Regression Models Project

Lee Bunce

22 October 2017

Executive Summary

This report seeks to answer two questions:

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

After exploring the data and fitting a number of models we settle on a model that predicts mpg in terms of transmission type, weight and number of cyliders. Our model suggests that manual transmissions have an improved mpg of 0.17, although our model suggest the difference is not statistically significant.

Exploratory Data Analysis

First we investigate the relationship between mpg and the other variables in the mtcars data set. The mtcars data set contains a number of variables and we first create a number of graphs that plot these variables against mpg. Some of these plots can be seen in the appendix.

From the plots we can see that a number of variables appear to be correlated with mpg. We shall therefore use this information in building and selecting our model, although we might expect to find that many of these variables are themselves closely correlatedalthough we might expect to find that many of these variables are themselves closely correlated.

Perhaps most importantly for our current purposes however is that manual cars appear to have higher mpg, before we take into account other variables.

Model

To build a model we start with a model that predicts mpg using am and at each step create new models by adding additional variables and then comparing the performance of the models.

```
mod1 <- lm(mpg~as.factor(am), mtcars)
```

Our inital model then has an r^2 value of 0.33. To try and improve on this we add additional variables based on the results of our exploratory data analysis. At each we check to see if the additional variable improves the adjusted r squared (though due to space restriction the comparisons are not presented here).

```
mod2 <- lm(mpg~as.factor(am) + wt, mtcars) # Major improvement
mod3 <- lm(mpg~as.factor(am) + wt + cyl, mtcars) # Major improvement
mod4 <- lm(mpg~as.factor(am) + wt + cyl + disp, mtcars) # No major improvement
mod5 <- lm(mpg~as.factor(am) + wt + cyl + carb, mtcars) # Some improvement
mod6 <- lm(mpg~as.factor(am) + wt + cyl + carb + vs, mtcars) # No major improvement
```

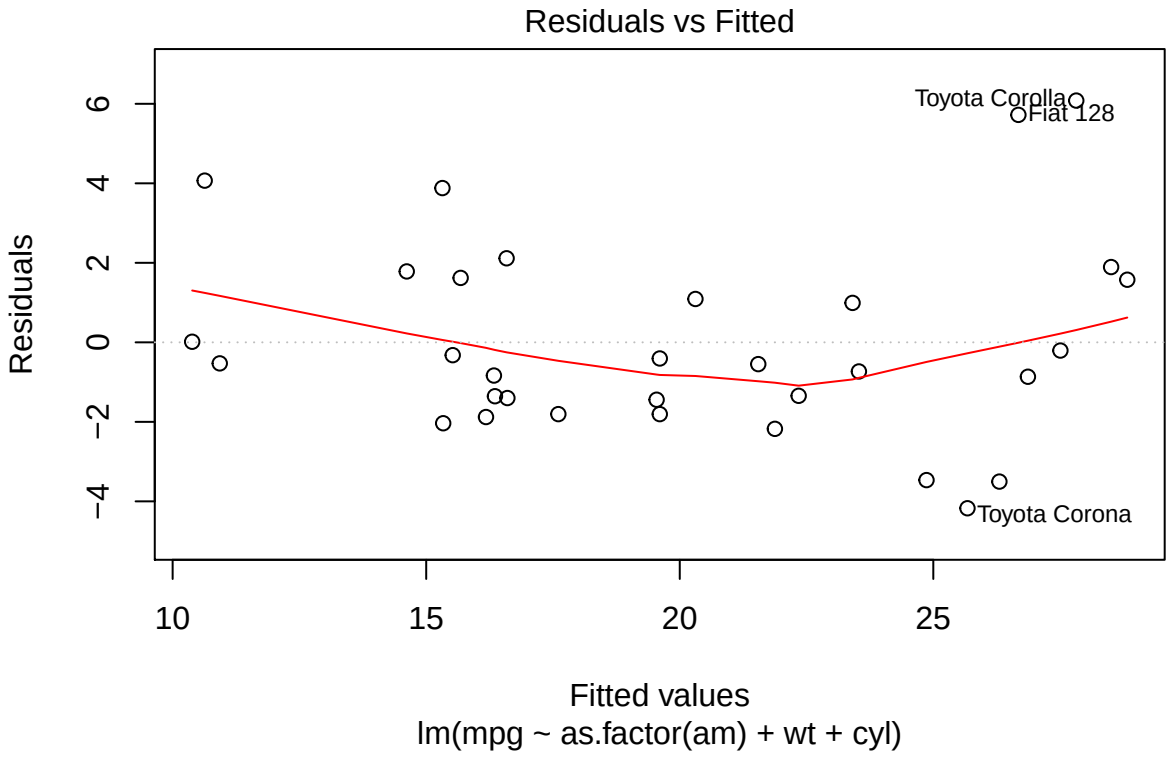
At the end of the process we settle on model 5. To check this we now compare all the models using an anova test, which can be found in the appendix.

