Analysis of predictor factors in vehicle miles per gallon

Naser Ameen

March 3, 2016

Executive Summary

This report details the analysis performed to answer two questions:

1. Whether an automatic or manual transmission will result in lowering miles per gallon (MPG) in a vehicle, and
2. What other explanatory factors explain the MPG consumption in a vehicle. The results indicate that in general a manual transmission result in lower MPG consumption.

In a simple linear regression model using automatic/manual as categorical predictors and MPG as the dependent variable, the independent variable explained 34% of the variation in MPG. ANOVA analysis revealed that the coefficient of the automatic/manual predictor was 7.254, meaning that manual cars consumes 7.254 more MPG than automatic cars. This may have do with the fact that the weight of manual cars higher at 3.768 tons versus automatic cars whose weight is 2.411 tons.

However, when a multi-factor model was utilized to predict MPG, the explanatory power increased. The final model incorporated weight, horsepower, and cylinder in addition to transmission. The model explained 87% of the variation in MPG. The weight coefficient of -2.5 means that holding all other factors constant, a 1 ton increase in weight will lower mpg by 2.5.

In conclusion, the single most important factor that determines MPG is weight, and this can be seen by doing a forward stepwise regression. All other predictors will only provide an incremental increase in predictive power. Indeed, most of the predictive power of other independents is already incorporated in weight. For example, 8 cylinder vehicles are heavier than 6 cylinders, which are in turn heavier than 4 cylinder cars, and the more horsepower, the more the number of cylinders and the more the weight.

Report

This report is divided into five sections. The *Data Setup* section details the transformation required the analyze the data. The *Automatic versus Manual* section analyzes whether or manual transmission will result in lowering MPG in a vehicle, the *Best Model* section explains other independent variables that have more explanatory power in determining MPG consumption, and the *Conclusions* provide some more insight into the predictors. Finally, the *Appendix* section shows supplementary charts and graphs detailing model specification and goodness of fit.

Data Setup

The mtcars model is a data frame of 32 observations on 11 variables. Brief descriptions of each variable are provided below. In order to ensure that categorical variables are not treated as numerical variables, factor conversions are done on some of the variables.

```
mpg   <- cars$mpg                # mpg Miles/(US) gallon - numerical
cyl   <- factor(cars$cyl)        # Number of cylinders - categorical
disp  <- cars$disp               # Displacement (cu.in.) - numerical
hp    <- cars$hp                 # Gross horsepower - numerical
```

```
drat  <- cars$drat                # Rear axle ratio - numerical
wt    <- cars$wt                  # Weight (1000 lbs or 1 ton) - numerical
qsec  <- cars$qsec                # 1/4 mile time -numerical
vs    <- cars$vs                  # V/S -numerical
am    <- factor(cars$am)          # Transmission (0 = automatic, 1 = manual) - categorical
gear  <- factor(cars$gear)        # Number of forward gears - categorical
carb  <- factor(cars$carb)        # Number of carbureators -categorical
```

Automatic versus Manual

In order to check whether automatic or manual transmission yields a lower MPG, a simple linear regression with `mpg` as the independent variable and `am` as the dependent variable is done as shown below:

$$mpg = \beta_0 + \beta_1 \times \ am + \epsilon$$

The summary information indicates that:

$$\beta_0 = 17.1473684$$

$$\beta_1 = 7.2449393$$

$$R^2 = 0.3597989$$

Thus the average MPG for automatic cars is 17.1473684, and the difference between the average MPG for automatic cars and manual vehicles is 7.2449393 can be interpreted as the increment above automatic transmission that can be achieved by manual transmission vehicles. 35.9798943% is fit is explained by the `am`. A p-value of ‘r summary(fit1)$coefficients[2,4] indicates that we can reject the null hypothesis that there is no difference in the average MPG of automatic and manual cars. The residual plot shown in the *Appendix* indicate that residuals are linear, independent and do not violate the homoscedasticity assumption. The summary statistics are also reported in the *Appendix*.

Best Model

The full model using all predictors for MPG is:

$$m\hat{p}g = \beta_0 + \beta_1 \times \ cyl + \beta_2 \times \ disp + \beta_3 \times \ hp + \beta_4 \times \ drat + \beta_5 \times \ wt + \beta_6 \times \ qsec + \beta_7 \times \ VS + \beta_8 \times \ am + \beta_9 \times \ gear + \beta_{10} \times \ carb$$

The model has a high adjusted R^2 with a value of 0.8066423. However, the problem is that the p-values are not significant at the 5% level. The lowest p-value is 0.0632522 attributable to the `wt` coefficient. In order to generate a more parsimonious model with better p-values, a backward stepwise regression is performed. The best model is the one with the lowest AIC value.

