Regression Models - Course Project

Executive Summary

This study aims to explore the relationship between a set of variables and miles per gallon (MPG) outcome. We are particularly interested in knowing which transmission (automatic or manual) is better for MPG, quantitatively speaking. Starting from the model in which MPG is a function of transmission, covariates were tested and the most significant ones (hp and weight) were included. The resulting model explains 82% of total variation. In this context, manual transmission showed better fuel consumption performance.

Exploratory Data Analyses

Based on documentation, the **mtcars** dataset has 32 observations on 11 variables. It comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models). The first thing that we might want is to use all these variables to predict fuel consumption (mpg: Miles/(US) gallon). Nevertheless, the summary of adjusted estimates (Table 1) shows that t-tests are not significant (adj.p > .05). In order to improve understanding, marginal estimates were also determined.

Table 1 - Adjusted and marginal estimates relative to mpg.

variable	adj.estimate	adj.t	adj.p	marg.estimate	marg.t	marg.p	adj.r.squared
(Intercept)	12.3034	0.6573	0.5181	NA	NA	NA	NA
am	2.5202	1.2254	0.2340	7.2449	4.1061	2.850207e-04	0.3385
carb	-0.1994	-0.2406	0.8122	-2.0557	-3.6157	1.084446e-03	0.2803
cyl	-0.1114	-0.1066	0.9161	-2.8758	-8.9197	6.112690e-10	0.7171
disp	0.0133	0.7468	0.4635	-0.0412	-8.7472	9.380330e-10	0.7090
drat	0.7871	0.4813	0.6353	7.6782	5.0960	1.776240e-05	0.4461
gear	0.6554	0.4389	0.6652	3.9233	2.9992	5.400948e-03	0.2050
hp	-0.0215	-0.9868	0.3350	-0.0682	-6.7424	1.787835e-07	0.5892
qsec	0.8210	1.1234	0.2739	1.4121	2.5252	1.708199e-02	0.1478
vs	0.3178	0.1510	0.8814	7.9405	4.8644	3.415937e-05	0.4223
wt	-3.7153	-1.9612	0.0633	-5.3445	-9.5590	1.293960e-10	0.7446

Marginal estimates have significant t-tests (marg.p < .05). A first approach should focus on the issues to be addressed, which are related to the discrete factor variable **am** (0 - automatic, 1 - manual). So, let's start by looking at the basic behavior of **mpg** as a function of **am** (see Figure 1 in Appendix).

Model Selection

An intuitive analysis of Figure 1 suggests that vehicles with manual transmission have the potential to maximize gas mileage. Technically speaking, the estimated expected increase in **mpg** comparing manual to automatic is **7.2449** (Table 1) and this difference is statistically signicant (p = 0.00028502). However, the adjusted R-squared (Table 1) shows that only about 34% of total variation is explained by this model (mpg ~ am). Hence, covariate adjustment is needed for robustness (discrete variables were coerced to factors).

The approach for adding covariates was to (1) individually include the other variables to the basic model, (2) comparatively test each new nested model by ANOVA, (3) definitely include the most significant variable into the model, and (4) repeat this procedure to each new model until there was no more significant t-test. Table 2 shows that the final model (mpg ~ am + hp + wt) was found after three iteractions.

Table 2 - Iteractive approach for including covariates.

mpg ~ am	Model 1: mpg ~ am + variable		Model 2: mpg ~ am + hp + variable		Model 3: mpg ~ am + hp + wt + variable	
variable	p.value	r.squared	p.value	r.squared	p.value	r.squared
cyl	8.010109e-07	0.7399447	0.052115599	0.7989306	0.09999822	0.8400875
disp	5.747528e-07	0.7149405	0.132991837	0.7776925	0.81222292	0.8165613
hp	2.920375e-08	0.7670025	NA	0.7670025	NA	0.8227357
drat	1.069548e-02	0.4554386	0.219518178	0.7715509	0.48234129	0.8195618
wt	1.867415e-07	0.7357889	0.003574031	0.8227357	NA	0.8227357
qsec	6.270759e-06	0.6652425	0.376552464	0.7654448	0.07573120	0.8367919
vs	6.500962e-06	0.6644330	0.075900991	0.7847948	0.18968524	0.8277202
am	NA	0.3384589	NA	0.7670025	NA	0.8227357
gear	3.732942e-02	0.4395670	0.504195895	0.7621212	0.93357109	0.8101067
carb	5.269993e-04	0.6551976	0.406850569	0.7693724	0.82002768	0.8028125

Below we have the summary of our best fit model. We also examine the model **mpg ~ hp + wt** controlled by **am** for comparative analysis. Figure 2 (see Appendix) shows the plot for residual checking and diagnostics.

```
## [1] "best fit model: mpg ~ am + hp + wt    adj.R-squared: 0.823"
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 34.00287512 2.642659337 12.866916 2.824030e-13
## am1          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
```

```
## [1] "controlled model: mpg ~ am:hp + am:wt    adj.R-squared: 0.805"
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 36.21665031 2.40961737 15.030042 1.226118e-14
## am0:hp      -0.03845534 0.01583981 -2.427766 2.213547e-02
## am1:hp      -0.03300443 0.01406085 -2.347256 2.649636e-02
## am0:wt      -3.36820050 0.95786817 -3.516351 1.566198e-03
## am1:wt      -3.29638285 1.48605669 -2.218208 3.514086e-02
```

