Regression Models Course Project

Executive Summary

Motor Trends wants to explore the effect of transmission (manual versus automatic) on mileage (miles per gallon, MPG). Thirty-two observations were analyzed from the *mtcars* data and it was found that mileage does not appear to be affected by transmission type. The main factors that affect mileage are car weight and horsepower.

Analysis and Modeling

Exploratory Analysis

The first step is to look at the data to become familiar with it. The command **?mtcars** reveals that:

- the response variable is **mpg**
- the main predictor we are interested in is **am**: Transmission (0 = automatic, 1 = manual)
- The other factors that could possibly be included as predictors in a model are **cyl**: Number of cylinders; **disp**: Displacement (cu.in.); **hp**: Gross horsepower; **drat**: Rear axle ratio; **wt**: Weight (lb/1000); **qsec**: 1/4 mile time; **vs**: V/S; **gear**: Number of forward gears; **carb**: Number of carburetors.

Examining the data we see that here are 32 observations of 11 variables. Several variables (cyl, vs, am, gear, carb) are first transformed into factors rather than left as numerical variables.

Figure 1 (see Appendix) shows scatter plots between all pairs of factors, as well as the correlation between them. The plots show that mpg does have a bivariate relationship with a lot of the other variables; however many of those variables seem to be related to one another. We will have to be careful in choosing a model due to possible confounding factors and multi-collinearity.

Models of the data

The simplest regression model that relates factor *am* to *mpg* is a linear model with only *am* as predictor. This yields the model $mpg = 17.1 + 7.24 \cdot am$. However, this model explains only 36% of the variation present in mpg. This means that transmission, by itself, explains little of the variation, although the coefficient 7.24 seems to indicate that it does. Other factors must be responsible for the unexplained variation, and, once they are taken into account, the coefficient for *am* will likely change.

There are several ways of choosing factors to include in the model. One may proceed by intuition or use a systematic approach. We use the following systematic approach: we will start with a complete model with all factors, and remove the factor with highest p-value (the least statistically significant factor in explaining the variation.) This will be repeated, removing one factor after another, until all factors are statistically significant at the 95% confidence level. This can be done manually or using the step function.

The complete model yields the coefficients shown in Table 1 and explains 89% of the variation present in mpg, a high value. In order to get a more parsimonious model, we can now remove factors. We start by removing *qsec* since it really is not an intrinsic property of a car but the result of other car characteristics. The full model without *qsec* explains 89% of the variation (essentially the same as the full model) so that it can safely be removed.

We now continue by removing the factor with the highest p-value and checking how much variation remains explained. This is done in the following order:

- Remove gear: $R^2 = 89\%$
- Remove carb: $R^2 = 87\%$
- Remove drat: $R^2 = 87\%$
- Remove vs: $R^2 = 87\%$
- Remove disp: $R^2 = 87\%$
- Remove cyl: $R^2 = 84\%$

Other than *am*, there are only 2 predictors left (*hp,wt*) and both are statistically significant at 95% and will not be removed from the model. The model with predictors *am,hp,wt* appears to be the best model so far: it is parsimonious and explains almost as much variation (84%) as a model with all factors (89%).

However, we can still do better with these factors, by including powers of 2 or more to the factors *hp* and *wt*. It turns out that only powers of 2 are statistically significant. The final resulting model is:

$$mpg = 20.2 - 0.276 \cdot am - 12.7 \cdot hp + 5.55 \cdot hp^2 - 19.1 \cdot wt + 6.5 \cdot wt^2$$

The confidence interval for the *am* coefficient is [-3.097, 2.546] at the 95% confidence level. Since the confidence interval includes zero, we cannot reject the null hypothesis that transmission (am) has **no effect** on mpg.

Diagnostics of possible problems with the model

Non-linearity of the response-predictor relationships: The plot of residuals against predicted values (Figure 2) shows no discernible pattern.

Normality of the residuals: The QQ-Plot in Figure 2 shows that the residual quantiles follow the normal quantiles closely, meaning that the residuals are reasonably close to normally distributed.

Non-constant variance of error terms: Figure 3 shows no change in variance of residuals (funnel shape, for example).

Outliers: Any point with a studentized residual over 3 is a suspected outlier that should be investigated. Figure 4 shows that all points have a studentized residual lower than 3.

High-leverage points: We would like to identify high leverage points. Figure 5 shows the "hatvalue" of the data points. The Lotus Europa and the Maserati Bora have the largest leverage on the regression. They are marginally suspect since the hatvalue is slightly over the usual limit of 3 times the mean hatvalue.

Collinearity: None of the Variance Inflation Factor (VIF) scores are greater than 4, suggesting no evidence of collinearity between the factors *am, hp, wt*.

Conclusions

- **Is an automatic or manual transmission better for MPG?** There is no evidence of a difference between the two.
- **Quantify the MPG difference between automatic and manual transmissions**: The confidence interval for the *am* coefficient is [-3.097, 2.546] at the 95% confidence level. Since the confidence interval includes zero, we cannot reject the null hypothesis that transmission (am) has no effect on mpg. Car weight and horsepower are sufficient predictors of mpg.

