Peer Graded Assignment Regression Models Course

Rogelio Caballero

June 23, 2017

Abstract

We conduct a brief exploratory analysis of the mtcars dataset and analyze some models to explain the relationship between the mileage per gallon and the type of transmission of the cars.

Exploratory Analysis

First let's load the dataset and take a look at the summary of the miles per galon variable `mpg`.

```
data(mtcars)
summary(mtcars$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   15.42   19.20   20.09   22.80   33.90
```

In the Appendix, there is a boxplot that roughly shows how `mpg` is related to `am`, a variable that indicates the type of transmission (0 represents automatic as we found here). We've defined a new categorical variable `auto` that indicates the type of transmission.

The boxplot suggests that a manual transmission gives more miles per gallon. Let's check some intervals:

```
df <- aggregate(mpg~auto, data = mtcars,
          FUN = function(x)c(mn = mean(x), sd = sd(x)))

library(xtable)
```

```
## Warning: package 'xtable' was built under R version 3.3.2
```

```
xt <- xtable(do.call(data.frame, c(df, check.names = FALSE)))
print(xt, type = "latex")
```

% latex table generated in R 3.3.1 by xtable 1.8-2 package % Sat Jun 24 12:18:07 2017

	auto	mpg.mn	mpg.sd
1	automatic	17.15	3.83
2	manual	24.39	6.17

The intervals that are less than one standard deviation of the mean for each transmission type overlap, so the difference in mpg is worth exploring while accounting for variations in other parameters.

Regression

Let's fit a linear model with `mpg` as outcome and only `auto` as predictor.

```
a0 <- lm(mpg~auto, data = mtcars)
summary(a0)$coef
```

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

As we can see the Intercept coincides with the mean for automatic transmission shown in the table (up to rounding error), and the coefficient `automanual` is the difference between the means of each transmission type. The very low p-values suggest that the coefficients are significantly different from zero.

In the Appendix you'll be able to find one residuals plot of this model. It basically shows no outliers.

Let's add the variable `hp` (horsepower) to the model (going blind here, I'm profoundly car-ignorant).

```
a1 <- lm(mpg ~ auto + hp, data = mtcars)
summary(a1)$coef
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 26.5849137 1.425094292 18.654845 1.073954e-17
## automanual   5.2770853 1.079540576  4.888270 3.460318e-05
## hp          -0.0588878 0.007856745 -7.495191 2.920375e-08
```

Even though the coefficient of `hp` is small, it's statistically significant, which means that there is a small but definitely non-zero effect of horsepower on miles per gallon. This coefficient is the variation of `mpg` due to the increase of one unit of `hp` keeping auto fixed. Plus, the coefficient of `automanual` is positive, which reinforces the conclusion that manual transmissions give more `mpg`. Now, we are going to add weight `wt`:

```
a2 <- lm(mpg ~ auto + hp + wt, data = mtcars)
summary(a2)$coef
```

```
##                 Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 34.00287512 2.642659337 12.866916 2.824030e-13
## automanual   2.08371013 1.376420152  1.513862 1.412682e-01
## hp          -0.03747873 0.009605422 -3.901830 5.464023e-04
## wt          -2.87857541 0.904970538 -3.180850 3.574031e-03
```

Again, the positive coefficient of `automanual` says that a manual transmission gives more miles per gallon. The coefficient of `wt` is the change in `mpg` per unit of weight keeping the other variables constant.

