Transmission Type and Fuel Economy in the MTCARS Data Set

Executive Summary

We analyzed the *mtcars* data set in R which is taken from the 1974 *Motor Trends* magazine and includes information on eleven variables for 32 automobiles. In particular, we attempted to establish the impact of transmission type (automatic vs. manual) on fuel economy measured in miles per gallon. Our final regression model included vehicle weight and quarter mile time as additional predictors. In the context of this data set, there appears to be an advantage of about 2.9 miles per gallon for manual over automatic transmissions. The 95% confidence interval on this value, though, is fairly wide, ranging from 0.05 to 5.8, just barely failing to include 0. It is important to note, though, that without knowing how the 32 automobiles in the data set were selected or what additional variables might be of importance in determining fuel economy, generalizing the specific impact of transmission type on fuel economy beyond this data set is most likely problematic.

Exploratory Data Analysis

Box plots for the seven variables with more than three values are show in Figure A1 in the Appendix while frequencies for the four binary or trinary variables are shown in Figure A2. The Maserati Bora appears to be an outlier with respect to horse power and number of carburetors while the Mercedes 230 has a notably longer quarter mile time. We retain these two automobiles as they are, but note that a separate analysis indicated that removing both cars from the data set had only a small effect on the final model estimates.

Correlations were computed between all pairs of variables. The resulting correlation matrix is shown in Figure A3. It is clear that there are a number of large correlations that will have to be considered in any regression analysis.

Scatterplots of mpg versus each of the other variables are shown in Figure A4. Although we do not consider quadratric terms in the current regression analysis, the curvature in the scatterplots for displacement and horsepower suggest that such terms might be worth considering in a future analysis.

Regression Analysis

We first fit the single-predictor model regressing mpg on only transmission type (am).

```
fit<-lm(mpg~am,data=mtcars)
```

	Estimate	Std.Error	t value	p value	2.5%	97.5%
(Intercept)	17.1	1.1	15.2	0.00000	14.9	19
am	7.2	1.8	4.1	0.00029	3.6	11

```
## [1] R-squared =  0.36       Adj.R-sq =  0.338       Res.Std.Error =  4.902
```

The coefficient for transmission type in this model is 7.2, suggesting a 7.2 miles-per-gallon advantage of a manual transmission (coded as 1) over an automatic transmission (coded as 0). This model, though, controls for no other variables. We next fit the model with all ten predictors.

```
fit10<-lm(mpg~cyl+disp+hp+drat+wt+qsec+vs+am+gear+carb,data=mtcars)
```

.

```
##              Int   cyl  disp     hp drat     wt qsec   vs   am gear  carb
## Estimate   12.30 -0.11 0.013 -0.021 0.79 -3.715 0.82 0.32 2.52 0.66 -0.20
## Std. Error 18.72  1.05 0.018  0.022 1.64  1.894 0.73 2.10 2.06 1.49  0.83
## t value     0.66 -0.11 0.747 -0.987 0.48 -1.961 1.12 0.15 1.23 0.44 -0.24
## Pr(>|t|)    0.52  0.92 0.463  0.335 0.64  0.063 0.27 0.88 0.23 0.67  0.81
```

```
## [1] R-squared =  0.869       Adj.R-sq =  0.807       Res.Std.Error =  2.65
```

The R-squared for the full model is 0.87, much higher than the 0.36 when transmission type is the only predictor in the model. Note, though, that the coefficient for transmission type (am) has dropped from 7.2 to 2.5 and is no longer statistically significantly different from 0.

Variance inflation factors for this model are shown below. It is clear that a number of the values are quite high. This is not suprising given the considerable number of large correlations among the variables.

```
##  cyl disp   hp drat   wt qsec   vs   am gear carb
## 15.4 21.6  9.8  3.4 15.2  7.5  5.0  4.6  5.4  7.9
```

We would like to reduce the number of predictors while still accounting for as much of the variance in MPG as we can. Using the *regsubsets* function in the *leaps* package, all subsets regression was run. With 10 potential predictors, there are 2^10=1024 possible models, half of which will include transmission type. The "best" model emerging from this analysis includes transmission type (am), weight (wt), and quarter mile time (qsec) as predictors (see Figure A5). We therefore fit the model with only these three predictors.

```
fit3<-lm(mpg~am+wt+qsec,data=mtcars)
```

