_mpg)<-c('mpg','cylinders','displacement','horsepower','weight','acceleration','modelyear','origin')  
auto\_mpg$horsepower <- as.numeric(auto\_mpg$horsepower)

auto\_mpg<-na.omit(auto\_mpg)  
auto\_mpg$origin<-as.factor(auto\_mpg$origin)

#Pairwise scatterplot   
pairs(auto\_mpg)

library(faraway)  
  
mod.1 = lm(mpg ~ 1, data = auto\_mpg[,-8])  
mod.1.upper = lm(mpg ~ ., data = auto\_mpg[,-8])  
  
step(mod.1, scope = list(lower = mod.1, upper = mod.1.upper))

#Variable selection  
library(leaps)  
mod.2 = regsubsets(auto\_mpg[,-c(1,8)], auto\_mpg$mpg)  
summary(mod.2)$which

summary(mod.2)$adjr2

summary(mod.2)$cp

#Residual Analysis  
mod.2.selected = lm(mpg ~ weight + modelyear + origin , data = auto\_mpg)  
yhat.2.selected= fitted(mod.2.selected)  
e.2.selected = auto\_mpg$mpg - yhat.2.selected  
plot(yhat.2.selected, e.2.selected, xlab = 'Fitted value', ylab = 'Residual', main = 'Residual vs Fitted')  
abline(h = 0, lty = 2)

#QQ-plot  
qqnorm(e.2.selected)  
qqline(e.2.selected)

#Shapiro-Wilk Test  
shapiro.test(e.2.selected)

#boxcox  
library(MASS)  
boxcox(mod.2.selected)

boxcox(mod.2.selected, lambda = seq(-0.5,0, length = 10))

boxcox(mod.2.selected, lambda = seq(-0.3,-0.25, length = 10))

#Residual Analysis after boxcox transformation  
mod.2.trans = lm(mpg^(-0.285) ~ weight + modelyear + origin , data = auto\_mpg)  
yhat.2.trans= fitted(mod.2.trans)  
e.2.trans = (auto\_mpg$mpg)^(-0.285) - yhat.2.trans  
plot(yhat.2.trans, e.2.trans, xlab = 'Fitted value', ylab = 'Residual', main = 'Residual vs Fitted')  
abline(h = 0, lty = 2)

#QQ-plot  
qqnorm(e.2.trans)  
qqline(e.2.trans)

#Shapiro-Wilk Test after boxcox transformation  
shapiro.test(e.2.trans)

summary(mod.2.trans)

#studentized deleted residuals  
sort(abs(rstudent(mod.2.trans)), decreasing = TRUE )[c(1,2,3,4,5)]

#Cook's distance  
sort(cooks.distance(mod.2.trans), decreasing =TRUE)[c(1,2,3)]

#Residual Analysis with deleted data   
mod.deleted= lm(mpg^(-0.285) ~ weight + modelyear + origin , data = auto\_mpg[-c(165,124,111), ] )  
yhat.deleted= fitted(mod.deleted)  
e.deleted = (auto\_mpg[-c(165,124,111), ]$mpg)^(-0.285) - yhat.deleted  
  
#QQ-plot with deleted data  
qqnorm(e.deleted)  
qqline(e.deleted)

#Shapiro-Wilk Test with deleted data  
shapiro.test(e.deleted)

#Find interaction  
mod.3.reduced = lm(mpg^(-0.285) ~ weight + modelyear + origin , data = auto\_mpg)  
mod.3.full =lm(mpg^(-0.285) ~ weight + modelyear + origin + weight\*modelyear + weight\*origin + modelyear\*origin , data = auto\_mpg)  
anova(mod.3.reduced, mod.3.full)

summary(mod.3.full)

#Final model  
mod.final = lm(mpg^(-0.285) ~ weight + modelyear + origin , data = auto\_mpg )  
summary(mod.final)

#Confidence Interval  
  
new\_ave = data.frame(weight = mean(auto\_mpg$weight), modelyear = max(auto\_mpg$modelyear), origin = as.factor(1) )  
  
ci\_ave = predict(mod.final , newdata = new\_ave, interval = 'confidence', level = 0.95)  
ci\_ave^(-1/0.285)