Regression Models Course Project

Han Yunxi

18 February 2015

Executive Summary

Motor Trend, a magazine about the automobile industry, are interested in exploring the relationship between a set of automobile variables and miles per gallon *(mpg)*. They are particularly interested in the following two questions:

Is an automatic or manual transmission better for MPG?

Quantify the MPG difference between automatic and manual transmissions.

The statistical inference shows that from the results, we see that manual cars average 24.4 *mpg*, automatic cars average 17.1 *mpg*. We also fit 4 multiple regression models, to find the model with the highest R^2 value. The final regression equation (model 4) is **mpg = 9.72 + (-2.93)wt + 1.01qsec + 14.08am + (-4.14)wt:am**. The conclusion is that cars that are lighter in weight with a manual transmission will have higher MPG values.

Exploratory Data Analysis (EDA)

From the `?mtcars` help file, we see that the dataset comprises fuel consumption *(outcome)* and 10 aspects *(predictors)* of automobile design and performance for 32 automobiles (1973–74 models).

```
library(data.table)
library(knitr)
data(mtcars)
```

EDA Findings A scatterplot matrix of *mpg* and other 10 predictors is generated. The top row shows the relationship between *mpg* and each of the predictors. From the scatterplot, there appears to be some relationship between *mpg* and *disp* , *hp*, *wt*. There are a great deal of random variables present in other predictors. In other words, the model is likely to consist of *disp* , *hp* and *wt*. This requires further investigation. *Refer to appendix for scatterplot*

Statistical Inference

Here, we calculate the 95% confidence intervals on *mpg* for automatic and manual cars. A t-test is used to evaluate the if there is no difference in average *mpg* between automatic(μ_0) and manual(μ_a) cars at 95% confidence interval. $H_0 : \mu_0 = \mu_a$ vs. $H_a : \mu_0 < \mu_a$

```
  t.test(mpg ~ am, data = mtcars, alternative = c("less"),paired=FALSE, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.0006868
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -3.913256
## sample estimates:
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

From the results, we see that manual cars average 24.4 *mpg*, automatic cars average 17.1 *mpg*. Since the p-value, 0.0006 is lesser than 0.05, we reject H_0 and conclude that manual cars have a better average *mpg* than automatic cars.

Multiple Regression

We will start with the predictors given in EDA (model 1), then with full model (model 2) that predicts *mpg* based on all predictors given in mtcars dataset. Thereafter we use **step** function to eliminate predictors with the highest p-value and refit the model to arrive at the final model (model 3). *All results can be found in appendix*

Model 1

```r
fit_EDA <- lm(mpg ~ disp + hp + wt, data = mtcars)  #model 1
```

Using the EDA results, we fit **mpg = disp + hp + wt** and test if it is a good fit. The adjusted R^2 is 0.8. So about 80% of the variation in the scores can be predicted using the model. Let's see if the model can be improved further.

Model 2

```r
fit_Full <- lm(mpg ~ ., data=mtcars)  # model 2
```

The results for model 2 shows the adjusted R^2 is also 0.8. Similar to Model 1, about 80% of the variation in the scores can be predicted using the model. There is no improvement to fitting all predictors.

Model 3

```r
fit_Stepwise <- step(lm(mpg ~ ., data=mtcars)) # model 3
```

The results shows that adjusted R^2 value 0.83 - an improvement of 3% over the full model. P-values are also significant at 95%. This model **mpg = wt + qsec + am** seems to be a good fit.

Interactions (model 4)

To overcome the limitation of **step** function to account for Interactions, we will introduce (wt*am) term to the regression model (model 3). It would be useful to add an interaction term to the model if we wanted to test the hypothesis that the relationship between the weight(wt) on the mpg was different for manual rather than automatic cars.

```
fit_Interaction <- lm(mpg ~ wt + qsec + am + wt*am , data = mtcars)
```

The adjusted R^2 for model 4 is 0.88, which is the best by far. Hence the final regression equation (model 4) will be **mpg = 9.72 + (-2.93)wt + 1.01qsec + 14.08am + (-4.14)wt:am**.

Anova

We can use the anova() function to compare the 4 models and obtain the significance in the change in R^2.

```
anova(fit_Full, fit_EDA, fit_Stepwise, fit_Interaction)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## Model 2: mpg ~ disp + hp + wt
## Model 3: mpg ~ wt + qsec + am
## Model 4: mpg ~ wt + qsec + am + wt * am
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     21 147.49
## 2     28 194.99 -7   -47.496 0.9661 0.48031
## 3     28 169.29  0    25.705
## 4     27 117.28  1    52.010 7.4050 0.01279 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The significant F-statistic informs us that the change in R^2 is significant, i.e. model 4 is a better fit to the data than model 1,2,3.