Comparing the residual sum of squares (see appedix) we see major improvements when we add the wt variable, the cyl variable and then the carb variable. However the addition of the carb variable is not so significant, and so in the interest of parsimony we now settle on model 3.

Interpreting the summary of mod3 above we see that the model predicts that manual cars will have a an increased mpg of 0.17, holding wt and cyl constant. However the estimate is not significant and so we cannot reject the hypothesis that there is no difference in mpg between autmatic and manual transmission cars in this data set.

As a final check we look at the residuals of our model.

```
plot(mod3, which = 1:1)
```



The plot above shows that the residuals are mostly small and evenly distributed, with the exception of a couple of outliers. As I results we can conlude the model is a reasonably good fit.

Appendix

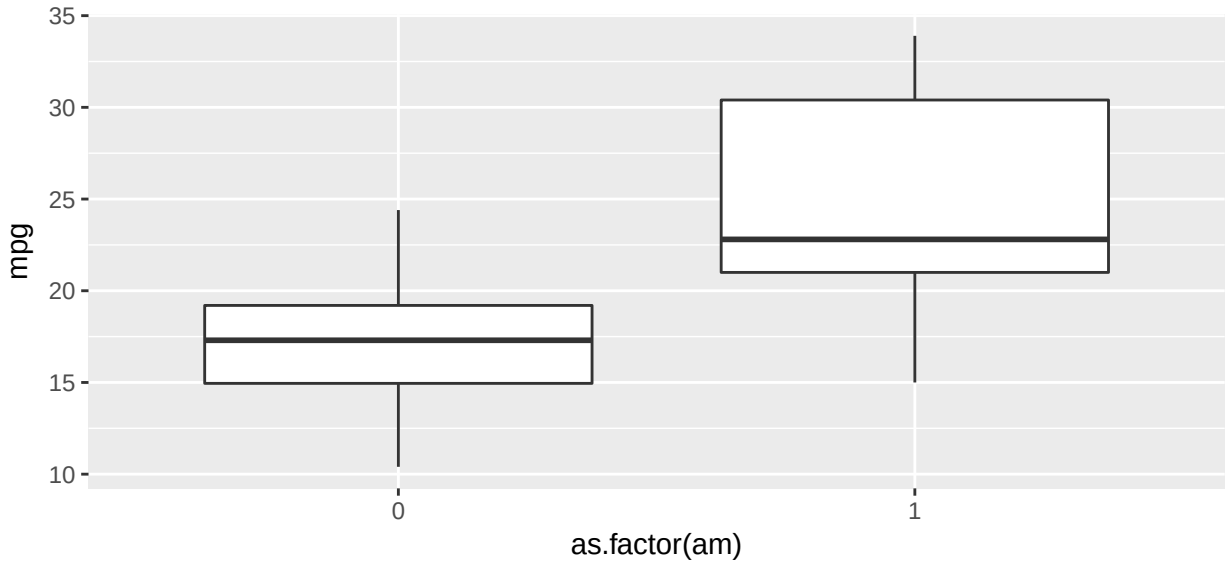
This section presents some of the graphs created in our exploratory data analysis. Each plot looks at the relationship between mpg and some other variable in the data set.

