Cars-Predictions.R

wilhe

2023-12-11

auto\_df <- read.csv("C:/Users/wilhe/Downloads/auto-mpg(1).csv")  
#load all necessary libraries  
library(ggplot2)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(psych)

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

library(MASS)

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

#looking at the first few rows and a summary of the dataframe  
head(auto\_df)

## mpg cylinder displacement horsepower weight acceleration model.year origin  
## 1 18 8 307 130 3504 12.0 70 1  
## 2 15 8 350 165 3693 11.5 70 1  
## 3 18 8 318 150 3436 11.0 70 1  
## 4 16 8 304 150 3433 12.0 70 1  
## 5 17 8 302 140 3449 10.5 70 1  
## 6 15 8 429 198 4341 10.0 70 1  
## car.name  
## 1 chevrolet chevelle malibu  
## 2 buick skylark 320  
## 3 plymouth satellite  
## 4 amc rebel sst  
## 5 ford torino  
## 6 ford galaxie 500

summary(auto\_df)

## mpg cylinder displacement horsepower   
## Min. : 9.00 Min. :3.000 Min. : 68.0 Length:398   
## 1st Qu.:17.50 1st Qu.:4.000 1st Qu.:104.2 Class :character   
## Median :23.00 Median :4.000 Median :148.5 Mode :character   
## Mean :23.51 Mean :5.455 Mean :193.4   
## 3rd Qu.:29.00 3rd Qu.:8.000 3rd Qu.:262.0   
## Max. :46.60 Max. :8.000 Max. :455.0   
## weight acceleration model.year origin   
## Min. :1613 Min. : 8.00 Min. :70.00 Min. :1.000   
## 1st Qu.:2224 1st Qu.:13.82 1st Qu.:73.00 1st Qu.:1.000   
## Median :2804 Median :15.50 Median :76.00 Median :1.000   
## Mean :2970 Mean :15.57 Mean :76.01 Mean :1.573   
## 3rd Qu.:3608 3rd Qu.:17.18 3rd Qu.:79.00 3rd Qu.:2.000   
## Max. :5140 Max. :24.80 Max. :82.00 Max. :3.000   
## car.name   
## Length:398   
## Class :character   
## Mode :character   
##   
##   
##

#horsepower is a character type instead of the expected numeric as the others  
  
#when the unique method is run, it shows some values were inputed as question marks  
unique(auto\_df$horsepower)

## [1] "130" "165" "150" "140" "198" "220" "215" "225" "190" "170" "160" "95"   
## [13] "97" "85" "88" "46" "87" "90" "113" "200" "210" "193" "?" "100"  
## [25] "105" "175" "153" "180" "110" "72" "86" "70" "76" "65" "69" "60"   
## [37] "80" "54" "208" "155" "112" "92" "145" "137" "158" "167" "94" "107"  
## [49] "230" "49" "75" "91" "122" "67" "83" "78" "52" "61" "93" "148"  
## [61] "129" "96" "71" "98" "115" "53" "81" "79" "120" "152" "102" "108"  
## [73] "68" "58" "149" "89" "63" "48" "66" "139" "103" "125" "133" "138"  
## [85] "135" "142" "77" "62" "132" "84" "64" "74" "116" "82"

#replaced all the question marks with NA and then change to column type to numeric and replaced NA with the mean  
auto\_df$horsepower <- replace(auto\_df$horsepower, auto\_df$horsepower == "?", NA)  
auto\_df$horsepower <- as.numeric(auto\_df$horsepower)  
unique(auto\_df$horsepower)

## [1] 130 165 150 140 198 220 215 225 190 170 160 95 97 85 88 46 87 90 113  
## [20] 200 210 193 NA 100 105 175 153 180 110 72 86 70 76 65 69 60 80 54  
## [39] 208 155 112 92 145 137 158 167 94 107 230 49 75 91 122 67 83 78 52  
## [58] 61 93 148 129 96 71 98 115 53 81 79 120 152 102 108 68 58 149 89  
## [77] 63 48 66 139 103 125 133 138 135 142 77 62 132 84 64 74 116 82

auto\_df$horsepower <- replace(auto\_df$horsepower, is.na(auto\_df$horsepower),   
 mean(auto\_df$horsepower, na.rm = TRUE))  
  
#check the summary again  
summary(auto\_df)

