Motor Trend Analysis

atops

July 10, 2015

Context

This analysis is performed on behalf of Motor Trend, a magazine about the automobile industry. We are interested in the relationship between a set of variables and miles per gallon (MPG) (outcome). In particular, we are interested in the following two questions:

1. “Is an automatic or manual transmission better for MPG”
2. "Quantify the MPG difference between automatic and manual transmissions"

Executive Summary

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Approach

In order to fit a model to answer the questions above, we will follow a process as follows:

1. Create some scatter plots showing relationships between predictor (and response) variables
2. In addition to transmission type (the predictor in question), make judgments about which variables may be most relevant to mpg, based on the scatter plots and what we know about cars.
3. Start with the simplest and most naive model of mpg as a function of transmission type. Call this the base model.
4. Seek to improve the base model through the addition (or subtraction) of variables based on F-statistics from ANOVA analysis of adjusted and unadjusted models (the impact of adding the variable).
5. Validate no outliers are having undue influence on the model based on hat values (measures of leverage) and df betas (change in individual coefficients when the ith point is deleted in fitting the model).
6. Settle on the best predicting yet most parsimonious model for mpg based on the available predictor variables. Draw conclusions on impact of transmission type by the coefficient, which shows the impact on mpg with different transmission types, holding all other model variables constant. The affect of all other unmodeled variables is captured in the error term, which we will try to keep as manageable as possible.

Exploratory Data Analysis

```r
data(mtcars)
library(ggplot2)
library(pander)

lm_justification <- c('left', 'right', 'right', 'right', 'right')
options(scipen=7)

panderOptions('round', 4)
panderOptions('keep.trailing.zeros', TRUE)
```

The Motor Trend data set has 32 observations on 11 variables:

var	Variable Description
mpg	Miles/(US) gallon
cyl	Number of cylinders
disp	Displacement (cu.in.)
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (lb/1000) *
qsec	1/4 mile time *
vs	V/S (V engine or straight engine)
am	Transmission (0 = automatic, 1 = manual) *
gear	Number of forward gears
carb	Number of carburetors

A pairs plot is included in the appendix. Based on inspection, the following variables appear significant:

1. am: The predictor in question, transmission type (manual or automatic) should have an impact on gas mileage.
2. cyl: Cars with smaller engines (fewer cylinders) get better gas mileage. Also, there appears to be a clear relationship in the scatter plot to support this. mpg goes down with an increase from 4 to 6 to 8 cylinders.
3. wt: Weight logically should impact gas mileage as it takes more energy to move a heavier vehicle. The data appear to support this with a strong downward trend in mpg with an increase in weight.

Several other variables appear to have a relationship with mpg, but perhaps not as strong. The only variable I would exclude off the bat is gear, which seems to have little or no relationship with mpg. We will have to consider which ones truly add value to the model.

Model Selection

The base model coefficients are as follows:

```r
basemodel <- lm(mpg ~ am, mtcars)
pander(summary(basemodel), justify = lm_justification)
```

	Estimate	Std. Error	t value	Pr(>|t|)
am	7.245	1.764	4.106	0.0003
(Intercept)	17.147	1.125	15.248	0.0000

