Differences in Fuel Efficiency Between Automatic and Manual Transmission Vehicles in 1974

Wayne Witzke

Synopsis

We attempt to establish whether an automatic or manual transmission provides better fuel efficiency, and quantify that fuel savings if a difference exists, using vehicle data collected by Motor Trend magazine for 1973-1974 models. A minimal adequate linear regression model is fitted against the data, including transmission type as a factor variable to attempt to isolate the transmission contribution to fuel efficiency. No significant contribution is found.

Preliminaries

This section prepares the analysis to run. For more information on replicating this analysis, see the appendix.

```
library(ggplot2); library(GGally);
data(mtcars);
options(digits=4, width=125);
```

Exploratory Analysis

Summary of Data

This analysis uses the `mtcars` data set included natively with most R distributions. This data includes 11 aspects of automobile design and performance for 32 different models of automobile from 1973 and 1974. These aspects include: miles per gallon (`mpg`); the number of cylinders (`cyl`, either 4, 6, or 8); the displacement (`disp`, in cubic inches); the gross horsepower (`hp`); the rear axle ratio (`drat`); the weight (`wt`, in 1000's of pounds); the quarter-mile time (`qsec`); the engine type (`vs`, either v-engine or straight engine); the transmission type (`am`, automatic or manual); the number of gears (`gear`, either 3, 4 or 5); and the number of carburetors (`carb`, either 1, 2, 3, 4, 6, or 8). The structure of the data set is:

```
dim(mtcars);
sapply(mtcars,class);
```

```
## [1] 32 11
##       mpg       cyl      disp        hp      drat        wt      qsec        vs        am      gear      carb
## "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
```

Exploratory Graphs

We are primarily interested in characterizing the relationship between miles per gallon and transmission type. However, there is correlation between these variables and other variables in the dataset. We can see this with a pairs plot. We can get a better overall characterization of the data by tidying it, specifically by ensuring that factor variables are properly expressed.

```
mtc = mtcars; mtc$cyl = factor(mtc$cyl); mtc$gear = factor(mtc$gear);
mtc$carb = factor(mtc$carb);
mtc$am = factor(mtc$am, labels = c("auto", "man"));
mtc$vs = factor(mtc$vs, labels = c("v", "str"));
```

Figure 1 in the appendix shows the pairs plot with this corrected data, using box plots and faceted density plots to show relationships between continuous and discrete variables, and faceted bar plots to show relationships between factors.

From this pair plot, we can see that fitting a simple, single-variable, meaningful linear regression is likely just not possible. That is, `mpg` appears to correlate with many of the variables available. In addition, it appears very likely that there are confounding variables. For instance, the pairs `mpg/wt`, `mpg/disp`, and `wt/disp` are all strongly correlated.

Regression Modeling

Because of the strong correlations between `mpg` and the other variables, and because of the strong possibility of confounding, we use a backward selection strategy with `anova` to attempt to find a multivariable regression model that isolates the impact of transmission on mpg. This involves removing variables from a complete model until we have found the minimum adequate fit, using both F-test p-values and Akaike information criterion (`AIC`) for selection. Interactions are also tested. The details of and code for this procedure can be found in Figure 2 in the appendix.

