Effect of car factors on miles per gallon

Jerry Harber

August 4, 2017

Executive Summary

Using the ***mtcars*** data set that is supplied with base R, the relationship between the variable ***mpg*** (i.e., outcome) and various predictors contained in the data set are explored using regression analysis. In particular, the following questions will be answered;

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

Results indicated that the manual transmission is better for MPG if the ***mpg*** variable is the only variable used in the regression model. Estimates for the average ***mpg*** increase were 7.2 mpg. Estimates for the 95% lower/upper confidence intervals for increases in ***mpg*** for manual transmission cars were 3.6 ***mpg*** / 10.8 ***mpg*** respectively.

Data Set

The ***mtcars*** data set from the base R system is used for the analysis. The data set contains the following variables;

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

See ***Figure 1 - appendix*** for some summary statistics for this data.Although the ***mpg*** variable has a possible lower limit of zero (i.e., 0 ***mpg***), it is not practical to assume any car would have a ***mpg*** of zero. Also the summary information indicates that ***mpg*** has a minimum value of 10.4 . It is possible that the ***mpg*** variable could be "bounded" at 0. If so, a ***Poisson Regression*** might be appropriate. However, since ***mpg*** has a lower limit of 10.4 in the dataset, a ***Linear Regression*** will be used to fit a model to demonstrate the effect of ***am*** (i.e., transmission type) on ***mpg***.

Exploratory Data Analysis

Pvalues for correlations of all mtcar variables with mpg

The variable ***mpg*** is significantly correlated with all the other variables in the data set. See ***Figure 2 - appendix***.

Linear regression model results

3 models were fitted. They are;

1. fit1 = mpg~am
   The effect of ***am*** transmission type (i.e. 0 = automatic, 1 = manual) on ***mpg***. The beta value for ***am*** is 7.245 (standard error is 1.764) and was significant at $p < 0.001$. The 95% confidence interval is computed as;

- df = 30
- alpha = 0.05
- t(1-alpha/2, df) = 2.042272
- lower CI = Beta(estimate) + t * se = 7.245 - 2.042272 * 1.764 = 3.642432
- upper CI = Beta(estimate) + t * se = 7.245 + 2.042272 * 1.764 = 10.84757

So it is estimated that a "manual" transmission car would have an increase in mpg of between 3.6 and 10.8 ***mpg***, assuming all other factors equal.

2. fit2 = mpg~am+wt The effect of ***am*** and ***wt*** (weight). The beta value for ***am*** was no longer significant at $p < 0.998$. ***wt*** was significant at $p < 0.001$. The beta value for ***wt*** was -5.353 indicating that for every 1000 lbs, a loss of about 5 mpg would occur.

3. fit3 = mpg~am+wt+cyl+disp+hp+drat+qsec+vs+gear+carb The effect of all regression variables on ***mpg***. Only ***wt*** was near significance at $p < .06$. ***am*** was no longer significant at $p < 0.234$

ANOVA analysis comparing fit1, fit2, and fit3 indicated that fit2 was significantly different than fit 1 at $p < 0.001$. Fit3 versus fit2 was not significantly different at $p < 0.06$. Therefore, it seems that fit2 seems to be the better model. See ***Figure 3 - appendix*** for the regression analysis results.

Residual Analysis

Residual analysis indicated the residuals seem to be normally distributed for the fit2 model. See ***Figure 4 - appendix*** for the residual analysis results.

Appendix

Figure 1 - Summary statistics for the mtcars data set.

```
library(datasets)
data("mtcars")
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```
summary(mtcars$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   15.42   19.20   20.09   22.80   33.90
```

Figure 2 - Pvalues for correlations with mpg for all mtcar variables

```
library(Hmisc, quietly = TRUE, warn.conflicts = TRUE)
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
```

```
corrs <- rcorr(as.matrix(mtcars, type="pearson"))
round(corrs$P,3)
```

```
##        mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
## mpg     NA 0.000 0.000 0.000 0.000 0.000 0.017 0.000 0.000 0.005 0.001
## cyl  0.000    NA 0.000 0.000 0.000 0.000 0.000 0.000 0.002 0.004 0.002
## disp 0.000 0.000    NA 0.000 0.000 0.000 0.013 0.000 0.000 0.001 0.025
## hp   0.000 0.000 0.000    NA 0.010 0.000 0.000 0.000 0.180 0.493 0.000
## drat 0.000 0.000 0.000 0.010    NA 0.000 0.620 0.012 0.000 0.000 0.621
## wt   0.000 0.000 0.000 0.000 0.000    NA 0.339 0.001 0.000 0.000 0.015
## qsec 0.017 0.000 0.013 0.000 0.620 0.339    NA 0.000 0.206 0.243 0.000
## vs   0.000 0.000 0.000 0.000 0.012 0.001 0.000    NA 0.357 0.258 0.001
## am   0.000 0.002 0.000 0.180 0.000 0.000 0.206 0.357    NA 0.000 0.754
## gear 0.005 0.004 0.001 0.493 0.000 0.000 0.243 0.258 0.000    NA 0.129
## carb 0.001 0.002 0.025 0.000 0.621 0.015 0.000 0.001 0.754 0.129    NA
```

Figure 3 - Linear regression model

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

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am             7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

```
fit2 <- update(fit1, mpg~am+wt)
summary(fit2)
```

```
##
```

```
## Call:
## lm(formula = mpg ~ am + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5295 -2.3619 -0.1317  1.4025  6.8782
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.32155    3.05464  12.218 5.84e-13 ***
## am          -0.02362    1.54565  -0.015    0.988
## wt          -5.35281    0.78824  -6.791 1.87e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.098 on 29 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7358
## F-statistic: 44.17 on 2 and 29 DF,  p-value: 1.579e-09
```

```
fit3 <- update(fit2, mpg~am+wt+cyl+disp+hp+drat+qsec+vs+gear+carb)
summary(fit3)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + cyl + disp + hp + drat + qsec +
##     vs + gear + carb, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## am           2.52023    2.05665   1.225   0.2340
## wt          -3.71530    1.89441  -1.961   0.0633 .
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

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

```
## Analysis of Variance Table
##
```

```
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + wt + cyl + disp + hp + drat + qsec + vs + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 63.0133 9.325e-08 ***
## 3     21 147.49  8    130.83  2.3283   0.05774 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4 - Residual analysis

```
plot(fit2, which=2)
```