Observations	Residual Std. Error	R^2	Adjusted R^2
32	4.902	0.3598	0.3385

Fitting linear model: mpg ~ am

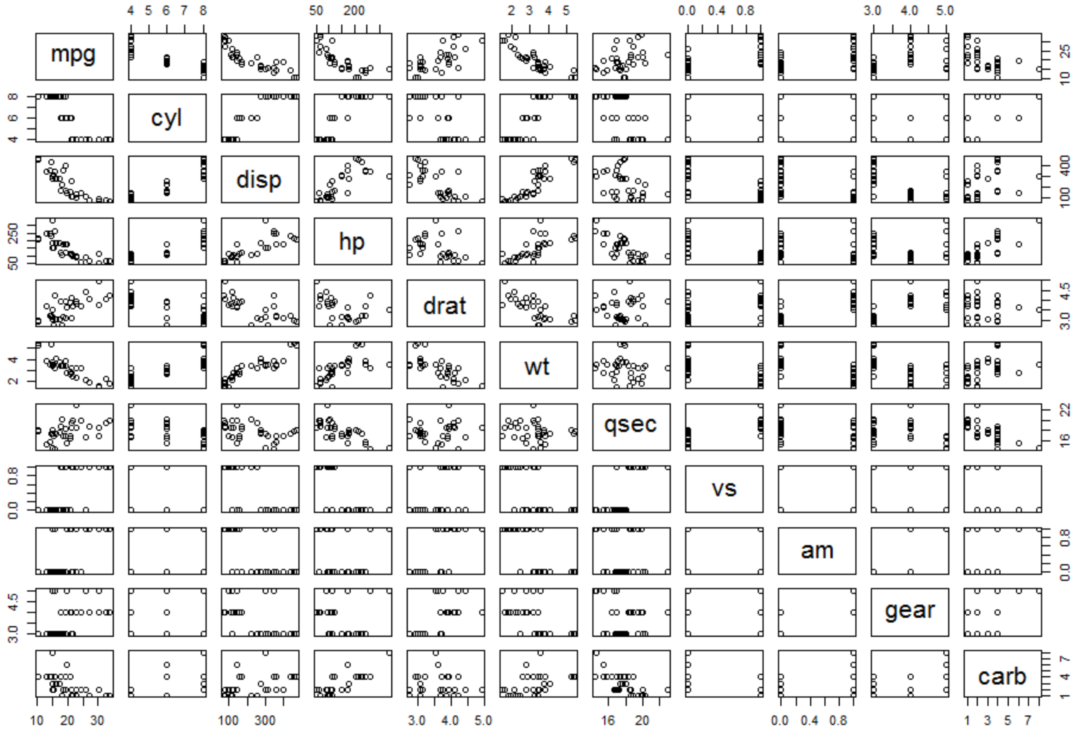
Based on p-values below 0.05, these predictor variables are significant so we reject the hypothesis that their coefficients are 0.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Appendix

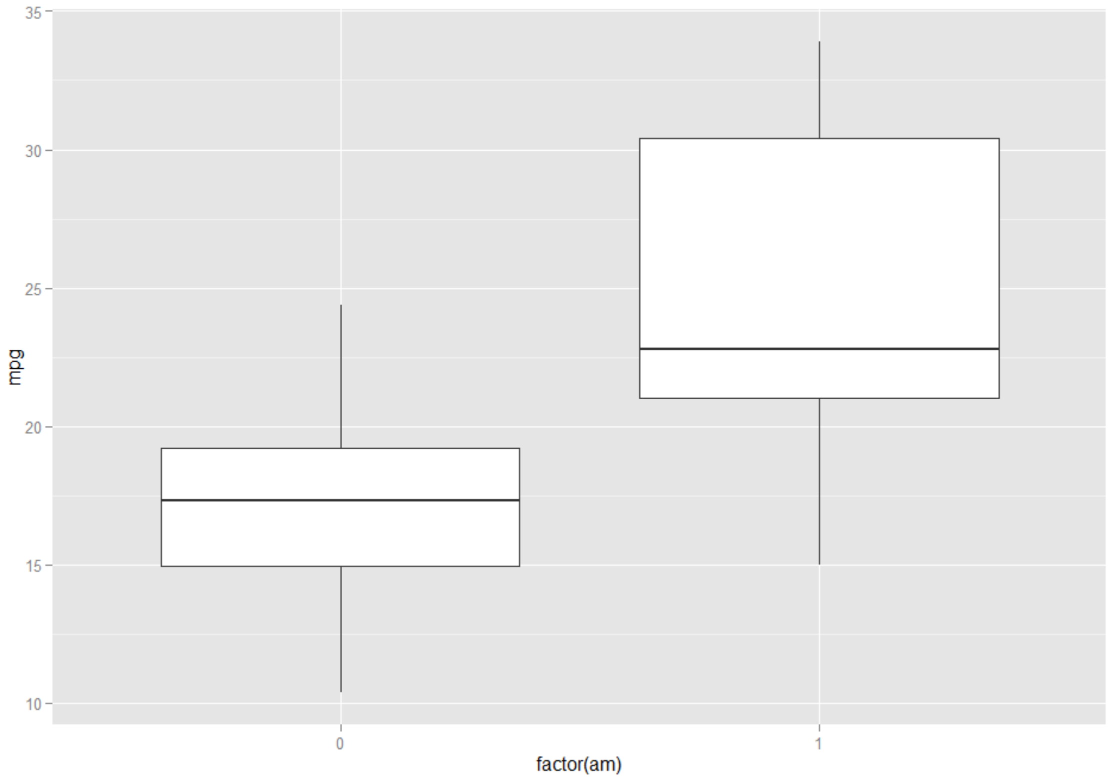
Exploratory Analysis: Pairs Plot

```r
pairs(mtcars)
```



