Predictive Modeling Exercises Week 2

Amol Gote

7/18/2020

#Problem 9

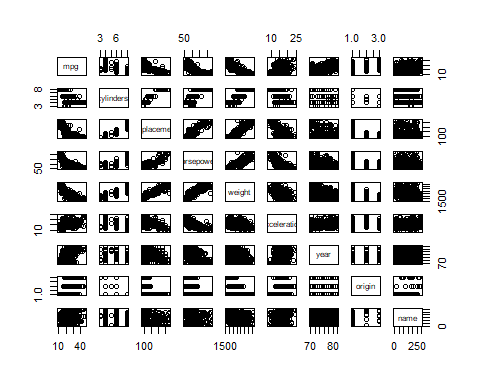
names(Auto)

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

#head(Auto)

# 9a - Produce a scatterplot matrix which includes all of the variables in the data set.

pairs(Auto)



auto\_data <- select (Auto,-c(name))  
#auto\_data

# 9b - Compute the matrix of correlations between the variables using the function cor(). You will need to exclude the name variable, cor() which is qualitative.

cor(auto\_data)

## 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

# 9c - Use the lm() function to perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Use the summary() function to print the results.Comment on the output

multiple\_linear\_regression = lm(mpg ~., data = auto\_data)  
summary(multiple\_linear\_regression)

##   
## Call:  
## lm(formula = mpg ~ ., data = auto\_data)  
##   
## 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

9c.i. Is there a relationship between the predictors and the response?

F-Test indicates a very low p-value, hence there is relationship between the predictors and the response.

9c.ii. Which predictors appear to have a statistically significant relationship to the response?

Following variables **have** statistically significant relationship to mpg

1. weight  
2. year  
3. origin  
4. displacement

Following variables **do not have** have statistically significant relationship to mpg

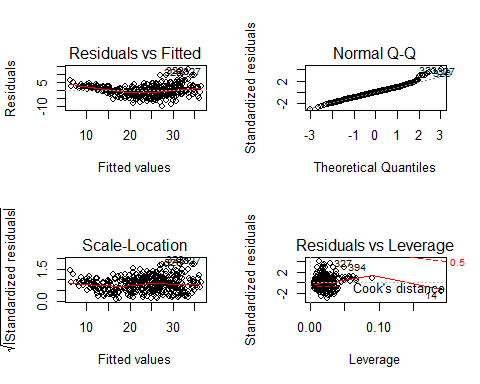
1. cylinders  
2. horsepower  
3. acceleration

9c.iii. What does the coefficient for the year variable suggest?

Coefficient is **0.750773**, which indicates for every passing year, mpg increases by 0.750773, so in summary cars become more fuel efficient every year by 0.750773 miles per gallon.

# 9d Use the plot() function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

par(mfrow=c(2,2))  
plot(multiple\_linear\_regression)



1. Residual v/s predicted (fitted) shows strong pattern in the residuals which indicates non-linearity in the data
2. Residual plot does show values in top right corner with values ranging from 323 to 327 which are unusally large outliers.
3. Leverage plot has point labelled as 14 which can serve as potential high leverage point.

# 9e Use the \* and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

interactions\_linear\_regression\_fit <- lm(mpg ~ . \* ., data = auto\_data)  
summary(interactions\_linear\_regression\_fit)

