Regression Models Course Project

Joseph S. Tabadero, Jr.

2017-10-01

1 Executive Summary

This is a report of the analysis of modeling the variation in the values of mileage per gallon (`mpg`) as a function of the variables in the `mtcars` data set. The following are the findings.

- Disregarding the effect of other variables, manual transmission gives better mileage performance by 7.24 mpg over automatic transmission. However, in the presence of other variables, this difference isnot extremely large, suggesting that transmission type is a confounding variable.
- A very good and parsimonious multivariate linear model explains `mpg` in terms of the weight (`wt`), number of cylinders (`cyl`), horsepower (`hp`), and type of transmission (`am`) of a vehicle.

The source for this project can be found at https://github.com/josephuses/coursera_regression_project.

2 Exploratory Data Analysis

```
library(tidyverse)
library(ggfortify)
```

Let us first conduct some exploratory data analysis to familiarize ourselves with the `mtcars` data and look at the behavior of some variables, specially `am`.

```
# dataset
data(mtcars)
# Mean and five number summaries of mpg variable:
summary(mtcars$mpg)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 10.400000 15.425000 19.200000 20.090625 22.800000 33.900000
```

```
# Convert am, cyl, gear, and vs as factor variables
mtcarsf <- mutate(mtcars, am = factor(am, labels = c("automatic", "manual")),
    cyl = factor(cyl), gear = factor(gear), vs = factor(vs, levels = c(0, 1),
        labels = c("V", "S")))
```

The boxplot of `mpg` according to `am` clearly shows that manual type transmission outperforms automatic transmissions in mileage. We shall now test this hypothesis by building linear models explaining the variation in the values of `mpg`.

3 Analyses of Regression Models

The boxplot in Appendix 4.1 suggests that manual has better milleage per gallon than automatic. Although we can conduct a t test to determine if this difference did not happen by chance alone, we use linear models and regression to explain this difference.

```r
mod1 <- lm(mpg ~ am, data = mtcarsf)
summary(mod1) %>% broom::tidy()
```

term	estimate	std.error	statistic	p.value
(Intercept)	17.14736842105	1.12460254124	15.2474921514	0.000000000000
ammanual	7.24493927126	1.76442163164	4.1061269831	0.000285020744

```r
summary(mod1) %>% broom::glance()
```

r.squared	adj.r.squared	sigma	statistic	p.value	df
0.359798943425	0.338458908206	4.90202882893	16.8602788013	0.000285020744	2

We have quantified this difference to be an increase of 7.25 miles per gallon of manual over automatic transmission vehicles, and this is more extreme to be explained by chance occurrence alone ($p = 2.85020743935e-04$) at a significance level of 0.05. However, the R-sqaured of 0.359798943425 shows that this univariate model explains only 35.98% of the variation in `mpg`. We can look at other multivariate linear models in order to see if we can improve the model fit. We can try explaining the variations in `mpg` with all of the other variables in `mtcars` but we can speed up our analysis by including only those variables that have a correlation higher than that of `mpg` and `am`.

```r
cors <- cor(mtcars$mpg, mtcars)
cors[, order(-abs(cors[1, ]))] %>% as_tibble()
```

	value
mpg	1.000000000000
wt	-0.867659376517
cyl	-0.852161959427
disp	-0.847551379262
hp	-0.776168371827
drat	0.681171907807
vs	0.664038919128
am	0.599832429455
carb	-0.550925073902
gear	0.480284757339
qsec	0.418684033922

```r
submtcars <- mtcarsf %>% select(mpg, wt, cyl, disp, hp, drat, vs, am)
mod2 <- lm(mpg ~ wt + cyl + disp + hp + drat + vs + am, data = mtcarsf)
summary(mod2) %>% broom::tidy()
```

term	estimate	std.error	statistic	p.value
(Intercept)	29.829969134145	6.744467882345	4.422879559147	0.000196207375
wt	-2.594622673563	1.201295382097	-2.159854031099	0.041448570735
cyl6	-2.055523435075	1.803107893903	-1.139989149859	0.266023824613
cyl8	-0.023304442903	3.816510169477	-0.006106217950	0.995180628128
disp	0.004360162924	0.013036105045	0.334468225684	0.741057132785
hp	-0.035794755877	0.014634234706	-2.445960215557	0.022513820233
drat	0.388141032956	1.466060243736	0.264751080056	0.793559398226
vsS	2.004897599760	1.829948485451	1.095603300148	0.284592667270

term	estimate	std.error	statistic	p.value
ammanual	2.558988827793	1.743021265173	1.468134026202	0.155611776081

```
summary(mod2) %>% broom::glance()
```

r.squared	adj.r.squared	sigma	statistic	p.value	df
0.873313066274	0.829248045848	2.49046412136	19.8187373528	1.3507e-08	9

`mod2` explains 87.33% of the variance in `mpg`. In this model, the weight (`wt`) and horsepower (`hp`) are the only ones showing a significant effect on the variation in `mpg`.

Looking at the diagnostic plots (Appendix 4.2), the Normal Q-Q plot and Residuals vs Fitted are somewhat okay for `mod2`, however the Scale-Location plot is showing some causes for worry of violations of homoscedasticity. Overall, `mod2` seems to be a good model.

