Transmission Type and Fuel Efficiency

Alyssa Goldberg

January 10, 2016

Executive Summary

Hypotheses and Questions

Using the mtcars dataset from a 1974 Motor Trend US magazine, we examine the following hypotheses: H_0 = There is no significant difference in fuel efficiency between cars with **automatic** and **manual** transmission. H_{01} = There are no confounding variables that contribute to the acceptance or rejection of H_0.

1. Is an automatic or manual transmission better for MPG ?
2. If so, quantify the MPG difference between automatic and manual transmissions

Summary of Findings

All figures are in the appendix 1. Cars with manual transmission get **7.24** more mpg than automatic transmission when only transmission type is considered

2. Exploring regression models using multivariate analysis strongly suggests that both weight `wt`, number of cylinders `cyl` and `hp` have a significant confounding effect on fuel efficiency as it relates to transmission type.

With other variables factored in, cars with manual transmission get an additional **1.8** mpg better than those with automatic transmission. In a model with an R^2 value of 84%, `am` explains only 13% of that 84% or at best 11% of the differnce in mpg when normalized.

Data

The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).

Note: The 1/4 mile time `qsec` is a measure of performance rather than a manufacturing variable and thus we remove qsec from the dataset for the purposes of analysis.

Factors and Measurements

`mpg` Miles/(US) gallon (the measure of, performance we are interested in), `disp` Displacement (cu.in.), `hp` Gross horsepower, `drat` Rear axle ratio, `wt` Weight (lb/1000), `qsec` 1/4 mile time, `cyl` Number of cylinders, `vs` V/S, `am` Transmission (0 = automatic, 1 = manual), `gear` Number of forward gears, `carb` Number of carburetors.

Figure 1 and Table 1 (below) include only `mpg` and `am` indicate that manual transmission has a mean mpg 7.245 automatic.

A t-test of `mpg` and `am` can reveal both the confidence interval and p-values for the two factors in the `am` variable:

```
t.test(mtcars[,-7]$mpg~mtcars[,-7]$am)
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars[, -7]$mpg by mtcars[, -7]$am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group Automatic    mean in group Manual
##                17.14737                24.39231
```

The CI, -11.2801944, -3.2096842, does not contain zero *and* a p-value of 0.0013736, significantly better than our 95% confidence level alpha = 0.05. If the only two variable measured were fuel efficiency and transmission type, we could safely reject H_0.

There are 8 additional manufacturing variables in the data set and the adjusted R^2 for `mpg~am` is low, only 0.338, so transmission alone accounts for only 33.85% of the total effect. We need to find a better model.

```
kable(xt, caption = "Table 1: Linear Regression Coefficients for Automatic vs. Manual Transmission")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.147368	1.124602	15.247492	0.000000
amManual	7.244939	1.764422	4.106127	0.000285

Table 1: Linear Regression Coefficients for Automatic vs. Manual Transmission

Figure 2 shows that cyl, disp, wt and hp all have a negative effect on fuel efficiency.

Find the Best Model

Step Analysis for AIC

We employ R's step() function to find the model with the lowest AIC. (results hidden to limit length of report) and produce a table comparing our three models, mpg ~ am, mpg ~ . and the result of the step analysis, mpg ~ cyl + hp + wt + am.

```
fitAIC=extractAIC(lm(mpg ~ am, mtcars[,-7]))
allAIC=extractAIC(lm(mpg ~ ., mtcars[,-7]))
bestAIC=extractAIC(lm(mpg ~ cyl + hp + wt + am, mtcars[,-c(7)]))
tblAIC<-as.data.frame(rbind("mpg ~ am"= fitAIC[2],"mpg ~ ." = allAIC[2],"mpg ~ cyl + hp + wt + am" = bestAIC[2]))
names(tblAIC)<-"AIC"
kable(tblAIC, caption = "Comparison of AIC Values of Three Models")
```

	AIC
mpg ~ am	103.67231
mpg ~ .	74.73161
mpg ~ cyl + hp + wt + am	61.65483

Comparison of AIC Values of Three Models

The model which includes: mpg~cyl + hp + wt + am has the lowest AIC value of 61.654829 and is recommended as the best fit by the step analysis. A summary of the best fit model coefficients confirm we are moving in the right direction. The net influence of the am variable when am = 1 (manual transmission) is 1.809. We will continue to test this model against the others.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	33.7083239	2.6048862	12.940421	0.0000000
cyl6	-3.0313445	1.4072835	-2.154040	0.0406827
cyl8	-2.1636753	2.2842517	-0.947214	0.3522509
hp	-0.0321094	0.0136926	-2.345025	0.0269346
wt	-2.4968294	0.8855878	-2.819404	0.0090814
amManual	1.8092114	1.3963045	1.295714	0.2064597

Analysis of Variance

An analysis of variance shows us that number of cyl, wt and disp are significant at a 95% confidence level, while the p-value of hp is higher than .05 indicating we *might* reject it as a significant factor.