	Estimate	Std.Error	t value	p value	2.5%	97.5%
(Intercept)	9.6	6.96	1.4	1.8e-01	-4.638	23.9
am	2.9	1.41	2.1	4.7e-02	0.046	5.8
wt	-3.9	0.71	-5.5	7.0e-06	-5.373	-2.5
qsec	1.2	0.29	4.2	2.2e-04	0.635	1.8

```
## [1] R-squared =  0.85       Adj.R-sq =  0.834       Res.Std.Error =  2.459
```

The coefficient for transmission type is 2.9, suggesting a 2.9 miles-per-gallon advantage of manual over automatic transmissions. The 95% confidence interval for the coefficient is 0.05 to 5.8 and the p value testing the null hypothesis of a zero coefficient is $< .05$. The R-square value of 0.85 is nearly as large as that for the full ten-predictor model and the adjusted R-square value is slightly higher than for the full model.

An ANOVA test was run comparing the ten-predictor model to the three-predictor model (using *anova(fit3,fit10)*.) The F value with 7 and 21 degrees of freedom was 0.44, yielding a p value of 0.86. These results suggest that adding the seven variables to the three-predictor model contributes very little.

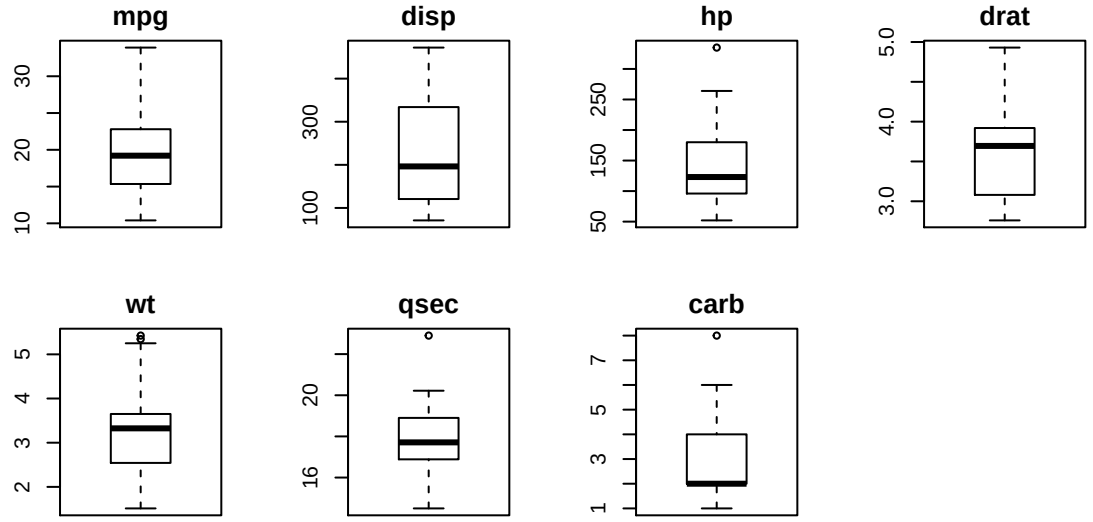
Residuals and Diagnostics

Residual plots for the three-predictor model are presented in Figure A6. Nothing in these plots points to notable problems with our final model.

Note: The full R markdown document is available at github.com/kbiolsi/RegressionProject/RegProject.Rmd

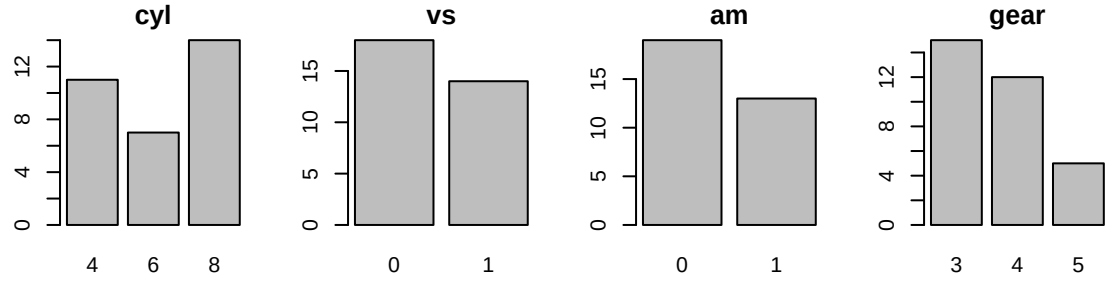
Appendix

Figure A1. Box plots for the seven numeric variables that take on at least four values.



