CMDA 3654: Assignment #3

Due on Friday, Oct 07, 2016

Jung Choi

"I have neither given nor received unauthorized assistance on this assignment."

October 7, 2016

1 MTCARS

1. (10 points) Load the **mtcars** dataset in R, and describe the dataset in your own words, in 23 lines.
Feel free to use help(mtcars) or ? mtcars for this problem and subsequent problems.

The **mtcars** is a data set that contains 32 different model vehicles comparing the fuel consumption and 10 other aspects that have relation to the fuel consumption for each vehicle. These 32 vehicles are set between the model years from 1973 to 1974. Each vehicle not only contains the MPG which is the biggest measurement for the fuel consumption but also important aspects that have influence on the overall fuel consumption.

2 LINEAR REGRESSION

2. (50 points) Linear regression. Consider **mpg** to be the response variable, and all other variables as features.

(a) Compute the correlation coefficient between mpg and all other features in the dataset. What are the two features most strongly correlated with **mpg**? (**Hint**: a strong correlation can be either positive or negative, use abs(x) to obtain the absolute value of a number x.)
The correlation coefficient between mpg and for all 10 variables, not including the mpg itself, came out in the following table below. The two highest coefficient came out to be wt and cyl.

wt	cyl	disp	hp	drat	vs	am	carb	gear	qsec
0.86766	**0.85216**	0.84755	0.77617	0.68117	0.66404	0.59983	0.55093	0.48029	0.41868

```
corr.cyl= cor(mtcars$mpg,mtcars$cyl)
corr.disp= cor(mtcars$mpg, mtcars$disp)
corr.hp= cor(mtcars$mpg, mtcars$hp)
corr.drat= cor(mtcars$mpg, mtcars$drat)
corr.wt= cor(mtcars$mpg, mtcars$wt)
corr.qsec= cor(mtcars$mpg, mtcars$qsec)
corr.vs= cor(mtcars$mpg, mtcars$vs)
corr.am= cor(mtcars$mpg, mtcars$am)
corr.gear= cor(mtcars$mpg, mtcars$gear)
corr.carb= cor(mtcars$mpg, mtcars$carb)
print(corr.cyl)
print(corr.disp)
print(corr.hp)
print(corr.drat)
print(corr.wt)
print(corr.qsec)
print(corr.vs)
print(corr.am)
print(corr.gear)
print(corr.carb)
sort(abs(cor(mtcars)[1,]), decreasing=TRUE)[2:11] #Ignore the correlation between
    the mpg to mpg.
```

(b) Fit two simple linear regression models: model 1 using the strongest feature from (a) and model 2 using the second strongest feature from (a). Report the linear regression formula (i.e., report the line equation) and the value of R^2 from the two models. If you had to choose between these two models, which one would you choose and why?

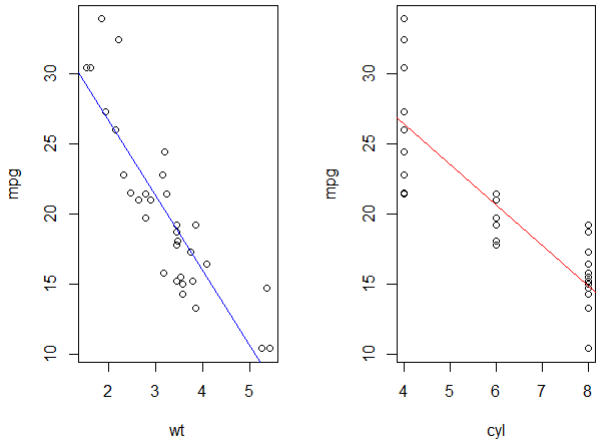

Figure 1: Two simple linear regression models

Both graphs have similar results for the R^2 values. Model1(mpg vs wt) contains the value of adjusted R^2 of 0.7446 and Model2(mpg vs cyl) contains the adjusted R^2 value of 0.7171. Which means that the Model1 has 74% of the variation in mpg that can be explained by the variable wt. Vice versa for the cyl as well. So, the best choice from the two would be model1 since the R^2 value is little higher and the

```
model1= lm(mpg~wt, data=mtcars)
model2= lm(mpg~cyl, data=mtcars)
summary(model1)
summary(model2)

# The regression equation for model 1 is
#         mpg = 37.2851 -5.3445*wt.
#The value of R^2 is:
#         Multiple R-squared:  0.7528,    Adjusted R-squared:  0.7446
# The regression equation for model 2 is
#         mpg = 37.8846 -2.8758*cyl.
#The value of R^2 is:
#         Multiple R-squared:  0.7262,    Adjusted R-squared:  0.7171

par(mfrow=c(1,2))
# plot(model2)
plot(mpg~wt,mtcars)
abline(model1,col="blue")
plot(mpg~cyl, mtcars)
abline(model2,col="red")
```

```
par(mfrow=c(2,2))
plot(model1)
plot(model2)
```

(c) Fit a multiple linear regression model with all features. Which features are significant in this model? What is the value of R^2 in this model?

