Regression Models Final Project

Fernando Gonzalez Prada

21 de agosto de 2015

Motor Trend is a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). 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 data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).

A data frame with 32 observations on 11 variables.

- **mpg:** Miles/(US) gallon
- **cyl:** Number of cylinders
- **disp:** Displacement (cu.in.)
- **hp:** Gross horsepower
- **drat:** Rear axle ratio
- **wt:** Weight (lb/1000)
- **qsec:** 1/4 mile time
- **vs:** V/S
- **am:** Transmission (0 = automatic, 1 = manual)
- **gear:** Number of forward gears
- **carb:** Number of carburetors

The first step is loading the data and performing an Exploratory Data Analysis

```r
suppressMessages(library(datasets))
suppressMessages(library(pastecs))
suppressMessages(library(ggplot2))
suppressMessages(library(car))

df <- mtcars
df$am <- factor(df$am, labels = c("Automatic", "Manual"))
```

The data is pretty clean. No missing values neither outliers are present. The data follows a nearly normal distribution.

But, for the numeric variables, “disp” and “hp” the range of values is much wider than for the other variables. So, before fitting the models, we must center and scale the variables.

See the Appendix for the details.

Scale and Center the numeric variables

```r
df2 = df
df2$mpg = scale(df2$mpg, center = TRUE, scale = TRUE)
df2$disp = scale(df2$disp, center = TRUE, scale = TRUE)
df2$hp = scale(df2$hp, center = TRUE, scale = TRUE)
```

```
df2$drat = scale(df2$drat, center = TRUE, scale = TRUE)
df2$wt = scale(df2$wt, center = TRUE, scale = TRUE)
df2$qsec = scale(df2$qsec, center = TRUE, scale = TRUE)
df2$vs = scale(df2$vs, center = TRUE, scale = TRUE)
df2$gear = scale(df2$gear, center = TRUE, scale = TRUE)
df2$carb = scale(df2$carb, center = TRUE, scale = TRUE)
```

Simple Linear Regression Model

```
simple <- lm(mpg ~ am, data = df2)
summary(simple)$coefficients
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -0.4883494  0.1865957 -2.617153 0.0137564004
## amManual     1.2020909  0.2927554  4.106127 0.0002850207
```

```
summary(simple)$r.squared
```

```
## [1] 0.3597989
```

If we only analize the relationship between “mpg” and “am”, the model says that **with manual transmition, we have 1.2 more milles per galllon**. However, we need to take into account the other variables. The R squared is only 0.3598, which explains only 36% of the variance, quite poor.

So, the next step is to perform Multiple Linear Regression, using Stepwise to obtain the best combination of vairables.

Multiple Linear Regression Model and Stepwise

```
multi <- step(
                lm(mpg ~ ., data = df2)
                    , trace = 0, steps = 10000, direction = "both");
summary(multi)$coefficients
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -0.1978918 0.11935630 -1.657993 1.084823e-01
## wt          -0.6358330 0.11546151 -5.506882 6.952711e-06
## qsec         0.3634657 0.08558828  4.246676 2.161737e-04
## amManual     0.4871184 0.23409933  2.080819 4.671551e-02
```

```
summary(multi)$r.squared
```

```
## [1] 0.8496636
```

With this second model,

R squared 0.8497 amManual 0.4871184 wt -0.6358330 qsec 0.3634657

Anova Analisys

```r
anova(simple, multi)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + qsec + am
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1     30 19.8462
## 2     28  4.6604  2    15.186 45.618 1.55e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix

Summary Statistics before centering and scaling the variables

```r
summary(df)
```

```
##       mpg             cyl             disp             hp
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0
##       drat             wt             qsec             vs
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000
##          am          gear            carb
##  Automatic:19   Min.   :3.000   Min.   :1.000
##  Manual   :13   1st Qu.:3.000   1st Qu.:2.000
##                 Median :4.000   Median :2.000
##                 Mean   :3.688   Mean   :2.812
##                 3rd Qu.:4.000   3rd Qu.:4.000
##                 Max.   :5.000   Max.   :8.000
```

Summary Statistics after centering and scaling the variables

