Regression Models - Course Project

Philip Mateescu

December 26, 2015

Impact of Transmission Type of Automobile Mileage

Synopsis/Executive Summary

Fuel efficiency, expressed in miles-per-gallon, or *mpg* in short, is a common selection criteris when comparing cars.

Using a dataset extracted from the 1974 *Motor Trend US* magazine, and comprising of fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models), we will explore the relationship between automatic or manual transmission types and the fuel consumption to answer the following questions:

- "Is an automatic or manual transmission better for MPG"
- "Quantify the MPG difference between automatic and manual transmissions"

Using simple linear regression analyses, we have determined that there is a significant difference between automatic and manual transmissions in cars. The cars equiped with manual transmission obtain an average of 7.97 more miles per gallon when compared to cars equiped with automatic transmission, when accounting for all other factors impacting fuel consumption.

Exploratory Data Analysis

In order to make better use of the `mtcars` dataset I have changed the type of some of the columns to better reflect their underlying types, for example changing the `cyl` column, the number of cylinders, from a `numeric` to a `factor`, to indicate that this is *not* a continuous measurement and changing the factors of the `am` column from `0, 1` to `A, M`.
See **Appendix - Table 1** for all changes made and **Appendix - Table 2** for a sample dataset.

Appendix - Table 3 shows there is a difference between the average mpg of automatic cars – `17.1473684`mpg, with a standard deviation of `3.8339664` – and the **higher** average mpg of cars equipped with manual transmission, at `24.3923077`mpg, with a standard deviation of `6.1665038`.
Appendix - Figure 1 shows a box-plot of the transmission/mpg relationship.

Fitting a linear model with `am` as the regressor and `mpg` as the outcome, shows there is a significant relationship between the two (*p-value* < 0.001).
See **Appendix - Table 4 - Summary of base linear model**.

```
model.base <- lm(mpg ~ am, mtcars)
```

However, an *Adjusted R-squared* of 0.34 indicates only about a third of the variation in mileage can be explained by the choice of transmission.

Statistical Inference

Under the assumption that the `mpg` has a normal distribution, w we the null hypothesis, *H0*, as: manual and automatic transmission come from the same population.

```
h0 <- t.test(mpg ~ am, data = mtcars)
```

Using an unpaired, two-sided T-test at 95% confidence level, yields a p-value of `0.0013736`. Since the *p-value* is less than 0.05, we reject our null hypothesis, thus the automatic and manual transmissions come from different populations.

Model Selection

We will use the Akaike information criterion and the step function to choose the model best fitted to explain the variation in mpg.

```
model.best <- step(lm(mpg ~ ., mtcars), k=log(nrow(mtcars)), trace = 0)
```

According to **Appendix - Table 5 - Summary of best linear model**, the model using `wt` - weight, `qsec` - quarter-mile time, and `am` - transmission type, explains 83% percent of the variance of the `mpg` variable., and all variables are significant at the 0.05 level or smaller.

However, in consulting with our imaginary engine experts, we have learned that the quarter-mile time is likely inversely proportional to the horsepower of the car (when weight is constant).

Horsepower itself is a complex relationship of engine configuration (size, number of cylinders, engine geometry, number of carburetors) and gearing - `gear`. All there variables are likely confounding for the quarter-mile time. Let's examine a model that regresses on horsepower and compare it to a model that uses the individual variables.

```
model.amwt <- lm(mpg ~ am + wt, mtcars)
model.hp <- update(model.amwt, mpg ~ am + wt + hp)
model.engine <- update(model.amwt, mpg ~ am + wt + disp + gear + cyl + carb + vs)
df.comp1 <- data.frame(
    arsq(model.amwt),
    arsq(model.hp),
    arsq(model.engine))
kable(df.comp1, col.names = c('am+wt', 'am+wt+hp', 'am+wt+disp+gear+cyl+carb+vs'), caption = 'Adjusted
```

Table 1: Adjusted R-squared values for 3 models

am+wt	am+wt+hp	am+wt+disp+gear+cyl+carb+vs
0.7357889	0.8227357	0.7701516

We notice that while adding the engine configuration explains the variation better than weight + transmission type alone, adding the horsepower to the `am + wt` model explains 82% of the variance, close to the best model chosen by the step function, but considerably easier to explain even to novice readers.