Once run, the procedure selects `mpg ~ am + wt + hp + cyl` as the best model. The details of this model, including coefficients, confidence intervals, leverage and influence tests, can be seen here.

```r
mtc.fit = lm(mpg ~ am + wt + hp + cyl, mtc);
summary(mtc.fit)$coef;
confint(mtc.fit);
head(hatvalues(mtc.fit)[order(hatvalues(mtc.fit),decreasing=TRUE)],6);
head(dffits(mtc.fit)[order(abs(dffits(mtc.fit)),decreasing=TRUE)],6);
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 33.70832    2.60489 12.9404 7.733e-13
## amman        1.80921    1.39630  1.2957 2.065e-01
## wt          -2.49683    0.88559 -2.8194 9.081e-03
## hp          -0.03211    0.01369 -2.3450 2.693e-02
## cyl6        -3.03134    1.40728 -2.1540 4.068e-02
## cyl8        -2.16368    2.28425 -0.9472 3.523e-01
##                2.5 %    97.5 %
## (Intercept) 28.35390 39.062744
## amman       -1.06093  4.679356
## wt          -4.31718 -0.676478
## hp          -0.06025 -0.003964
## cyl6        -5.92406 -0.138632
## cyl8        -6.85902  2.531671
##       Maserati Bora Lincoln Continental       Toyota Corona   Chrysler Imperial       Mazda RX4 Wag  Cadillac Fleetwood
##              0.4714              0.2937              0.2778              0.2611              0.2496              0.2496
## Chrysler Imperial    Toyota Corolla     Toyota Corona          Fiat 128        Volvo 142E     Maserati Bora
##            1.1759            0.9378           -0.9094            0.8370           -0.7683            0.7033
```

Residual disgnostic plots can be seen in Figure 3 in the appendix.

Conclusions

The quality of the best linear regression fit is fairly good. The residual plots (Figure 3) do not show any marked deviations from what might be expected from a good model, and examining the highest leverage and influence points does not reveal problematic outliers. Most of the p-values from the model satisfy $\alpha = 0.05$, indicating that they are likely significant contributing variables in the analysis. The coefficients seem reasonable as well. That is, it is reasonable that for every 1000 pound increase in weight, you lose about 2.5 miles per gallon, or that as horsepower increases by 1, you lose about 0.03 miles per gallon.

Unfortunately, the transmission type was *not* one of the coefficients that appeared to contribute significantly to the model. At $p \approx 0.2$, it fails to reject the null hypothesis that a model including transmission type is identical to a model that does not include it. Either there is no contribution from transmission or there is not enough data to detect an existing significant contribution. This makes quantifying any such contribution impossible with this data set.

Appendix

Figure 1: Pairs Plot

```r
tpplot = ggpairs(mtc[, c("qsec","drat","hp","disp","wt","mpg","am","cyl","vs","gear","carb")],
    lower=list(continuous=wrap("smooth", size=0.2),combo="facetdensity",discrete="blank"),
    upper=list(continuous=wrap("cor", size=2.5, color="black"),combo="box",discrete="facetbar"), axisLabels="none");
suppressMessages(print(tpplot, left = 0.75, bottom = 0.75));
```

