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)

## Parsed with column specification:  
## cols(  
## X1 = col\_double(),  
## X2 = col\_integer(),  
## X3 = col\_double(),  
## X4 = col\_character(),  
## X5 = col\_double(),  
## X6 = col\_double(),  
## X7 = col\_integer(),  
## X8 = col\_integer(),  
## X9 = col\_character(),  
## X10 = col\_character(),  
## X11 = col\_character()  
## )

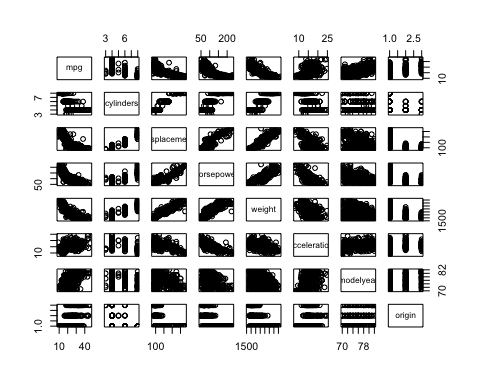
## Warning in rbind(names(probs), probs\_f): number of columns of result is not  
## a multiple of vector length (arg 1)

## Warning: 246 parsing failures.  
## row # A tibble: 5 x 5 col row col expected actual file expected <int> <chr> <chr> <chr> <chr> actual 1 3 <NA> 11 columns 10 columns 'auto-mpg.txt' file 2 5 <NA> 11 columns 10 columns 'auto-mpg.txt' row 3 7 <NA> 11 columns 10 columns 'auto-mpg.txt' col 4 9 <NA> 11 columns 10 columns 'auto-mpg.txt' expected 5 14 <NA> 11 columns 12 columns 'auto-mpg.txt'  
## ... ................. ... .................................................. ........ .................................................. ...... .................................................. .... .................................................. ... .................................................. ... .................................................. ........ ..................................................  
## See problems(...) for more details.

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)

## Warning: NAs introduced by coercion

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

## Start: AIC=1611.93  
## mpg ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + weight 1 16497.8 7321.2 1151.5  
## + displacement 1 15440.2 8378.8 1204.4  
## + horsepower 1 14433.1 9385.9 1248.9  
## + cylinders 1 14403.1 9415.9 1250.1  
## + modelyear 1 8027.7 15791.3 1452.8  
## + acceleration 1 4268.5 19550.5 1536.5  
## <none> 23819.0 1611.9  
##   
## Step: AIC=1151.49  
## mpg ~ weight  
##   
## Df Sum of Sq RSS AIC  
## + modelyear 1 2752.3 4569.0 968.66  
## + horsepower 1 327.4 6993.8 1135.56  
## + acceleration 1 168.3 7152.9 1144.37  
## + displacement 1 150.9 7170.3 1145.33  
## + cylinders 1 115.1 7206.1 1147.28  
## <none> 7321.2 1151.49  
## - weight 1 16497.8 23819.0 1611.93  
##   
## Step: AIC=968.66  
## mpg ~ weight + modelyear  
##   
## Df Sum of Sq RSS AIC  
## <none> 4569.0 968.66  
## + acceleration 1 10.5 4558.5 969.77  
## + cylinders 1 5.0 4564.0 970.24  
## + horsepower 1 3.3 4565.7 970.38  
## + displacement 1 0.0 4568.9 970.66  
## - modelyear 1 2752.3 7321.2 1151.49  
## - weight 1 11222.4 15791.3 1452.81

##   
## Call:  
## lm(formula = mpg ~ weight + modelyear, data = auto\_mpg[, -8])  
##   
## Coefficients:  
## (Intercept) weight modelyear   
## -14.347253 -0.006632 0.757318

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

## (Intercept) cylinders displacement horsepower weight acceleration  
## 1 TRUE FALSE FALSE FALSE TRUE FALSE  
## 2 TRUE FALSE FALSE FALSE TRUE FALSE  
## 3 TRUE FALSE FALSE FALSE TRUE TRUE  
## 4 TRUE FALSE TRUE FALSE TRUE TRUE  
## 5 TRUE TRUE TRUE FALSE TRUE TRUE  
## 6 TRUE TRUE TRUE TRUE TRUE TRUE  
## modelyear  
## 1 FALSE  
## 2 TRUE  
## 3 TRUE  
## 4 TRUE  
## 5 TRUE  
## 6 TRUE

