Regression Models Project

Josep Anton Mir Tutusaus

Friday, November 06, 2015

1. Executive summary

Motor Trend, an automobile trend magazine is interested in exploring the relationship between a set of variables and miles per gallon (MPG) outcome. In this project, we will analyze the mtcars dataset from the 1974 Motor Trend US magazine to answer the following questions:

-Is an automatic or manual transmission better for miles per gallon (MPG)?

-How different is the MPG between automatic and manual transmissions?

We fitted a simple model that correlates automatic/manual transmission and weight with miles per gallon.

2. Exploratory analysis

The data set has 32 observations (models of cars) on 11 variables.

We are particularly interested in the realtionship between automatic/manual transmission [am] and Miles/(US) gallon [mpg].

```
summary(lm(mpg ~ factor(am) - 1, data = mtcars))$coef
```

```
##              Estimate Std. Error  t value     Pr(>|t|)
## factor(am)0 17.14737   1.124603 15.24749 1.133983e-15
## factor(am)1 24.39231   1.359578 17.94109 1.376283e-17
```

As we can see in this fit between mpg and am, with omitted intercept, the mean mpg for automatic and manual are, respectively, 17.15 and 24.39, both significally superior to 0.

```
summary(lm(mpg ~ factor(am), data = mtcars))$coef
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## factor(am)1  7.244939   1.764422  4.106127 2.850207e-04
```

If we compare the two groups (automatic and manual transmission), we observe that manual transmission has significally higher mpg mean than automatic transmission.

We suspect, nonetheless, that other variables could affect mpg.

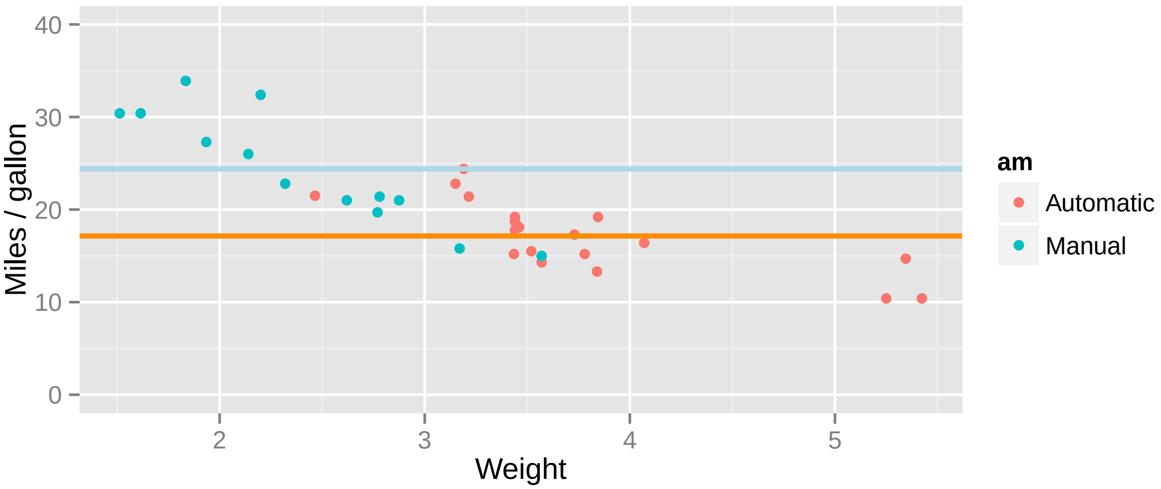
```
summary(lm(mpg ~ . , data = mtcars))$coef
```

```
##                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
```

Weight, in particular, seems to be negatively correlated with mpg, and that makes sense: every 1000 lbs we move up, the mpg diminishes by 3.71.

Let's visualise it:

```
mtcars$am <- factor(mtcars$am)
levels(mtcars$am) <- c("Automatic", "Manual")
g = ggplot(data = mtcars, aes(y = mpg, x = wt))
g = g + geom_point(aes(colour=am)) + ylim(0,40)
g = g + xlab("Weight") + ylab("Miles / gallon")
g = g + geom_hline(yintercept = as.numeric(summarise(group_by(mtcars, am), mean(mpg))[1,2]), color = "da
g = g + geom_hline(yintercept = as.numeric(summarise(group_by(mtcars, am), mean(mpg))[2,2]), color = "li
g
```



We can see that the cars with automatic transmission, which have inferior mean (horizontal line at 17.15 mpg) also weight more, as can be visualized by the size of the dots. The cars with manual transmission have higher mpg mean (24.39) and weight less. We suspect that weight plays an important role in mpg.