Appendix

An R Markdown file is available to fully reproduce the results.

Figure 1: Scatterplot and correlation between mtcars variables

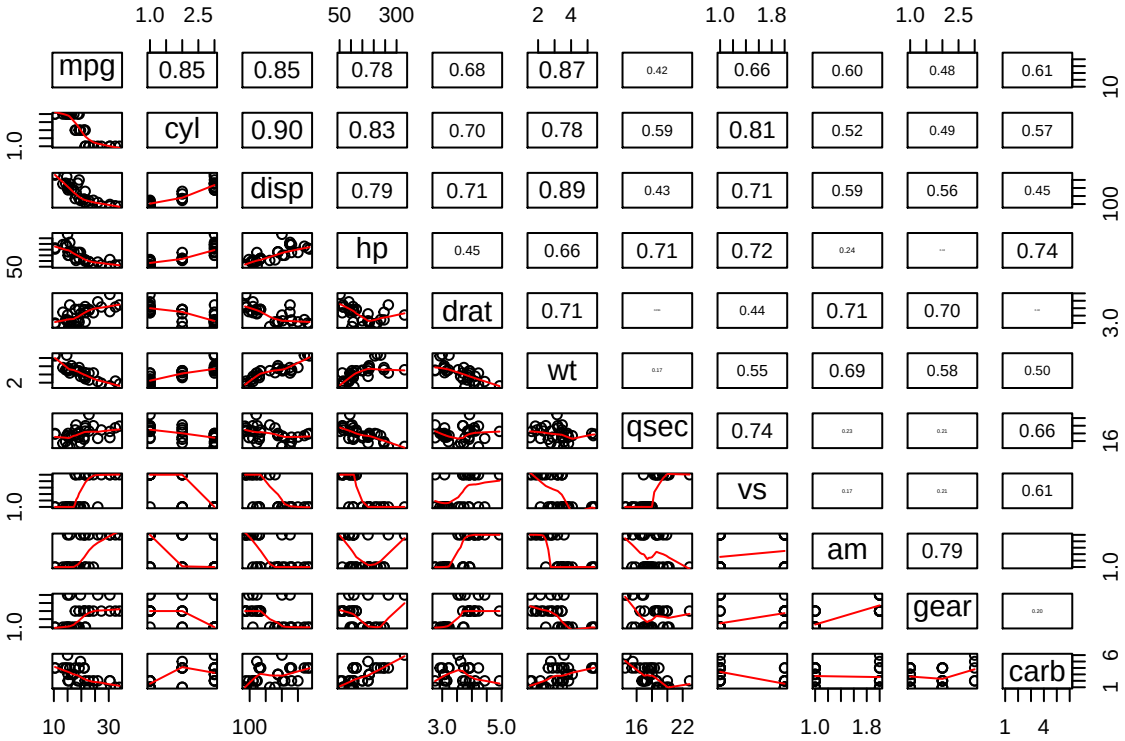

Table 1: coefficients of the initial complete model

```
## (Intercept)        cyl6        cyl8        disp          hp        drat
## 23.87913244 -2.64869528 -0.33616298  0.03554632 -0.07050683  1.18283018
##          wt        qsec         vs1         am1       gear4       gear5
## -4.52977584  0.36784482  1.93085054  1.21211570  1.11435494  2.52839599
##       carb2       carb3       carb4       carb6       carb8
## -0.97935432  2.99963875  1.09142288  4.47756921  7.25041126
```

