Regression Models Course Project

Andrew Januszewski

Setup

```
library(tidyverse)
library(GGally)
knitr::opts_chunk$set(echo = TRUE)
data("mtcars")
```

Executive Summary

Moter Trend, a magazine about the automobile industry, is interested in examining the relationship between a set of variables and miles per gallon (MPG). They are particularly interested in answering the following questions:
    * Is an automatic or manual transmission better for MPG?
    * Can we quantify the difference between automatic and manual transmissions?
This analysis concludeds manual transmissions are better for MPG than automatic. The three best predictors of MPG were transmission type, vehicle weight, and 1/4 mile time. Holding the other two variables constant, manual transmissions are ~2.94 more MPG efficient than automatics.

Exploratory Data Analysis

As shown in appendix figure one, transmission type (am, 0 = automatic, 1 = manual) and MPG share a moderately strong, positive correlation. However. we also see other variables having similar or stronger correlation values with MPG (bottom row of figure one). As such, these predictors should be controlled for during any predictive model fitting.

```
ggcorr(mtcars, name = 'Correlation', label = T)
```

For instance, consider the strong negative correlation between horsepower (hp) and MPG. It suggests as horsepower increases, MPG decreases. Figure two in the appendix confirms this notion (black line). However, notice that our prediction changes significantly when we include transmission type (am) as a variable (blue - manual, red - automatic). Also, take note that the relationship between horsepower and MPG appears to be non-linear.

```
ggplot(mtcars, aes(y = mpg, x = hp, color = factor(am))) +
  geom_point() +
  geom_smooth(method = 'lm', formula = y ~ x, color = 'black', se = F) +
  geom_smooth(method = 'lm', se = F) +
  labs(y = 'MPG', x = 'Horsepower') +
  theme(legend.position = 'none')
```

We can further prove these observations by examining the residuals of each model. Appendix figure three displays a reduction in Residual Standard Error (RSE) from 3.863 to 2.909. Additionally, explained error (R-squared) increased with the addition of am as a predictor. As we know, R-squared will continually grow with every variable added. However, Adjusted R-squared accounts for that and increased from 0.5892 to 0.767. Furthermore, both predictors are siginificant and a nested model confirms that in figure four.

```
fit_mpg_hp <- lm(data = mtcars, mpg ~ hp - 1)
fit_mpg_hp_am <- lm(data = mtcars, mpg ~ hp + factor(am) - 1)
summary(fit_mpg_hp)
summary(fit_mpg_hp_am)
anova(fit_mpg_hp, fit_mpg_hp_am)
```

A plot of the residuals in appendix figure five also shows much less dispersion around zero with the addition of am as a predictor.

```
plot(fit_mpg_hp$residuals)
plot(fit_mpg_hp_am$residuals)
```

Data Modeling

We will fit and compare two linear models to answer the questions. The first will be extremely simple while the second will leverage more variables choosen with some automobile knowledge.
For the simple model, we use transmission type as the sole predictor of MPG. It appears transmission type is significant, but this simple model only explains ~34% of variance.

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

```
##
## Call:
## lm(formula = mpg ~ factor(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 ***
## factor(am)1    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
```

For the second model, we will use some subject matter knowledge to include additional predictors. This knowledge allows us to discount some combinations of highly correlated variables such as displacement and cylinders as they are semi-interchangeable (displacement is the volume of all the pistons inside the cylinders, more cylinders = higher displacement). Both are also highly correlated with horsepower. Using this logic, we can exclude several other combinations. These could also be verified with the Variance Inflation Factor (VIF), but we're running out of allotted pages for this assignment so we'll use the step function to find the most crucial variables.

