Regression models assignment

Jason Collins

2 June 2017

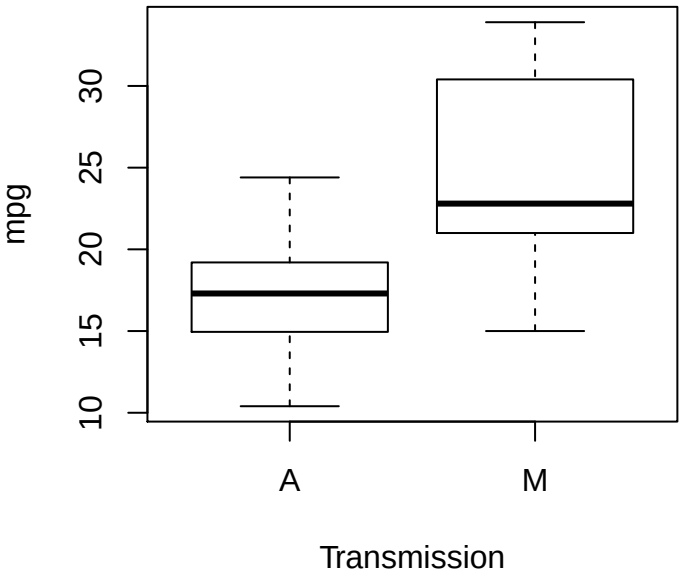
Summary

The mtcars dataset contains 11 variables on 32 cars, including mileage and transmission type. In this dataset, the automatic cars have lower mileage than the cars with manual transmissions, with a mean mileage per gallon of 17.15 and 24.39 respectively. However, this relationship is largely due to the correlation of the transmission type with other car features such as weight and horsepower. A model accounting for these features suggests that transmission type itself has no effect on mileage.

Data exploration

The mtcars dataset comprises 32 observations across 11 variables.

On a raw comparison, there is a clear relationship between the the car transmission type and miles per gallon. This is reflected in the mean mileage, with cars with automatic transmission having a mean mileage per gallon of 17.15 compared to the manual cars' mean mileage of 24.39. A t-test of the difference (-7.24) returns a p value of 0.001, suggesting we reject the null hypothesis of equality.



However, a visual inspection of the relationship between the type of transmission and the other variables in the dataset (see Figure 1 in the Appendix) shows a range of other factors that could affect mileage are correlated with transmission type. This includes autmatic cars tending to be both heavier and have higher horsepower. The model will need to control for these factors.

Modelling

Including all variables in the model will sacrifice too many degrees of freedom (there are only 32 observations), so we will select a first model based on a subset of variables likely to affect mpg - horsepower and weight - in addition to transmission type.

$$mpg = \beta_0 + \beta_1 am + \beta_2 hp + \beta_3 wt$$

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  34.0029     2.6427 12.8669   0.0000
## amM           2.0837     1.3764  1.5139   0.1413
## hp           -0.0375     0.0096 -3.9018   0.0005
## wt           -2.8786     0.9050 -3.1808   0.0036
```

A plot of the residuals (see Figure 2 in the Appendix) suggests that there is an omitted variable, with its shape suggesting a quadratic term. Trialing several variables suggests that a quadratic of weight may provide the most improvement to the model.

$$mpg = \beta_0 + \beta_1 am + \beta_2 hp + \beta_3 wt + \beta_4 wt^2$$

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  47.0354     5.5522  8.4715   0.0000
## amM           0.2788     1.4319  0.1947   0.8471
## hp           -0.0282     0.0094 -2.9869   0.0059
## wt          -10.4460     3.0201 -3.4588   0.0018
## I(wt^2)       0.9475     0.3638  2.6045   0.0148
```

The residual plot for this second model (see Figure 3 in the Appendix) suggests no remaining pattern in the residuals.

An F-test comparing the two models indicates that the addition of the quadratic variable results in an improved fit, with a probability of 0.0148.

This model suggests that mileage decreases with horespower and weight, although in the case of weight this occurs at a decreasing rate (the effect of the positive coefficient on the quadratic term). A 95% confidence interval for the coefficient on the manual transmission dummy is between -2.66 and 3.22, which includes zero. Despite the positive coefficient on transmission, this effect of manual transmission is not significantly different from zero.

Considering the other variables in the dataset, no model with an additional variable passes as F-test at the 5% significance level relative to the second model above. The following table shows the p-value for an F-test of a series of models with each remaining variable added to Model 2. None demonstrate a better fit.

```
##           cyl   disp   drat   qsec     vs   gear  carb
## Pr(>F) 0.4646 0.6657 0.6151 0.0783 0.2978 0.5349 0.758
```