Exploratory Data Analysis

am vs. mpg

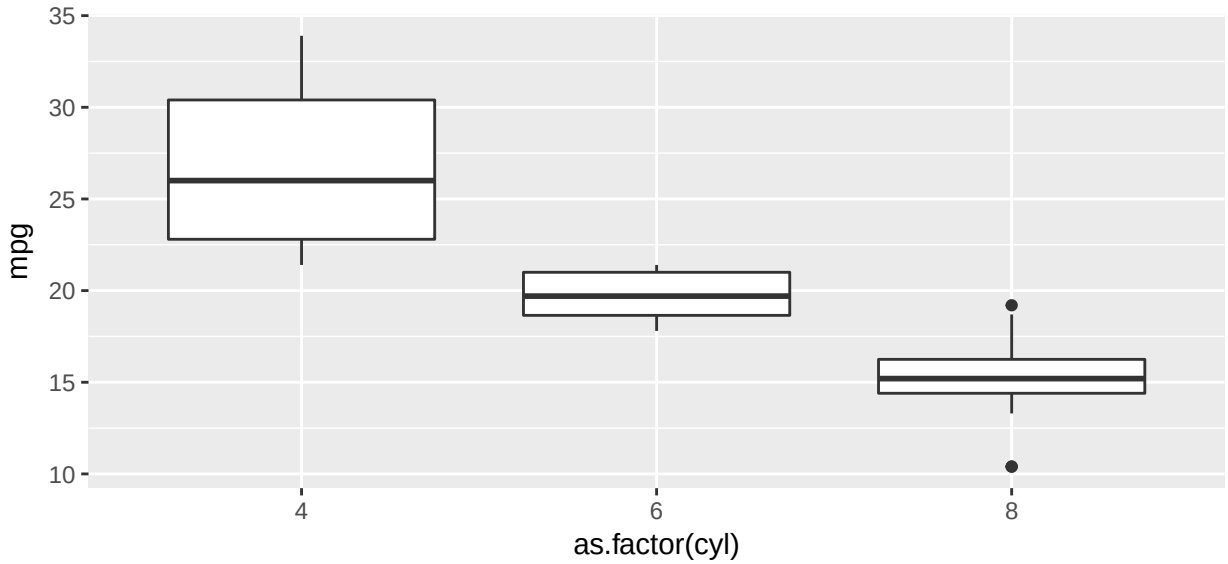
This chart suggests that manual cars have a higher mpg.

```
ggplot(mtcars, aes(as.factor(am), mpg)) +
  geom_boxplot()
```



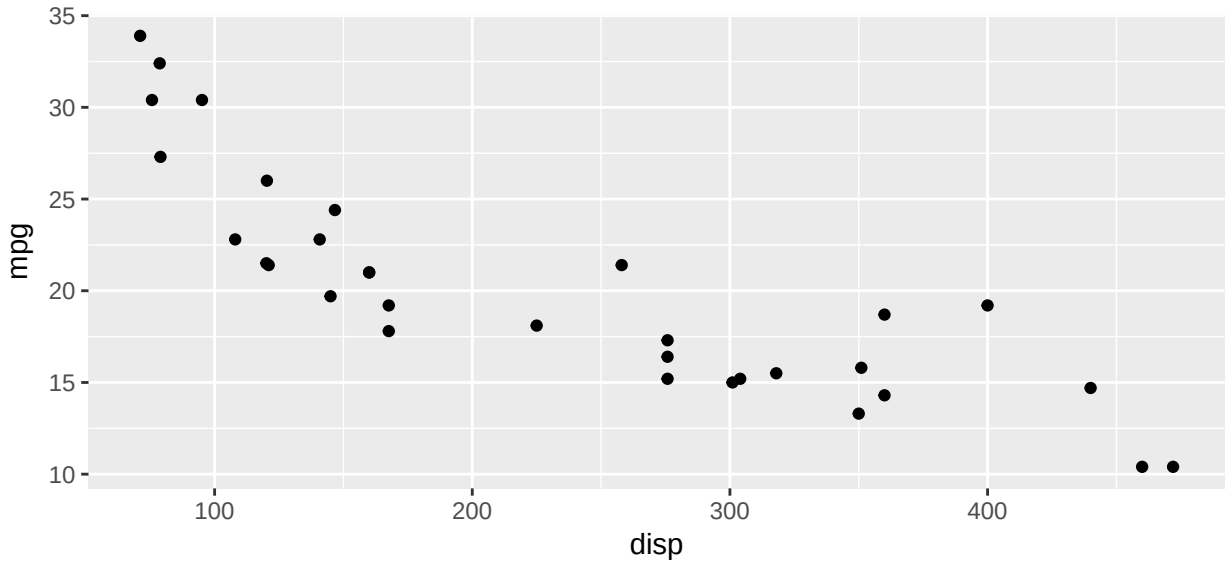
mpg vs. cyl

```
ggplot(mtcars, aes(as.factor(cyl), mpg)) +
  geom_boxplot()
```