```
kable(as.data.frame(summary(aov(mpg~., data))[[1]]))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cyl	2	824.784590	412.392295	54.2425208	0.0000001
disp	1	57.642804	57.642804	7.5818366	0.0141316
hp	1	18.502205	18.502205	2.4336203	0.1383175
drat	1	11.914476	11.914476	1.5671272	0.2286170
wt	1	55.786898	55.786898	7.3377268	0.0154894
vs	1	1.391962	1.391962	0.1830867	0.6744376
am	1	13.368787	13.368787	1.7584147	0.2034425
gear	2	3.390542	1.695271	0.2229813	0.8025747
carb	5	17.620932	3.524186	0.4635410	0.7976639
Residuals	16	121.643991	7.602749	NA	NA

The next step in finding out how closely correlated these confounding variables are. A correlation matrix (figure X) shows that `disp` and `cyl` have a correlation of 90% while `wt` and `disp` are 89%. Both show significantly lower p-values so it is possible that including `disp` to a model might result in "overfit."

```
kable(c4, caption = "Correlation among variables with lowest p-value in analysis of variance model")
```

	correlation
wt & cyl	0.7824958
wt & disp	0.8879799
cyl & disp	0.9020329

Correlation among variables with lowest p-value in analysis of variance model

We compare several models to determine which one is truly "best" and the fourth model which retains `hp` does have the highest R^2: 0.8400875, accounting for 84% of the variability in fuel efficiency.

```
kable(fitsummary)
```

	Adj. R-squared
mpg ~ .	0.8015942
mpg ~ am	0.3384589
mpg ~ cyl + wt + am	0.8121603
mpg ~ cyl + hp + wt + am	0.8400875

An ANOVA test of the best fit model and the original model provides a p-value of 1.6910^{-8} indicating that we can reject H_{01}- that number of cylinders, weight and horsepower are not confounding factors (more cylinders = lower `mpg`, higher weight = lower mpg, higher horsepower=lower mpg) to transmission type as regards fuel efficiency.

When number of cylinders, weight and horsepower are included as confounding variables, the net increase in fuel efficiency for an automobile with automatic transmission is ** 1.809**

The table below generated with R's bootstrap function lists the relative importance of each variable and how much it contributes to the final R^2 value of 84%.

```
kable(btbl,caption = "% contribution to R-squared value of best fit model" )
```

	% Contribution
cyl	32.12
hp	29.59
wt	24.80
am	13.49

% contribution to R-squared value of best fit model

Residuals

The Residual Plots (Figure 4) confirm our model in the following ways: 1. Residuals vs. Fitted plot points are randomly distributed confirming independence. 2. Points on the Normal Q-Q plot hug the normal line with some outliers in the tails confirming normal distribution of the residuals. 3. The Scale-Location plot displays a regular band pattern of points confirming constant variance.

Appendix

figure 1: Boxplot mpg ~ am (mpg vs transmission type)