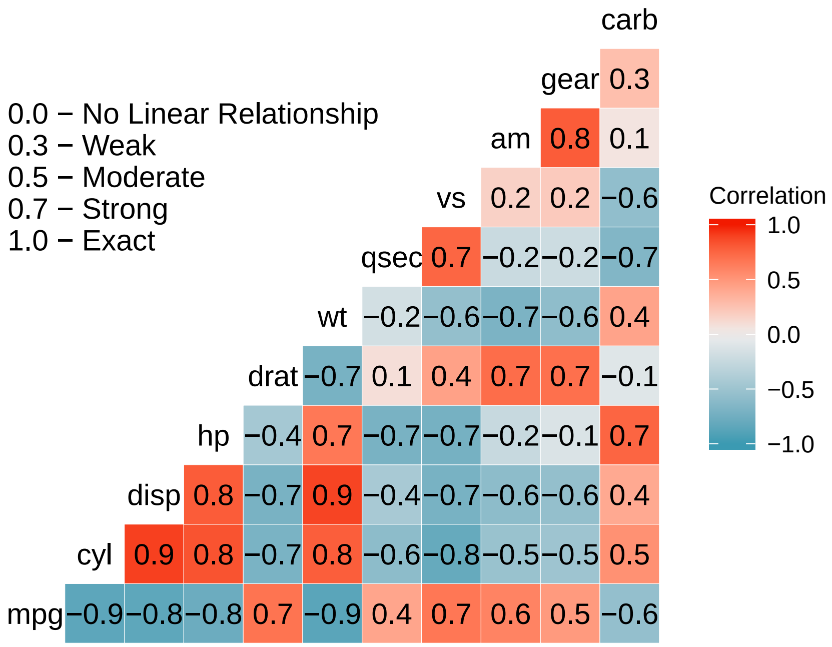
Again, transmission type is significant and the model now explains 83% of variance. This means 17% of unexplained variance could come from variables outside the model and dataset. Keeping this uncertainty in mind, we can conclude that manual transmissions are better for MPG. Holding the other two most explanatory variables constant (wt and qsec), manual transmissions are ~2.94 more MPG efficient.

```r
fit_lm_vars <- step(lm(data = mtcars, mpg ~ . - cyl - disp - gear - drat), trace = 0)
summary(fit_lm_vars)
```

```
## 
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.4811 -1.5555 -0.7257  1.4110  4.6610 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   9.6178     6.9596   1.382 0.177915    
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am            2.9358     1.4109   2.081 0.046716 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336 
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

Appendix

Figure 1:



	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
gear										0.3
am									0.8	0.1
vs								0.2	0.2	−0.6
qsec							0.7	−0.2	−0.2	−0.7
wt						−0.2	−0.6	−0.7	−0.6	0.4
drat					−0.7	0.1	0.4	0.7	0.7	−0.1
hp				−0.4	0.7	−0.7	−0.7	−0.2	−0.1	0.7
disp			0.8	−0.7	0.9	−0.4	−0.7	−0.6	−0.6	0.4
cyl		0.9	0.8	−0.7	0.8	−0.6	−0.8	−0.5	−0.5	0.5
mpg	−0.9	−0.8	−0.8	0.7	−0.9	0.4	0.7	0.6	0.5	−0.6

Figure 2:

```
Black = MPG predicted by horsepower
Blue = MPG predicted by horsepower for manual transmission
Red = MPG predicted by horsepower for automatic transmission
```

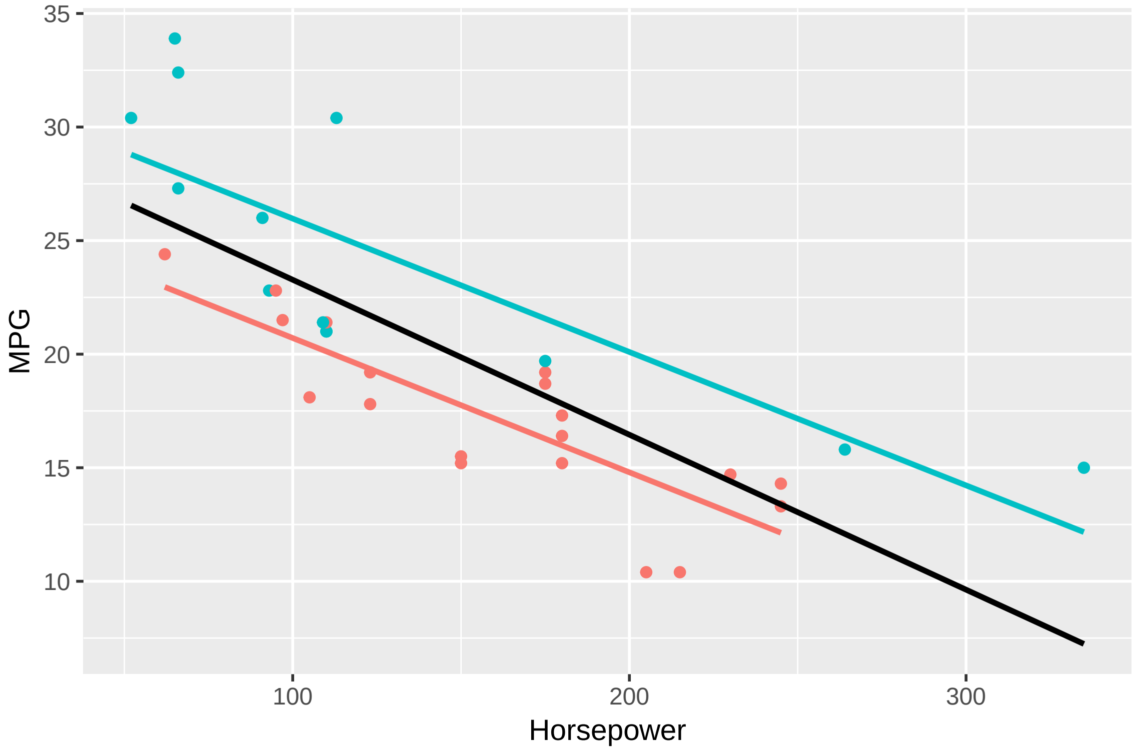


Figure 3:

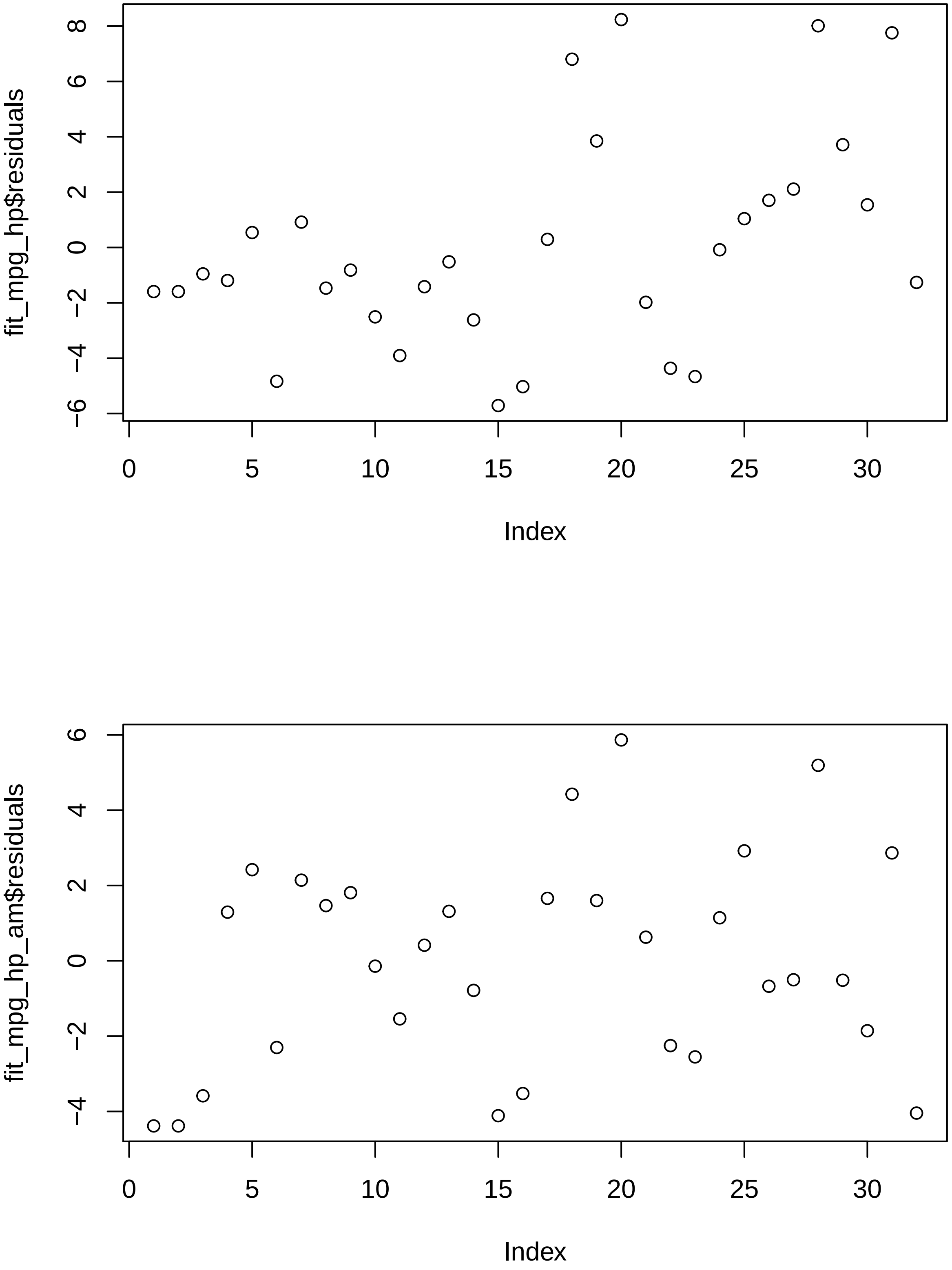
```
##
## Call:
## lm(formula = mpg ~ hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7121 -2.1122 -0.8854  1.5819  8.2360
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.09886    1.63392  18.421  < 2e-16 ***
## hp          -0.06823    0.01012  -6.742 1.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.863 on 30 degrees of freedom
## Multiple R-squared:  0.6024, Adjusted R-squared:  0.5892
## F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

```
##
## Call:
## lm(formula = mpg ~ hp + factor(am), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3843 -2.2642  0.1366  1.6968  5.8657
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.584914   1.425094  18.655  < 2e-16 ***
## hp          -0.058888   0.007857  -7.495 2.92e-08 ***
## factor(am)1  5.277085   1.079541   4.888 3.46e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.909 on 29 degrees of freedom
## Multiple R-squared:  0.782,  Adjusted R-squared:  0.767
## F-statistic: 52.02 on 2 and 29 DF,  p-value: 2.55e-10
```

Figure 4:

```
## Analysis of Variance Table
##
## Model 1: mpg ~ hp
## Model 2: mpg ~ hp + factor(am)
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     30 447.67
## 2     29 245.44  1    202.24 23.895 3.46e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 5:


fit_mpg_hp$residuals
8
6
4
2
0
−2
−4
−6
0
5
10
15
20
25
30
Index
fit_mpg_hp_am$residuals
6
4
2
0
−2
−4
0
5
10
15
20
25
30
Index