Regression Models Project

E.T. Moseley

January 28, 2017

Executive Summary

In this analysis, we will generate two models to explain differences in mileage per gallon using the mtcars data set. Our first linear model includes only the transmission to predict mileage per gallon. Our second linear fit includes weight in addition to the transmission type.

Furthermore, we have determined that a manual transmission is associated with an increase in gas mileage on average of 7.25 mpg (disregarding all other variables).

Exploratory Data Analysis

In this analysis, we will be exploring the "mtcars" dataset to determine what variables are associated with increased or decreased mileage per gallon. Precisely, we will be investigating what relationship (if any) the transmission, automatic or manual, has on the miles per gallon.

Per ?mtcars documentation:

A data frame with 32 observations on 11 variables.

[, 1] mpg Miles/(US) gallon

[, 2] cyl Number of cylinders

[, 3] disp Displacement (cu.in.)

[, 4] hp Gross horsepower

[, 5] drat Rear axle ratio

[, 6] wt Weight (lb/1000)

[, 7] qsec 1/4 mile time

[, 8] vs V/S

[, 9] am Transmission (0 = automatic, 1 = manual)

[,10] gear Number of forward gears

[,11] carb Number of carburetors

```r
attach(mtcars)#Load data

str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
```

```
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

Let's see if there may be a relationship between miles per gallon and transmission:

```
auto <- mtcars[(mtcars$am == 0),]#0 = automatic
manu <- mtcars[(mtcars$am == 1),]#1 = manual
summary(auto$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   14.95   17.30   17.15   19.20   24.40
```

```
summary(manu$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   21.00   22.80   24.39   30.40   33.90
```

Manual transmissions have higher Median and Mean mileage per gallon, with Medians of 22.80 vs 17.30 (manual vs automatic) and Means of 24.39 vs 17.15.

Fitting Multiple Models

We can run a linear regression with mpg predicted by transmission type to determine if the relationship has statistical significance.

```
fit <- lm(mpg ~ am, data = mtcars)
summary(fit)$coefficients
```

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

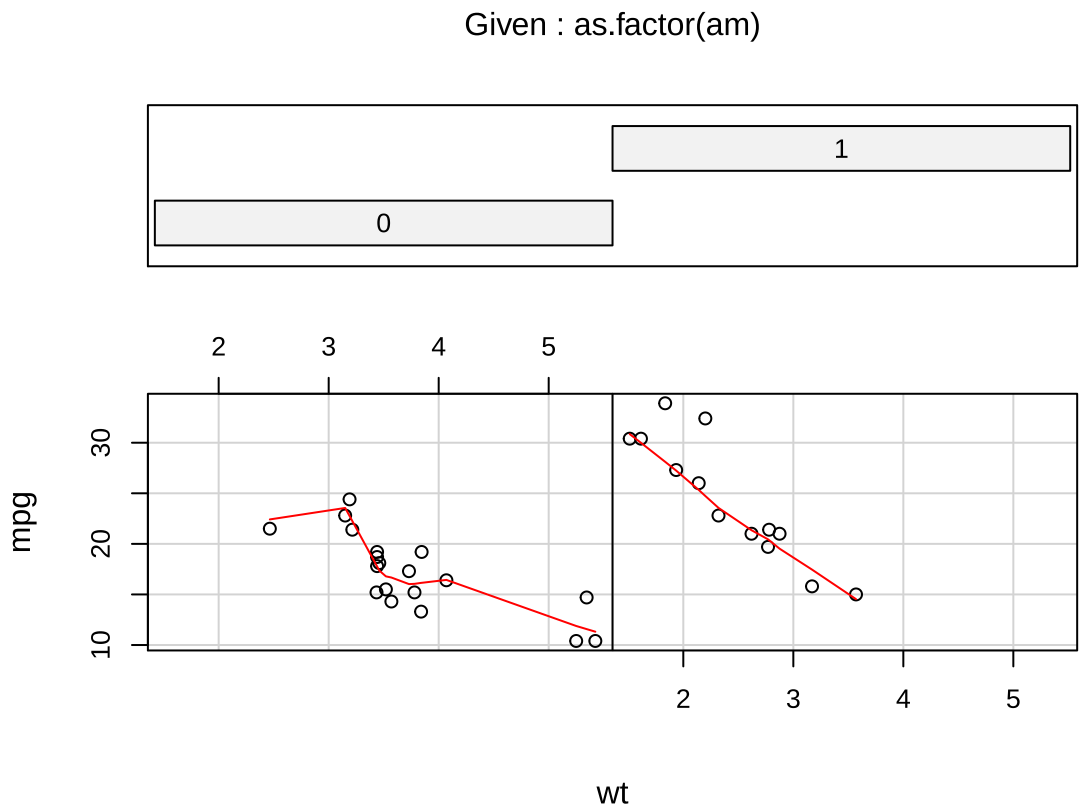
The probability that the transmission is not significant to mileage per gallon is 2.85e-04, which is much less than zero. Furthermore, we see that the slope of this model, 7.245 (the difference in mileage associated with a manual vs an automatic transmission) is precisely identical to the difference in means of both respective data:

```
mean(manu$mpg) - mean(auto$mpg)
```

```
## [1] 7.244939
```

It is safe to reject the null hypothesis and investigate further. We will also begin to explore if weight affects gas mileage:

```r
coplot(mpg ~ wt | as.factor(am), data = mtcars, panel = panel.smooth, rows = 1)
```



Visually, it would appear that, for both manual and automatic transmissions, weight affects fuel mileage.

We will investigate briefly if weight is a factor by adding it to our regression model and seeing if it improves the fit.

```r
fit2 <- lm(mpg ~ am + wt, data = mtcars)
summary(fit2)$coefficients
```

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

Here, weight seems to be more predictive of mileage per gallon than does the transmittion type, with P-values of 1.87e-07 and 9.88e-01, respectively.

We can now test to determine if these two models are distinct.

```r
anova(fit, fit2, test="Chisq")
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
##   Res.Df    RSS Df Sum of Sq  Pr(>Chi)
```

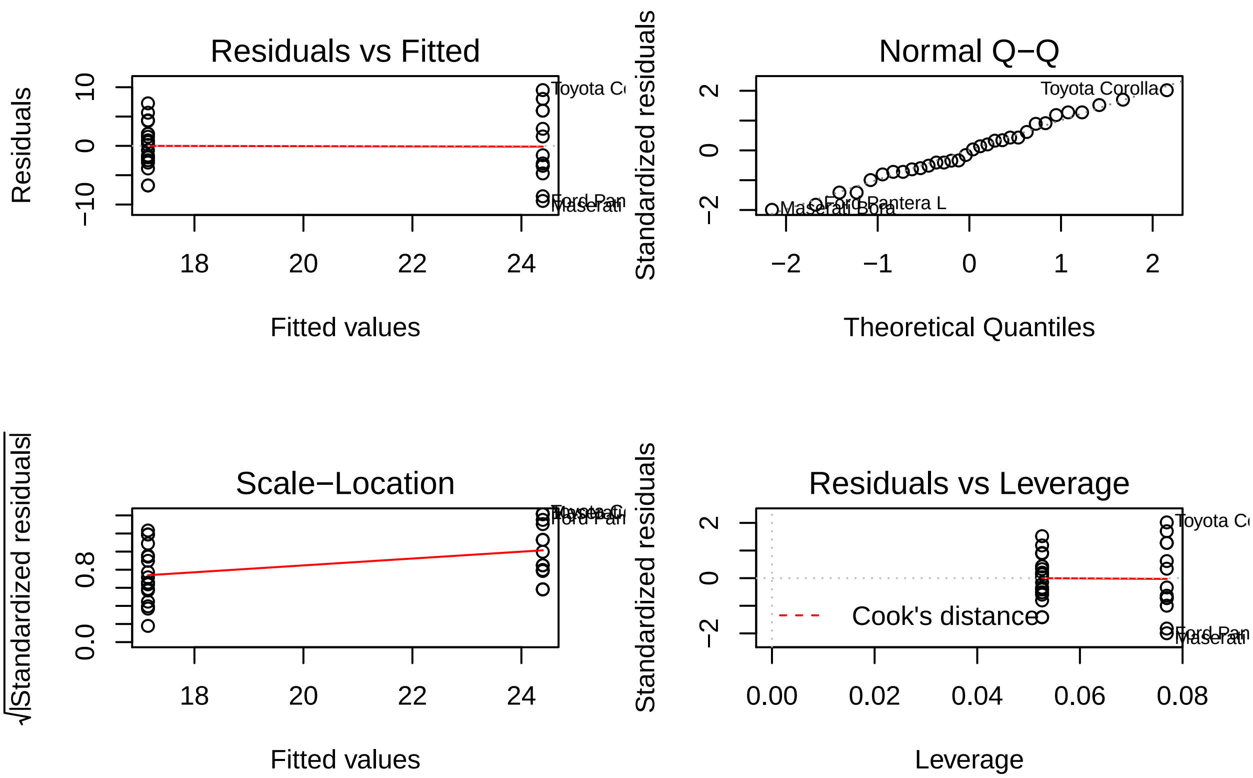
```
## 1     30 720.90
## 2     29 278.32  1    442.58 1.115e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this case, with the addition of wt to the model, the probability that the variation in predictions in either model is due only to chance is 1.12e-11, which is insignificant.