Going back to our cars experts with our findings, we found them keen to point out that there is an interaction between the transmission type and the weight of the car, given that automatic transmissions tend to simply weigh more.

Armed with this new knowledge, let's consider a model where the `wt` and `am` interact.

```
model.final <- lm(mpg ~ am + wt + am:wt + hp, mtcars)
```

Surprisingly (or perhaps not), a collaboration between the statistics student and the car experts produce a model that explains the variance even better than the model selected by the step function: a whooping **85%** and all variables statistically significant at level 0.05 or lower! See **Appendix - Table 6 - Summary of final linear model**.

(Note: we'd be disingenuous if we didn't point out that if we take `step`'s best model and include the `am:wt` interaction we get an even better `0.8804219` adjusted R-squared).

This model shows that when weight and horsepower remain constant, cars with manual transmission get `11.55 + (-3.58) = 7.97` more miles per gallong than cars equiped with automatic transmission.

Residuals and Diagnostics

In this section, we'll perform a few diagnostics of our model, examine the residuals and leverage variables in order to detect any potential problems with our model.

We start with a diagnostic plot of our model, shows in the **Appendix - Figure 2**. From these four chart can observe that:

- There is no consistent pattern when examining *Fitted Values vs Residuals*, thus supporting the independence of our chosed predictors;
- The *Normal Q-Q* plot indicates the standardized residuals are normally distributed and close to out fitted line;
- The *Scale-Location* chart shows scatter, thus confirming constant variance,
- Finally, the *Residuals vs Leverage* chart, shows that while we have outliest, none have considerable leverage (all within the 0-0.5 band).

The `hatvalues` function gives us the leverage points in the model. The top 3 points are:

```
leverage <- hatvalues(model.final)
tail(sort(leverage), n = 3)
```

```
## Lincoln Continental        Lotus Europa       Maserati Bora
##           0.3140138           0.3833796           0.4825868
```

The top 3 most influential cars can be found using the `dfbetas` function:

```
influence <- dfbetas(model.final)
tail(sort(influence[,2]), n = 3)
```

```
##          Fiat 128    Toyota Corolla Chrysler Imperial
##         0.1664201         0.5194461         0.5906497
```

Appendix

Tables

```
kable(old.to.new.columns, caption = '')
```

	Old.Types	New.Types
mpg	numeric	numeric
cyl	numeric	factor
disp	numeric	numeric
hp	numeric	numeric
drat	numeric	numeric
wt	numeric	numeric
qsec	numeric	numeric
vs	numeric	numeric
am	numeric	factor
gear	numeric	factor
carb	numeric	factor

Table 1 - Changes in variable types

```
print(mtcars[2:5,])
```

Table 2 - Sample data from *mtcars*

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  M    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  M    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  A    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  A    3    2
```

```
avgsd <- data.frame(Automatic = c(a.avg, a.sd), Manual = c(m.avg, m.sd))
row.names(avgsd) <- c('Average mpg', 'Std dev')
kable(avgsd, row.names = TRUE)
```

	Automatic	Manual
Average mpg	17.147368	24.392308
Std dev	3.833966	6.166504

Table 3 - Averages and standard deviations of fuel consumptions for automatic and manual cars

```r
summary(model.base)
```

Table 4 - Summary of base linear model

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## amM            7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

```r
summary(model.best)
```

Table 5 - Summary of best linear model

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## amM           2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