Residual Diagnostics

After selecting regression values and fitting the regression model 4, it is necessary to plot the residuals to investigate good/poor model fit. The results are 1. The Residuals vs. Fitted plot shows no systematic patterns and the variation in the residuals does not seem to change with the size of the fitted values. 2. The histogram of residuals is normally distrubuted.

Results

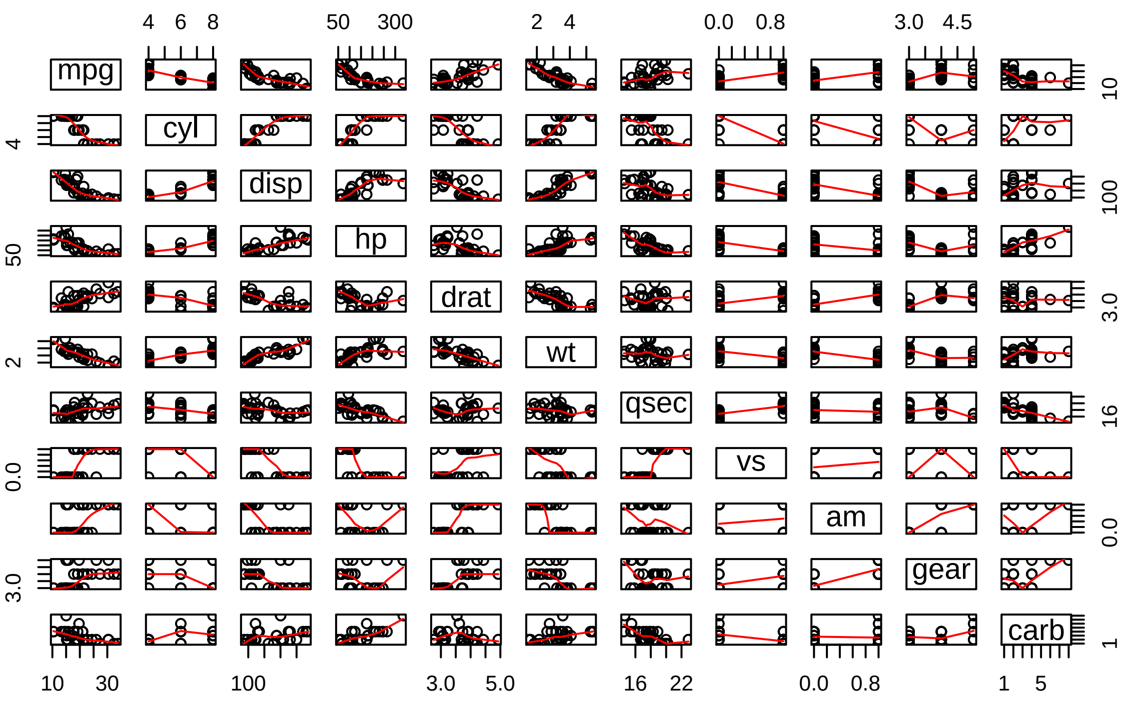
To interpret the results,

1. For every 1000lb increase in weight, the *mpg* for manual transmission shows a $14.08(1) - 4.14(wt)(1) = 9.94$ improvement over automatics.
2. Manual transmission is better for *mpg*.