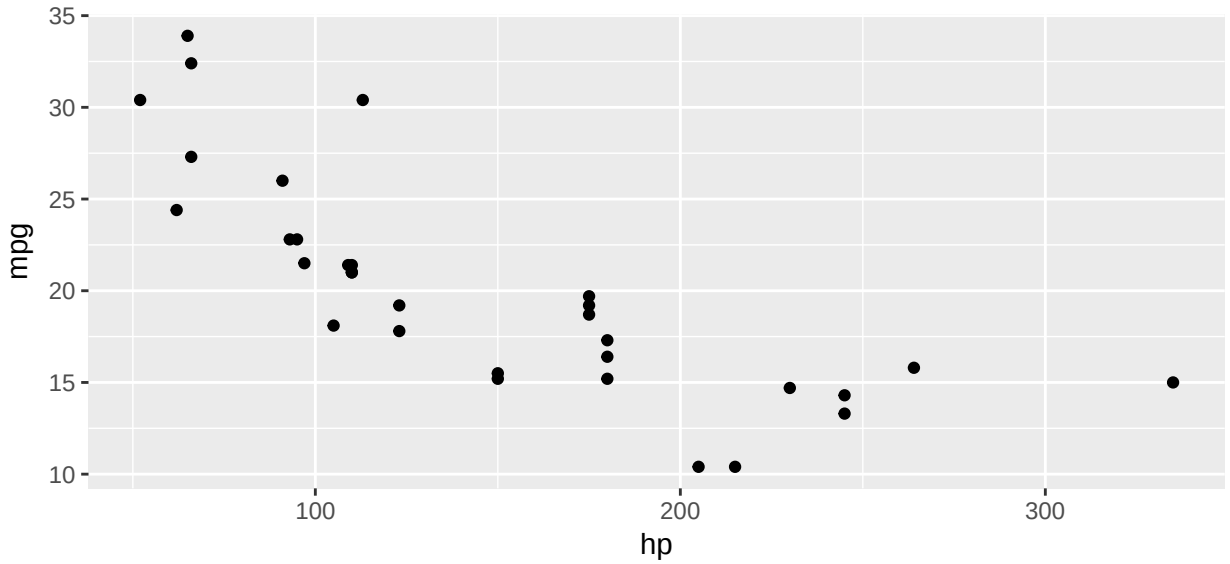
disp vs. mpg

```
ggplot(mtcars, aes(disp, mpg)) + # keep
  geom_point()
```



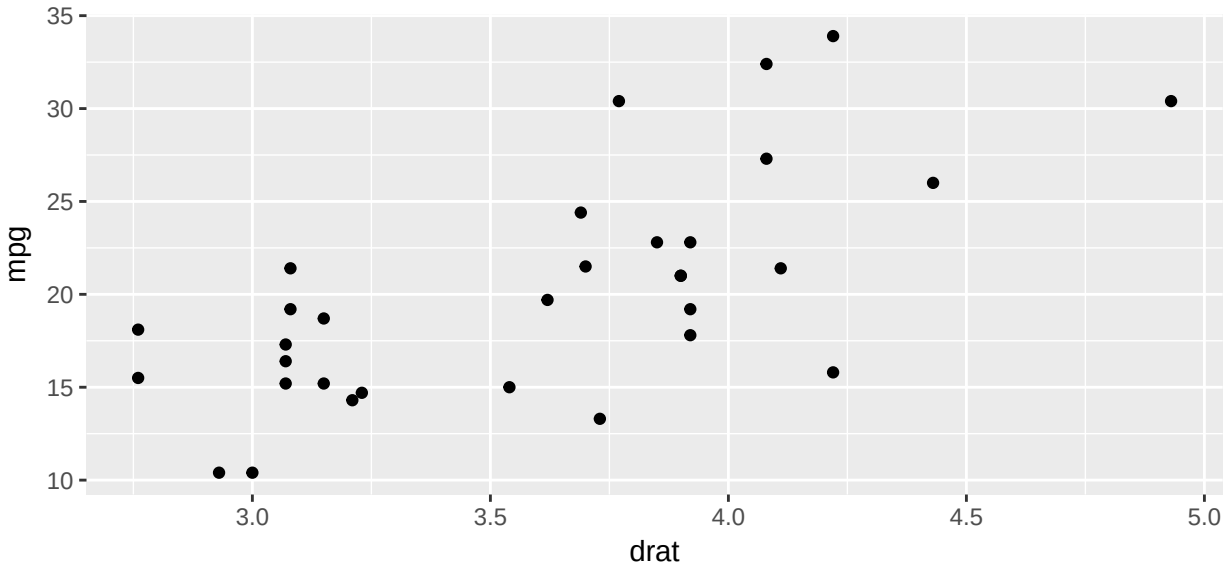
hp vs mpg

```
ggplot(mtcars, aes(hp, mpg)) +
  geom_point()
```