Let's see whether the addition of variables to our model gives us more detail:

```
anova(a0, a1, a2, test = "Chisq")
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ auto
## Model 2: mpg ~ auto + hp
## Model 3: mpg ~ auto + hp + wt
##   Res.Df    RSS Df Sum of Sq  Pr(>Chi)
## 1     30 720.90
## 2     29 245.44  1    475.46 < 2.2e-16 ***
## 3     28 180.29  1     65.15  0.001468 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The low p-values show that the two models built on top of `a0` are significantly different from `a0` and, hence, they make a good selection.

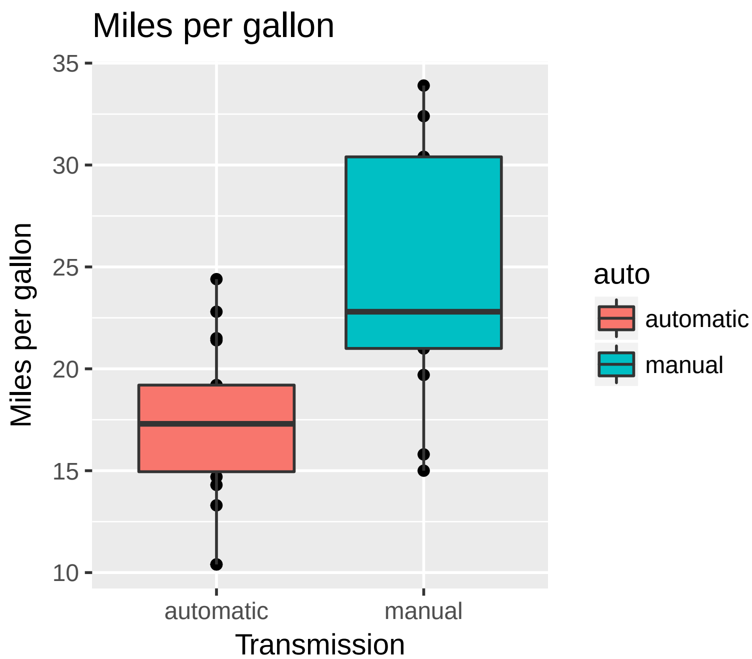
Conclusions

Each one of the three nested models that we used reduced the residual sum of squares with respect to its predecessor, which means that each variable that we chose added more detail to our overall understanding of the variations of `mpg`. In each one of the models, we saw that a manual transmission gives more miles per gallon, a conclusion drawn from the sign of the coefficient of `auto`.

Appendix

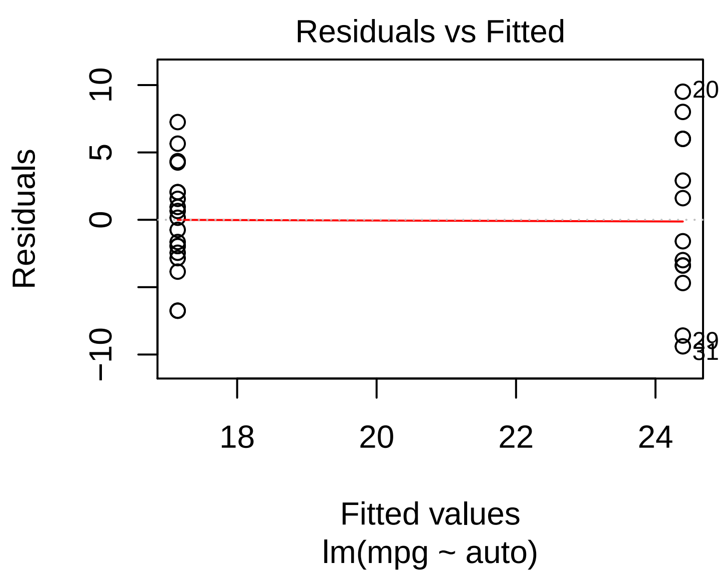
Full code can be found in my Github repository (here).

Boxplot of `mpg` vs. transmission type:



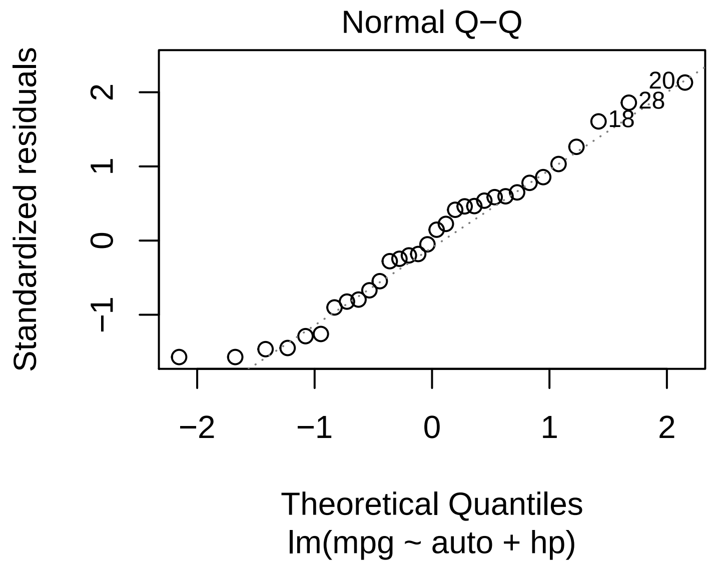
Residuals of the `mpg ~ auto` model:

```r
plot(a0, which = 1)
```



Residuals of the `mpg ~ auto + hp` model:

```r
plot(a1, which = 2)
```



Residuals of the `mpg ~ auto + hp + wt` model:

```r
plot(a2, which = 3)
```