Fig 2: Normal Q−Q

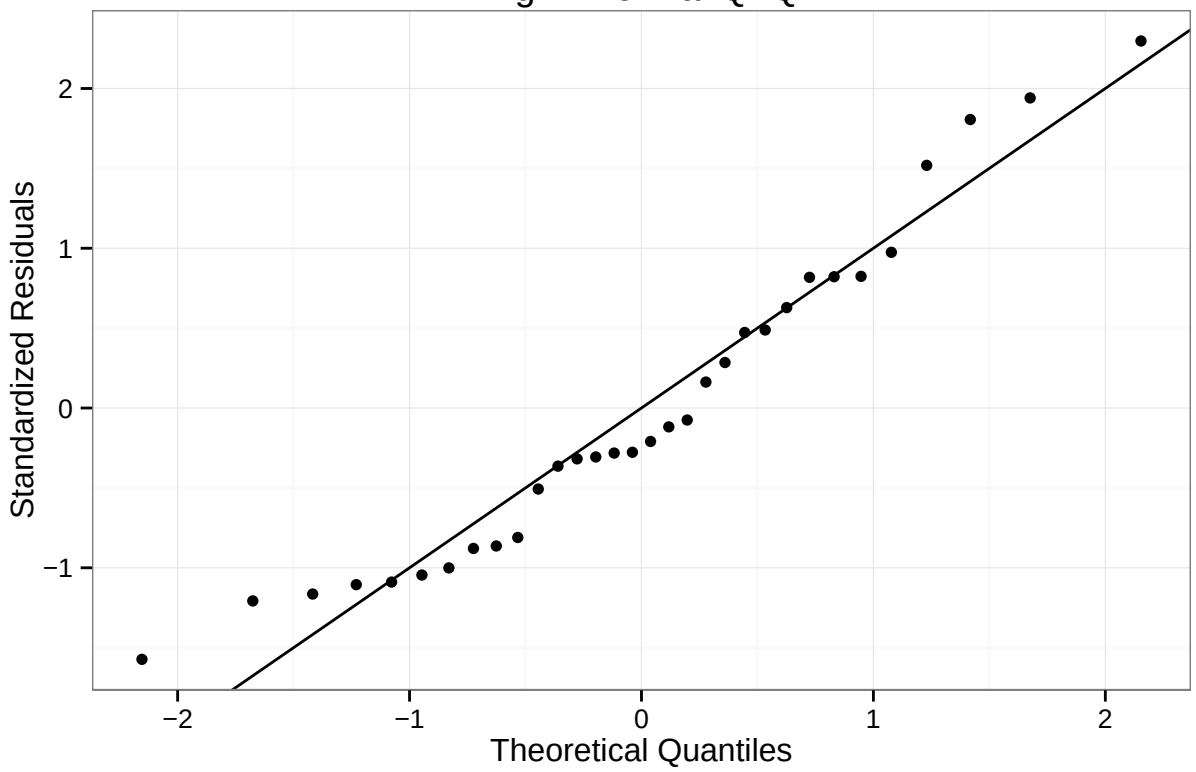

Fig. 3: Residual vs Fitted

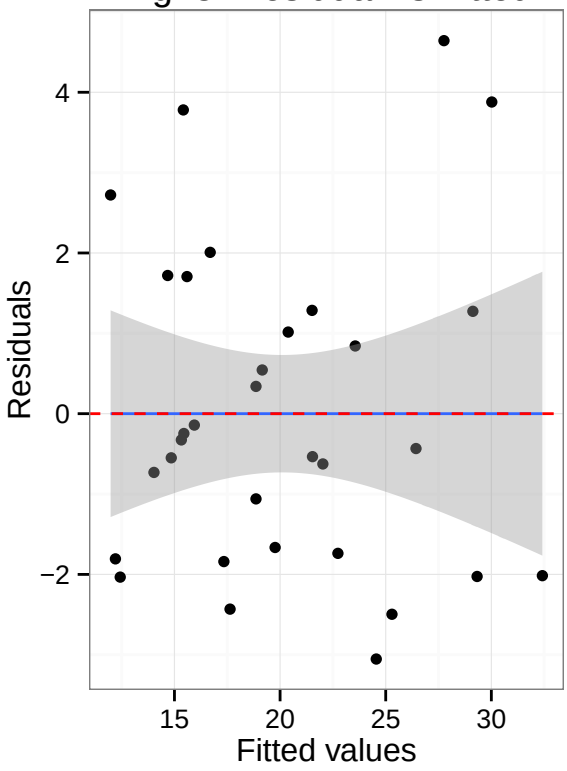

Fig. 4: Studentized Resid. vs Fitted

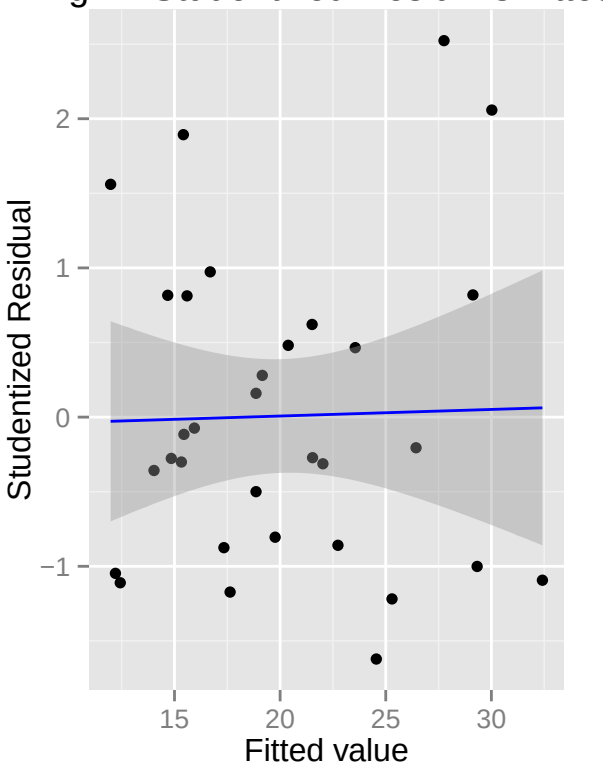

Figure 5: Hat Values of Data Points