Using the anova method to the likehood ratio, I figured the value for p-value to be 0.03742. Now, I can claim that the multiple linear regression model is more significant than the simple model. The R^2 for the multivariate model comes out to be 86.9%.

```
> mModel=lm(mpg~., data=mtcars)
> summary(mModel)

Call:
lm(formula = mpg ~ ., data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4506 -1.6044 -0.1196  1.2193  4.6271

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.30337   18.71788   0.657   0.5181
cyl         -0.11144    1.04502  -0.107   0.9161
disp         0.01334    0.01786   0.747   0.4635
hp          -0.02148    0.02177  -0.987   0.3350
drat         0.78711    1.63537   0.481   0.6353
wt          -3.71530    1.89441  -1.961   0.0633 .
qsec         0.82104    0.73084   1.123   0.2739
vs           0.31776    2.10451   0.151   0.8814
am           2.52023    2.05665   1.225   0.2340
gear         0.65541    1.49326   0.439   0.6652
carb        -0.19942    0.82875  -0.241   0.8122
---
Signif. codes:  0 ?**?0.001 ?*?0.01 ??0.05 ??0.1 ??1

Residual standard error: 2.65 on 21 degrees of freedom
Multiple R-squared:  0.869,     Adjusted R-squared:  0.8066
F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07

> anova(model1, model2, mModel)
Analysis of Variance Table

Model 1: mpg ~ wt
Model 2: mpg ~ cyl
Model 3: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     30 278.32
2     30 308.33  0   -30.012
3     21 147.49  9   160.840 2.5445 0.03742 *
---
Signif. codes:  0 ?**?0.001 ?*?0.01 ??0.05 ??0.1 ??1
```

(d) Using **stepAIC** , identify the best subset of features. Fit a multiple linear regression model using the best subset of features. Write down the regression formula and R^2 for this model. Are any of the features from (a) included in this model? Do they have the same coefficients as they had in model 1 or model 2 from (b)? If the coefficient values have changed, explain why.

mpg = 38.75179 -3.16697(wt) -0.94162(cyl) -0.01804(hp)
Multiple R-squared: 0.8431, Adjusted R-squared: 0.8263

As you can on the results below, wt and wt + cyl are both included as part of the search result, but it shows that the mMOdel from **part(b)** to be the most significant value out of all.

