Regression analysis of mtcars

Tyler Burleigh

4/14/2019

Executive summary

In this document I aim to answer two questions: (1) Is an automatic or manual transmission better for MPG, and (2) what is the MPG difference between automatic and manual transmissions?

I found that transmission did not account for a significant difference in MPG after controlling for other factors that were associated with both MPG and transmission. The estimated difference was 0.17, with a 95% confidence interval of -2.97 to 3.32.

Load libraries and data

Also recode `am` to a factor with meaningful labels.

```
data(mtcars)
library(ggplot2)
library(reshape2)
library(tidyverse)
library(RColorBrewer)
mtcars$am <- factor(mtcars$am, labels = c("automatic", "manual"))
```

Models

Model 1

The outcome (`mpg`) is continuous, so we can use a linear regression. We'll start with a simple model containing the 1 variable. The significance criteria we'll use is $p < .05$.

This model is significant ($p < .05$). The coefficient is 7.25. Normally we would interpret this coefficient as "for every 1 unit increase in X, Y changes by this amount". Because `am` is a binary factor, this just means that the `mpg` for a manual transmission (`am` = 1) is 7.25 units greater than an automatic transmission (`am` = 0).

```
mdl <- lm(mpg ~ am, data = mtcars)
summary(mdl)$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## ammanual     7.244939   1.764422  4.106127 2.850207e-04
```

Model 2 - Controls

Now let's add the control variables. These control variables were selected because they had a correlation of r > 0.5 with both the outcome (`mpg`) and predictor (`am`) variables of interest. (See `Correlation Matrix` in the `Appendix`).

```r
mdl2 <- lm(mpg ~ am + factor(cyl) + disp + drat + wt, data = mtcars)
summary(mdl2)$coefficients
```

```
##                   Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  34.796706680 6.70238133  5.1916931 2.271364e-05
## ammanual      0.261365828 1.53969659  0.1697515 8.665717e-01
## factor(cyl)6 -4.374626281 1.58014542 -2.7684960 1.045306e-02
## factor(cyl)8 -6.407637752 2.75347514 -2.3271094 2.835383e-02
## disp          0.001425009 0.01407923  0.1012136 9.201883e-01
## drat         -0.255596739 1.56599820 -0.1632165 8.716602e-01
## wt           -3.251684420 1.27325003 -2.5538459 1.713143e-02
```

Having controlled for these variables, the effect of transmission on `mpg` is much smaller (coefficient = 0.26, p > .05).

We'll compare this model to the first one, to see if it was "better" (i.e., if it accounted for more variance). We see that it was better (p < .05), because the residual sums of squares is much smaller.

```r
anova(mdl, mdl2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + factor(cyl) + disp + drat + wt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     25 182.67  5    538.22 14.732 9.059e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Quantify the mpg difference

So what is the MPG difference between automatic and manual transmissions? Well, our second model with controls is better, so we will use that. We'll use the coefficient from model 2 and generate a 95% confidence interval around it. It wasn't significant, so we can expect this interval to include zero. The estimated difference is 0.17 (-2.91 to 3.43).

```r
round(confint(mdl2), 2)
```

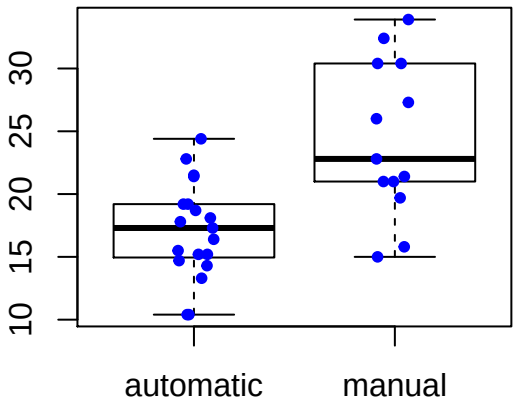
```
##               2.5 % 97.5 %
## (Intercept)   20.99  48.60
## ammanual      -2.91   3.43
## factor(cyl)6  -7.63  -1.12
## factor(cyl)8 -12.08  -0.74
## disp          -0.03   0.03
## drat          -3.48   2.97
## wt            -5.87  -0.63
```

Appendix

Diagnostics

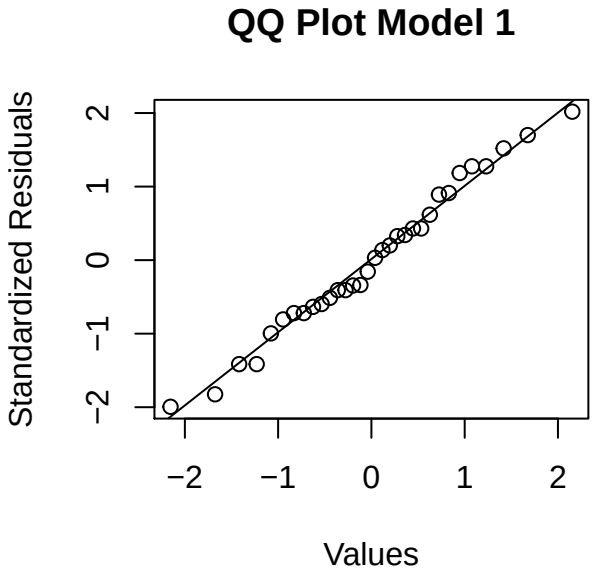
Box plot with scatter. There don't appear to be any outliers, because the dots fall within the boxplot ranges.