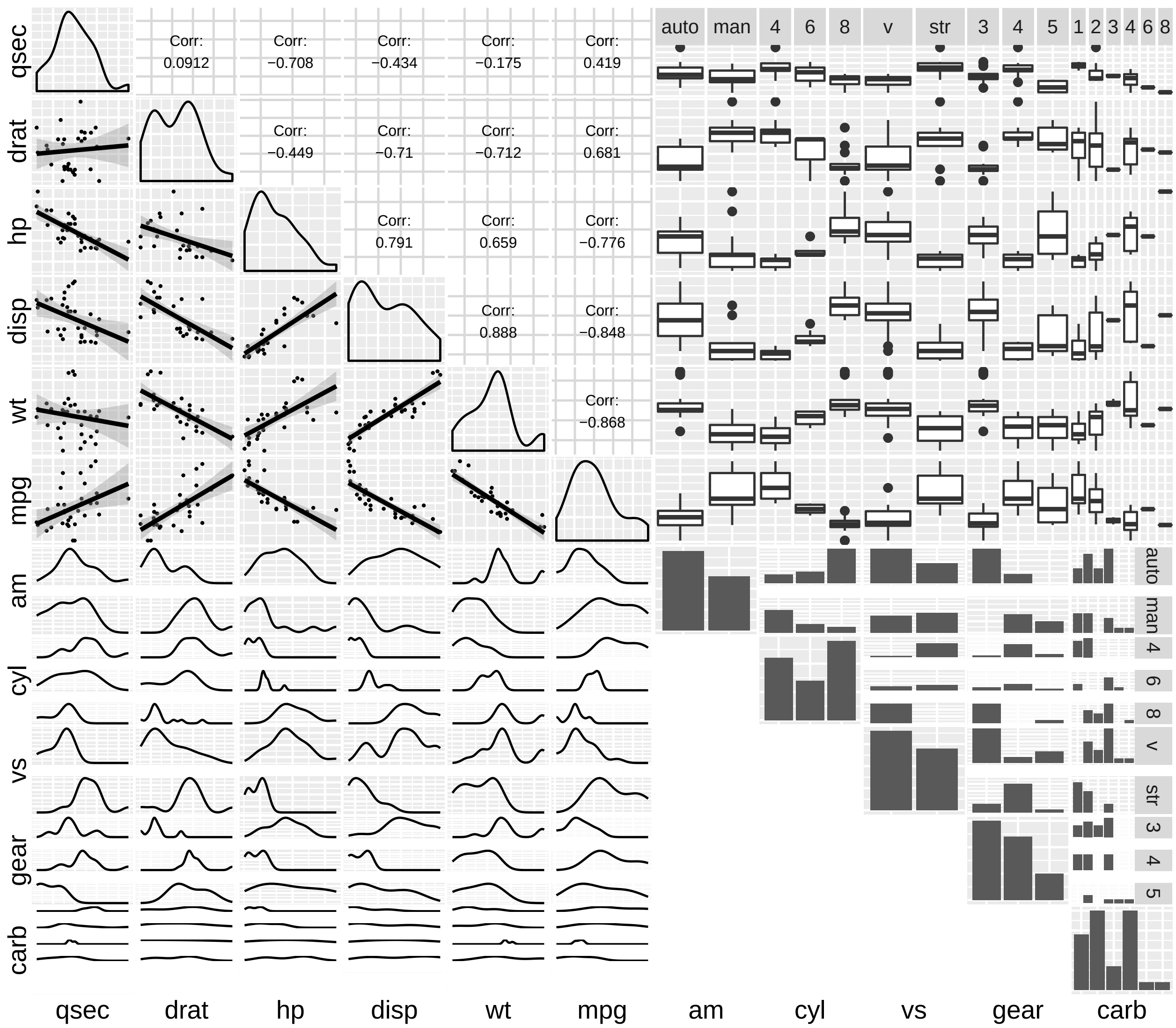


Figure 2: Backward Selection + Anova to Find Minimal Adequate Fit

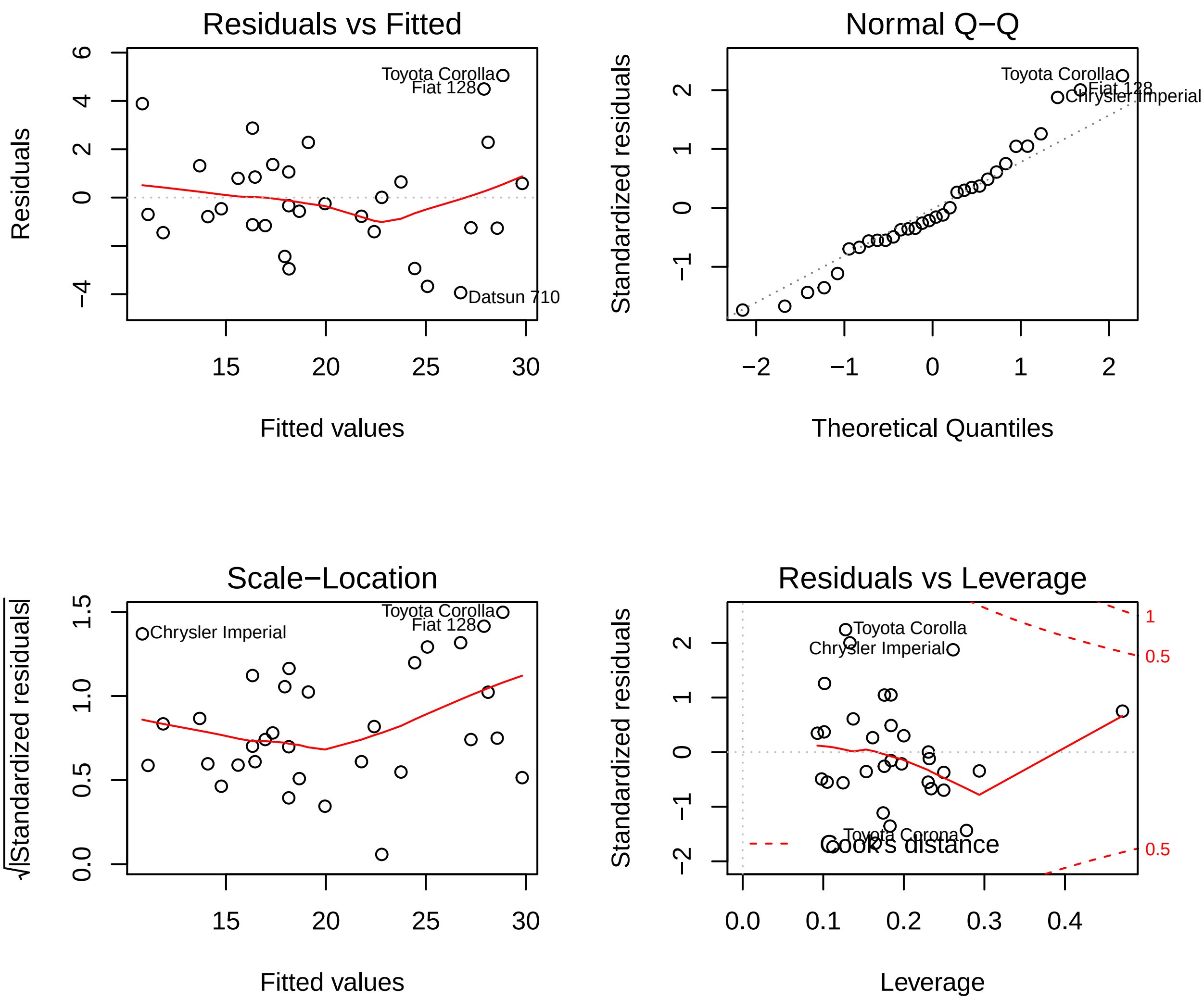
```r
fit = list(lm(mpg ~ am+wt+disp+hp+drat+qsec+cyl+vs+gear+carb, mtc)); # Start with full model.
fit[[2]] = update(fit[[1]], ~.-wt); fit[[3]] = update(fit[[1]], ~.-disp); fit[[4]] = update(fit[[1]], ~.-hp);
fit[[5]] = update(fit[[1]], ~.-drat); fit[[6]] = update(fit[[1]], ~.-qsec); fit[[7]] = update(fit[[1]], ~.-cyl);
fit[[8]] = update(fit[[1]], ~.-vs); fit[[9]] = update(fit[[1]], ~.-gear); fit[[10]] = update(fit[[1]], ~.-carb);
round1 = rbind(sapply(fit[1:10], function(n) anova(fit[[1]],n)[,"Pr(>F)"][2]),sapply(fit[1:10], AIC));
colnames(round1) = c("base","wt","disp","hp","drat","qsec","cyl","vs","gear","carb");
round1[, order(round1[2,])]; # Select "carb" to remove for high p-value and low AIC.
fit = list(lm(mpg ~ am+wt+disp+hp+drat+qsec+cyl+vs+gear, mtc)); # Start with full model.
fit[[2]] = update(fit[[1]], ~.-wt); fit[[3]] = update(fit[[1]], ~.-disp); fit[[4]] = update(fit[[1]], ~.-hp);
fit[[5]] = update(fit[[1]], ~.-drat); fit[[6]] = update(fit[[1]], ~.-qsec); fit[[7]] = update(fit[[1]], ~.-cyl);
```

```r
fit[[8]] = update(fit[[1]], ~.-vs); fit[[9]] = update(fit[[1]], ~.-gear);
round2 = rbind(sapply(fit[1:9], function(n) anova(fit[[1]],n)[,"Pr(>F)"][2]),sapply(fit[1:9], AIC));
colnames(round2) = c("base","wt","disp","hp","drat","qsec","cyl","vs","gear");
round2[, order(round2[2,])]; # Select "gear" to remove for high p-value and low AIC.
fit = list(lm(mpg ~ am+wt+disp+hp+drat+qsec+cyl+vs, mtc)); # Start with full model.
fit[[2]] = update(fit[[1]], ~.-wt); fit[[3]] = update(fit[[1]], ~.-disp); fit[[4]] = update(fit[[1]], ~.-hp);
