Statistical Anaylsis of Motor Trends Car Dataset

By: Wesley Small (smallwesley) - Date: June 2016

Executive Summary

This report will examine the 1974 Motor Trends automobile dataset, and look specifically at the relationship between the fuel consumption, listed as "mpg", as compared with it's relationship between the numerous additional variables. There shall be emphasis on examining the Motor Trends MTCARS dataset to answer these specific queries:

1. "Is an automatic car better for MPG than a manual car? Which is better overall."
2. "Quantify the MPG difference between automatic vs manual transmissions.

Utilizing linear regression modeling and multivariate anaylsis, we are able to say with a bit more certainty that manual cars have better overall millage than automatic transmission cars.

Initially, using only transmission as a predictor, were able to describe that manual cars received 7.24 more mile per gallon. With multivariate anaylsis (possibly confounded), we still see slight increase of mpg for manual vehicles over automatic cars. Finally, we generated a best fit model containing a subset of predictors (horsepower, weight and transmission), better explaining the variability in MPG statistics. While not explaining the variable only 84%, we this bestfit model answered that manual cars recieved about 1.8 to 2.0 more MPG than automatic cars.

Exploratory Data Anaylsis

```
data(mtcars); dfData <- mtcars;
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Maserati Bora	15.0	8	301.0	335	3.54	3.570	14.60	0	1	5	8

```
dfData$am <- as.factor(ifelse(dfData$am == 1,"Manual", "Automatic"))
dfData$vs <- as.factor(ifelse(dfData$vs == 1,"straight", "v-engine"))
dfData$cyl <- as.factor(dfData$cyl); dfData$carb <- as.factor(dfData$carb);
names(dfData)[names(dfData)=="am"] <- "transmission"
names(dfData)[names(dfData)=="wt"] <- "weight"
```

In FIGURE 2 (see appendix), we chart out a histogram of the MPG data alone. We find that it is emprically normal in distribution.

In FIGURE 3 (see appendix), we graph out boxplot using the formula (mpg ~ transmission). We can with using only this variable to filter fuel consumption, we see the manual cars perform better than automatic cars.

Statistical Inference:

Using hypothetical testing we'll define our criterial for a null hypothesis and the alternative.

1. H_0: Avg.MPG(Automatic-Cars) == Avg.MPG(Manual)
2. H_A: Avg.MPG(Automatic-Cars) != Avg.MPG(Manual)

```
myTee <- t.test(mpg ~ transmission, data = dfData, var.equals=FALSE, paired = FALSE)
```

Mean/T-Estimate	Confidence Intervals	T-Statistic	P-Value
17.147 vs 24.392	-11.28 <=> -3.21	-3.767	0.001

From this T-TEST answer, we see a low p-value of 0.0013736. Thus, we would reject the null hypothesis in favour of the alternative as stated above. We can see given the means of each transmission type group that on average, manual transmission out perform automatic vehicles by 7.24 miles/gallon.

Per Contra!

We cannot summarily accept this mid-point conclusion. We know this vehicle dataset has a number of specific variables that may affect our MPG totals. Given the variables such as weight, displacement, cylinders, horsepower, gear ratio, etc, there variables affect how the each engine type perform. Given these addtional variables we must evaluate how they compare along side transmission type (automatic vs manual), to see if one is truly better than the other for mileage. We should be clear that including variables into model can increase standard errors within the regression modeling. We also should make note that one of two or more variables may have strong relationships (highly correlated) to the other, which could skew our model if added.

In FIGURE 4 (see appendix), we take an initial snaphot of correlation between all columns in the dataset. We find that MPG is strongly correlated to Cylinders, Displacement and Weight. Cylinders is closely associated to displayment engine block type (vs), appears that there is a deep set of colinearity in this content. We shall the a producedual approach constructing a set of linear models utilizing the variables present us in the MTCARS dataset

Modeling (Linear Regression)

A) BASE MODEL: SIMPLE LINEAR REGRESSION

We'll look at the inital model, the relationship between MPG and TRANSMISSION, holding all other variables constant.