## mpg cylinder displacement horsepower weight   
## Min. : 9.00 Min. :3.000 Min. : 68.0 Min. : 46.0 Min. :1613   
## 1st Qu.:17.50 1st Qu.:4.000 1st Qu.:104.2 1st Qu.: 76.0 1st Qu.:2224   
## Median :23.00 Median :4.000 Median :148.5 Median : 95.0 Median :2804   
## Mean :23.51 Mean :5.455 Mean :193.4 Mean :104.5 Mean :2970   
## 3rd Qu.:29.00 3rd Qu.:8.000 3rd Qu.:262.0 3rd Qu.:125.0 3rd Qu.:3608   
## Max. :46.60 Max. :8.000 Max. :455.0 Max. :230.0 Max. :5140   
## acceleration model.year origin car.name   
## Min. : 8.00 Min. :70.00 Min. :1.000 Length:398   
## 1st Qu.:13.82 1st Qu.:73.00 1st Qu.:1.000 Class :character   
## Median :15.50 Median :76.00 Median :1.000 Mode :character   
## Mean :15.57 Mean :76.01 Mean :1.573   
## 3rd Qu.:17.18 3rd Qu.:79.00 3rd Qu.:2.000   
## Max. :24.80 Max. :82.00 Max. :3.000

#split the data into train and test sets  
auto\_train <- auto\_df[1:300, 1:8]  
auto\_test <- auto\_df[301:398, 1:8]  
  
#simple linear regression models  
simplemodel1 <- lm(mpg ~ cylinder, data=auto\_train)  
summary(simplemodel1)

##   
## Call:  
## lm(formula = mpg ~ cylinder, data = auto\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.0593 -2.2405 -0.3613 1.7595 16.9803   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 37.8781 0.7224 52.44 <2e-16 \*\*\*  
## cylinder -2.9396 0.1191 -24.69 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.607 on 298 degrees of freedom  
## Multiple R-squared: 0.6717, Adjusted R-squared: 0.6706   
## F-statistic: 609.7 on 1 and 298 DF, p-value: < 2.2e-16

simplemodel2 <- lm(mpg ~ displacement, data=auto\_train)  
summary(simplemodel2)

##   
## Call:  
## lm(formula = mpg ~ displacement, data = auto\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.8432 -1.9571 -0.4975 1.9047 16.2304   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 31.250836 0.429861 72.70 <2e-16 \*\*\*  
## displacement -0.048680 0.001786 -27.26 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.368 on 298 degrees of freedom  
## Multiple R-squared: 0.7138, Adjusted R-squared: 0.7129   
## F-statistic: 743.3 on 1 and 298 DF, p-value: < 2.2e-16

simplemodel3 <- lm(mpg ~ horsepower, data=auto\_train)  
summary(simplemodel3)

##   
## Call:  
## lm(formula = mpg ~ horsepower, data = auto\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.7972 -2.7863 -0.3347 2.4683 14.2993   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 34.807601 0.645963 53.88 <2e-16 \*\*\*  
## horsepower -0.125144 0.005432 -23.04 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.775 on 298 degrees of freedom  
## Multiple R-squared: 0.6404, Adjusted R-squared: 0.6392   
## F-statistic: 530.7 on 1 and 298 DF, p-value: < 2.2e-16

simplemodel4 <- lm(mpg ~ weight, data=auto\_train)  
summary(simplemodel4)

##   
## Call:  
## lm(formula = mpg ~ weight, data = auto\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.1077 -1.8842 -0.0333 1.7275 15.1232   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 40.3879027 0.6368804 63.41 <2e-16 \*\*\*  
## weight -0.0062524 0.0001957 -31.96 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.992 on 298 degrees of freedom  
## Multiple R-squared: 0.7741, Adjusted R-squared: 0.7733   
## F-statistic: 1021 on 1 and 298 DF, p-value: < 2.2e-16

simplemodel5 <- lm(mpg ~ acceleration, data=auto\_train)  
summary(simplemodel5)

##   
## Call:  
## lm(formula = mpg ~ acceleration, data = auto\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.2441 -4.1160 -0.9237 3.0894 16.2186   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.5588 1.8243 2.499 0.013 \*   
## acceleration 1.0641 0.1177 9.044 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.577 on 298 degrees of freedom  
## Multiple R-squared: 0.2154, Adjusted R-squared: 0.2127   
## F-statistic: 81.8 on 1 and 298 DF, p-value: < 2.2e-16

simplemodel6 <- lm(mpg ~ model.year, data=auto\_train)  
summary(simplemodel6)

##   
## Call:  
## lm(formula = mpg ~ model.year, data = auto\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.967 -4.978 -1.447 4.249 20.023   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -26.7513 9.3692 -2.855 0.0046 \*\*   
## model.year 0.6388 0.1258 5.079 6.71e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.04 on 298 degrees of freedom  
## Multiple R-squared: 0.07966, Adjusted R-squared: 0.07657   
## F-statistic: 25.79 on 1 and 298 DF, p-value: 6.712e-07

simplemodel7 <- lm(mpg ~ origin, data=auto\_train)  
summary(simplemodel7)