Start: AIC=115.94

The coefficient value will be different since we know that our models contain outliers and values that show significant changes on the variants.So, in order to create a best fit model, stepwise will search between the scope arguments and outputs the best model according to AIC.

```
> fit1 <- lm(mpg~1,data=mtcars)
> foo1 <- stepAIC(fit1,direction="both",
+                 scope=list(upper=mModel,lower=fit1))

Start:  AIC=115.94
mpg ~ 1

       Df Sum of Sq     RSS     AIC
+ wt    1    847.73  278.32  73.217
+ cyl   1    817.71  308.33  76.494
+ disp  1    808.89  317.16  77.397
+ hp    1    678.37  447.67  88.427
+ drat  1    522.48  603.57  97.988
+ vs    1    496.53  629.52  99.335
+ am    1    405.15  720.90 103.672
+ carb  1    341.78  784.27 106.369
+ gear  1    259.75  866.30 109.552
+ qsec  1    197.39  928.66 111.776
<none>              1126.05 115.943

Step:  AIC=73.22
mpg ~ wt

       Df Sum of Sq     RSS     AIC
+ cyl   1     87.15  191.17  63.198
+ hp    1     83.27  195.05  63.840
+ qsec  1     82.86  195.46  63.908
+ vs    1     54.23  224.09  68.283
+ carb  1     44.60  233.72  69.628
+ disp  1     31.64  246.68  71.356
<none>              278.32  73.217
+ drat  1      9.08  269.24  74.156
+ gear  1      1.14  277.19  75.086
+ am    1      0.00  278.32  75.217
- wt    1    847.73 1126.05 115.943

Step:  AIC=63.2
mpg ~ wt + cyl

       Df Sum of Sq    RSS    AIC
+ hp    1    14.551 176.62 62.665
+ carb  1    13.772 177.40 62.805
<none>              191.17 63.198
+ qsec  1    10.567 180.60 63.378
+ gear  1     3.028 188.14 64.687
+ disp  1     2.680 188.49 64.746
+ vs    1     0.706 190.47 65.080
+ am    1     0.125 191.05 65.177
```

```
+ drat   1     0.001 191.17 65.198
- cyl    1    87.150 278.32 73.217
- wt     1   117.162 308.33 76.494

Step:  AIC=62.66
mpg ~ wt + cyl + hp

       Df Sum of Sq    RSS    AIC
<none>              176.62 62.665
- hp     1    14.551 191.17 63.198
+ am     1     6.623 170.00 63.442
+ disp   1     6.176 170.44 63.526
- cyl    1    18.427 195.05 63.840
+ carb   1     2.519 174.10 64.205
+ drat   1     2.245 174.38 64.255
+ qsec   1     1.401 175.22 64.410
+ gear   1     0.856 175.76 64.509
+ vs     1     0.060 176.56 64.654
- wt     1   115.354 291.98 76.750
```

3 LOGISTIC REGRESSION.

3. (40 points) Logistic regression. Consider **am** to be the response variable, and all other variables as features.

(a) Describe the variable **am** in one sentence.

The variable **am** refers to the transmission(0 = automatic, 1 = manual) of each vehicle.

(b) Construct a plot of **hp** (x-axis) and **wt** (y-axis), with different colors for automatic and manual transmission. From the plot, do you think automatic and manual transmission can be distinguished by weight and horsepower?

The plot for the wt vs hp can easily be distinguishable by the type of transmissions. It's is quiet obvious to see that the automatic cars are generally more heavier than the manual cars.

(c) Fit a logistic regression model with wt as the only feature. Using this model, explain whether heavier cars are more likely or less likely to have manual transmission. If weight increases by 1000 lbs, what is the change in odds of a car having manual transmission?

Blue indicates automatic transmission and red indicates manual transmission cars. Heavier cars are LESS likely to have manual transmission according to the logis plots on Figure 3.Both automatic and manual transmission cars have couple outliers which they seem to create larger standard deviation for the Scale-Location plot. But overall, given data sets are fairly distinct in terms of the ratio between the wt and the transmission. When the weight is increased by 1000lbs, the odds that an increased 1000lbs is approximately 0.1788183 times as likely to have manual transmission as an automatic transmission.

```
> summary(logis)
```

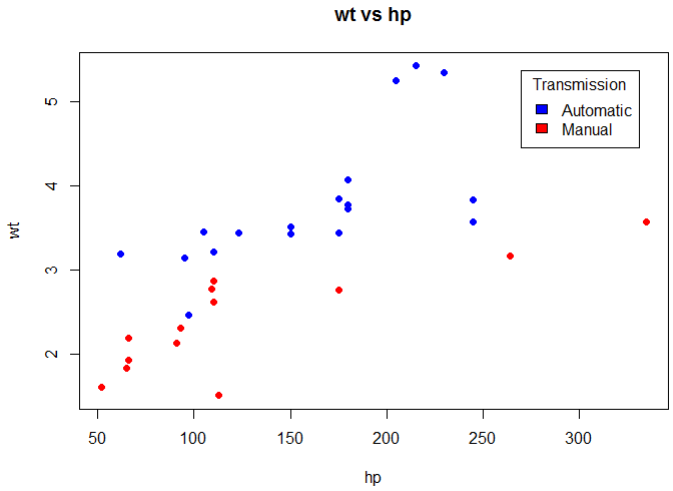

Figure 2: am Transmission (0 = automatic(blue), 1 = manual(red))