```
baseModelFit <- lm(mpg ~ transmission, dfData);
knitr::kable(summary(baseModelFit)$coef[c("(Intercept)","transmissionManual"),], format = "markdown")
```

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

The adjusted R^2 value is at 0.3384589 means only 34% of differences (or variability) in mileage (mpg) can be explained by the transmission type. Hence, we should look into add more variables into our model to learn how strong one transmission type is over the other.

B) ALL VARIABLES MODEL: MULTIVARIABLE LINEAR REGRESSION

```
allVariablesModelFit <- lm(mpg ~ ., dfData);
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	22.320741	18.840464	1.1847235	0.2534269

	Estimate	Std. Error	t value	Pr(>|t|)
transmissionManual	1.123761	2.616354	0.4295142	0.6732771

We see that the R^2 value 0.7927971 provides insight that usage of more of the variables leads to a better explanation of the variability of the mileage. The p-values for hp and weight are < 0.10 significant code, thus suggest there is evidence form the dataset sample, that there is an effect to mileage. We'll explore these variables alongside with transmission in a new model fit.

C) BETTER MODEL FIT: MULTIVARIABLE REGRESSION

In this 3rd model fit, only use 3 predictors: HorsePower Weight and Transmission(required).

```
betterModelFit <- lm(mpg ~ hp + weight + transmission, dfData);
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	34.0028751	2.6426593	12.866916	0.0000000
hp	-0.0374787	0.0096054	-3.901830	0.0005464
weight	-2.8785754	0.9049705	-3.180850	0.0035740
transmissionManual	2.0837101	1.3764202	1.513862	0.1412682

We see that the R^2 value at82.27% is much better. The coeficient slope for transmissionManual indicates that manual cars get 2.08 more mpg than automatic cars, when we factor in weight and horsepower.

D) STEP-WISE REGRESSION

From further research on optimizing linear regression modeling step-wise regression offer a way to obtain a strong model. This method not an exact science or favoured approache, however it provides a sanity check on finding a better linear regression model fit. The Step-AIC method R coding example was reference to from this website: [http://www.statmethods.net/stats/regression.html].

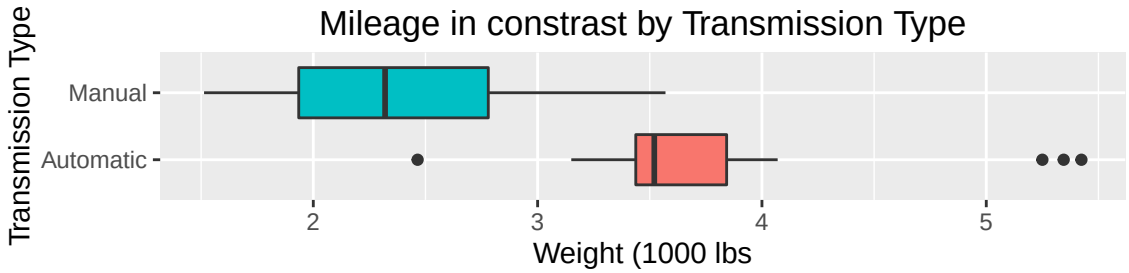
```
stepwiseModelFit <- stepAIC(allVariablesModelFit, direction="both",trace=FALSE)
```

	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
weight	-2.4968294	0.8855878	-2.819404	0.0090814
transmissionManual	1.8092114	1.3963045	1.295714	0.2064597