##   
## Call:  
## lm(formula = mpg ~ . \* ., data = auto\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.6303 -1.4481 0.0596 1.2739 11.1386   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.548e+01 5.314e+01 0.668 0.50475   
## cylinders 6.989e+00 8.248e+00 0.847 0.39738   
## displacement -4.785e-01 1.894e-01 -2.527 0.01192 \*   
## horsepower 5.034e-01 3.470e-01 1.451 0.14769   
## weight 4.133e-03 1.759e-02 0.235 0.81442   
## acceleration -5.859e+00 2.174e+00 -2.696 0.00735 \*\*  
## year 6.974e-01 6.097e-01 1.144 0.25340   
## origin -2.090e+01 7.097e+00 -2.944 0.00345 \*\*  
## cylinders:displacement -3.383e-03 6.455e-03 -0.524 0.60051   
## cylinders:horsepower 1.161e-02 2.420e-02 0.480 0.63157   
## cylinders:weight 3.575e-04 8.955e-04 0.399 0.69000   
## cylinders:acceleration 2.779e-01 1.664e-01 1.670 0.09584 .   
## cylinders:year -1.741e-01 9.714e-02 -1.793 0.07389 .   
## cylinders:origin 4.022e-01 4.926e-01 0.816 0.41482   
## displacement:horsepower -8.491e-05 2.885e-04 -0.294 0.76867   
## displacement:weight 2.472e-05 1.470e-05 1.682 0.09342 .   
## displacement:acceleration -3.479e-03 3.342e-03 -1.041 0.29853   
## displacement:year 5.934e-03 2.391e-03 2.482 0.01352 \*   
## displacement:origin 2.398e-02 1.947e-02 1.232 0.21875   
## horsepower:weight -1.968e-05 2.924e-05 -0.673 0.50124   
## horsepower:acceleration -7.213e-03 3.719e-03 -1.939 0.05325 .   
## horsepower:year -5.838e-03 3.938e-03 -1.482 0.13916   
## horsepower:origin 2.233e-03 2.930e-02 0.076 0.93931   
## weight:acceleration 2.346e-04 2.289e-04 1.025 0.30596   
## weight:year -2.245e-04 2.127e-04 -1.056 0.29182   
## weight:origin -5.789e-04 1.591e-03 -0.364 0.71623   
## acceleration:year 5.562e-02 2.558e-02 2.174 0.03033 \*   
## acceleration:origin 4.583e-01 1.567e-01 2.926 0.00365 \*\*  
## year:origin 1.393e-01 7.399e-02 1.882 0.06062 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.695 on 363 degrees of freedom  
## Multiple R-squared: 0.8893, Adjusted R-squared: 0.8808   
## F-statistic: 104.2 on 28 and 363 DF, p-value: < 2.2e-16

interactions\_linear\_regression\_fit <- lm(mpg ~ displacement:weight + weight:year + displacement:year, data=auto\_data)  
summary(interactions\_linear\_regression\_fit)

##   
## Call:  
## lm(formula = mpg ~ displacement:weight + weight:year + displacement:year,   
## data = auto\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.3640 -3.1823 -0.5201 2.4878 17.1727   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.719e+01 1.716e+00 21.672 < 2e-16 \*\*\*  
## displacement:weight -5.950e-06 2.314e-06 -2.572 0.01049 \*   
## weight:year -3.292e-05 1.050e-05 -3.136 0.00184 \*\*   
## displacement:year -1.633e-04 1.539e-04 -1.061 0.28916   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.7 on 388 degrees of freedom  
## Multiple R-squared: 0.6401, Adjusted R-squared: 0.6374   
## F-statistic: 230.1 on 3 and 388 DF, p-value: < 2.2e-16

interactions\_linear\_regression\_fit <- lm(mpg ~ displacement \* weight + weight \* year + displacement \* year, data=auto\_data)  
summary(interactions\_linear\_regression\_fit)

##   
## Call:  
## lm(formula = mpg ~ displacement \* weight + weight \* year + displacement \*   
## year, data = auto\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.7299 -1.6773 -0.0834 1.2071 13.5557   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.250e+01 1.889e+01 -2.250 0.025 \*   
## displacement 4.021e-02 7.820e-02 0.514 0.607   
## weight -4.230e-03 1.086e-02 -0.390 0.697   
## year 1.269e+00 2.438e-01 5.205 3.17e-07 \*\*\*  
## displacement:weight 1.880e-05 2.319e-06 8.107 7.00e-15 \*\*\*  
## weight:year -7.481e-05 1.429e-04 -0.524 0.601   
## displacement:year -1.455e-03 1.070e-03 -1.359 0.175   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.958 on 385 degrees of freedom  
## Multiple R-squared: 0.8586, Adjusted R-squared: 0.8564   
## F-statistic: 389.6 on 6 and 385 DF, p-value: < 2.2e-16

Lower p-value indicates significances, so in this case following interactions appear to be significant

1. displacement and year
2. acceleration and year
3. acceleration and origin
4. displacement and weight

# 9f Try a few different transformations of the variables, such as log(X), √X, X2. Comment on your findings.

log\_transformation <- lm(mpg ~ log(horsepower) + log(weight) + log(acceleration), data = auto\_data)  
summary(log\_transformation)

##   
## Call:  
## lm(formula = mpg ~ log(horsepower) + log(weight) + log(acceleration),   
## data = auto\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.8237 -2.5240 -0.2389 2.0105 15.3681   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 190.152 8.255 23.035 < 2e-16 \*\*\*  
## log(horsepower) -11.799 1.933 -6.103 2.53e-09 \*\*\*  
## log(weight) -12.306 1.820 -6.762 5.03e-11 \*\*\*  
## log(acceleration) -5.363 1.970 -2.723 0.00677 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.961 on 388 degrees of freedom  
## Multiple R-squared: 0.7445, Adjusted R-squared: 0.7425   
## F-statistic: 376.8 on 3 and 388 DF, p-value: < 2.2e-16

