Exploring miles per (US) gallon relationship with type of transmission

Wildson B B Lima

25/10/2020

Packages

We are going to need some R packages to assist the analysis.

Executive summary

We are interested here in exploring a dataset to find the relationship between miles per gallon (MPG) and some other variables. The following questions are of particular interest here:

- Is automatic or manual transmission better for MPG?
- Quantifying the MPG difference between automatic and manual transmissions.

We were able to model a linear regression from which we found that on average a manual car can runs 1.8 more miles per (US) gallon than a automatic car. This result is very uncertain from a 95% significance level, though, with a confidence interval ranging from -1.06 to 4.68 mpg.

Data

We are going to use here the mtcars dataset. Its data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). We have the following variables in the dataset:

[, 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 Engine (0 = V-shaped, 1 = straight) [, 9] am Transmission (0 = automatic, 1 = manual) — [,10] gear Number of forward gears — [,11] carb Number of carburetors

Exploratory Data Analysis

First, let's take a look at the data.

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
```

```
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

We can see from here data is all loaded as numeric, even though some variables are categorical. We need to change this right away before doing more with the data because this affects all types of analysis.

Now, let's take a look at the summary of the data.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   15.43   19.20   20.09   22.80   33.90
```

Ok, that seems good enough. We can see mpg mean here is 20.09 and data ranges from 15.43 to 33.9 mpg. A histogram better illustrates how data is distributed. So we do one here [appendix].

We can see it is a little bit skewed, but because we have sample size of 32, we are fine about normality assumptions.

Modeling

For the model selection, we're gonna use a backward elimination, starting with all variables. Variables are removed one at a time, till we can no longer improve the adjusted R^2. We chose to improve adjusted R^2 because it better describes the strength of a model fit, since this metric is more responsive to explanatory variables that add more explanation about the variability of the response variable to the model[1]. We start like this:

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5087 -1.3584 -0.0948  0.7745  4.6251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.57171   19.56616   1.358   0.1945
## cyl.L       -0.23770    5.06256  -0.047   0.9632
## cyl.Q        2.02541    2.14952   0.942   0.3610
## disp         0.03555    0.03190   1.114   0.2827
## hp          -0.07051    0.03943  -1.788   0.0939 .
## drat         1.18283    2.48348   0.476   0.6407
## wt          -4.52978    2.53875  -1.784   0.0946 .
## qsec         0.36784    0.93540   0.393   0.6997
## vsS          1.93085    2.87126   0.672   0.5115
## ammanual     1.21212    3.21355   0.377   0.7113
## gear.L       1.78785    2.64200   0.677   0.5089
## gear.Q       0.12235    2.40896   0.051   0.9602
## carb.L       6.06156    6.72822   0.901   0.3819
```

[1] Open Intro Statistics

```
## carb.Q       1.78825    2.80043   0.639   0.5327
## carb.C       0.42384    2.57389   0.165   0.8714
## carb^4       0.93317    2.45041   0.381   0.7087
## carb^5      -2.46410    2.90450  -0.848   0.4096
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.833 on 15 degrees of freedom
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.779
## F-statistic:  7.83 on 16 and 15 DF,  p-value: 0.000124
```

We have good adjusted R-squared of 0.78 for a start, but there is a lot more of variability accounted in R-squared, 0.89, which suggests we have unnecessary variables at play. So let's do the model selection.

Now check a summary of the result.