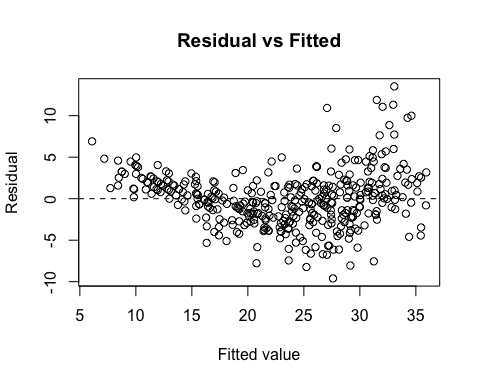
summary(mod.2)$adjr2

## [1] 0.6918423 0.8071941 0.8071393 0.8067872 0.8067841 0.8062826

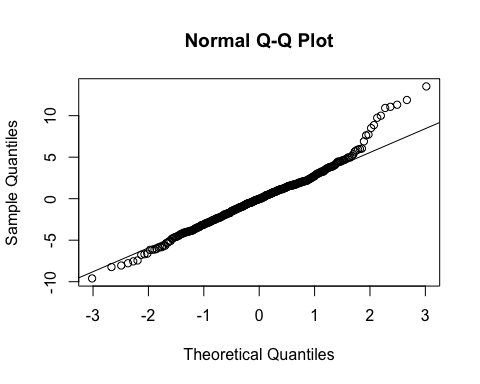
summary(mod.2)$cp

## [1] 232.396144 1.169751 2.284200 3.992019 5.000800 7.000000

#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)



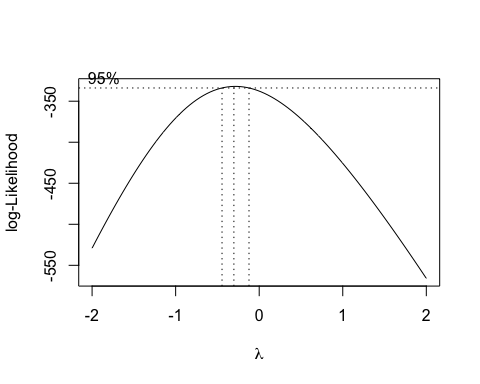
#QQ-plot  
qqnorm(e.2.selected)  
qqline(e.2.selected)



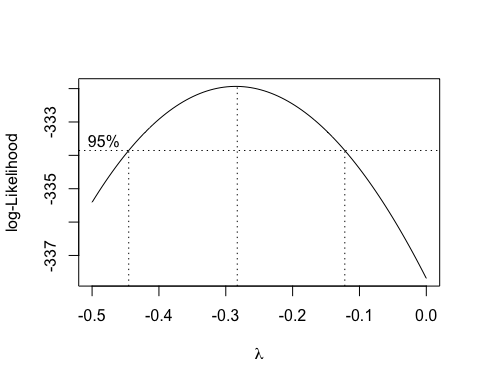
#Shapiro-Wilk Test  
shapiro.test(e.2.selected)

##   
## Shapiro-Wilk normality test  
##   
## data: e.2.selected  
## W = 0.97513, p-value = 2.951e-06

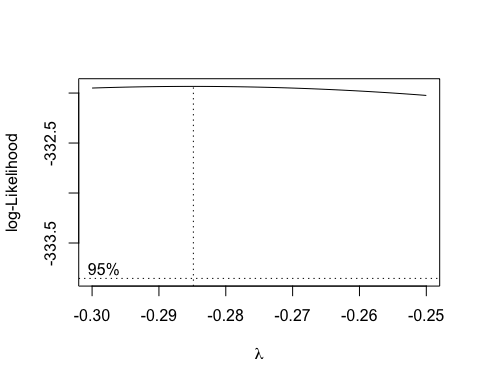
#boxcox  
library(MASS)  
boxcox(mod.2.selected)



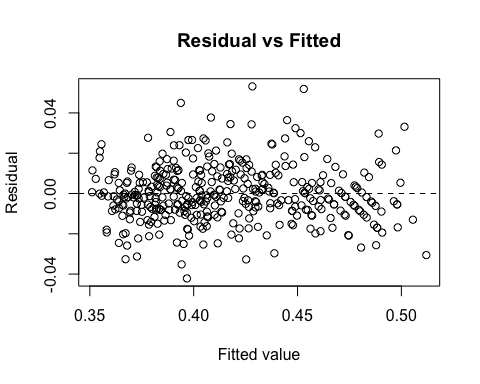
boxcox(mod.2.selected, lambda = seq(-0.5,0, length = 10))



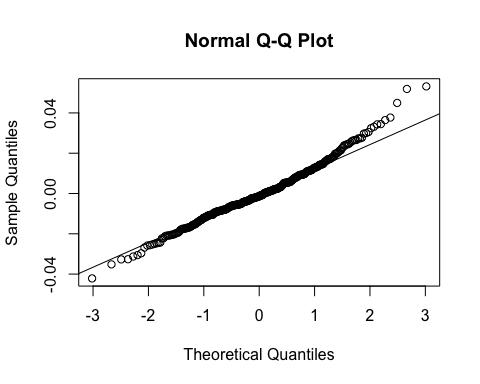
boxcox(mod.2.selected, lambda = seq(-0.3,-0.25, length = 10))



#Residual Analysis  
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  
shapiro.test(e.2.trans)

##   
## Shapiro-Wilk normality test  
##   
## data: e.2.trans  
## W = 0.98356, p-value = 0.0001951

summary(mod.2.trans)