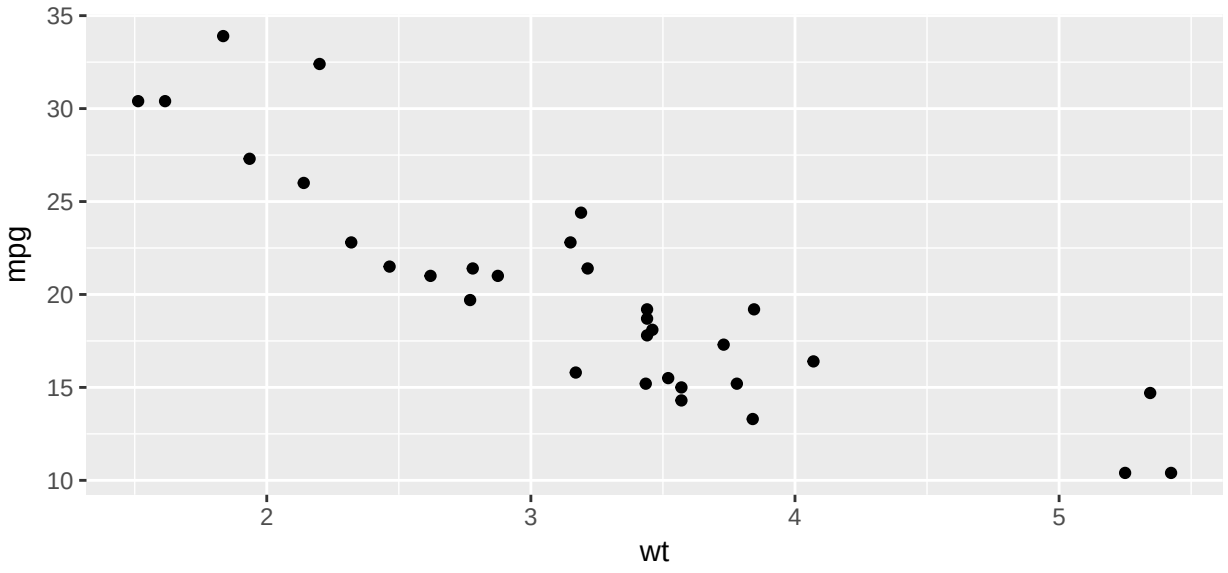
drat vs. mpg

```
ggplot(mtcars, aes(drat, mpg)) +
  geom_point()
```



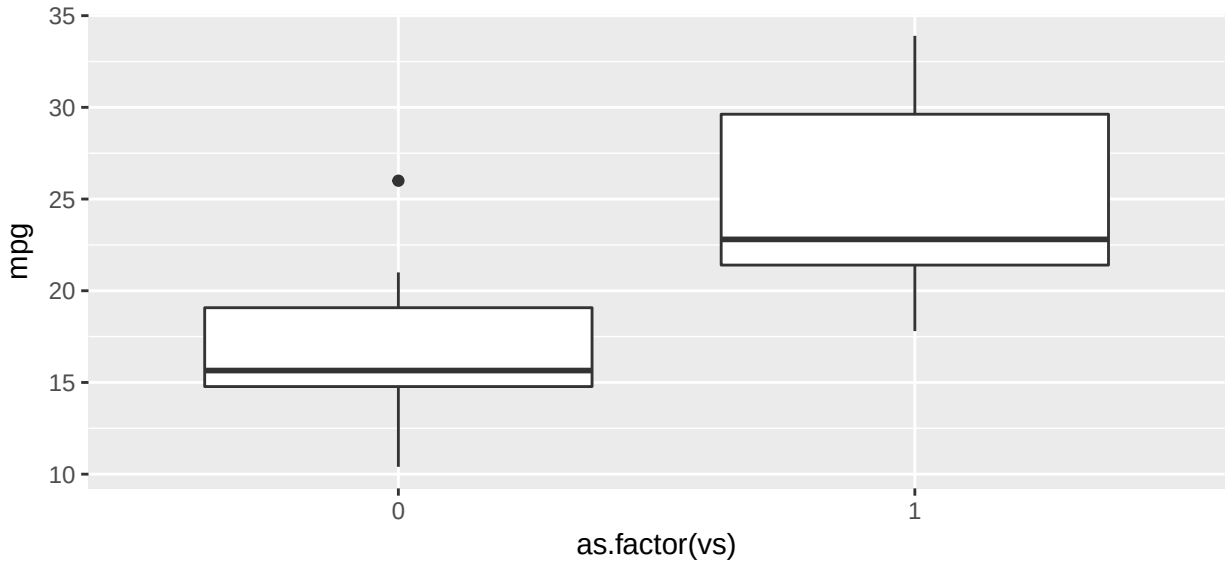
wt vs. mpg

```
ggplot(mtcars, aes(wt, mpg)) +
  geom_point()
```



