Maximizing MPG: Automatic or Manual?

Kairsten Fay

2/5/2017

Executive Summary

Motor Trend magazine has recently been interested in vehicles' fuel consumption. The objective of this analysis was to answer the following questions about a car's miles per gallon (MPG):

- Is an automatic or manual transmission better for MPG?
- What is the quantified MPG difference between automatic and manual transmissions?

At the 95% confidence level, there was not enough evidence to conclude that there was any signficant difference in the MPG of manual versus automatic cars. The selected linear model showed an average 2.80 MPG increase in manual versus automatic cars when all other variables were held constant, but the standard error (1.92) was high enough such that the null hypothesis (mean = 0) was within our inferential range.

Exploration and Model Building

I began by exploring the `mtcars` dataset and took a closer look at the `am` variable.

```
str(mtcars$am)
```

```
##  num [1:32] 1 1 1 0 0 0 0 0 0 0 ...
```

The input variable of interest was `am`, a binary, "dummy" variable coded 0 for automatic transmission and 1 for manual transmission. The output variable was `mpg`. I explored the relationship by building a simple graph of `mpg` versus `am` and by using `ggplot2`'s `smooth` function (Appendix i). At first glance, it appeared that manual cars had higher MPG than automatic ones. However, because this model only included one predictor, I knew I must be careful of any bias introduced by neglecting important variables that also influence MPG. I cautiously created a linear model using this simplest scenario and named it `fit1`.

```
fit1 <- lm(mpg~am, data=mtcars)
```

Then, I turned to multivariable regression, which adjust regression estimates with respect to other variables. In multivariable regression, the effect of other variables has been removed from the predictor and the response, reducing bias. The interpretation of a multivariate regression coefficient is the expected change in the response per unit change in the regressor, holding all other variables fixed. I used the `corrgram` package to visualize which variables were strongly correlated with the input variable, `am` (Appendix ii). The variable with the strongest correlation with `am` was `gear`.

I created a linear model named, `fit2`, that included all of the available input variables in `mtcars` except for `gear`, which was the variable most strongly correlated to `am`. Then, I defined a third linear model, `fit3`, which contained all of the input variables as predictors of `mpg`.

```
fit2 <- lm(mpg~cyl+disp+hp+drat+wt+qsec+vs+factor(am)+carb, data=mtcars)
fit3 <- lm(mpg~., data=mtcars)
```

Analysis and Model Evaluation

Then, I looked at residuals in all of my models (Appendix iii). All of the residuals appear uncorrelated with the model fits, which is desirable. Then I measured the averages of the residuals to see if they converged at 0.

```r
mean(fit1$residuals)
```

```
## [1] -6.591949e-17
```

```r
mean(fit2$residuals)
```

```
## [1] 1.526557e-16
```

```r
mean(fit3$residuals)
```

```
## [1] 7.459311e-17
```

Indeed, the residuals were evenly distributed for all models and had a mean of 0. Because my models were nested, I performed an ANOVA by loading the models in increasing order of total variables.

```r
anova(fit1, fit2, fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + factor(am) +
##     carb
## Model 3: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     22 148.85  8    572.05 10.1809 1.031e-05 ***
## 3     21 147.49  1      1.35  0.1926    0.6652
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Only the second model, `fit2`, showed significance for all of the variables used. I also looked at variance inflation for `fit2` and `fit3`.

```r
vif <- car::vif
vif(fit2)
```

```
##        cyl       disp         hp       drat         wt       qsec
##  13.462909  21.486868   9.753363   3.355679  14.666167   7.478622
##         vs factor(am)       carb
##   4.956449   4.190894   6.484582
```

```r
vif(fit3)
```

```
##       cyl      disp        hp      drat        wt      qsec        vs
## 15.373833 21.620241  9.832037  3.374620 15.164887  7.527958  4.965873
##        am      gear      carb
##  4.648487  5.357452  7.908747
```

`fit2` had less standard error for the `am` variable than if it were orthogonal to all other regressors than did `fit3`. Therefore, `fit2` was an overall better model, and I proceeded with it for the rest of the analysis.