```
summary(model.final)
```

Table 6 - Summary of final linear model

```
##
## Call:
## lm(formula = mpg ~ am + wt + am:wt + hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0639 -1.3315 -0.9347  1.2180  5.0822
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.947333   2.723411  11.363 8.55e-12 ***
## amM         11.554813   4.023277   2.872  0.00784 **
## wt          -2.515586   0.844497  -2.979  0.00605 **
## hp          -0.026949   0.009796  -2.751  0.01048 *
## amM:wt      -3.577910   1.442796  -2.480  0.01968 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.332 on 27 degrees of freedom
## Multiple R-squared:  0.8696, Adjusted R-squared:  0.8503
## F-statistic: 45.01 on 4 and 27 DF,  p-value: 1.451e-11
```

Charts

Figure 1 - Transmission Type vs MPG .

```
ggplot(mtcars, aes(am, mpg, col = am)) + geom_boxplot() + geom_jitter(aes(col=am)) +
    labs(x='Transmission Type', y = 'mpg') +
    theme(legend.position='none')
```

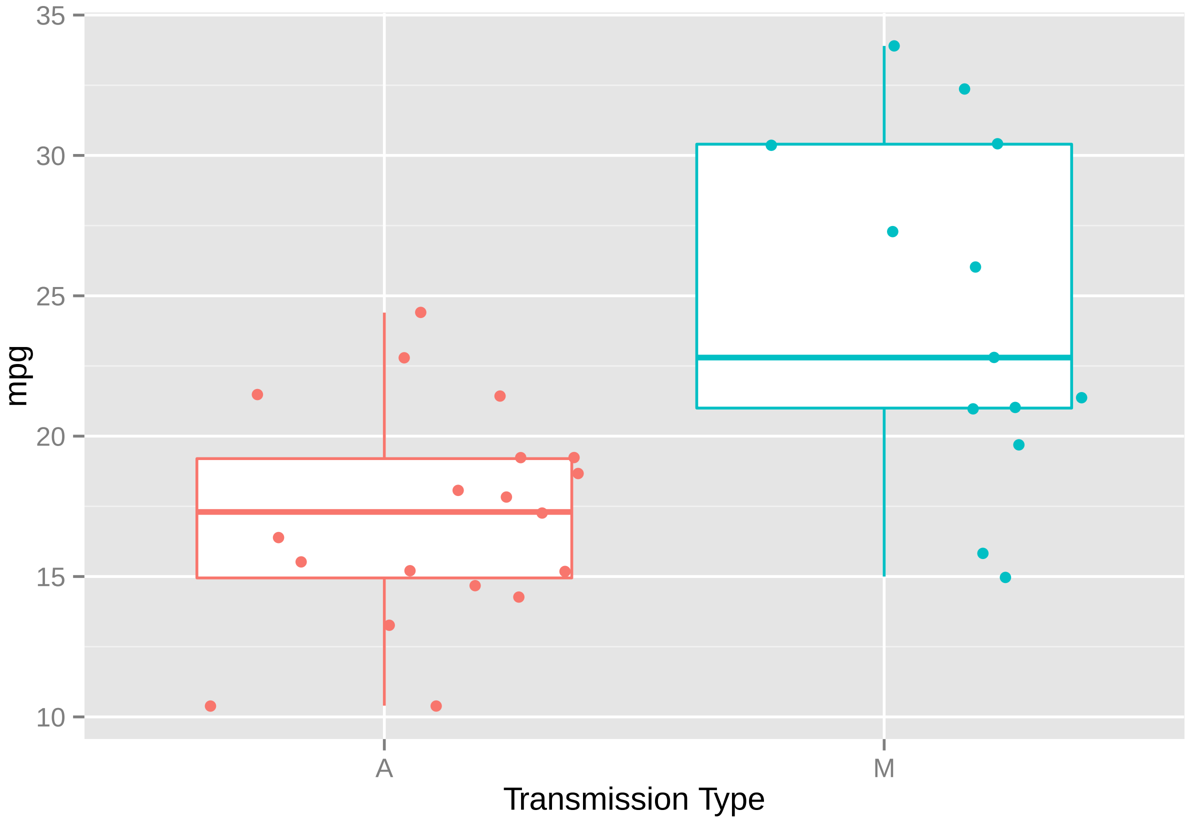
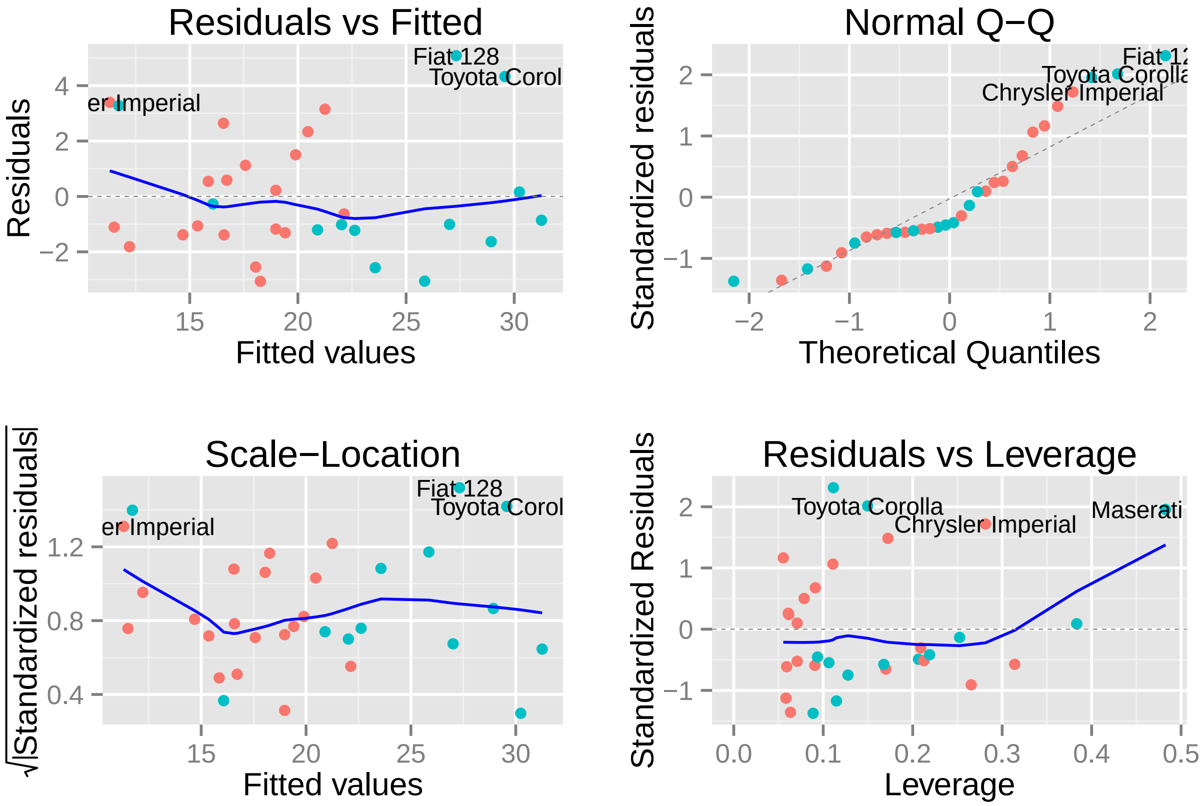


Figure 2 - Diagnostic Plots for model ***mpg ~ am + wt + am:wt + hp*** **.**

```r
# using ggplot and ggfortiy to create a diagnostic plot similar to the one plot.lm produces
# but prettier
# source: http://rpubs.com/sinhrks/plot_lm
autoplot(model.final, data = mtcars, colour = 'am', label.size=3, main='Model mpg ~ am + wt + am:wt + hp
    theme(legend.position='none')
```



Legend: red dots - Automatic transmission, teal dots - Manual transmission