```r
boxplot(mpg ~ am, data = mtcars)
stripchart(mpg ~ am, data = mtcars, vertical = TRUE,
    method = "jitter", add = TRUE, pch = 20, col = 'blue')
```



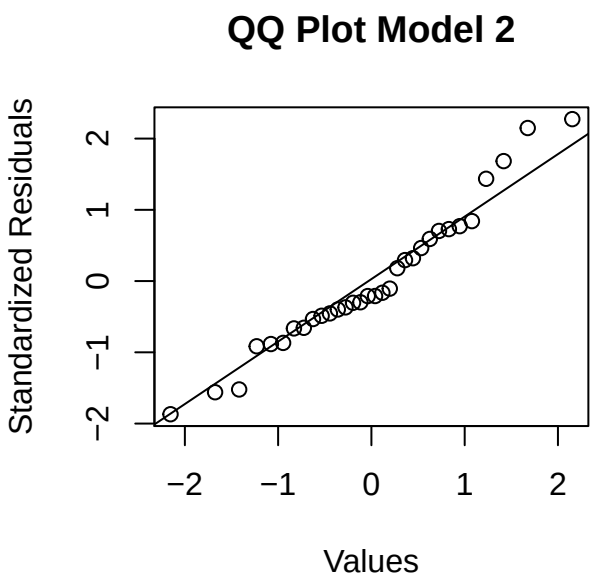
QQ plot to assess normality of the standardized residuals from our first model. The residuals look normally distributed because the dots are all pretty close to the fit line.

```r
mdl_res <- rstandard(mdl)
qqnorm(mdl_res, ylab="Standardized Residuals", xlab="Values", main="QQ Plot Model 1")
qqline(mdl_res)
```



QQ plot to assess normality of the standardized residuals from our second model. The residuals are not terribly non-normal, but there is some departure at the upper quantiles.