```r
summary(df2)
```

```
##         mpg.V1             cyl              disp.V1
##  Min.   :-1.6078826   Min.   :4.000   Min.   :-1.2879099
##  1st Qu.:-0.7741273   1st Qu.:4.000   1st Qu.:-0.8867035
##  Median :-0.1477738   Median :6.000   Median :-0.2777331
##  Mean   : 0.0000000   Mean   :6.188   Mean   : 0.0000000
##  3rd Qu.: 0.4495434   3rd Qu.:8.000   3rd Qu.: 0.7687521
##  Max.   : 2.2912716   Max.   :8.000   Max.   : 1.9467538
##         hp.V1               drat.V1               wt.V1
##  Min.   :-1.3810318   Min.   :-1.5646078   Min.   :-1.7417722  
##  1st Qu.:-0.7319924   1st Qu.:-0.9661175   1st Qu.:-0.6500027  
##  Median :-0.3454858   Median : 0.1841059   Median : 0.1101223  
##  Mean   : 0.0000000   Mean   : 0.0000000   Mean   : 0.0000000  
##  3rd Qu.: 0.4858679   3rd Qu.: 0.6049193   3rd Qu.: 0.4013971  
##  Max.   : 2.7465668   Max.   : 2.4939041   Max.   : 2.2553357  
##        qsec.V1               vs.V1                 am    
##  Min.   :-1.8740103   Min.   :-0.8680278   Automatic:19  
##  1st Qu.:-0.5351317   1st Qu.:-0.8680278   Manual   :13  
##  Median :-0.0776466   Median :-0.8680278                 
##  Mean   : 0.0000000   Mean   : 0.0000000                 
##  3rd Qu.: 0.5882951   3rd Qu.: 1.1160357                 
##  Max.   : 2.8267546   Max.   : 1.1160357                 
##        gear.V1              carb.V1       
##  Min.   :-0.9318192   Min.   :-1.122152  
##  1st Qu.:-0.9318192   1st Qu.:-0.503034  
##  Median : 0.4235542   Median :-0.503034  
##  Mean   : 0.0000000   Mean   : 0.000000  
##  3rd Qu.: 0.4235542   3rd Qu.: 0.735203  
##  Max.   : 1.7789276   Max.   : 3.211677
```

Checking for normality:

Perform Shapiro Test

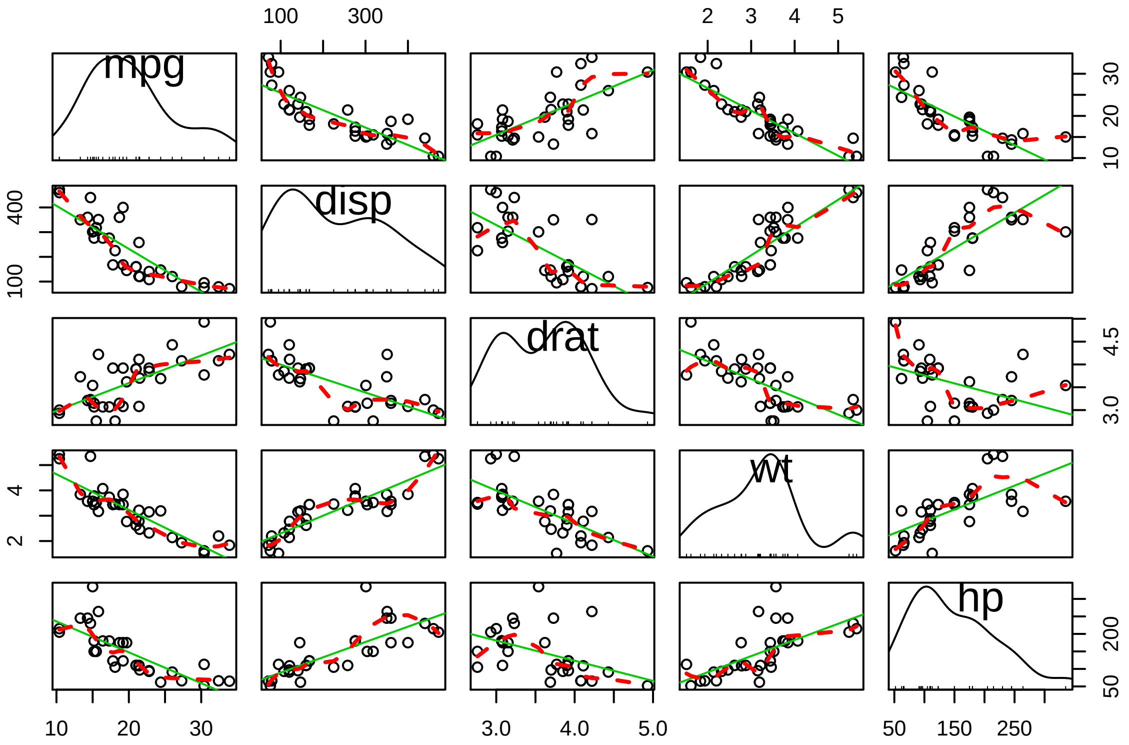
```
round(
    stat.desc(
        df[, c("mpg", "disp", "hp", "drat", "wt", "qsec", "vs", "gear", "carb")],
        basic = FALSE, norm = TRUE, desc=FALSE)
    ,digits = 3)
```