##   
## Call:  
## lm(formula = mpg^(-0.285) ~ weight + modelyear + origin, data = auto\_mpg)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.042143 -0.008286 -0.001333 0.008135 0.053164   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.975e-01 1.682e-02 35.526 < 2e-16 \*\*\*  
## weight 3.491e-05 1.088e-06 32.084 < 2e-16 \*\*\*  
## modelyear -3.731e-03 2.038e-04 -18.313 < 2e-16 \*\*\*  
## origin2 -7.769e-03 2.168e-03 -3.583 0.000383 \*\*\*  
## origin3 -5.412e-03 2.172e-03 -2.492 0.013123 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.01397 on 387 degrees of freedom  
## Multiple R-squared: 0.8823, Adjusted R-squared: 0.8811   
## F-statistic: 725.5 on 4 and 387 DF, p-value: < 2.2e-16

#leverage  
sort(hatvalues(mod.2.trans), decreasing = TRUE)[c(1,2,3,4,5)]

## 355 210 296 356 54   
## 0.02725322 0.02667348 0.02664503 0.02637170 0.02619988

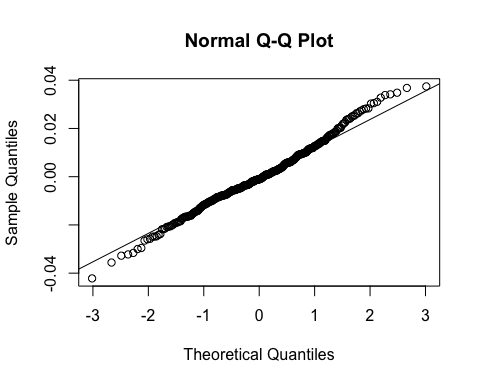
#studentized deleted residuals  
sort(abs(rstudent(mod.2.trans)), decreasing = TRUE )[c(1,2,3,4,5)]

## 165 124 111 382 274   
## 3.882467 3.788329 3.286674 3.067614 2.748952

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

## 111 274 276   
## 0.03684560 0.03020079 0.02938542

#delete outlier   
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  
qqnorm(e.deleted)  
qqline(e.deleted)



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

##   
## Shapiro-Wilk normality test  
##   
## data: e.deleted  
## W = 0.99213, p-value = 0.0378

#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)

## Analysis of Variance Table  
##   
## Model 1: mpg^(-0.285) ~ weight + modelyear + origin  
## Model 2: mpg^(-0.285) ~ weight + modelyear + origin + weight \* modelyear +   
## weight \* origin + modelyear \* origin  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 387 0.075545   
## 2 382 0.074799 5 0.00074532 0.7613 0.5782

summary(mod.3.full)

##   
## Call:  
## lm(formula = mpg^(-0.285) ~ weight + modelyear + origin + weight \*   
## modelyear + weight \* origin + modelyear \* origin, data = auto\_mpg)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.042715 -0.008672 -0.000952 0.008150 0.053045   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.161e-01 7.934e-02 7.766 7.52e-14 \*\*\*  
## weight 2.732e-05 2.427e-05 1.126 0.260878   
## modelyear -3.966e-03 1.050e-03 -3.778 0.000183 \*\*\*  
## origin2 2.639e-02 4.727e-02 0.558 0.576997   
## origin3 -2.630e-02 4.649e-02 -0.566 0.571966   
## weight:modelyear 9.728e-08 3.244e-07 0.300 0.764449   
## weight:origin2 1.194e-06 3.803e-06 0.314 0.753725   
## weight:origin3 7.847e-06 5.185e-06 1.513 0.131047   
## modelyear:origin2 -4.946e-04 6.258e-04 -0.790 0.429843   
## modelyear:origin3 4.002e-05 5.946e-04 0.067 0.946365   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.01399 on 382 degrees of freedom  
## Multiple R-squared: 0.8835, Adjusted R-squared: 0.8808   
## F-statistic: 321.9 on 9 and 382 DF, p-value: < 2.2e-16

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

##   
## Call:  
## lm(formula = mpg^(-0.285) ~ weight + modelyear + origin, data = auto\_mpg)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.042143 -0.008286 -0.001333 0.008135 0.053164   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.975e-01 1.682e-02 35.526 < 2e-16 \*\*\*  
## weight 3.491e-05 1.088e-06 32.084 < 2e-16 \*\*\*  
## modelyear -3.731e-03 2.038e-04 -18.313 < 2e-16 \*\*\*  
## origin2 -7.769e-03 2.168e-03 -3.583 0.000383 \*\*\*  
## origin3 -5.412e-03 2.172e-03 -2.492 0.013123 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.01397 on 387 degrees of freedom  
## Multiple R-squared: 0.8823, Adjusted R-squared: 0.8811   
## F-statistic: 725.5 on 4 and 387 DF, p-value: < 2.2e-16

#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)

## fit lwr upr  
## 1 25.92349 26.63265 25.23841