3. Modelling

Let's propose 2 models: one that takes into account both weight and type of transmission and a third with an interaction term. We saw before that a model with only am variable could not explain much of the variability, so we do not present that model in this section:

```
fit1 <- lm(mpg ~ wt + am, data = mtcars)
fit2 <- lm(mpg ~ wt * am, data = mtcars)

summary(fit1)$coef
```

```
##                Estimate Std. Error     t value     Pr(>|t|)
## (Intercept) 37.32155131  3.0546385 12.21799285 5.843477e-13
## wt          -5.35281145  0.7882438 -6.79080719 1.867415e-07
## amManual    -0.02361522  1.5456453 -0.01527855 9.879146e-01
```

```
summary(fit2)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 31.416055  3.0201093 10.402291 4.001043e-11
## wt          -3.785908  0.7856478 -4.818836 4.551182e-05
## amManual    14.878423  4.2640422  3.489277 1.621034e-03
## wt:amManual -5.298360  1.4446993 -3.667449 1.017148e-03
```

On the first model mpg = a + b*wt* + *c*am + E, b equals the slope, c is the difference in means between automatic and manual. E is the error term (everything we didn't measure).

On the second model, mpg = a + b*wt* + *c*am + d*wt*am + E, with interaction term, c is the change in the intercept for manual transmission, d is the change in slope for manual transmission. E is the error term (everything we didn't measure).

```
anova(fit1, fit2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt + am
## Model 2: mpg ~ wt * am
##   Res.Df    RSS Df Sum of Sq     F   Pr(>F)
## 1     29 278.32
## 2     28 188.01  1    90.312 13.45 0.001017 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We compare the models and choose the second one because has better RSS and explains 83.3% of the variability.

Refer to Apendix I.1 for the plots for the two models.

4. Residuals Analysis and diagnostics

According to the residual plots (Apendix I.2), we can verify the following underlying assumptions:

1. The Residuals vs. Fitted plot shows no pattern, so that verifies the independence condition.
2. The Normal Q-Q plot shows that the residuals are normally distributed (plotted around the line).
3. The Scale-Location plot confirms the constant variance assumption, as the points are randomly distributed.
4. The Residuals vs. Leverage show that there are some outliers.

```
m <- hatvalues(fit2)
tail(sort(m),3)
```

```
##   Chrysler Imperial Lincoln Continental       Maserati Bora
##           0.2809856           0.3044512           0.3709866
```

```
n <- dfbetas(fit2)
tail(sort(n[,4]),3)
```

```
##     Merc 240D Maserati Bora  Lotus Europa
##     0.2164611     0.3132730     0.3913157
```

Points with most leverage can be calculated via hatvalues and the ones that influence the model the most via dfbetas. Those are in accord with the points outliers in the residuals plots.

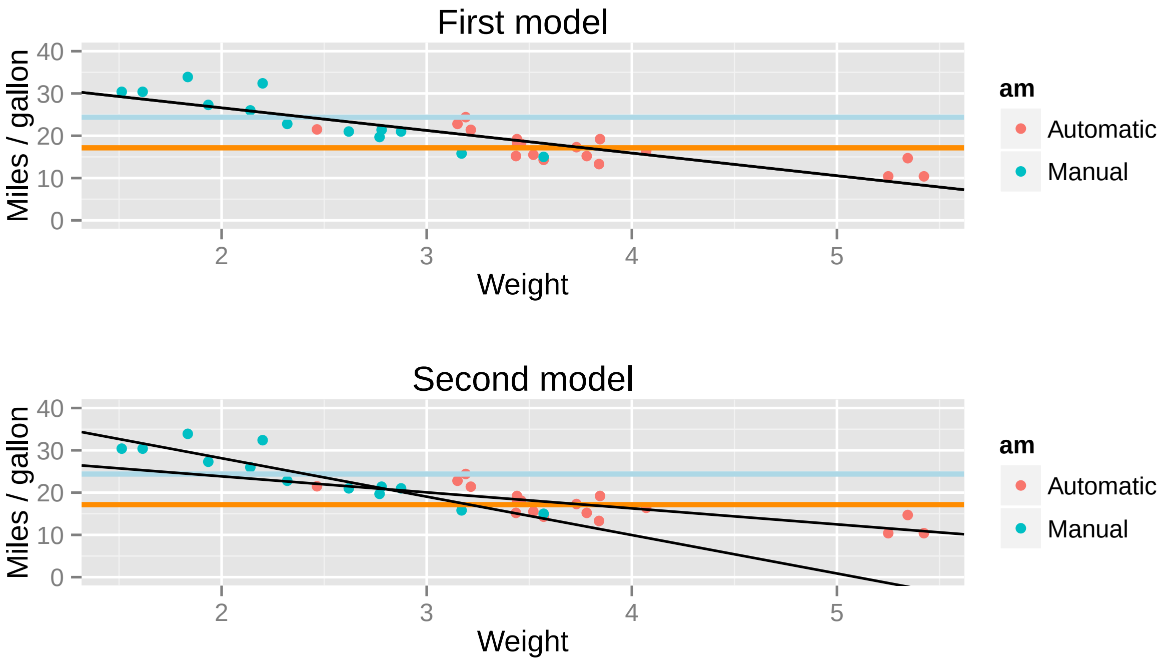
```
g1 <- g + geom_abline(intercept = fit1$coef[1], slope = fit1$coef[2]) + geom_abline(intercept = fit1$co
```