```
##                mpg   disp     hp   drat     wt  qsec     vs   gear  carb
## skewness    0.611  0.382  0.726  0.266  0.423 0.369  0.240  0.529 1.051
## skew.2SE    0.737  0.460  0.876  0.321  0.510 0.445  0.290  0.638 1.268
## kurtosis   -0.373 -1.207 -0.136 -0.715 -0.023 0.335 -2.002 -1.070 1.257
## kurt.2SE   -0.230 -0.746 -0.084 -0.442 -0.014 0.207 -1.237 -0.661 0.777
## normtest.W  0.948  0.920  0.933  0.946  0.943 0.973  0.632  0.773 0.851
## normtest.p  0.123  0.021  0.049  0.110  0.093 0.594  0.000  0.000 0.000
```

Observing the values of the “normtest.W” row, most of the variables have values close to 1, which is an indicator the the data is nearly normally distributed.

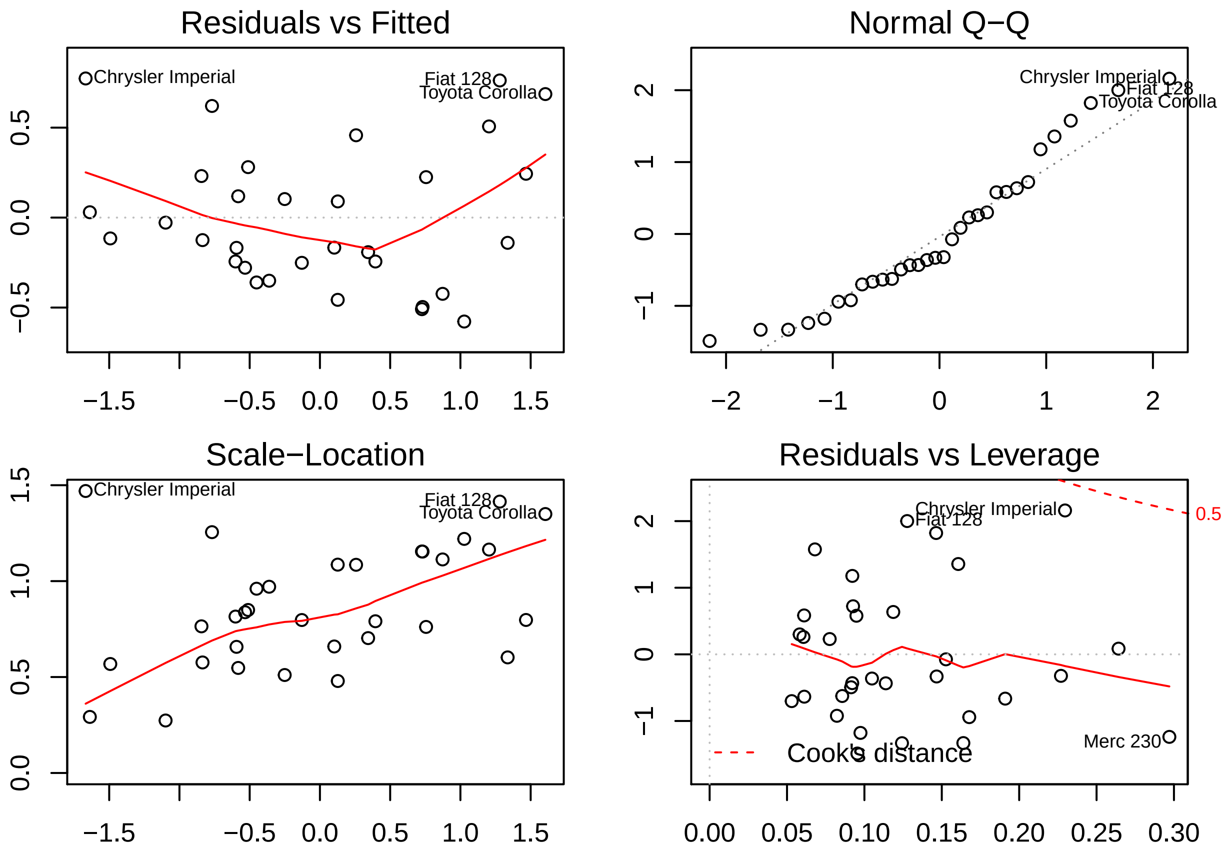
Bivariate relationship among the variables:

```
scatterplotMatrix(~mpg+disp+drat+wt+hp, data=df, spread=FALSE, smoother.args=list(lty=2))
```



Residuals

```
par(mfrow=c(2,2),mar=c(2,2,2,2))
plot(multi)
```


Residuals vs Fitted
Chrysler Imperial
Fiat 128
Toyota Corolla
−0.5
0.0
0.5
−1.5
−0.5
0.0
0.5
1.0
1.5
Normal Q−Q
Chrysler Imperial
Fiat 128
Toyota Corolla
−1
0
1
2
−2
−1
0
1
2
Scale−Location
Chrysler Imperial
Fiat 128
Toyota Corolla
0.0
0.5
1.0
1.5
−1.5
−0.5
0.0
0.5
1.0
1.5
Residuals vs Leverage
Chrysler Imperial
Fiat 128
Merc 230
0.5
Cook's distance
−1
0
1
2
0.00
0.05
0.10
0.15
0.20
0.25
0.30