```r
summary(fit2)$coefficients
```

```
##                Estimate  Std. Error     t value   Pr(>|t|)
## (Intercept) 15.64180646 16.78606059  0.93183307 0.36153694
## cyl         -0.27315026  0.95981077 -0.28458762 0.77862172
## disp         0.01395085  0.01747263  0.79844026 0.43315023
## hp          -0.02062744  0.02127976 -0.96934533 0.34290994
## drat         0.84088668  1.60057435  0.52536559 0.60458429
## wt          -3.86609068  1.82849991 -2.11435104 0.04605085
```

```
## qsec         0.79507208  0.71495481  1.11205920 0.27811774
## vs           0.35800211  2.06357126  0.17348667 0.86385526
## factor(am)1  2.80344698  1.91663440  1.46269261 0.15769184
## carb        -0.04506069  0.73653480 -0.06117932 0.95176880
```

The `fit2` model estimated an expected 2.80 increase in MPG on average for manual cars than for automatic ones with all other variables held constant. The standard error for this variable was 1.92. I then calculated confidence intervals for `am`.

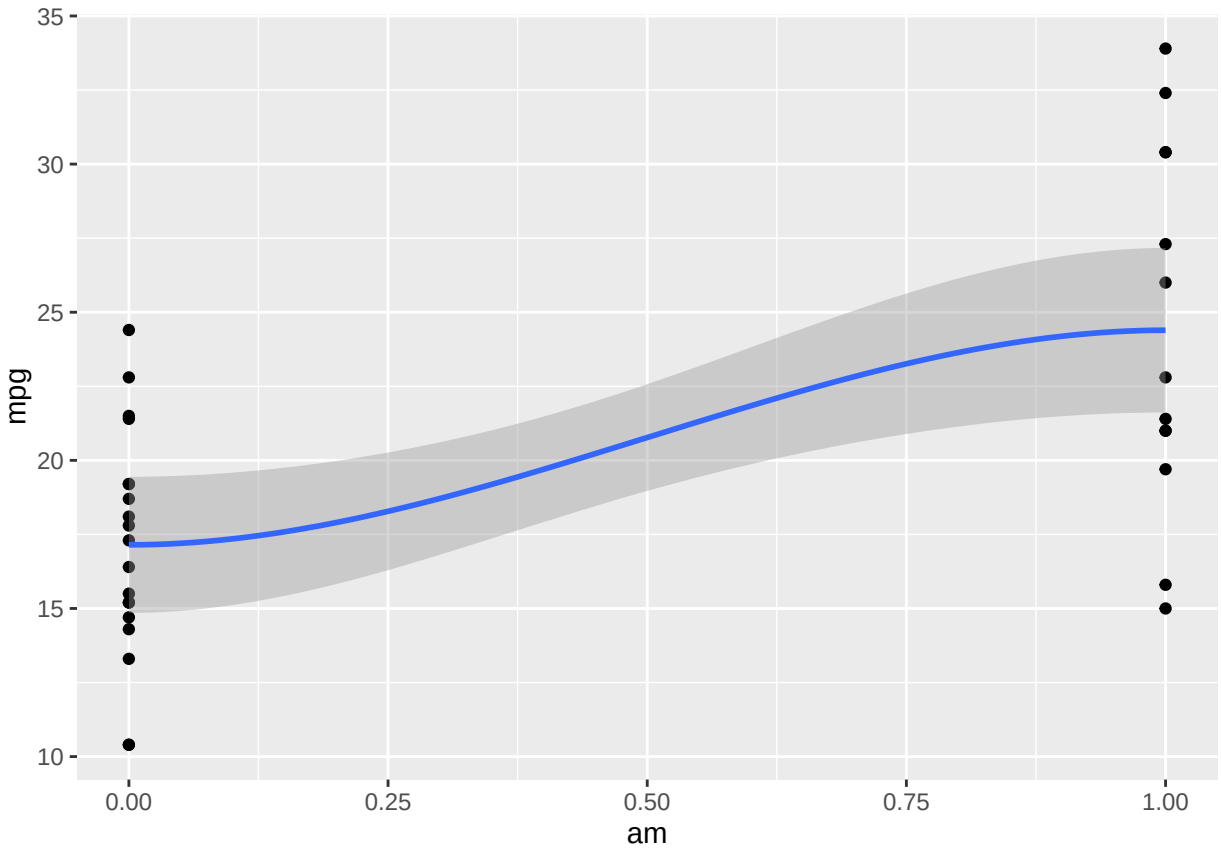
```
confint(fit2)
```