We now see if a bidirectional stepwise selection will yield a better model.

```
mod3 <- step(mod2, direction = "both", trace = FALSE)
summary(mod3) %>% broom::tidy()
```

term	estimate	std.error	statistic	p.value
(Intercept)	33.708323901280	2.604886184504	12.940421006417	0.000000000001
wt	-2.496829420354	0.885587793485	-2.819403608228	0.009081407558
cyl6	-3.031344490503	1.407283510716	-2.154039656843	0.040682717936
cyl8	-2.163675323179	2.284251724546	-0.947214048228	0.352250869148
hp	-0.032109429991	0.013692574260	-2.345025075769	0.026934605236
ammanual	1.809211382941	1.396304503014	1.295714064544	0.206459673770

```
summary(mod3) %>% broom::glance()
```

r.squared	adj.r.squared	sigma	statistic	p.value	df
0.865879872487	0.840087540273	2.41011962976	33.5712127658	1.51e-10	6

`mod3` explains about 86.59% of the the variation in `mpg` but is more parsimonious than `mod2`. In this model, apart from the type of transmission, horsepower, the type of cylinder and weight of the vehicle are used to explain the variation in the milleage of the vehicles. The Normal Q-Q plot and the Residuals vs Fitted plot (Appendix 4.3) look okay, but the Scale-Location plot still show some signs of violations of equality of variance. We will see if adding `disp`, `drat` and `vs` (`mod2`) significantly improves the model fit.

```
anova(mod3, mod2)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
26	151.025592373	NA	NA	NA	NA
23	142.655465415	3	8.37012695887	0.449831860033	0.719832917309

From the results, we see that `mod2` is not a great improvement over `mod3`. We can therefore choose `mod3` to explain the variation in the mileage per gallon of the vehicles.

Using `mod3` to explain variation in the `mpg`, we can say that when all other variables are held constant:

- A unit increase in the weight of a vehicle significantly (p = 0.0091) reduces milleage per gallon by 2.5 mpg.
- 6-cylinder vehicles have lower mileage per gallon by 3 mpg over 4-cylinder vehicles. 8-cylinder vehicles have even lower mileage per gallon compared to 6-cylinder vehicles by 2.16 mpg.
- An increase of 1 unit in horsepower results to a decrease of 0.03 mpg and this per unit increase is significant (p = 0.0269).
- Manual transmission vehicles have higher mileage than automatic transmission vehicles, although this is not significant (p = 0.2064).

4 Conclusions

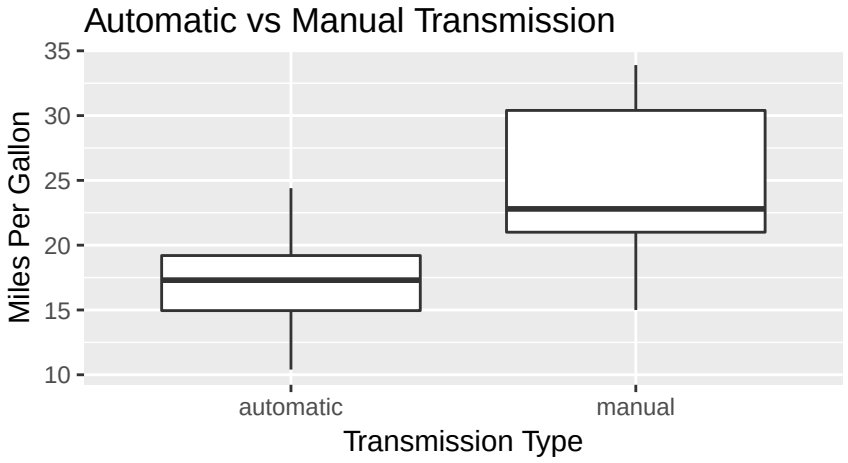
With the foregoing analyses, we therefore conclude that:

- Disregarding the effect of other variables, manual transmission gives better mileage performance by 7.24 mpg over automatic transmission. However, in the presence of other variables, this difference is not extremely large, suggesting that transmission type is a confounding variable.
- A very good and parsimonious multivariate linear model explainsmpgin terms of the weight (wt),number of cylinders (cyl), horsepower (hp), and type of transmission (am) of a vehicle.

5 Appendix

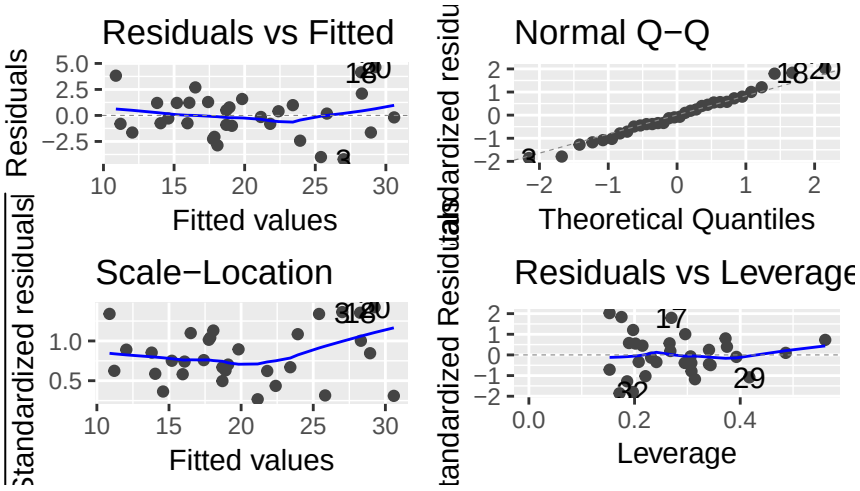
5.1 Boxplot of `mpg` by `am`

```
ggplot(mtcarsf, aes(am, mpg)) + geom_boxplot() + xlab("Transmission Type") +
    ylab("Miles Per Gallon") + ggtitle("Automatic vs Manual Transmission")
```



5.2 Diagnostic plots of `mod2`

```
autoplot(mod2)
```



5.3 Diagnostic plots of `mod3`

```
autoplot(mod3)
```