Some text.

```
ggplot(data=mtcars) + geom_boxplot(mapping=aes(x=factor(am), y=mpg))
```



Some text.

Model Selection: Base Model

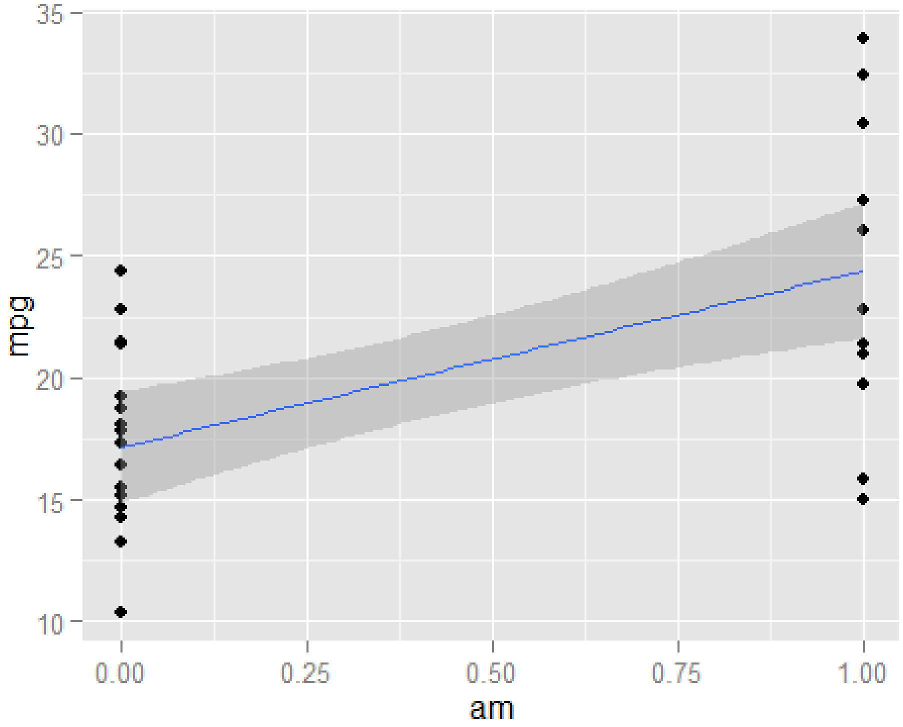
```r
basemodel <- lm(mpg ~ am, mtcars)
pander(summary(basemodel), justify = lm_justification)
```

	Estimate	Std. Error	t value	Pr(>|t|)
am	7.245	1.764	4.106	0.0003
(Intercept)	17.147	1.125	15.248	0.0000

Observations	Residual Std. Error	R^2	Adjusted R^2
32	4.902	0.3598	0.3385

Fitting linear model: mpg ~ am

```r
qplot(data=mtcars, x=am, y=mpg) + stat_smooth(method="lm")
```