```
g2 <- g + geom_abline(intercept = fit2$coef[1], slope = fit2$coef[2]) + geom_abline(intercept = fit2$co
```

Apendix I

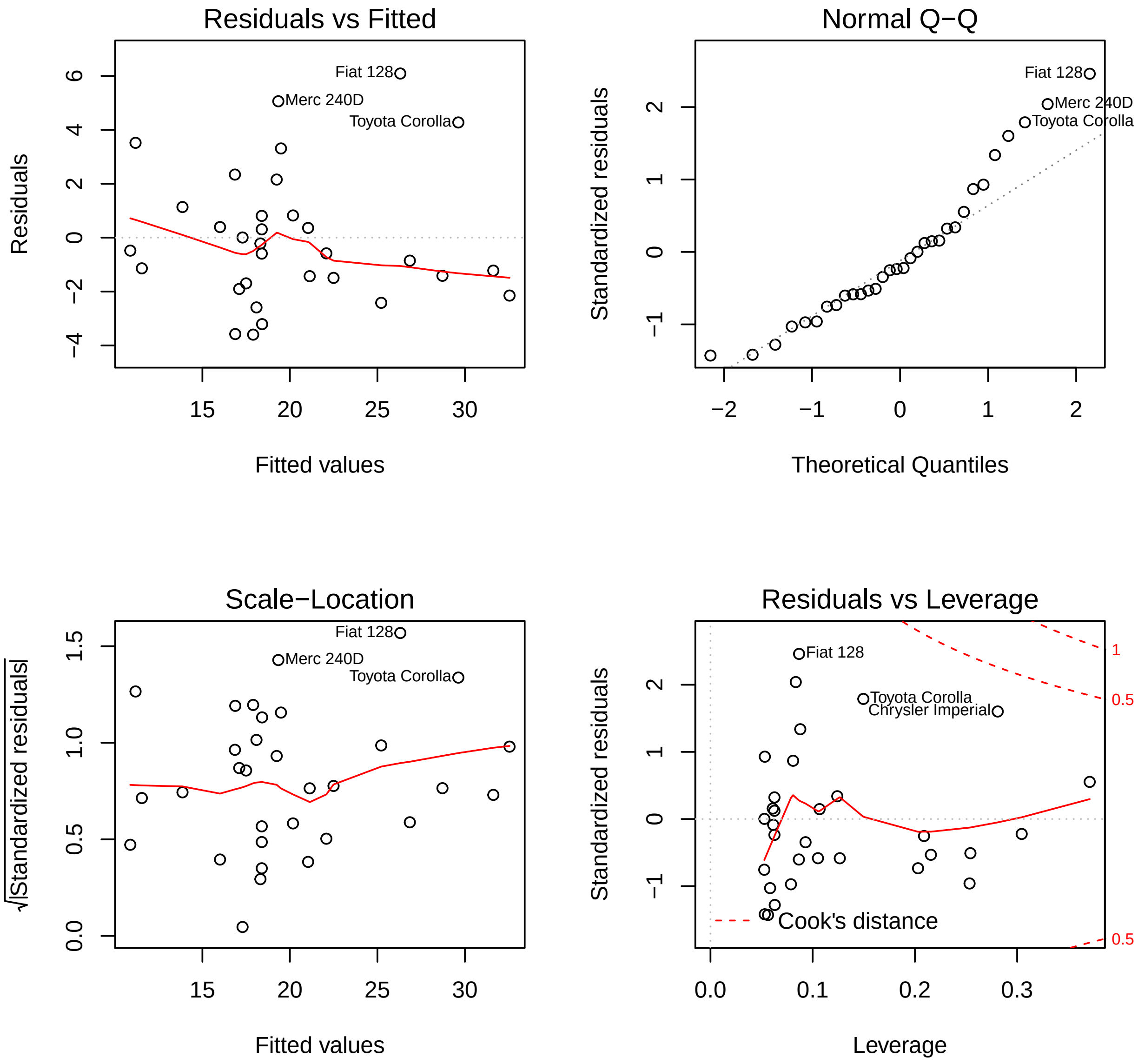
I.1 Models

```
vplayout <- function(x, y) viewport(layout.pos.row = x, layout.pos.col = y)

grid.newpage()
pushViewport(viewport(layout = grid.layout(2, 1)))

print(g1, vp = vplayout(1,1))
print(g2, vp = vplayout(2,1))
```



I.2 Residual plots

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


Residuals vs Fitted
Residuals
Fitted values
Fiat 128
Merc 240D
Toyota Corolla
−4
−2
0
2
4
6
15
20
25
30
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Fiat 128
Merc 240D
Toyota Corolla
−1
0
1
2
−2
−1
0
1
2
Scale−Location
√|Standardized residuals|
Fitted values
Fiat 128
Merc 240D
Toyota Corolla
0.0
0.5
1.0
1.5
15
20
25
30
Residuals vs Leverage
Standardized residuals
Leverage
Fiat 128
Toyota Corolla
Chrysler Imperial
Cook's distance
1
0.5
0.5
−1
0
1
2
0.0
0.1
0.2
0.3