```
##                    2.5 %      97.5 %
## (Intercept) -19.17035252 50.45396543
## cyl          -2.26367597  1.71737545
## disp         -0.02228516  0.05018686
## hp           -0.06475896  0.02350409
## drat         -2.47850136  4.16027472
## wt           -7.65816739 -0.07401397
## qsec         -0.68765345  2.27779762
## vs           -3.92158274  4.63758697
## factor(am)1  -1.17140949  6.77830345
## carb         -1.57254037  1.48241898
```

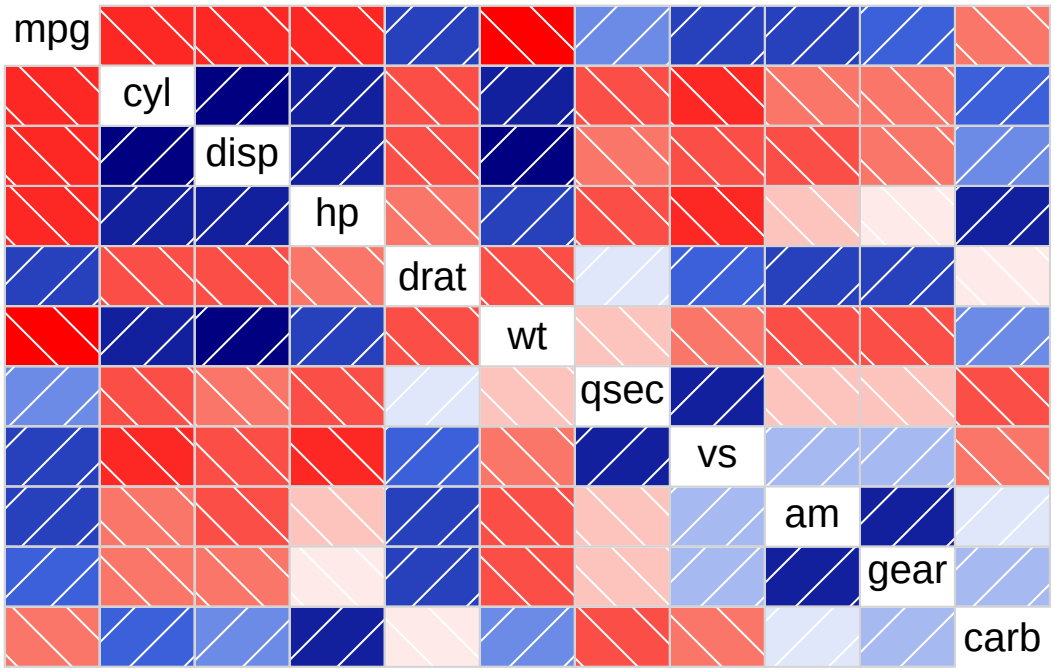