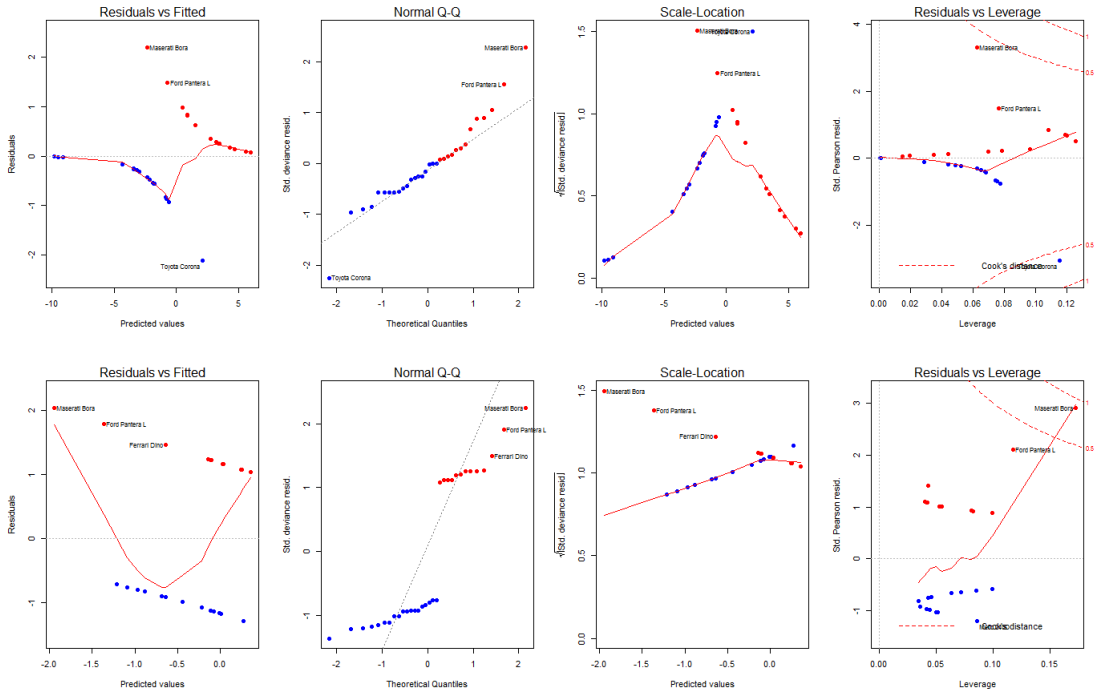

Figure 3: Top: am vs wt; Bottom: am vs hp

```
Call:
glm(formula = am ~ wt, family = "binomial")

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.11400  -0.53738  -0.08811   0.26055   2.19931

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   12.040      4.510   2.670  0.00759 **
wt            -4.024      1.436  -2.801  0.00509 **
---
Signif. codes:  0 ?**?0.001 ?*?0.01 ??0.05 ??0.1 ??1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 43.230  on 31  degrees of freedom
Residual deviance: 19.176  on 30  degrees of freedom
AIC: 23.176

Number of Fisher Scoring iterations: 6
> exp(predict(logis, data.frame(wt=2), type="link") - predict(logis, data.frame(wt
    =1), type="link") )
         1
0.01788183
```

(d) Fit a logistic regression model with hp as the only feature. Using this model, explain whether cars with higher horsepower are more likely or less likely to have manual transmission. If horsepower increases by 100, what is the change in odds of a car having manual transmission?

On average, automatic cars have higher horsepower. When the horsepower is increased by 100hp, the odds that an increased 100hp is approximately 0.4440971 times as likely to have a manual transmission as an automatic transmission

```
> summary(logis2)

Call:
glm(formula = am ~ hp, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2955  -0.9968  -0.7818   1.1630   2.0379

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.776614   0.915429   0.848    0.396
hp          -0.008117   0.006074  -1.336    0.181

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 43.230  on 31  degrees of freedom
Residual deviance: 41.228  on 30  degrees of freedom
AIC: 45.228

Number of Fisher Scoring iterations: 4
> exp(predict(logis2, data.frame(hp=200), type="link") - predict(logis2, data.
    frame(hp=100), type="link") )
```

```
        1
0.4440971
```

(e) If you had to choose between these two models, which one would you choose and why?

In Figure 2, both logistic models are represented using the different scales and methods. Normality for both models are significantly different. Top 4 graphs represent logistic model for am vs wt and the bottom 4 graphs represent logistic model for am vs hp. As you could tell, the residuals are normally distributed and easily distinguishable by the transmission types. Out best fit model from taking exponents of the odds when one is increased by 1000lbs and other is increased by 100hp, **model 1** from part **(C)** is not the best, but most optimal use among the two.