.
.

Figure A2. Bar plots for the four binary or trinary variables.



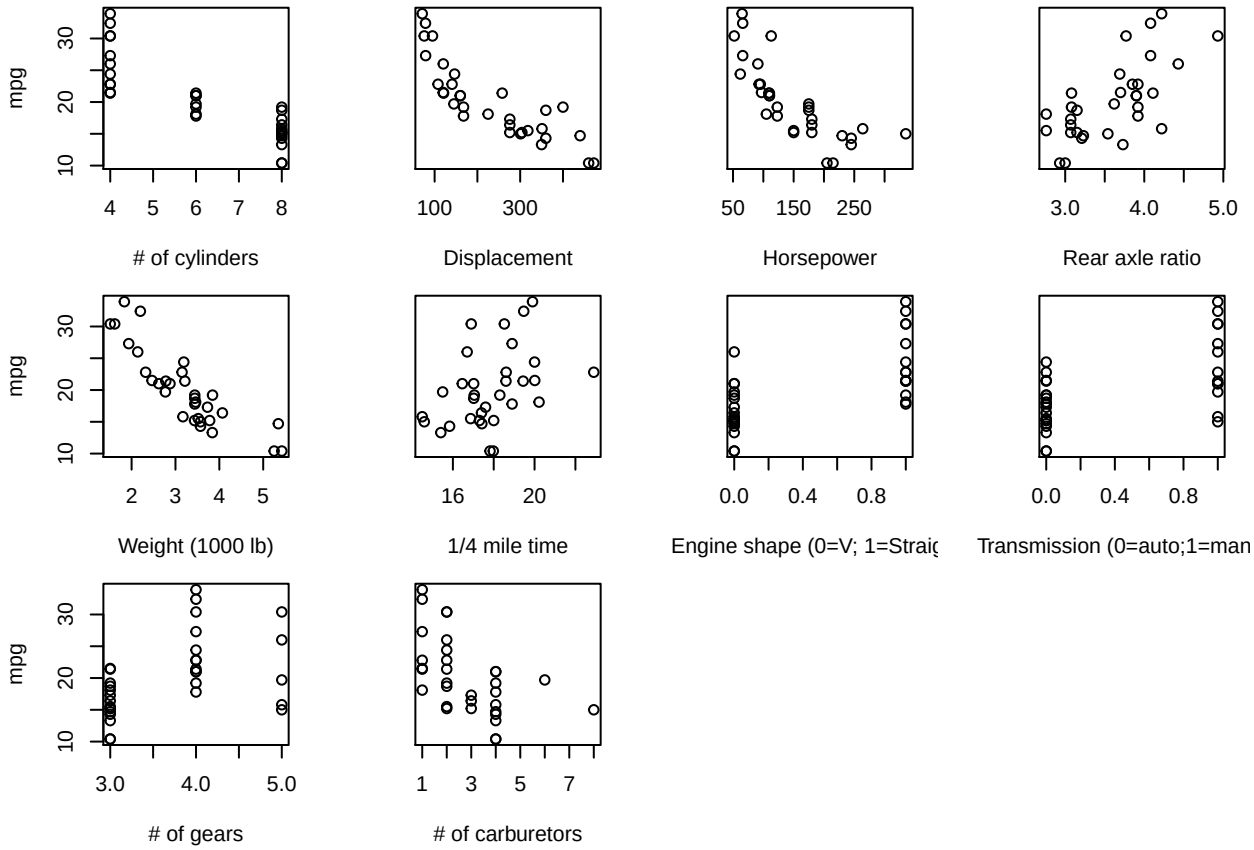
.
.