We see that the R^2 value at 84.01%. The coeficienct slope for transmissionManual indicates that manual cars get 1.81 more mpg than automatic cars, when we factor in weight and horsepower.

Concluding Remarks:

We see that manual cars have better fuel consumption by about 2.0 more mpg than automatics. This was obtained from the what we saw in the betterfit model and the stepwise model.



It appears fuel consumption has more to do with how well each engine performed given a certain weight, the horsepower + cylinders. As we see in the above plot, automatic cars are far heavier on average. An assumption, is that automatic cars have heavier set of parts and have a heavier chasis, due to on average their are more luxury sedans, etc using this transmission type. All this factors into how big an engine, how much horsepower is required to moved said vehicles around, and effectively down to how much fuel is consumed.

APPENDIX + FIGURES

FIGURE 1: About the MTCARS dataset

Description: This dataset comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). The mtcars dataset is a date-frame with 32 observations on 11 variables:

#	Name	Description	#	Name	Description
1	mpg	Miles/(US) gallon	7	qsec	1/4 mile time
2	cyl	Number of cylinders	8	vs	V/S (0=V-engine, 1=straight engine)
3	disp	Displacement (cu.in.)	9	am	transmission
4	hp	Gross horsepower			(0 = automatic, 1 = manual)
5	drat	Rear axle ratio	10	gear	Number of forward gears
6	wt	Weight (1000 lbs)	11	carb	Number of carburetors

FIGURE 2: Visualize the normalized MPG data

This histogram we have a histogram and density plot of MPG column from our dataset. It appears to be Unimodal and symmetric, thus we can rely on our mean and standard deviation in our linear model calculations.

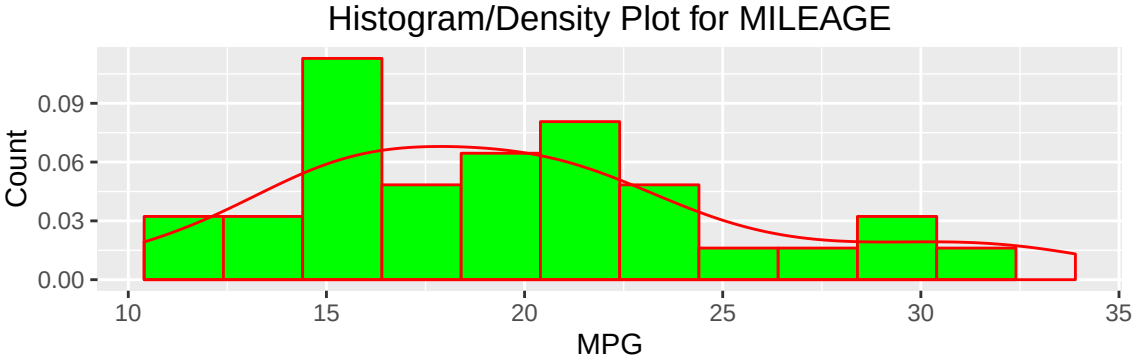


FIGURE 3: MPG vs. Transmission Type

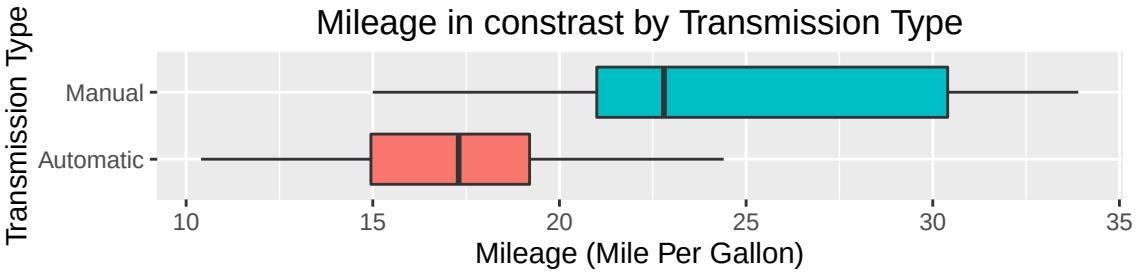


FIGURE 4: Corelation of variables

Since our original data (mtcars) is numerically tidy, we can examine a snapshot of correlation matrix, which provides insight to some of the relationshop between the variables.

```
data(mtcars); cov_matrix <- cor(mtcars); knitr::kable(symnum(cov_matrix, show.max = NULL), format = "markd
```

	m	cy	ds	h	dr	w	q	v	a	g	cr
mpg											
cyl	+										
disp	+	*									
hp	,	+	,								
drat	,	,	,	.							
wt	+	,	+	,	,						
qsec	.	.	.	,							
vs	,	+	,	,	.	.	,				
am	.	.	.		,	,					
gear	.	.	.		,	.			,		
carb	.	.	.	,		.	,	.			

FIGURE 4: Residual Plots and Diagnostics

Diagnostics are taken of the better fit line and see that their are not outliers in the date. It appear normal.