##   
## Call:  
## lm(formula = mpg ~ origin, data = auto\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.5637 -3.6457 -0.3916 2.6084 19.6223   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.3055 0.6520 20.41 <2e-16 \*\*\*  
## origin 5.0861 0.3958 12.85 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.051 on 298 degrees of freedom  
## Multiple R-squared: 0.3565, Adjusted R-squared: 0.3543   
## F-statistic: 165.1 on 1 and 298 DF, p-value: < 2.2e-16

#multiple linear regression  
multimodel\_all <- lm(mpg ~ ., data= auto\_train)  
summary(multimodel\_all)

##   
## Call:  
## lm(formula = mpg ~ ., data = auto\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.2958 -1.6377 0.1021 1.5690 13.5802   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.8264772 4.9557290 1.176 0.240670   
## cylinder -0.4555536 0.2979923 -1.529 0.127411   
## displacement 0.0100198 0.0066682 1.503 0.134018   
## horsepower -0.0166816 0.0117498 -1.420 0.156753   
## weight -0.0053422 0.0005648 -9.458 < 2e-16 \*\*\*  
## acceleration -0.0245665 0.0939317 -0.262 0.793864   
## model.year 0.4433173 0.0602869 7.353 1.96e-12 \*\*\*  
## origin 0.9840977 0.2949226 3.337 0.000957 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.674 on 292 degrees of freedom  
## Multiple R-squared: 0.8232, Adjusted R-squared: 0.819   
## F-statistic: 194.3 on 7 and 292 DF, p-value: < 2.2e-16

multimodel2 <- lm(mpg ~ weight + model.year + origin, data = auto\_train)  
summary(multimodel2)

##   
## Call:  
## lm(formula = mpg ~ weight + model.year + origin, data = auto\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.1750 -1.5586 0.0463 1.6506 13.6915   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.2289898 4.3438467 0.743 0.45786   
## weight -0.0056852 0.0002215 -25.664 < 2e-16 \*\*\*  
## model.year 0.4585332 0.0559760 8.192 7.82e-15 \*\*\*  
## origin 0.8495008 0.2646593 3.210 0.00147 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.676 on 296 degrees of freedom  
## Multiple R-squared: 0.8206, Adjusted R-squared: 0.8188   
## F-statistic: 451.3 on 3 and 296 DF, p-value: < 2.2e-16

multimodel3 <- lm(mpg ~ displacement + weight, data = auto\_train)  
summary(multimodel3)

##   
## Call:  
## lm(formula = mpg ~ displacement + weight, data = auto\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.4362 -1.8464 -0.1621 1.6470 15.2024   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 38.7466846 0.8832868 43.866 <2e-16 \*\*\*  
## displacement -0.0113429 0.0042751 -2.653 0.0084 \*\*   
## weight -0.0049513 0.0005273 -9.390 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.963 on 297 degrees of freedom  
## Multiple R-squared: 0.7793, Adjusted R-squared: 0.7778   
## F-statistic: 524.5 on 2 and 297 DF, p-value: < 2.2e-16

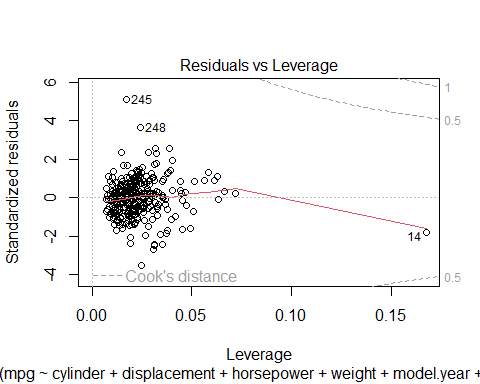
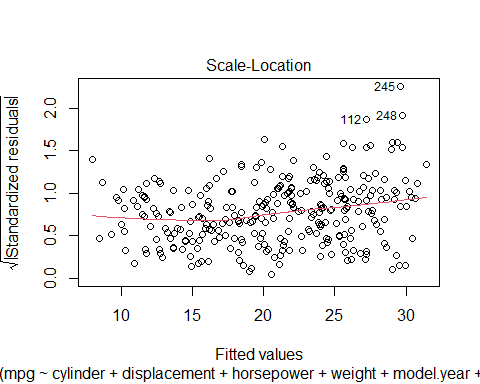
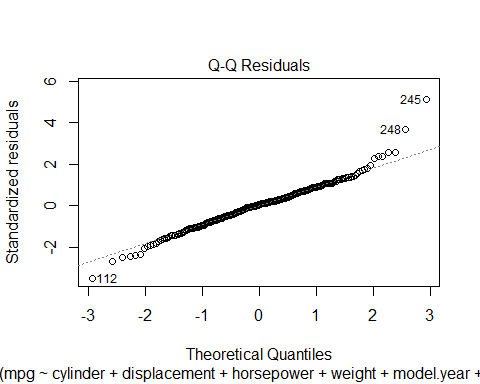
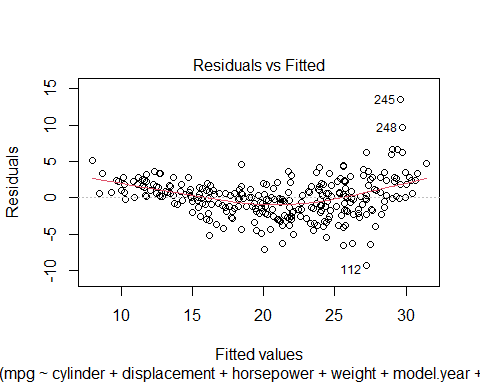
#stepwise model  
stepwise\_model <- stepAIC(multimodel\_all, direction="both")