Model Selection: Adding weight

```
model2 <- lm(mpg ~ am + wt, mtcars)
pander(summary(model2), justify = lm_justification)
```

	Estimate	Std. Error	t value	Pr(>|t|)
am	-0.0236	1.5456	-0.0153	0.9879
wt	-5.3528	0.7882	-6.7908	0.0000
(Intercept)	37.3216	3.0546	12.2180	0.0000

Observations	Residual Std. Error	R^2	Adjusted R^2
32	3.098	0.7528	0.7358

Fitting linear model: mpg ~ am + wt

```
pander(anova(basemodel, model2))
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
30	720.9	NA	NA	NA	NA
29	278.3	1	442.6	46.12	0

Analysis of Variance Table

FIX THIS: *The ANOVA output gives the maximum likelihood ratio test result of the significance of am (manual or automatic transmission) to the base model. The p-value (0.9089) is greater than 0.05 so we do not reject the hypothesis that this variable coefficient is not equal to zero so we can safely exclude it from the model. Frankly, this is surprising to me as I would have expected manual transmission vehicles to have* *better gas mileage than automatic. However, based on what I know of this data set, there are some exotic sports cars (Lotus, Ferrari, Maserati with manual transmissions) that may not be representative of the overall population. Nonetheless, this is the data set we have to work with.*

Model Selection: Adding cylinder

```r
model3 <- lm(mpg ~ am + factor(cyl), mtcars)
pander(summary(model3), justify = lm_justification)
```

	Estimate	Std. Error	t value	Pr(>|t|)
am	2.560	1.298	1.973	0.0585
factor(cyl)6	-6.156	1.536	-4.009	0.0004
factor(cyl)8	-10.068	1.452	-6.933	0.0000
(Intercept)	24.802	1.323	18.752	0.0000

Observations	Residual Std. Error	R^2	Adjusted R^2
32	3.074	0.7651	0.7399

Fitting linear model: mpg ~ am + factor(cyl)

```r
pander(anova(basemodel, model3))
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
30	720.9	NA	NA	NA	NA
28	264.5	2	456.4	24.16	0

Analysis of Variance Table

Renders transmission type not significant. Not helpful.

Model Selection: Adding horsepower

```r
model3 <- lm(mpg ~ am + hp, mtcars)
pander(summary(model3), justify = lm_justification)
```

	Estimate	Std. Error	t value	Pr(>|t|)
am	5.2771	1.0795	4.888	0
hp	-0.0589	0.0079	-7.495	0
(Intercept)	26.5849	1.4251	18.655	0