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars2)
##
## 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) 31.97665    3.06337  10.438 8.61e-11 ***
## cyl.L       -1.52995    1.61521  -0.947  0.35225
## cyl.Q        1.59177    0.88076   1.807  0.08231 .
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## ammanual     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
```

Look at that, now we have a good adjusted R-squared of 0.84 and R-squared of 0.87. It's pretty good.That means up to 86.59% of the response variable is explained by a model made of the explanatory variables cyl + hp + wt + am. Generally, multiple regression linear models depend on the following assumptions:

- the residual of the model are nearly normal
- the variability of the residuals is nearly normal
- the residual are independent, and
- each variable is linearly related to the outcome.

To check model assumptions, we are gonna need to look some graphs. First, the normal probability plot [appendix].

It seems like we might have some residual outliers, but generally it seems good enough. Now, to check if the variance is approximately constant, the absolute values of residual against fitted values [appendix].

Once again, there is two potential outliers but generally it seems good enough. A good way of checking for linear relationship between the response and explanatory variables is to see the how residual varies with each explanatory variables. That way, we account for the other variables in the model at each plot, and not just a bivariate relationship. We are looking for random scatters around 0 [appendix].

Once again, pretty good results. All residuals seem to be scattered around zero. For the last assumption to check, independence of residual, we only need the observations to be independent. We have no reason to doubt that is the case here. Note that individually we had some variables with p-value higher than the a significance level of 5%, but the checked assumptions of the model and the adjusted R-squared gives great strength to the model. This way, we can say the model conditions are in good shape.

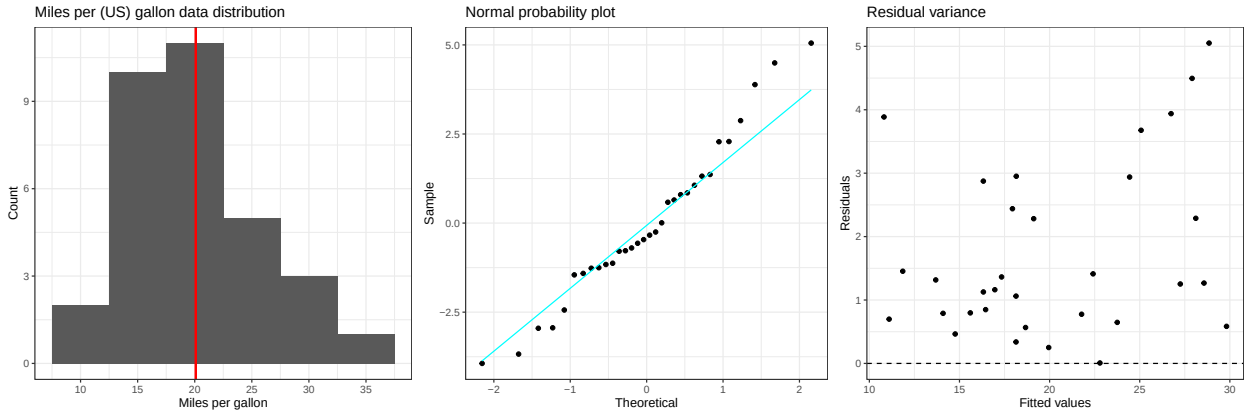
Results

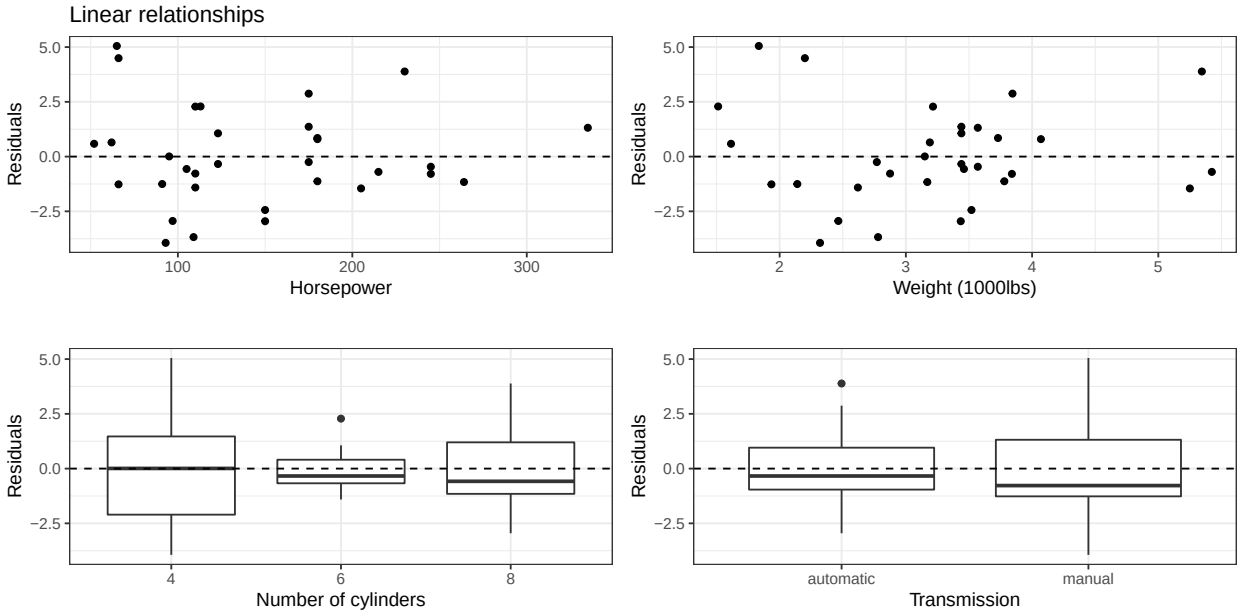
Let's find the answer to the questions we were interested. From this model, as we have seen, we can say that on average a car with manual transmissions runs 1.8 more miles per (US) gallon than an automatic. That seems like we are ready to say that manual transmissions cars are indeed better. Let's first look at a 95% confidence interval before we can confirm this.

```
##                2.5 %   97.5 %
## ammanual -1.060934 4.679356
```

Things are not as good as they seemed. At a 95% significance level, the average miles per (US) gallon a manual car runs ranges from -1.06 to 4.68. So, this is the level of uncertainty our answer has.

Appendix of figures




Linear relationships
Residuals
5.0
2.5
0.0
−2.5
100
200
300
Horsepower
2
3
4
5
Weight (1000lbs)
4
6
8
Number of cylinders
automatic
manual
Transmission