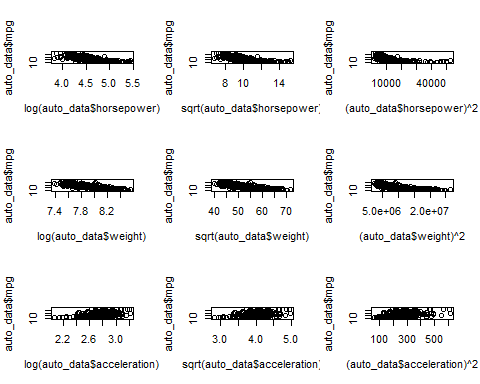
sqrt\_transformation <- lm(mpg ~ sqrt(horsepower) + sqrt(weight) + sqrt(acceleration), data = auto\_data)  
summary(sqrt\_transformation)

##   
## Call:  
## lm(formula = mpg ~ sqrt(horsepower) + sqrt(weight) + sqrt(acceleration),   
## data = auto\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.1090 -2.7005 -0.2846 2.1044 15.8554   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 76.1791 5.0416 15.110 < 2e-16 \*\*\*  
## sqrt(horsepower) -1.7276 0.3713 -4.653 4.50e-06 \*\*\*  
## sqrt(weight) -0.5358 0.0672 -7.973 1.75e-14 \*\*\*  
## sqrt(acceleration) -1.6296 1.0218 -1.595 0.112   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.088 on 388 degrees of freedom  
## Multiple R-squared: 0.7278, Adjusted R-squared: 0.7257   
## F-statistic: 345.9 on 3 and 388 DF, p-value: < 2.2e-16

square\_transformation <- lm(mpg ~ horsepower^2 + weight^2 + acceleration^2, data = auto\_data)  
summary(square\_transformation)

##   
## Call:  
## lm(formula = mpg ~ horsepower^2 + weight^2 + acceleration^2,   
## data = auto\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.079 -2.736 -0.331 2.170 16.262   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 45.6782929 2.4085431 18.965 < 2e-16 \*\*\*  
## horsepower -0.0474956 0.0159891 -2.970 0.00316 \*\*   
## weight -0.0057894 0.0005776 -10.024 < 2e-16 \*\*\*  
## acceleration -0.0020657 0.1233378 -0.017 0.98665   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.246 on 388 degrees of freedom  
## Multiple R-squared: 0.7064, Adjusted R-squared: 0.7041   
## F-statistic: 311.1 on 3 and 388 DF, p-value: < 2.2e-16

par(mfrow = c(3, 3))  
plot(log(auto\_data$horsepower), auto\_data$mpg)  
plot(sqrt(auto\_data$horsepower), auto\_data$mpg)  
plot((auto\_data$horsepower)^2, auto\_data$mpg)  
plot(log(auto\_data$weight), auto\_data$mpg)  
plot(sqrt(auto\_data$weight), auto\_data$mpg)  
plot((auto\_data$weight)^2, auto\_data$mpg)  
plot(log(auto\_data$acceleration), auto\_data$mpg)  
plot(sqrt(auto\_data$acceleration), auto\_data$mpg)  
plot((auto\_data$acceleration)^2, auto\_data$mpg)



1. Log resulted in highest F-Value and R Squared value.
2. Based on the plots below relation indicates significance
   1. MPG and Horse power
   2. MPG and weight

# Problem 10 - Carseats data set.

names(Carseats)

## [1] "Sales" "CompPrice" "Income" "Advertising" "Population"   
## [6] "Price" "ShelveLoc" "Age" "Education" "Urban"   
## [11] "US"

#head(Carseats)

# 10a Fit a multiple regression model to predict Sales using Price, Urban, and US.

multiple\_regression\_model1 <- lm(Sales ~ Price + Urban + US, data=Carseats)  
summary(multiple\_regression\_model1)

##   
## Call:  
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.9206 -1.6220 -0.0564 1.5786 7.0581   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.043469 0.651012 20.036 < 2e-16 \*\*\*  
## Price -0.054459 0.005242 -10.389 < 2e-16 \*\*\*  
## UrbanYes -0.021916 0.271650 -0.081 0.936   
## USYes 1.200573 0.259042 4.635 4.86e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.472 on 396 degrees of freedom  
## Multiple R-squared: 0.2393, Adjusted R-squared: 0.2335   
## F-statistic: 41.52 on 3 and 396 DF, p-value: < 2.2e-16

# 10b Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative!

attach(Carseats)  
str(data.frame(Price, Urban, US))

## 'data.frame': 400 obs. of 3 variables:  
## $ Price: num 120 83 80 97 128 72 108 120 124 124 ...  
## $ Urban: Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 2 2 1 1 ...  
## $ US : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 2 1 2 1 2 ...