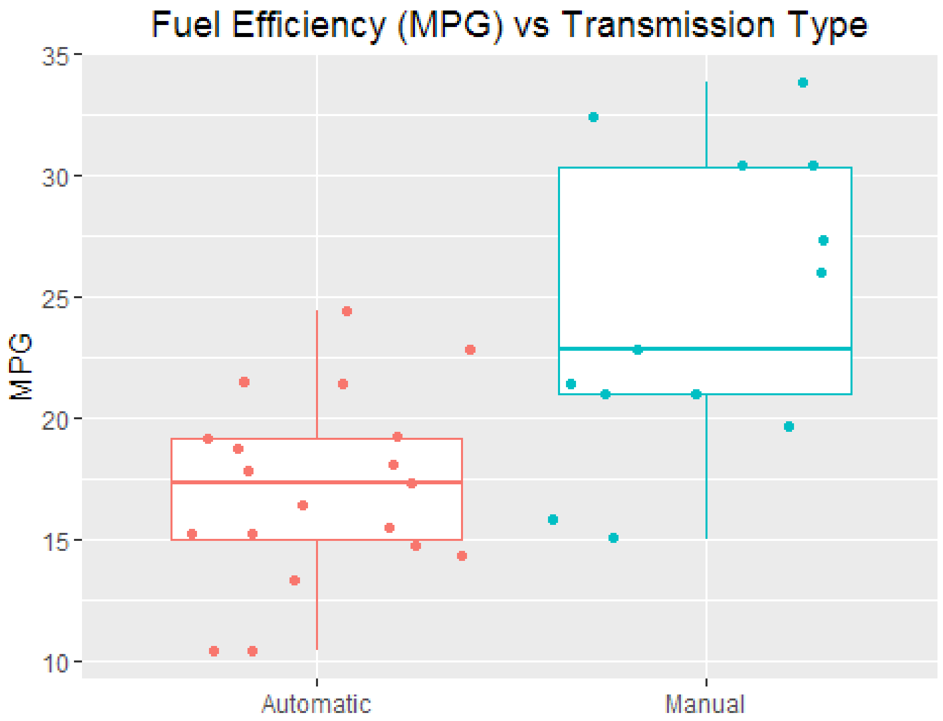


Figure 2: Pairs Plot

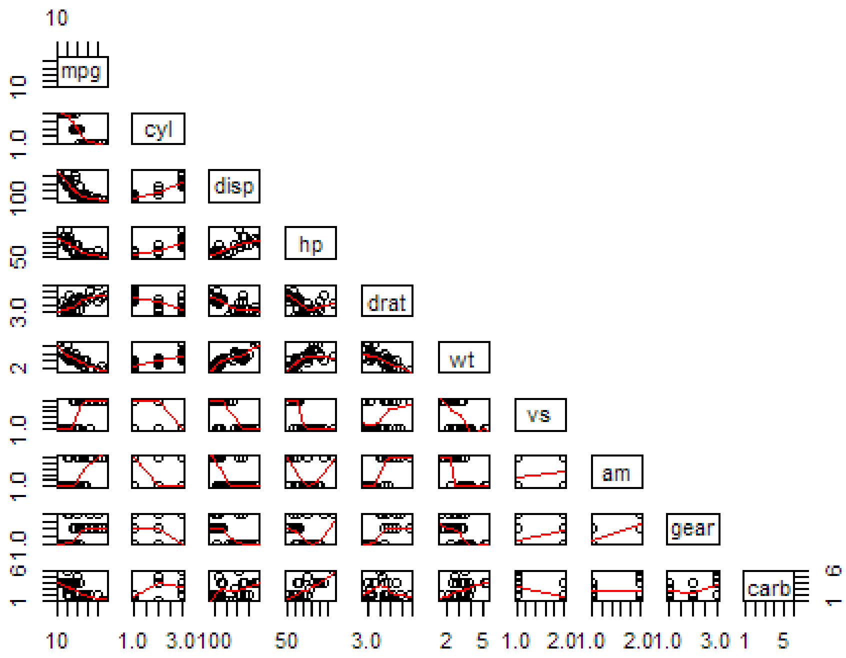
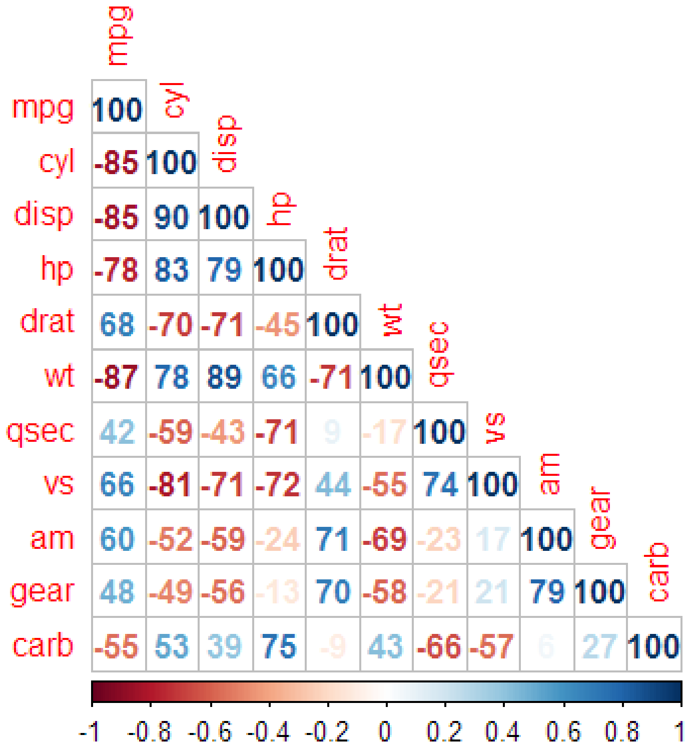
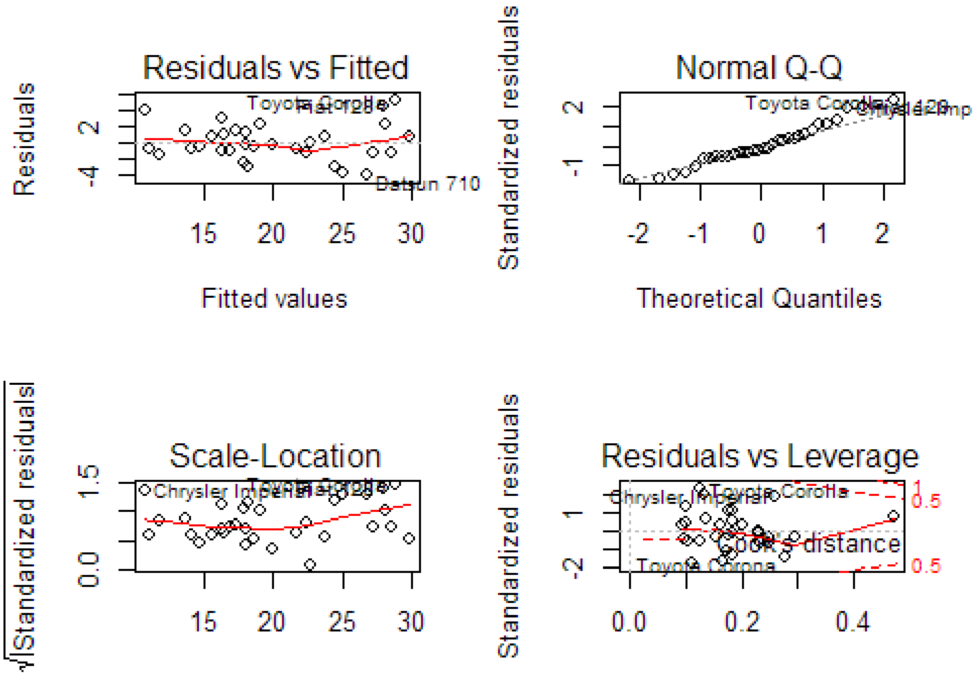


Figure 3: Correlation Matrix



	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	100										
cyl	-85	100									
disp	-85	90	100								
hp	-78	83	79	100							
drat	68	-70	-71	-45	100						
wt	-87	78	89	66	-71	100					
qsec	42	-59	-43	-71	9	-17	100				
vs	66	-81	-71	-72	44	-55	74	100			
am	60	-52	-59	-24	71	-69	-23	17	100		
gear	48	-49	-56	-13	70	-58	-21	21	79	100	
carb	-55	53	39	75	-9	43	-66	-57	6	27	100

Figure 4: Residual Plots



Residuals vs Fitted
Residuals
-4
2
Toyota Corolla
Fiat 128
Datsun 710
15
20
25
30
Fitted values
Normal Q-Q
Standardized residuals
-1
2
Toyota Corolla
Fiat 128
Chrysler Imp
-2
-1
0
1
2
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
0.0
1.5
Chrysler Imperial
Toyota Corolla
Fiat 128
15
20
25
30
Fitted values
Residuals vs Leverage
Standardized residuals
-2
1
Chrysler Imperial
Toyota Corolla
Cook's distance
Toyota Corona
1
0.5
0.5
0.0
0.2
0.4
Leverage