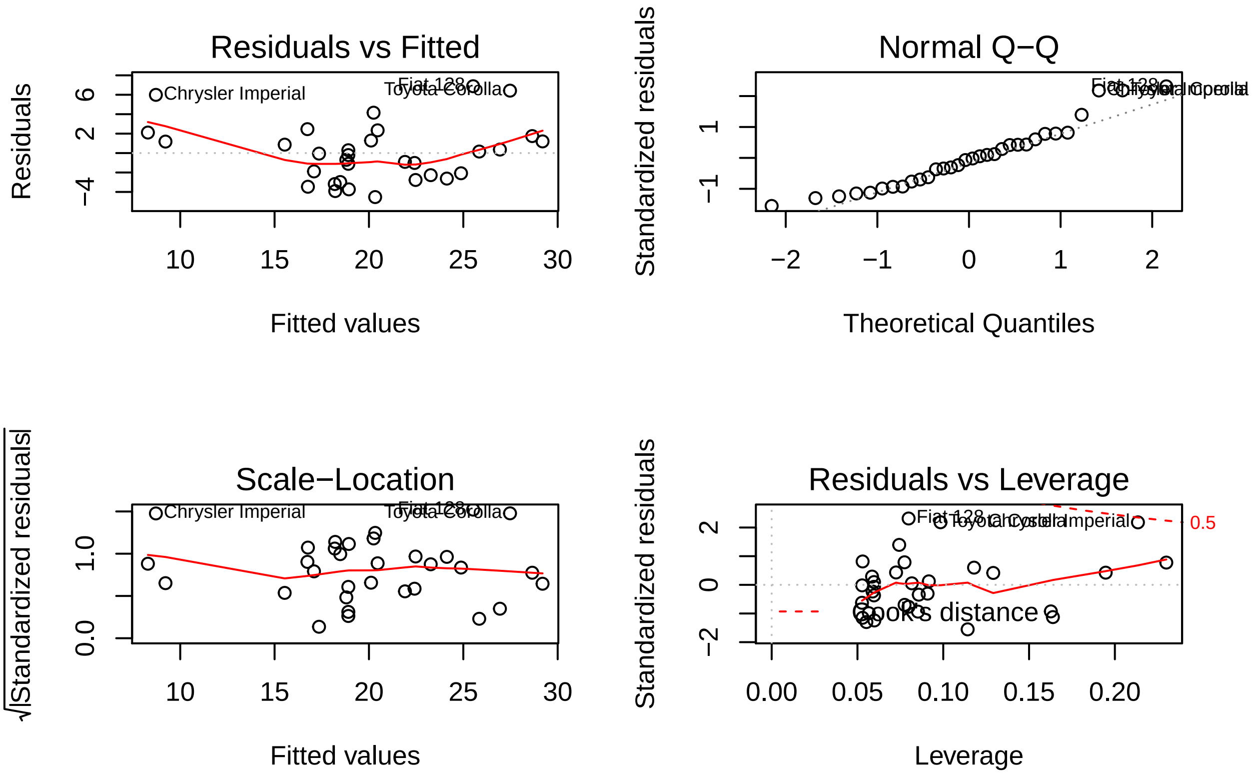
Apprendix

Residual plots for fit (mpg ~ am) and fit2 (mpg ~ am + wt)

```r
par(mfrow=c(2,2))
plot(fit)
```




```r
plot(fit2)
```


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Leverage
Cook's distance
0.5