```
mdl2_res <- rstandard(mdl2)
qqnorm(mdl2_res, ylab="Standardized Residuals", xlab="Values", main="QQ Plot Model 2")
qqline(mdl2_res)
```



Correlation matrix

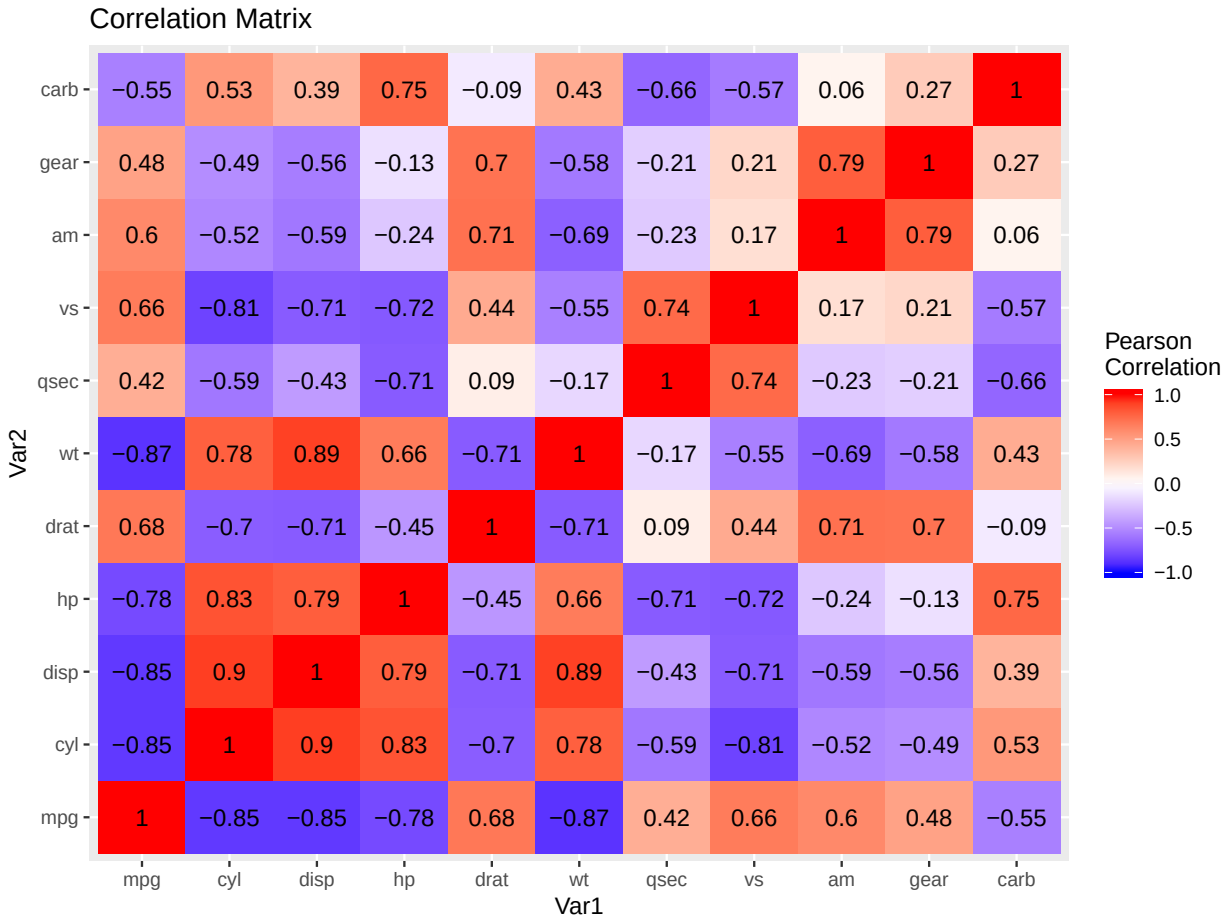
Correlate all variables with all other variables. If we see large correlations between our primary model variables and the other variables, we should "control for them" in our analysis.

We see that `mpg` is associated with all variables to some degree. But let's be conservative and only consider variables with correlations over an absolute value of 0.5. These are: `cyl`, `disp`, `hp`, `drat`, `wt`, `vs`, and `carb`.

We see that `am` is associated with: `cyl`, `disp`, `drat`, `wt`, and `gear` over the same 0.5 level of correlation.

The intersection of these is: `cyl`, `disp`, `drat`, and `wt`. We'll use these as controls.

```
cor(mtcars %>% mutate(am = as.numeric(am))) %>%
  round(., 2) %>%
  melt(.) %>%
  ggplot(., aes(x = Var1, y = Var2, fill = value)) +
    geom_tile() +
    ggtitle("Correlation Matrix") +
    scale_fill_gradient2(low = "blue", high = "red", mid = "white",
     midpoint = 0, limit = c(-1,1), space = "Lab",
     name="Pearson\nCorrelation") +
  geom_text(aes(Var2, Var1, label = value), color = "black", size = 4)
```


Correlation Matrix
carb −0.55 0.53 0.39 0.75 −0.09 0.43 −0.66 −0.57 0.06 0.27 1
gear 0.48 −0.49 −0.56 −0.13 0.7 −0.58 −0.21 0.21 0.79 1 0.27
am 0.6 −0.52 −0.59 −0.24 0.71 −0.69 −0.23 0.17 1 0.79 0.06
vs 0.66 −0.81 −0.71 −0.72 0.44 −0.55 0.74 1 0.17 0.21 −0.57
qsec 0.42 −0.59 −0.43 −0.71 0.09 −0.17 1 0.74 −0.23 −0.21 −0.66
wt −0.87 0.78 0.89 0.66 −0.71 1 −0.17 −0.55 −0.69 −0.58 0.43
drat 0.68 −0.7 −0.71 −0.45 1 −0.71 0.09 0.44 0.71 0.7 −0.09
hp −0.78 0.83 0.79 1 −0.45 0.66 −0.71 −0.72 −0.24 −0.13 0.75
disp −0.85 0.9 1 0.79 −0.71 0.89 −0.43 −0.71 −0.59 −0.56 0.39
cyl −0.85 1 0.9 0.83 −0.7 0.78 −0.59 −0.81 −0.52 −0.49 0.53
mpg 1 −0.85 −0.85 −0.78 0.68 −0.87 0.42 0.66 0.6 0.48 −0.55
Var2
mpg cyl disp hp drat wt qsec vs am gear carb
Var1
Pearson Correlation
1.0
0.5
0.0
−0.5
−1.0