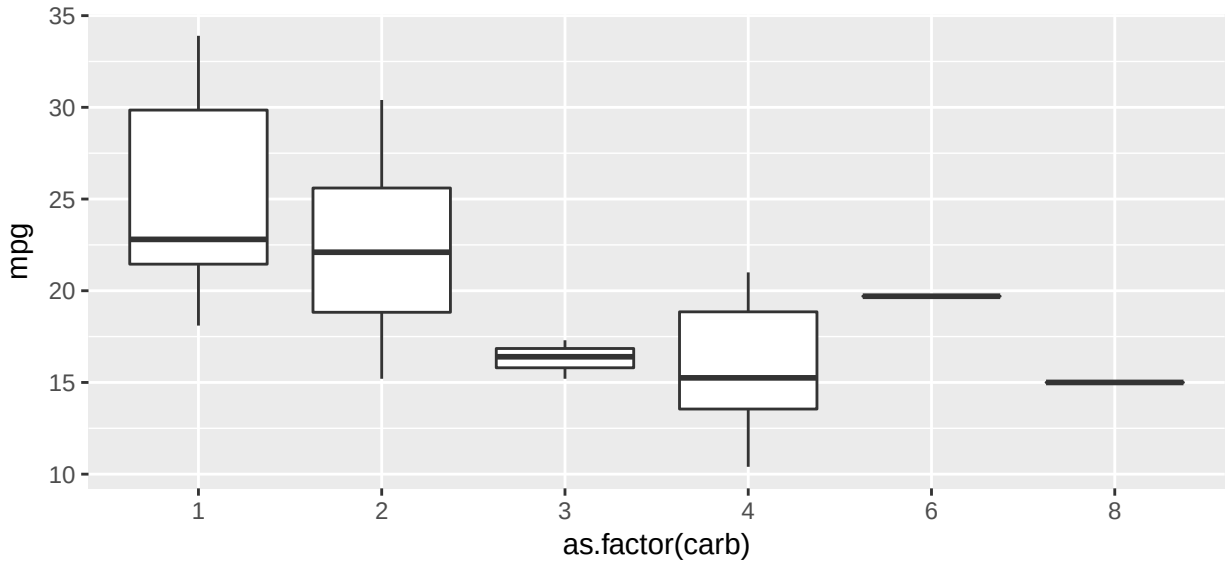
vs vs. mpg

```
ggplot(mtcars, aes(as.factor(vs), mpg)) +
  geom_boxplot()
```



carb vs. mpg

```
ggplot(mtcars, aes(as.factor(carb), mpg)) +
  geom_boxplot()
```



Model

Anova test

```
anova(mod1, mod2, mod3, mod4, mod5, mod6)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ as.factor(am)
## Model 2: mpg ~ as.factor(am) + wt
```

```
## Model 3: mpg ~ as.factor(am) + wt + cyl
## Model 4: mpg ~ as.factor(am) + wt + cyl + disp
## Model 5: mpg ~ as.factor(am) + wt + cyl + carb
## Model 6: mpg ~ as.factor(am) + wt + cyl + carb + vs
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 68.5760 9.202e-09 ***
## 3     28 191.05  1     87.27 13.5226  0.001078 **
## 4     27 188.43  1      2.62  0.4062  0.529500
## 5     27 168.71  0     19.72
## 6     26 167.80  1      0.91  0.1405  0.710870
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model summary

```
summary(mod3)
```

```
##
## Call:
## lm(formula = mpg ~ as.factor(am) + wt + cyl, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1735 -1.5340 -0.5386  1.5864  6.0812
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     39.4179     2.6415  14.923 7.42e-15 ***
## as.factor(am)1   0.1765     1.3045   0.135  0.89334
## wt              -3.1251     0.9109  -3.431  0.00189 **
## cyl             -1.5102     0.4223  -3.576  0.00129 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.612 on 28 degrees of freedom
## Multiple R-squared:  0.8303, Adjusted R-squared:  0.8122
## F-statistic: 45.68 on 3 and 28 DF,  p-value: 6.51e-11
```