Using a 95% confidence interval, the lower end of the estimated average change in MPG from manual to automatic cars was -1.17. On the upper end, it was 6.78. Therefore, because the confidence interval captured the null hypotheis (mean difference of 0), I could not reject the null hypothesis.

Apendix

i.

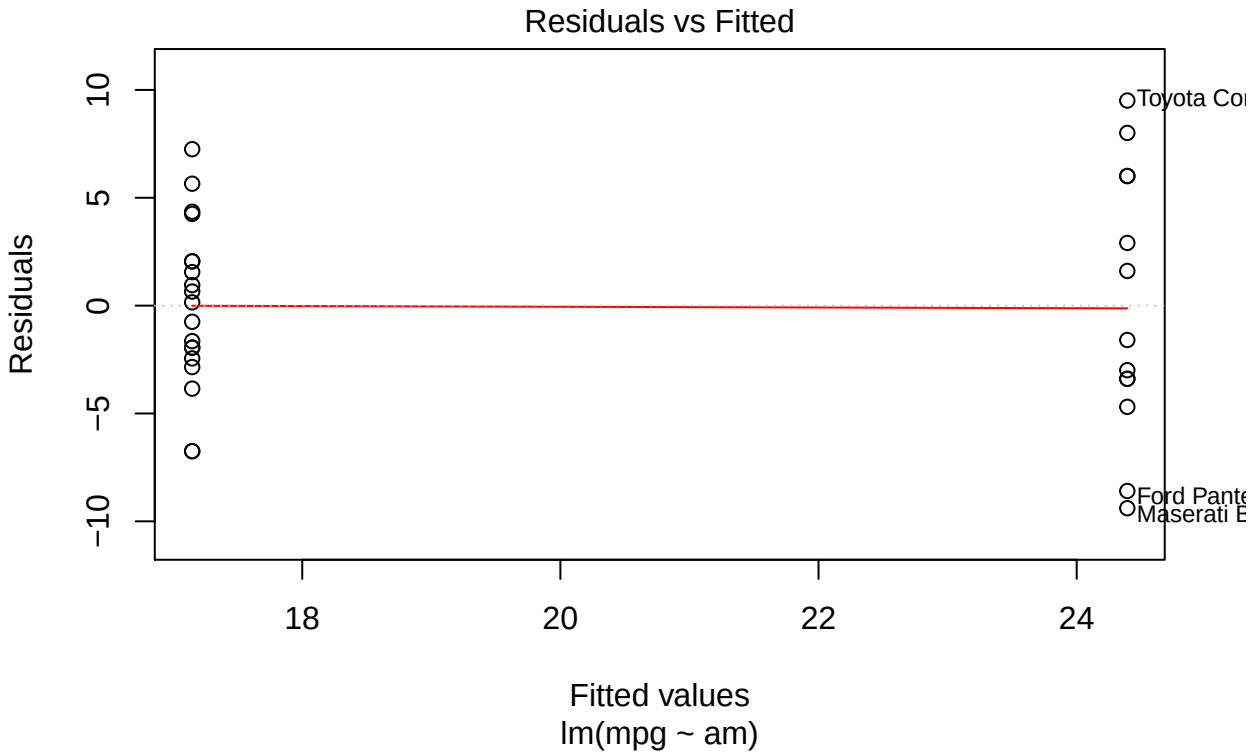

35
30
25
20
15
10
mpg
0.00
0.25
0.50
0.75
1.00
am

ii.

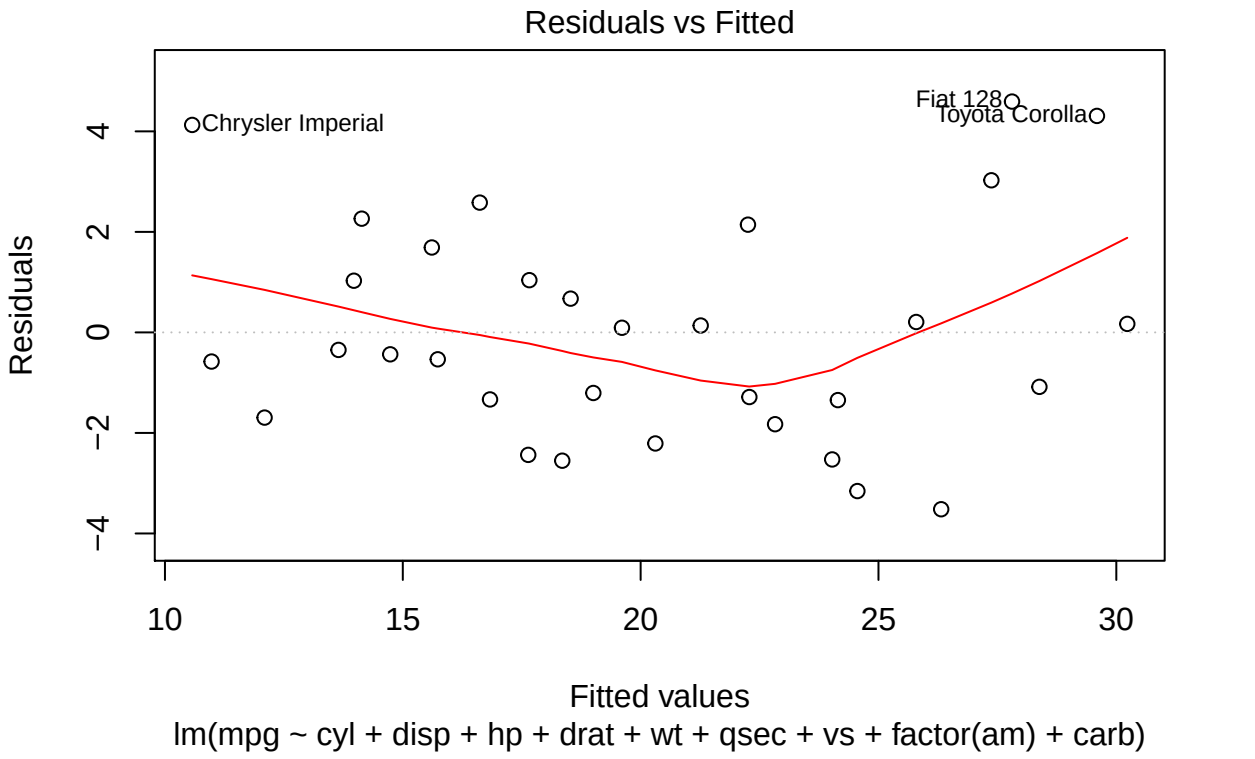

mpg
cyl
disp
hp
drat
wt
qsec
vs
am
gear
carb

iii.

```
plot(fit1, which=1)
```




```
plot(fit2, which=1)
```



```r
plot(fit3, which=1)
```