fit[[5]] = update(fit[[1]], ~.-drat); fit[[6]] = update(fit[[1]], ~.-qsec); fit[[7]] = update(fit[[1]], ~.-cyl);
fit[[8]] = update(fit[[1]], ~.-vs);
round2 = rbind(sapply(fit[1:8], function(n) anova(fit[[1]],n)[,"Pr(>F)"][2]),sapply(fit[1:8], AIC));
colnames(round2) = c("base","wt","disp","hp","drat","qsec","cyl","vs");
round2[, order(round2[2,])]; # Select "drat" to remove for high p-value and low AIC.
fit = list(lm(mpg ~ am+wt+disp+hp+qsec+cyl+vs, mtc)); # Start with full model.
fit[[2]] = update(fit[[1]], ~.-wt); fit[[3]] = update(fit[[1]], ~.-disp); fit[[4]] = update(fit[[1]], ~.-hp);
fit[[5]] = update(fit[[1]], ~.-qsec); fit[[6]] = update(fit[[1]], ~.-cyl); fit[[7]] = update(fit[[1]], ~.-vs);
round2 = rbind(sapply(fit[1:7], function(n) anova(fit[[1]],n)[,"Pr(>F)"][2]),sapply(fit[1:7], AIC));
colnames(round2) = c("base","wt","disp","hp","qsec","cyl","vs");
round2[, order(round2[2,])]; # Select "disp" to remove for high p-value and low AIC.
fit = list(lm(mpg ~ am+wt+hp+qsec+cyl+vs, mtc)); # Start with full model.
fit[[2]] = update(fit[[1]], ~.-wt); fit[[3]] = update(fit[[1]], ~.-hp); fit[[4]] = update(fit[[1]], ~.-qsec);
fit[[5]] = update(fit[[1]], ~.-cyl); fit[[6]] = update(fit[[1]], ~.-vs);
round2 = rbind(sapply(fit[1:6], function(n) anova(fit[[1]],n)[,"Pr(>F)"][2]),sapply(fit[1:6], AIC));
colnames(round2) = c("base","wt","hp","qsec","cyl","vs");
round2[, order(round2[2,])]; # Select "qsec" to remove for high p-value and low AIC.
fit = list(lm(mpg ~ am+wt+hp+cyl+vs, mtc)); # Start with full model.
fit[[2]] = update(fit[[1]], ~.-wt); fit[[3]] = update(fit[[1]], ~.-hp); fit[[4]] = update(fit[[1]], ~.-cyl);
fit[[5]] = update(fit[[1]], ~.-vs);
round2 = rbind(sapply(fit[1:5], function(n) anova(fit[[1]],n)[,"Pr(>F)"][2]),sapply(fit[1:5], AIC));