Appendix

Figure 1: Relationship between transmission type and other variables in dataset

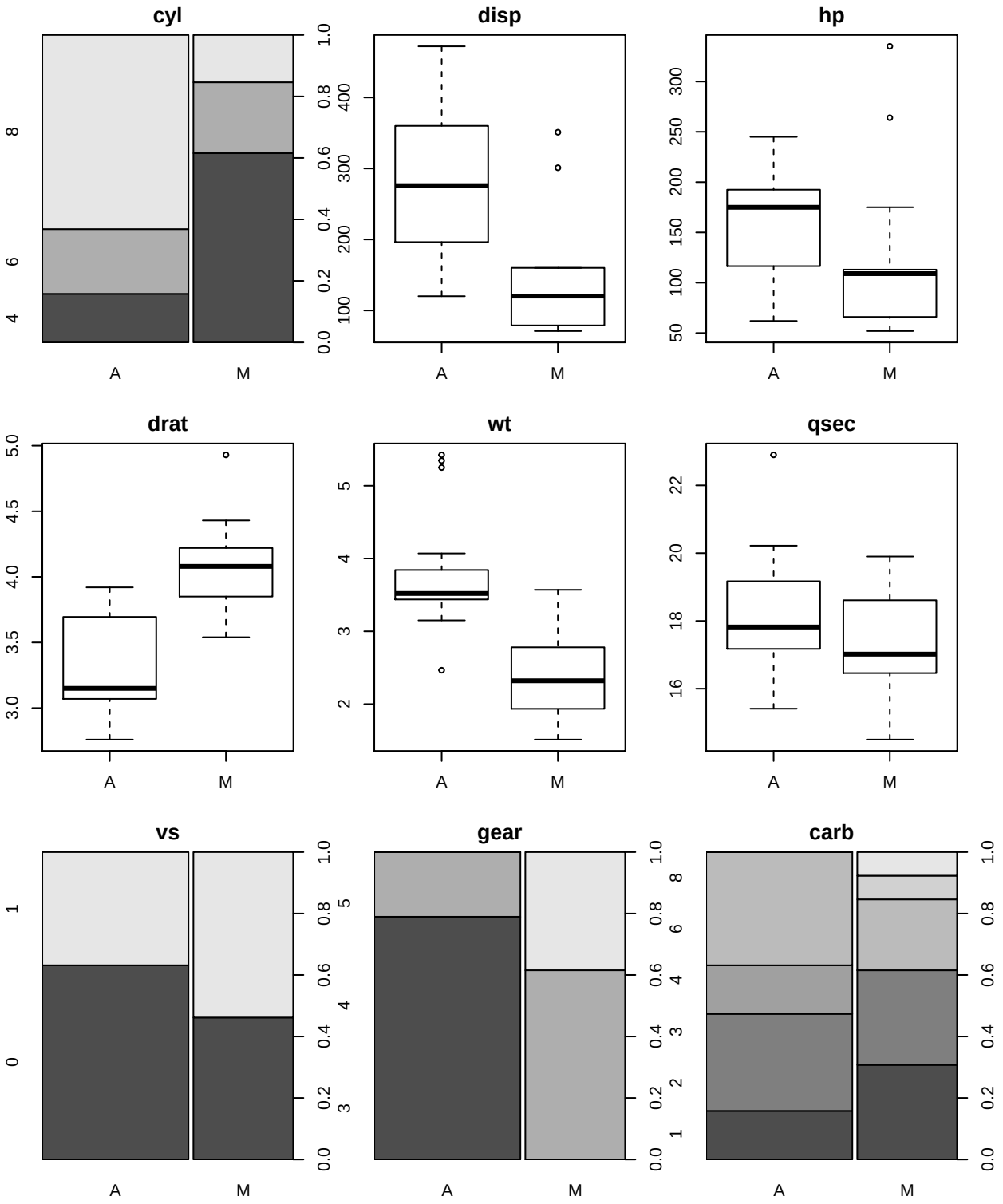


Figure 2: Residual plot for model 1

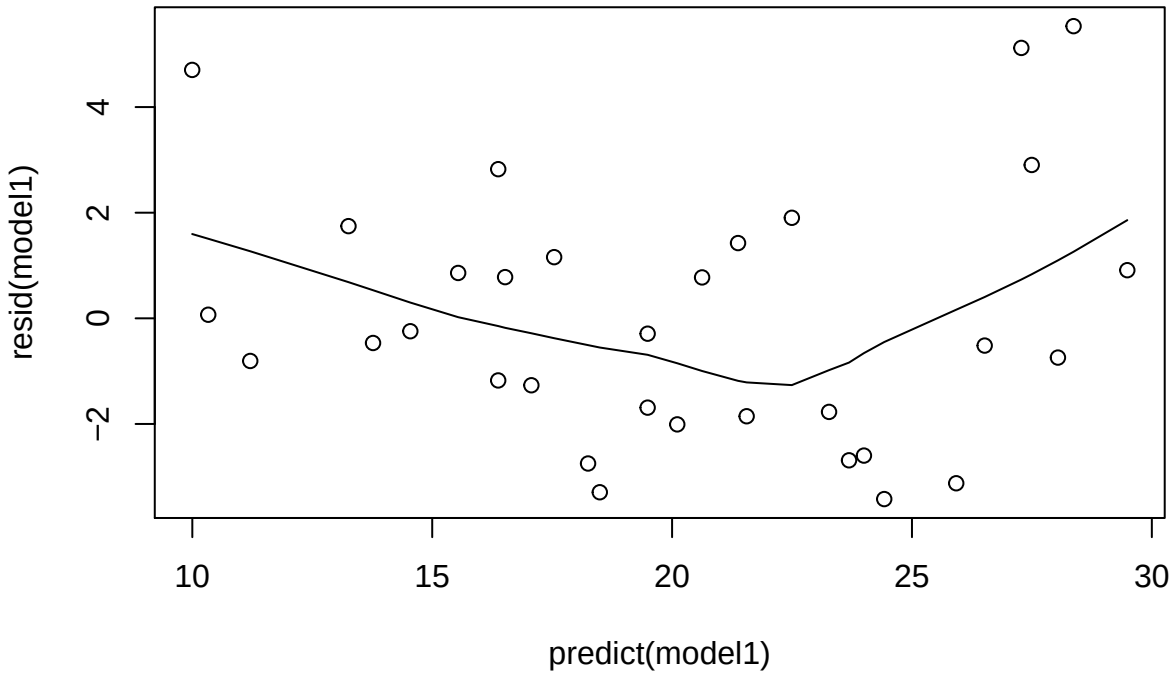


Figure 3: Residual plot for model 2