Observations	Residual Std. Error	R^2	Adjusted R^2
32	2.909	0.782	0.767

Fitting linear model: mpg ~ am + hp

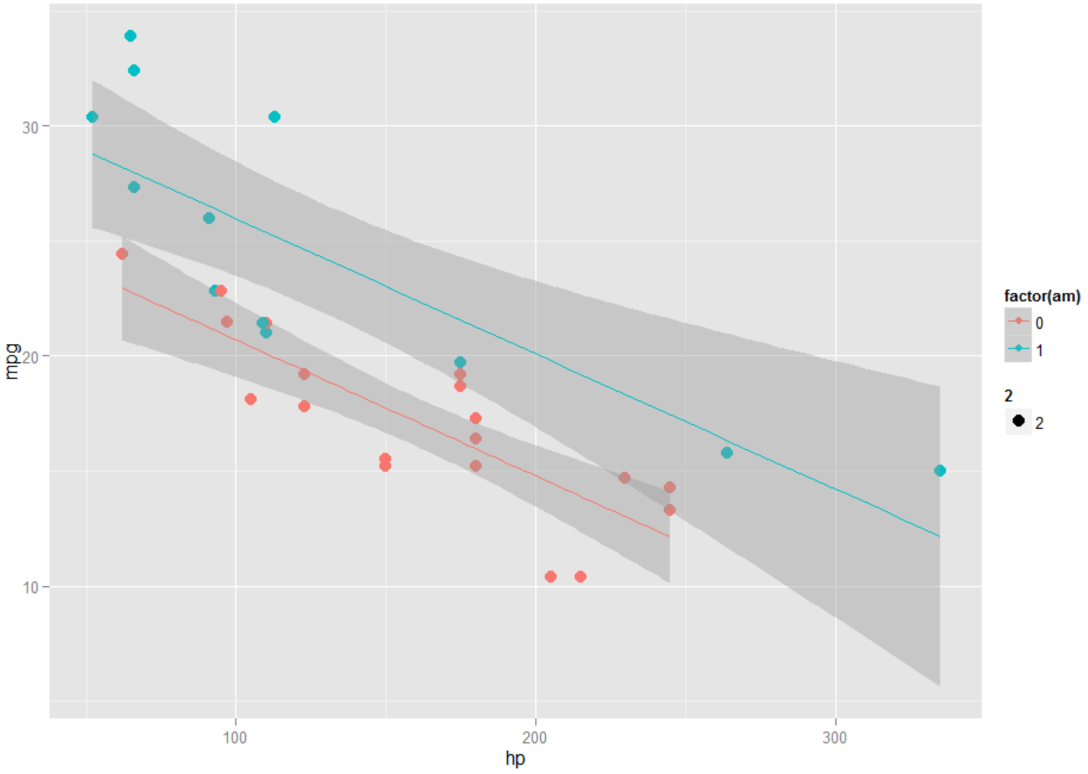
```r
pander(anova(basemodel, model3))
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
30	720.9	NA	NA	NA	NA
29	245.4	1	475.5	56.18	0

Analysis of Variance Table

Major improvement. R^2 goes from 0.3598 to 0.78203 compared with the base model.

```
qplot(x=hp, y=mpg, color=factor(am), data=mtcars, size=2) +
   stat_smooth(method="lm", size=0.5)
```



Model Selection: Adding transmission and horsepower interaction term

```
model4 <- lm(mpg ~ am * hp, mtcars)
pander(summary(model4), justify = lm_justification)
```

	Estimate	Std. Error	t value	Pr(>|t|)
am	5.2177	2.6651	1.9578	0.0603
hp	-0.0591	0.0129	-4.5684	0.0001
am:hp	0.0004	0.0165	0.0245	0.9806
(Intercept)	26.6248	2.1829	12.1968	0.0000

Observations	Residual Std. Error	R^2	Adjusted R^2
32	2.961	0.782	0.7587

*Fitting linear model: mpg ~ am * hp*

```
pander(anova(model3, model4))
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
29	245.4	NA	NA	NA	NA
28	245.4	1	0.0053	0.0006	0.9806

Analysis of Variance Table

Not an improvement. R^2 goes from 0.78203 to 0.78204, which is basically unchanged. The interaction term has p-value of 0.98 (not significant), and the p-value for transmission type is 0.06 (>0.05), which is not significant at the 95% level we would like to see. Revert to model without interaction term.

Checklist

[] Did the student interpret the coefficients correctly?

[] Did the student answer the questions of interest or detail why the question(s) is (are) not answerable?

[] Did the student do a residual plot and some diagnostics?

[] Did the student quantify the uncertainty in their conclusions and/or perform an inference correctly?

[] Written as a PDF printout of a compiled (using knitr) R markdown document.

[] Brief. Roughly the equivalent of 2 pages or less for the main text. Supporting figures in an appendix can be included up to 5 total pages including the 2 for the main report. The appendix can only include figures.

[] Was the report brief (about 2 pages long) for the main body of the report and no longer than 5 with supporting appendix of figures?

[x] Includes a first paragraph executive summary.

[x] Did the student do some exploratory data analyses?

[x] Did the student fit multiple models and detail their strategy for model selection?

[x] Did the report include an executive summary?

[x] Was the report done in Rmd (knitr)?