The best model uses the predictors *cyl6*, *cyl8*, *hp*, *wt*, and, *am1* with the following coefficients:

```
## (Intercept)        cyl6        cyl8          hp          wt         am1
## 33.70832390 -3.03134449 -2.16367532 -0.03210943 -2.49682942  1.80921138
```

The model has an R^2 of 0.8658799 which is lower than the R^2 of the full model, as expected. However, the *adjusted* $^{\mathrm{R}}$ is better at 0.8400875. But even in the final model some of the relationships are spurious as shown by the high p-values for the *am1* and *cyl8* predictors. The summary statistics are reported in the *Appendix*.

Conclusions

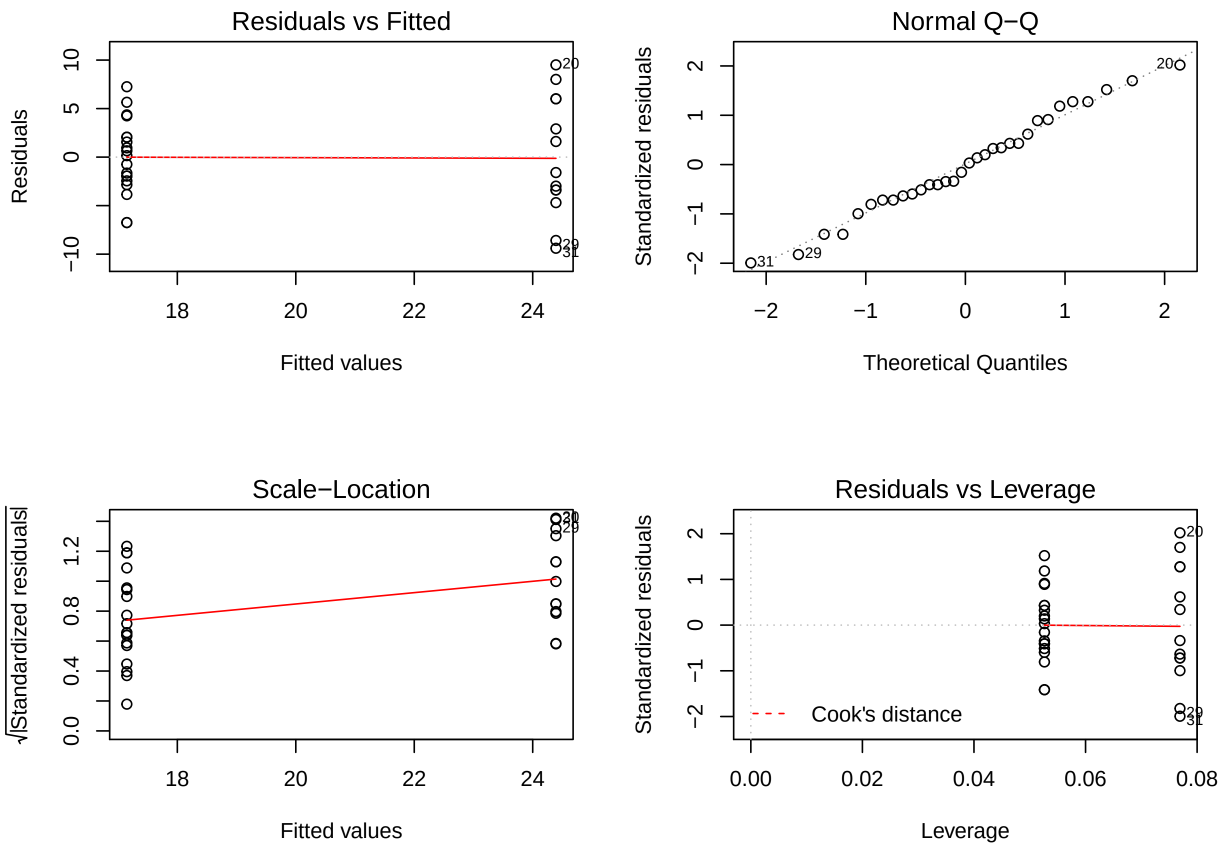
The adjusted R^2 of the best model is higher than both the full model and the single predictor automatic/manual model. The best model has less spurious predictors than the full model. But even in the best model, there are spurious predictors. The weight predictor is the single most important predictor of MPG, and any other variable simply provides incremental bumps in explanatory power. Take for example the number of cylinders - the higher the number of cylinders, the more the horsepower, and the more the weight. So weight essentially incorporates some of the predictive powers of cylinder and horsepower.