– end –

Appendix

Exploratory Data Analysis


Pair Graph of Motor Trend Car Road Tests

Results of Multiple Regression

```
summary(fit_EDA)$coef
```

```
##                  Estimate Std. Error     t value     Pr(>|t|)
## (Intercept) 37.1055052690 2.11081525 17.57875558 1.161936e-16
## disp        -0.0009370091 0.01034974 -0.09053451 9.285070e-01
## hp          -0.0311565508 0.01143579 -2.72447633 1.097103e-02
## wt          -3.8008905826 1.06619064 -3.56492586 1.330991e-03
```

```
summary(fit_Full)$coef
```

```
##                Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept) 12.30337416 18.71788443  0.6573058 0.51812440
## cyl         -0.11144048  1.04502336 -0.1066392 0.91608738
## disp         0.01333524  0.01785750  0.7467585 0.46348865
## hp          -0.02148212  0.02176858 -0.9868407 0.33495531
## drat         0.78711097  1.63537307  0.4813036 0.63527790
## wt          -3.71530393  1.89441430 -1.9611887 0.06325215
## qsec         0.82104075  0.73084480  1.1234133 0.27394127
## vs           0.31776281  2.10450861  0.1509915 0.88142347
## am           2.52022689  2.05665055  1.2254035 0.23398971
## gear         0.65541302  1.49325996  0.4389142 0.66520643
## carb        -0.19941925  0.82875250 -0.2406258 0.81217871
```

```r
summary(fit_Stepwise)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## am           2.935837  1.4109045  2.080819 4.671551e-02
```

```r
summary(fit_Interaction)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.723053  5.8990407  1.648243 0.1108925394
## wt          -2.936531  0.6660253 -4.409038 0.0001488947
## qsec         1.016974  0.2520152  4.035366 0.0004030165
## am          14.079428  3.4352512  4.098515 0.0003408693
## wt:am       -4.141376  1.1968119 -3.460340 0.0018085763
```

```r
#calculating R-squared values
a<-summary(fit_EDA)$adj.r.squared
b<-summary(fit_Full)$adj.r.squared
c<-summary(fit_Stepwise)$adj.r.squared
d<-summary(fit_Interaction)$adj.r.squared
R <- matrix( c("Model 1", "Model 2", "Model 3", "Model 4", a,b,c,d), nrow =2, ncol = 4, byrow =TRUE)
R
```

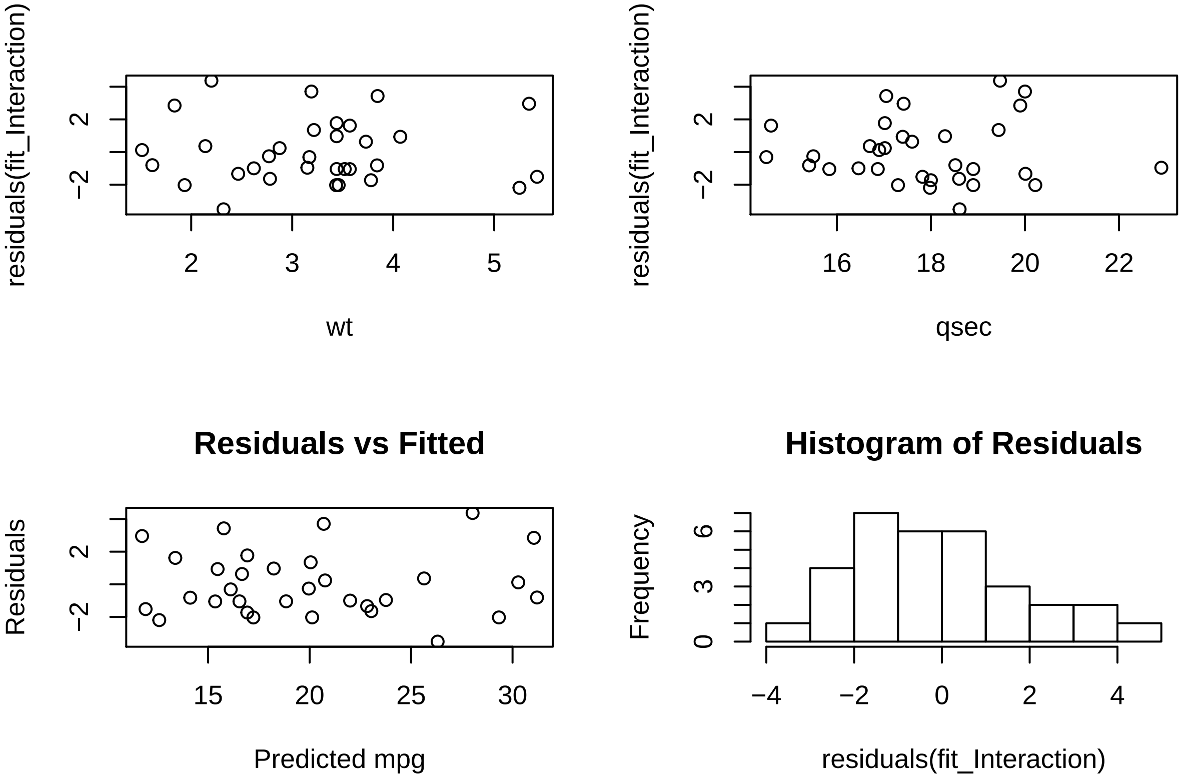
```
##      [,1]                [,2]                [,3]
## [1,] "Model 1"           "Model 2"           "Model 3"
## [2,] "0.808282872047642" "0.806642318990986" "0.833556080257604"
##      [,4]
## [1,] "Model 4"
## [2,] "0.880421944614729"
```