Figure A3. Correlation matrix for all pairs of variables.

```
##        mpg   cyl  disp    hp   drat    wt   qsec    vs     am  gear   carb
## mpg   1.00 -0.85 -0.85 -0.78  0.681 -0.87  0.419  0.66  0.600  0.48 -0.551
## cyl  -0.85  1.00  0.90  0.83 -0.700  0.78 -0.591 -0.81 -0.523 -0.49  0.527
## disp -0.85  0.90  1.00  0.79 -0.710  0.89 -0.434 -0.71 -0.591 -0.56  0.395
## hp   -0.78  0.83  0.79  1.00 -0.449  0.66 -0.708 -0.72 -0.243 -0.13  0.750
## drat  0.68 -0.70 -0.71 -0.45  1.000 -0.71  0.091  0.44  0.713  0.70 -0.091
## wt   -0.87  0.78  0.89  0.66 -0.712  1.00 -0.175 -0.55 -0.692 -0.58  0.428
## qsec  0.42 -0.59 -0.43 -0.71  0.091 -0.17  1.000  0.74 -0.230 -0.21 -0.656
## vs    0.66 -0.81 -0.71 -0.72  0.440 -0.55  0.745  1.00  0.168  0.21 -0.570
## am    0.60 -0.52 -0.59 -0.24  0.713 -0.69 -0.230  0.17  1.000  0.79  0.058
## gear  0.48 -0.49 -0.56 -0.13  0.700 -0.58 -0.213  0.21  0.794  1.00  0.274
## carb -0.55  0.53  0.39  0.75 -0.091  0.43 -0.656 -0.57  0.058  0.27  1.000
```

.
.
.

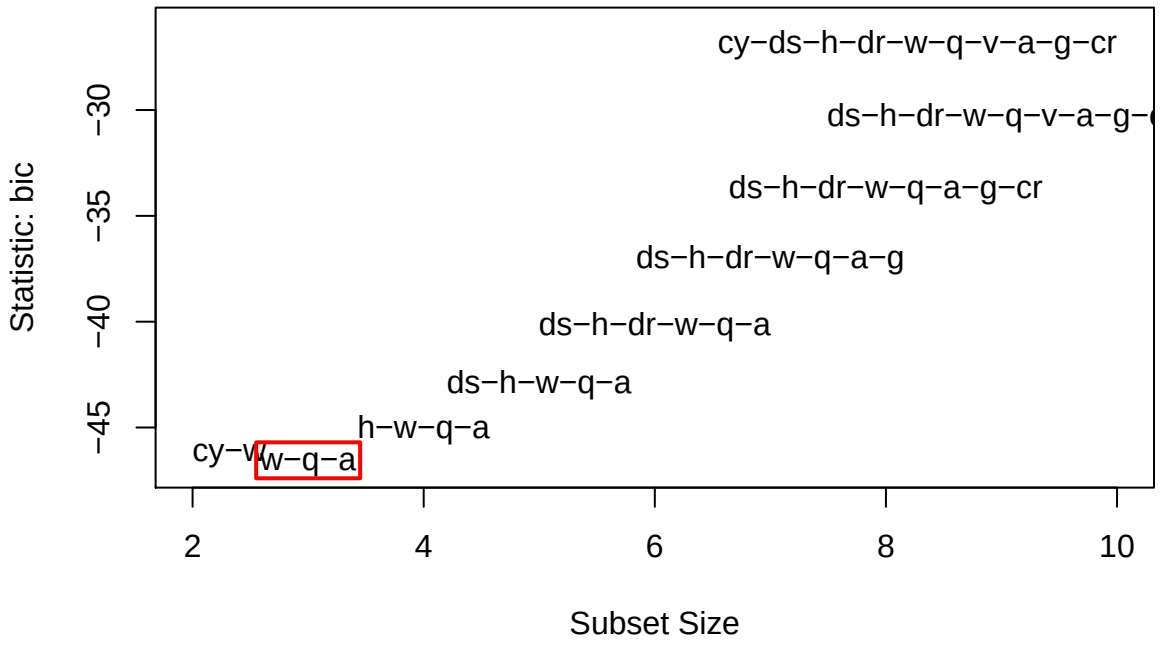
Figure A4. Scatterplots of MPG versus the ten potential predictors.



.

.

Figure A5. All subsets regression: models of each size with smallest Bayesian Information Criterion (BIC) value



.

.
.

Figure A6. Residual plots for three-predictor model.