colnames(round2) = c("base","wt","hp","cyl","vs");
round2[, order(round2[2,])]; # Select "vs" to remove for high p-value and low AIC.
fit = list(lm(mpg ~ am+wt+hp+cyl, mtc)); # Start with full model.
fit[[2]] = update(fit[[1]], ~.-wt); fit[[3]] = update(fit[[1]], ~.-hp); fit[[4]] = update(fit[[1]], ~.-cyl);
round2 = rbind(sapply(fit[1:4], function(n) anova(fit[[1]],n)[,"Pr(>F)"][2]),sapply(fit[1:4], AIC));
colnames(round2) = c("base","wt","hp","cyl");
round2[, order(round2[2,])]; # Additional removals will make the fit worse. So, done.
fit = lm(mpg ~ am + wt + hp + cyl, mtc);
fitI2 = update(fit, ~.^2);
fitI3 = update(fit, ~.^3);
fitI4 = update(fit, ~.^4);
anova(fit, fitI2)[,"Pr(>F)"][2]; # Checking for interactions (i.e. "am:wt").
anova(fit, fitI3)[,"Pr(>F)"][2]; # More interactions.
anova(fit, fitI4)[,"Pr(>F)"][2]; # Last interaction ("am:wt:hp:cyl"). Note no significant difference when including interactions.
```

```
##          carb     gear     qsec     drat      cyl       vs  base     disp        wt        hp
## [1,]   0.8814   0.7839   0.6997   0.6407   0.5211   0.5115    NA   0.2827   0.09462   0.09393
## [2,] 162.6398 166.2543 167.5437 167.6958 167.9963 168.1659 169.2 169.7605 173.37444 173.40019
##          gear     disp     drat       vs      cyl    qsec  base       wt        hp
## [1,]   0.6922   0.7043   0.6785   0.4675   0.4081   0.386    NA   0.1011   0.07998
## [2,] 159.8170 160.8761 160.9216 161.5058 161.5077 161.873 162.6 165.0490 165.66657
##          drat      cyl     disp       vs     qsec  base      hp        wt
## [1,]   0.6994   0.4514   0.6255   0.5627   0.4564    NA   0.104   0.03691
## [2,] 158.0388 158.1308 158.1714 158.3155 158.6424 159.8 161.749 164.29407
##         disp       