Residual Diagnostics

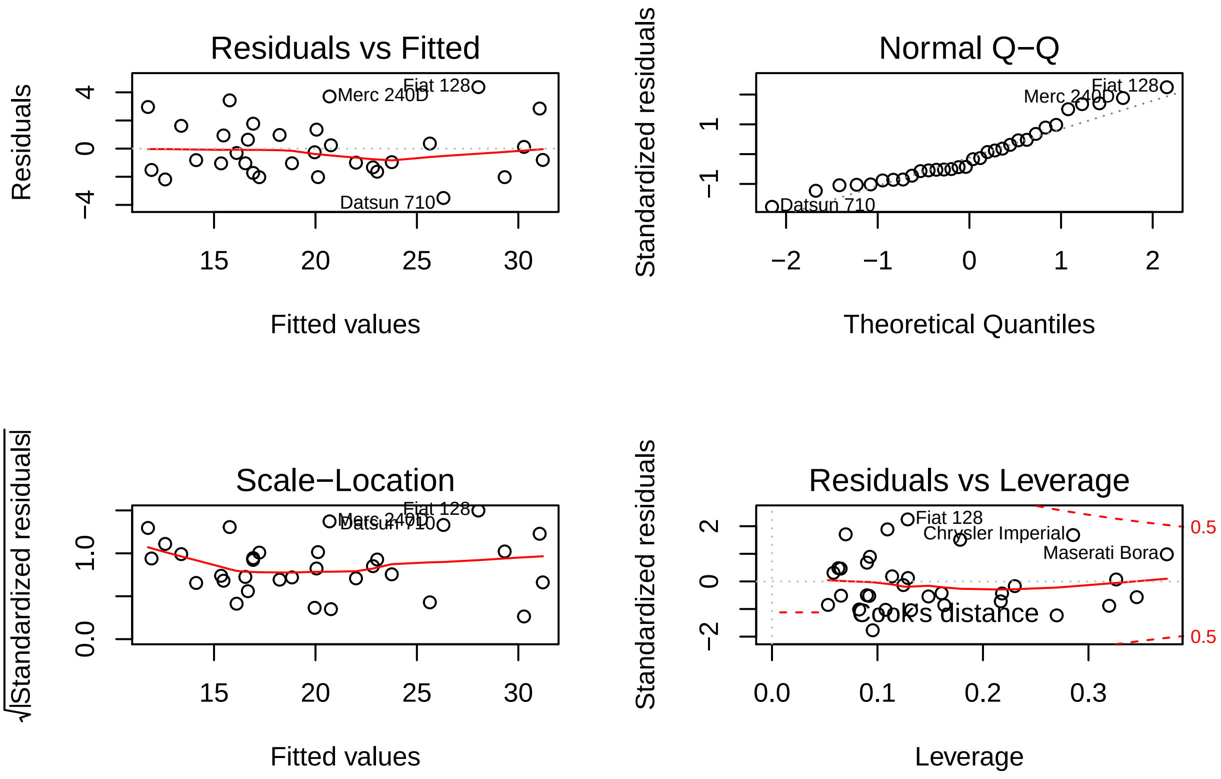
```r
par(mfrow=c(2,2))
plot(mtcars$wt, residuals(fit_Interaction), xlab="wt")
plot(mtcars$qsec, residuals(fit_Interaction), xlab="qsec")

plot(fitted(fit_Interaction), residuals(fit_Interaction),
     xlab = "Predicted mpg", ylab ="Residuals", main = "Residuals vs Fitted")

hist(residuals(fit_Interaction), main="Histogram of Residuals")
```



```r
plot(fit_Interaction)
```


Residuals vs Fitted
Residuals
−4
0
4
Merc 240D
Fiat 128
Datsun 710
15
20
25
30
Fitted values
Normal Q−Q
Standardized residuals
−1
1
Merc 240D
Fiat 128
Datsun 710
−2
−1
0
1
2
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
0.0
1.0
Merc 240D
Fiat 128
Datsun 710
15
20
25
30
Fitted values
Residuals vs Leverage
Standardized residuals
−2
0
2
Fiat 128
Chrysler Imperial
Maserati Bora
Cook's distance
0.5
0.5
0.0
0.1
0.2
0.3
Leverage