DAV\_Midterm\_25thMarch

Akhil Patil

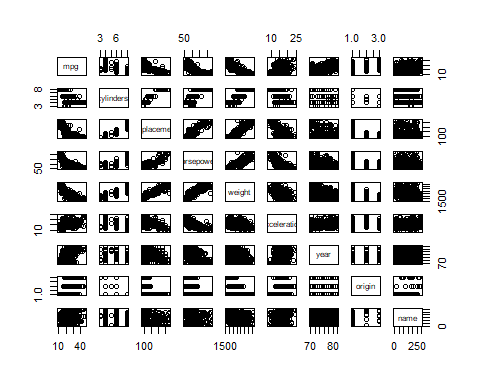
March 25, 2019

DAV Mid - Term (Part 2)

#install.packages("ISLR")  
library(ISLR)  
data(Auto)

A. Produce a scatterplot matrix which include all the variables in the data set.

pairs(Auto)



B.Compute the matrix of correlations between the variables using the function cor(). You will need to exclude the “name” variable, which is qualitative.

names(Auto)

## [1] "mpg" "cylinders" "displacement" "horsepower"   
## [5] "weight" "acceleration" "year" "origin"   
## [9] "name"

cor(Auto[1:8])

## mpg cylinders displacement horsepower weight  
## mpg 1.0000000 -0.7776175 -0.8051269 -0.7784268 -0.8322442  
## cylinders -0.7776175 1.0000000 0.9508233 0.8429834 0.8975273  
## displacement -0.8051269 0.9508233 1.0000000 0.8972570 0.9329944  
## horsepower -0.7784268 0.8429834 0.8972570 1.0000000 0.8645377  
## weight -0.8322442 0.8975273 0.9329944 0.8645377 1.0000000  
## acceleration 0.4233285 -0.5046834 -0.5438005 -0.6891955 -0.4168392  
## year 0.5805410 -0.3456474 -0.3698552 -0.4163615 -0.3091199  
## origin 0.5652088 -0.5689316 -0.6145351 -0.4551715 -0.5850054  
## acceleration year origin  
## mpg 0.4233285 0.5805410 0.5652088  
## cylinders -0.5046834 -0.3456474 -0.5689316  
## displacement -0.5438005 -0.3698552 -0.6145351  
## horsepower -0.6891955 -0.4163615 -0.4551715  
## weight -0.4168392 -0.3091199 -0.5850054  
## acceleration 1.0000000 0.2903161 0.2127458  
## year 0.2903161 1.0000000 0.1815277  
## origin 0.2127458 0.1815277 1.0000000

C. perform a multiple linear regression with “mpg” as the response and all other variables except “name” as the predictors.Comment on the following:

fit2 <- lm(mpg ~ . - name, data = Auto)  
summary(fit2)

##   
## Call:  
## lm(formula = mpg ~ . - name, data = Auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.5903 -2.1565 -0.1169 1.8690 13.0604   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -17.218435 4.644294 -3.707 0.00024 \*\*\*  
## cylinders -0.493376 0.323282 -1.526 0.12780   
## displacement 0.019896 0.007515 2.647 0.00844 \*\*   
## horsepower -0.016951 0.013787 -1.230 0.21963   
## weight -0.006474 0.000652 -9.929 < 2e-16 \*\*\*  
## acceleration 0.080576 0.098845 0.815 0.41548   
## year 0.750773 0.050973 14.729 < 2e-16 \*\*\*  
## origin 1.426141 0.278136 5.127 4.67e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.328 on 384 degrees of freedom  
## Multiple R-squared: 0.8215, Adjusted R-squared: 0.8182   
## F-statistic: 252.4 on 7 and 384 DF, p-value: < 2.2e-16

1. Is there a relationship between the predictors and the response ?

By testing the hypothesis  
 h0:bi=0. ha: bi != 0

Here h0 is Null Hypotheses ,ha is alternative hypotheses , bi are the predictor variables.

we get pvalue < 2.2e-16 which is less than signifance level of 0.01 1%. We can reject Null hypothesis. So , there is relationship between the mpg and other variables.

1. Which predictors appear to have a statistically significant relationship to the response ?

ANS :- when we check the p-values associated with each predictor variables t-statistic. We can confirm that except “cylinders”, “horsepower” and “acceleration” other predictors like “displacement” , “weight” , “year” , “origin” are statistically significant.

1. What does the coefficient for the “year” variable suggest ?

The variable coefficient of the “year” tells us that the average effect of an increase of 1 year is an increase of 0.7507727 in “mpg” and all other predictors remaining constant. That is cars become more fuel efficient every year by almost 1 mpg / year.

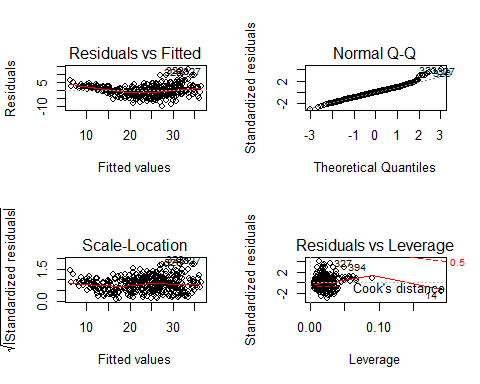
D. Obtain the 95% confidence interval of the coefficients

confint(fit2,level = 0.95)

## 2.5 % 97.5 %  
## (Intercept) -26.349864469 -8.087004775  
## cylinders -1.129001385 0.142248747  
## displacement 0.005119788 0.034671499  
## horsepower -0.044058392 0.010156103  
## weight -0.007756074 -0.005192013  
## acceleration -0.113769257 0.274920933  
## year 0.650551315 0.850994041  
## origin 0.879280169 1.973000822

E. Produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Does the residual plots suggest any unusually large outliers ? Does the leverage plots identify any observations with unusually high leverages ?

par(mfrow = c(2, 2))  
plot(fit2)



Plot of residuals versus fitted values indicates the presence of a bit non linearity in the data. The plot of residuals versus leverage shows that there a few outliers (higher than 2 or lower than -2) and one high leverage point(point 14).

F)set aside one-forth of the observations to predict using the above linear model. compute the mean square error of the predicted values.

auto\_sub = Auto[sample(nrow(Auto), 294), ]  
fit1=lm(mpg ~ . - name, data = auto\_sub)  
summary.fit1 = summary(fit1)  
MeanSquareError = mean(summary.fit1$residuals^2)  
MeanSquareError

## [1] 11.43106