vs      cyl     qsec base        hp        wt
## [1,]   0.655   0.5413   0.3789   0.4826   NA    0.1013   0.03458
## [2,] 156.323 156.5694 156.7394 156.7397  158 159.8565 162.38656
##          qsec       vs      cyl  base        hp        wt
## [1,]   0.5256   0.5012   0.2269    NA   0.09153   0.01586
## [2,] 154.8713 154.9385 156.2776 156.3 158.19821 162.23959
##           vs  base      cyl        hp        wt
## [1,]   0.269    NA   0.1318   0.01871   0.01302
## [2,] 154.467 154.9 156.0584 160.08735 160.91977
##       base   cyl        hp        wt
## [1,]    NA   0.1   0.02693 9.081e-03
## [2,] 154.5 156.1 158.60654 1.610e+02
## [1] 0.5496
## [1] 0.6211
## [1] 0.6211
```

Figure 3: Residual Diagnostics

```r
par(mfrow=c(2,2));
plot(mtc.fit);
```




```r
par(mfrow=c(1,1));
```

Figure 4: System Information

This analysis was performed using the hardware and software specified in this section.

```r
sessionInfo();
```

```
## R version 3.2.5 (2016-04-14)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 15.10
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C               LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8    LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C             LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] GGally_1.0.1    ggplot2_2.1.0   rmarkdown_0.9.5
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.4      codetools_0.2-14 reshape_0.8.5    digest_0.6.9     plyr_1.8.3       grid_3.2.5       gtable_0.2.0
##  [8] formatR_1.3      magrittr_1.5     evaluate_0.8.3   scales_0.4.0     stringi_1.0-1    reshape2_1.4.1   labeling_0.3
## [15] tools_3.2.5      stringr_1.0.0    munsell_0.4.3    yaml_2.1.13      colorspace_1.2-6 htmltools_0.3.5  knitr_1.12.3
```