```
fit <- lm(mpg ~ hp + weight + transmission, dfData)
par(mfrow = c(2, 2))
plot(fit)
```

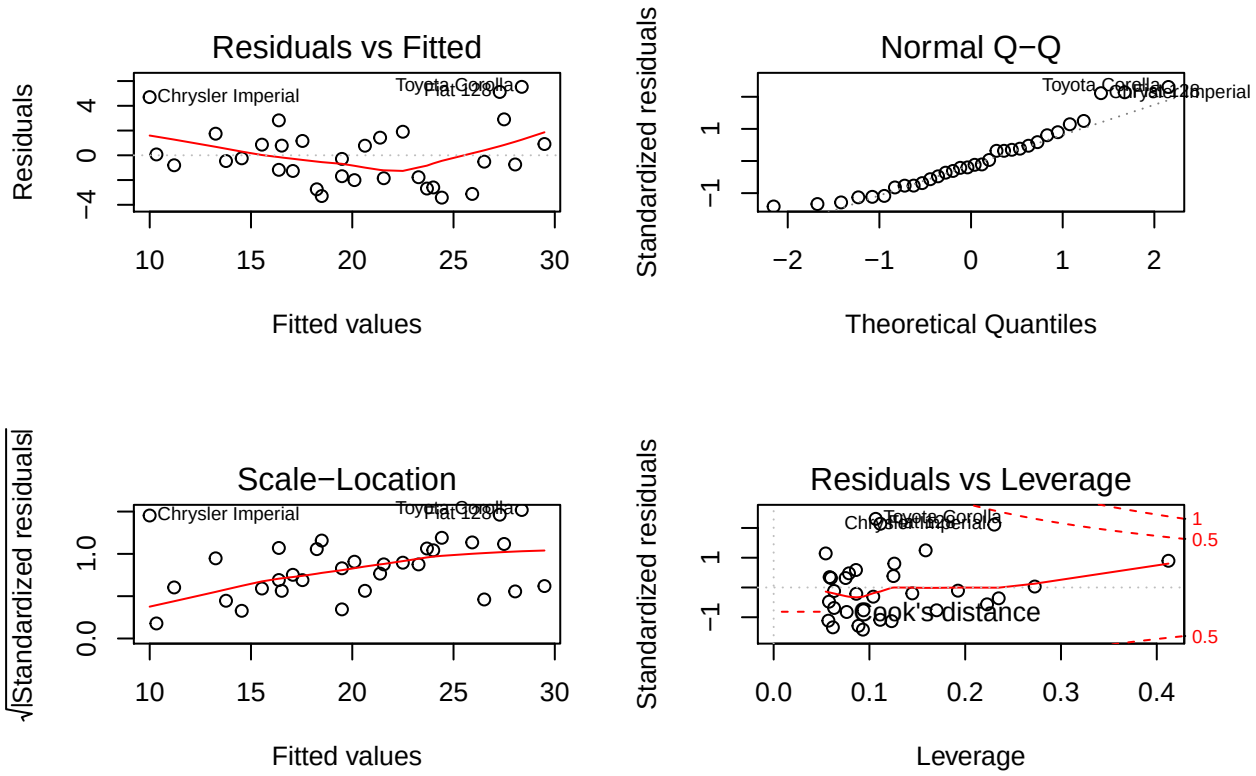


FIGURE 5: Additional tools and Diagnostics on models

From the class slides, lectures and website resources such as: [http://www.statmethods.net/stats/rdiagnostics.html], we can take a set of diagnotics on the models.

```
fit1 <- lm(mpg ~ transmission, mtcars)
fit2 <- lm(mpg ~ hp + weight + transmission, mtcars)
fit3 <- lm(mpg ~ ., mtars)

# !!!! DIAGNOSE OUT TURNED OFF DUE TO VERBOSITY !!!

coefficients(fit) # model coefficients
confint(fit, level=0.95) # CIs for model parameters
fitted(fit) # predicted values
residuals(fit) # residuals
anova(fit,fit2,fit3) # anova table
vif(fit3) # VIF Collinarity check
vcov(fit) # covariance matrix for model parameters
influence(fit) # regression diagnostics
```