## Start: AIC=598.06  
## mpg ~ cylinder + displacement + horsepower + weight + acceleration +   
## model.year + origin  
##   
## Df Sum of Sq RSS AIC  
## - acceleration 1 0.49 2088.6 596.14  
## <none> 2088.1 598.06  
## - horsepower 1 14.41 2102.5 598.13  
## - displacement 1 16.15 2104.2 598.38  
## - cylinder 1 16.71 2104.8 598.46  
## - origin 1 79.62 2167.7 607.29  
## - model.year 1 386.68 2474.8 647.03  
## - weight 1 639.75 2727.8 676.24  
##   
## Step: AIC=596.14  
## mpg ~ cylinder + displacement + horsepower + weight + model.year +   
## origin  
##   
## Df Sum of Sq RSS AIC  
## <none> 2088.6 596.14  
## - cylinder 1 16.35 2104.9 596.47  
## - horsepower 1 17.08 2105.6 596.58  
## - displacement 1 17.19 2105.8 596.59  
## + acceleration 1 0.49 2088.1 598.06  
## - origin 1 80.83 2169.4 605.53  
## - model.year 1 388.34 2476.9 645.29  
## - weight 1 829.98 2918.5 694.52

summary(stepwise\_model)

##   
## Call:  
## lm(formula = mpg ~ cylinder + displacement + horsepower + weight +   
## model.year + origin, data = auto\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.2375 -1.6420 0.1006 1.5828 13.4968   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.3264184 4.5647579 1.167 0.24422   
## cylinder -0.4489092 0.2964350 -1.514 0.13101   
## displacement 0.0102492 0.0065997 1.553 0.12151   
## horsepower -0.0149336 0.0096484 -1.548 0.12275   
## weight -0.0054099 0.0005014 -10.791 < 2e-16 \*\*\*  
## model.year 0.4439322 0.0601452 7.381 1.63e-12 \*\*\*  
## origin 0.9893078 0.2937809 3.368 0.00086 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.67 on 293 degrees of freedom  
## Multiple R-squared: 0.8232, Adjusted R-squared: 0.8196   
## F-statistic: 227.4 on 6 and 293 DF, p-value: < 2.2e-16

#predict for the test set  
predictions <- predict(stepwise\_model, newdata = auto\_test)  
  
#plotted the residuals to see how well the linear model works  
plot(stepwise\_model)



#summary of residuals of the model regression line  
summary(resid(stepwise\_model))

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -9.2376 -1.6420 0.1006 0.0000 1.5828 13.4968

#summary of residuals the actual minus the predicted values  
resids <- auto\_test$mpg - predictions  
summary(resids)

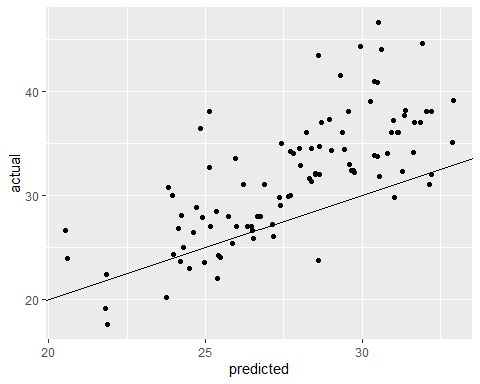
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -4.894 1.072 3.381 3.999 6.200 16.091

mean(resids)

## [1] 3.999132

#created a scatter plot to compare the predicted values to the actual values  
predict\_df <- data.frame(predicted = predictions, actual = auto\_test$mpg)   
ggplot(predict\_df, aes(x=predicted, y=actual)) +  
 geom\_point() +  
 geom\_abline(interecept=0, slope=1)

## Warning in geom\_abline(interecept = 0, slope = 1): Ignoring unknown parameters:  
## `interecept`



#created a histogram showing the predicted and actual values  
ggplot(predict\_df) +  
 geom\_histogram(aes(x=predicted, fill="predicted"), alpha=0.5) +  
 geom\_histogram(aes(x=actual, fill="observed"), alpha=0.5) +  
 xlab("mpg")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