Appendix

The summary statistics of automatic or manual transmission is provided below:

```
##
## Call:
## lm(formula = mpg ~ am)
##
## 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 ***
## am1            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
```

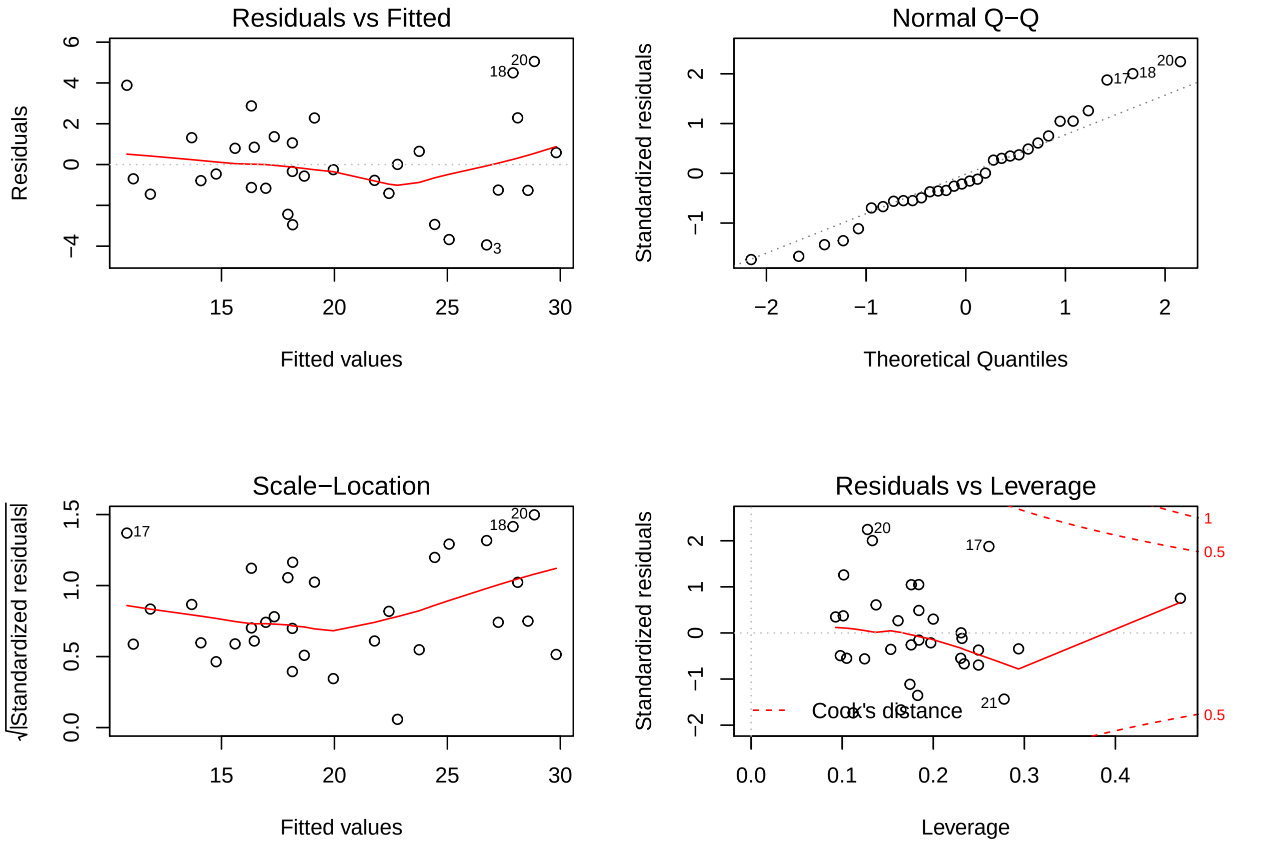
The diagnostic plot of regression residuals are shown below. Both the residual versus fit and the standardized residual versus fit show no major issues with linearity, normality, homoscedasticity, and independence.



The summary statistics of best fit model is provided below. Some of the p-values are not significant at the 5% level.

```
##
## Call:
## lm(formula = mydf$mpg ~ cyl + hp + wt + am, data = mydf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## am1          1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

The regression residuals shown below no major linearity, normally homoscedasticity, and independence issues, although there is a slight dip in the residuals in the middle of the data. However, there is no major clustering of the residuals around certain fit values. Thus the model does have some predictive power.



Finally, the correlation matrix shown below highlights the importance of weight as a predictor. The matrix shows that all three prdictors *wt*, *hp*, *cyl* are negatively correlated with *mpg*, with correlation of at least -77%. However, *hp* and *cyl* are also highly positively correlated with *wt*, with correlation factors of at least 65%.

```r
cor_df <- data.frame(mpg=mpg,wt=wt,hp=hp,cyl=as.numeric(cyl))
cor(cor_df)
```

```
##            mpg         wt         hp        cyl
## mpg  1.0000000 -0.8676594 -0.7761684 -0.8521620
## wt  -0.8676594  1.0000000  0.6587479  0.7824958
## hp  -0.7761684  0.6587479  1.0000000  0.8324475
## cyl -0.8521620  0.7824958  0.8324475  1.0000000
```