Urban and US and qualitative variables.

1. **Price:** Multiple regression model indicates relationship between price and sales, the t-statistic p-value is is low. The coefficient is negative which idicates that price increases sales decreases. The coefficient value is -0.054459, so if price increases by $1000, number of units sold decreases by 54.45
2. **UrbanYes:** p-value is relatively high so it is not significant as far sales concerned. There is not enough evidence for relationship between location of store in Urban area in US and sales.
3. **USYEs:** There is relationship between USYes and Sales. So if the store is in US sales increases, this positive relationship as coefficient is positive. Coefficient value is 1.200573, so if the store is in US the sales increases by 1201 units.

# 10c Write out the model in equation form, being careful to handle the qualitative variables properly.

sales = 13.043469 + (-0.054459 \* Price) + (-0.021916 \* UrbanYes) + (1.200573 \* USYes)

OR

sales = 13.043469 + (-0.054459 \* Price) + (-0.021916 \* Urban) + (1.200573 \* US) With Urban=1 if the store is in an urban location and 0 if not, and US=1 if the store is in the US and 0 if not.

# 10d For which of the predictors can you reject the null hypothesis H0 : βj = 0?

We can reject the null hypothesis for Price and USYes since the p-values are significantly low.

# 10e On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

multiple\_regression\_model2 <- lm(Sales ~ Price + US, data=Carseats)  
summary(multiple\_regression\_model2)

##   
## Call:  
## lm(formula = Sales ~ Price + US, data = Carseats)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.9269 -1.6286 -0.0574 1.5766 7.0515   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.03079 0.63098 20.652 < 2e-16 \*\*\*  
## Price -0.05448 0.00523 -10.416 < 2e-16 \*\*\*  
## USYes 1.19964 0.25846 4.641 4.71e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.469 on 397 degrees of freedom  
## Multiple R-squared: 0.2393, Adjusted R-squared: 0.2354   
## F-statistic: 62.43 on 2 and 397 DF, p-value: < 2.2e-16

# 10f How well do the models in (a) and (e) fit the data?

anova(multiple\_regression\_model1, multiple\_regression\_model2)

## Analysis of Variance Table  
##   
## Model 1: Sales ~ Price + Urban + US  
## Model 2: Sales ~ Price + US  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 396 2420.8   
## 2 397 2420.9 -1 -0.03979 0.0065 0.9357

Based on the Residual standard error and R-squared they fit the data similarly. P-Value of F-Statistic is also not different.

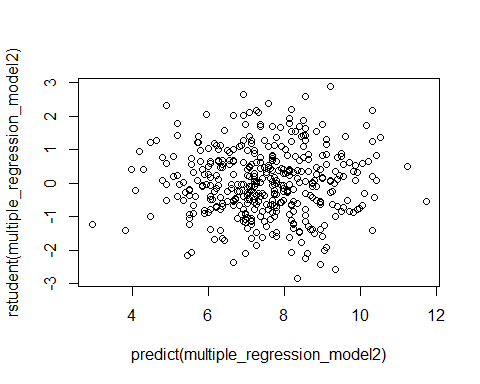
# 10g Using the model from (e), obtain 95% confidence intervals for the coefficient(s).

confint(multiple\_regression\_model2)

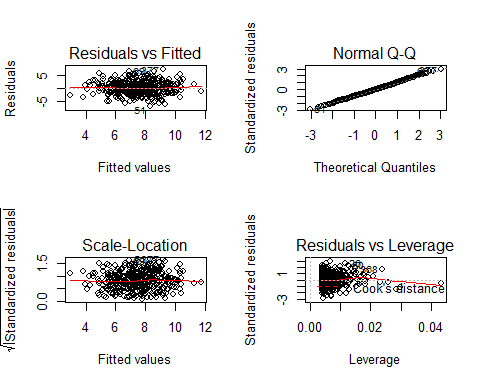
## 2.5 % 97.5 %  
## (Intercept) 11.79032020 14.27126531  
## Price -0.06475984 -0.04419543  
## USYes 0.69151957 1.70776632

# 10h Is there evidence of outliers or high leverage observations in the model from (e)?

plot(predict(multiple\_regression\_model2), rstudent(multiple\_regression\_model2))



par(mfrow=c(2,2))  
plot(multiple\_regression\_model2)



1. All r-Student residulas are within the bounding ranges of 3 and -3, so there are no potential outliers.
2. Residual vs Leverage plot does indicate the points that have high leverage (43)

hatvalues(multiple\_regression\_model2)[order(hatvalues(multiple\_regression\_model2), decreasing = T)][1]

## 43   
## 0.04333766