Discussion

The best fit model (BFM) captured 82.3% of total variance. Looking at the coefficents of BFM, we can see that the estimated expected increase in MPG comparing manual to automatic is **2.0837**. However, this difference is not statistically significant (p = 0.1412). In order to better understand this phenomenon, the controlled model (CM) was also considered. The CM coefficients are statistically significant and show that the increase of 1 HP results in MPG decreases of 0.039 miles for automatic cars and 0.033 miles for manual cars, which suggests better performance for manual transmission (about 16% better). Moreover, as weight increases 1000 lbs, MPG decreases 3.368 miles for automatic cars and 3.296 miles for manual cars, which favors manual transmissions again, but with a slight difference (about 2%). The Figure 3 shows that the group status **am** does not reverse the influence of **hp** but practically equalizes the effect of **wt**. The residual plot (Figure 2) does not reveal a pattern but shows the strong influence of three vehicles on the model, which is reflected in QQ Plot where standardized residuals don't seem to be normal. This fact deserves special treatment in future studies. In conclusion, this study suggests that manual transmition has better fuel consumption performance, considered the model limitations.

P.S.: The markdown file used to generate this report can be found in https://github.com/Fpschwartz1/RegressionModels/blob/master/CourseProject.Rmd

Appendix

Figure 1 - Linear model for mpg ~ am

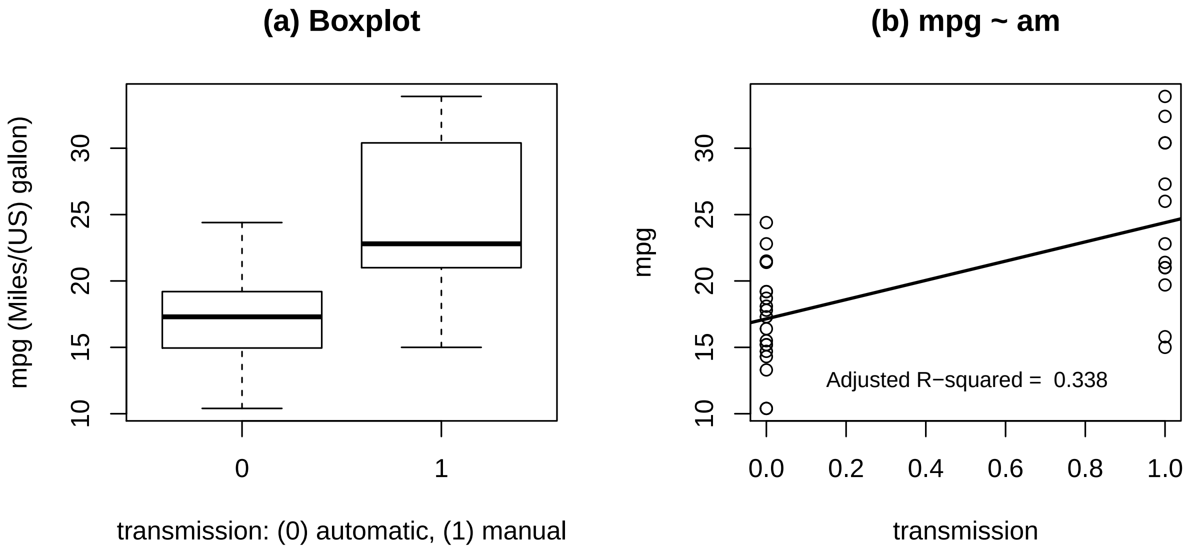


Figure 2 - Residual plot for the best fit model

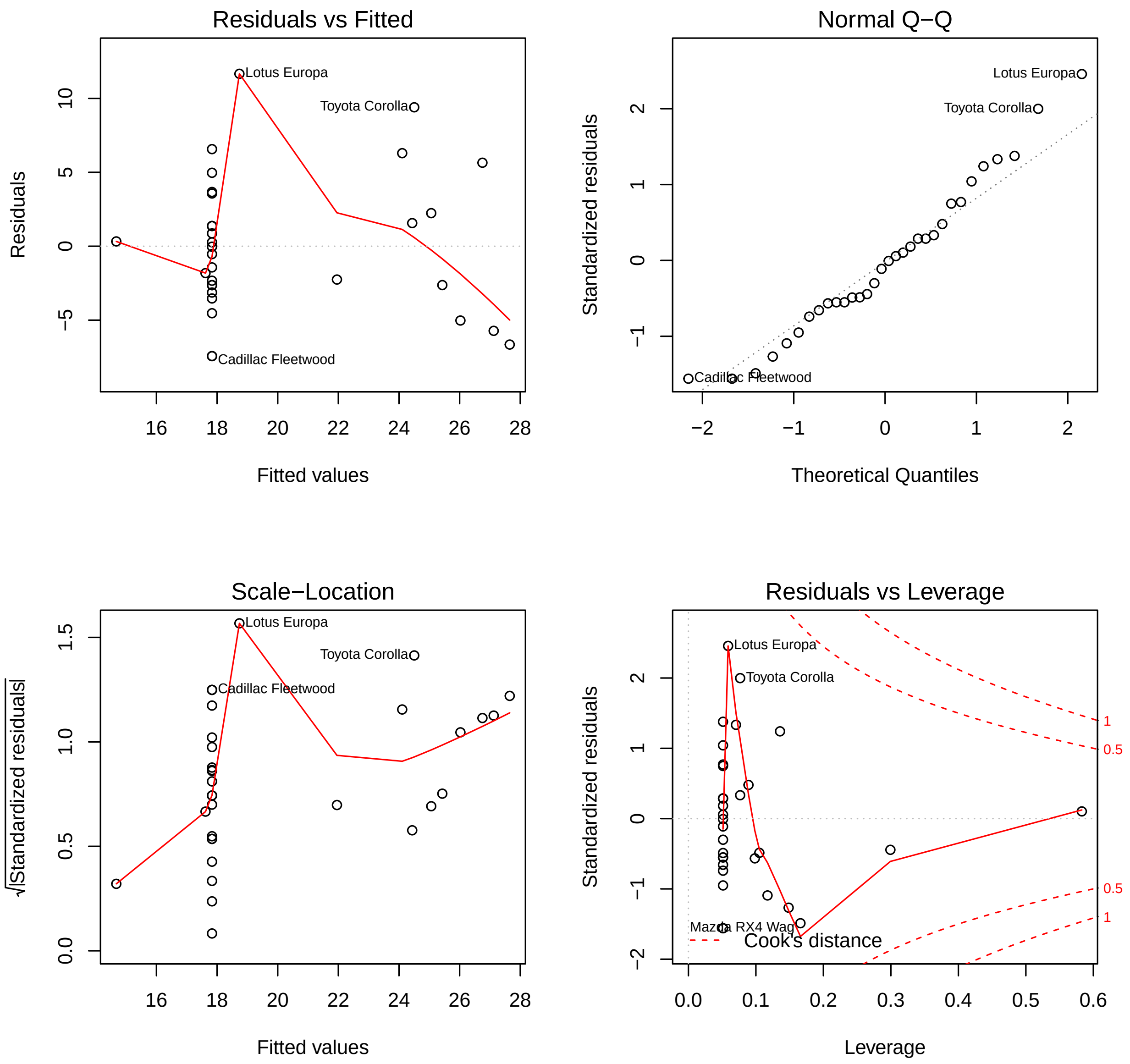
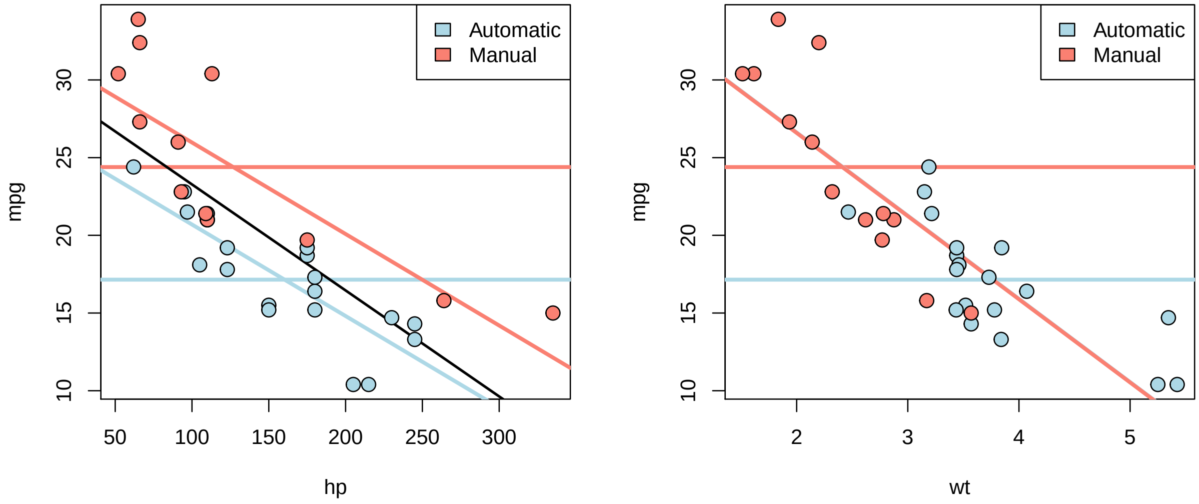


Figure 3 - Marginal plots


Automatic
Manual
mpg
hp
Automatic
Manual
mpg
wt