Regression Models Course Project - Kevin O'Leary

Executive summary

In this report we are interested in exploring the relationship between a set of variables and miles per gallon (MPG) in the mtcars dataset using regression models and exploratory data analyses. In particular we are interested in whether automatic or manual transmission is better for MPG. This report will quantify the MPG difference between automatic and manual transmissions within a 95% confidence interval.

Before we begin, we first need to manipulate our data somewhat and convert some of our numerical variables to factor variables.

Model selection strategy

One approach to model selection is to build up nested models and use the anova function to see if the contribution of each new variable is significant. According to this method, the most significant additions are the number of cylinders, engine displacement, horsepower and weight (cyl, disp, hp and wt, respectively).

Another way to simplify the variables is to create a maximal model, that includes all the predictors, and remove one term at a time. Here, we can use the step function to select the best model by AIC.

This method results in a model with one fewer variable (disp) than our first model.

Using the anova function we can investigate the difference between our models and find that the p-value isn't significant (0.75) so we can confidently disregard disp at this stage. Just to double-check, we can also use the vif function to measure how much the variance of the estimated regression coefficients are inflated.

Again, disp is highly correlated so we can disregard our first model in favour of of the second. Now that we are happy with our predictors, we have reached the minimal adequate model. Our model has an adjusted R squared of 0.84 which suggests satisfactory accuracy.

Residual Plots and Diagnostics

Figure 2 plots the contribution made by each of our predictors.

Controlling for all other factors, transmission is actually the weakest contributer in our model. Weight and horsepower are the major contributers but most of the horsepower contribution might be explained by weight as an increase in horsepower generally results in an increase in weight.

The residual plots in Figure 3 appear to exhibit homogeneity, normality, and independence. However, we must be careful about putting too much weight on these plots as they are based on quite a small data set.

Residuals vs Fitted

When a linear regression model is suitable for a data set, the residuals are more or less randomly distributed around the 0 line, which appears to be the case here. This suggests that the assumption that the relationship is linear is reasonable. The residuals also appear to roughly form a horizontal band around the 0 line. This suggests that the variances of the error terms are equal.

Normal Q-Q

A quantile normal plot is good for checking normality. The plot shows a little more variance than you would expect in a normal distribution but our model doesn't account for all the variance so this is understandable.

Scale-Location

For a good model, the values should be more or less randomly distributed. Like the first residuals v fitted plot, there is no discernable pattern to the plot.

Residuals vs Leverage

Note that the standardized residuals are more or less centered around zero and reach 2-3 standard deviations away from zero, and symmetrically so about zero, as would be expected for a normal distribution. No point has a large Cook's distance, that is >0.5.

Exploratory Data Analysis

Now that we know our major factors, we can explore the data more closely. Figure 4 shows a continuous and discrete pairwise plot that highlights how our predictors interplay in the data.

There is a negative, almost linear correlation between both weight, horsepower and mpg. That is, the higher the weight and horsepower, the lower the mpg.

Some observations; * Average fuel economy is higher for manual cars * A rise in the number of cylinders corresponds to lower mpg * Weight is positively correlated with horsepower * Horsepower appears to be greater for automatic cars, however, manual cars have high outliers

---

Conclusions

Figure 1 is a box plot of transmission and mpg which suggests that manual cars add 7.2 mpg to fuel economy but the most obvious vehicle design features affecting fuel economy such as vehicle weight are not accounted for.

Our new model shows manual transmission resulting in a predicted 1.8 increase in mpg when compared to automatic, holding all other variables constant. Our computed 95% confidence interval gives manual transmission an effect of between -1 and 4.7 mpg.

A key observation from our data analysis is that automatic cars tend to have more cylinders and manual tend to have fewer (Figure 5) and an increase in cylinders is correlated with an increase in weight which in turn is negatively correlated with mpg. Whatever tendency automatic cars have for being heavier than manuals in general is compounded by this bias in the data. We can't then answer "Is an automatic or manual transmission better for MPG" without limiting the statement to our data and not the general population.

Table 1: Model Summary

Coefficients	Estimate.	Std. Error	t value	Pr(>|t|)
Intercept	33.70832390	2.60488618	12.940421	7.733392e-13
Manual	1.80921138	1.39630450	1.295714	2.064597e-01
6 Cylinder	-3.03134449	1.40728351	-2.154040	4.068272e-02
8 Cylinder	-2.16367532	2.28425172	-0.947214	3.522509e-01
Horsepower	-0.03210943	0.01369257	-2.345025	2.693461e-02
Weight	-2.49682942	0.88558779	-2.819404	9.081408e-03

Bob	BOb
Residual standard error:	2.41 on 26 degrees of freedom
Multiple R-squared:	0.8659
Adjusted R-squared:	0.8401
F-statistic:	33.57 on 5 and 26 DF
p-value:	1.506e-10

Appendix: Figures

Figure 1: Manual v Automatic Transmission boxplot:

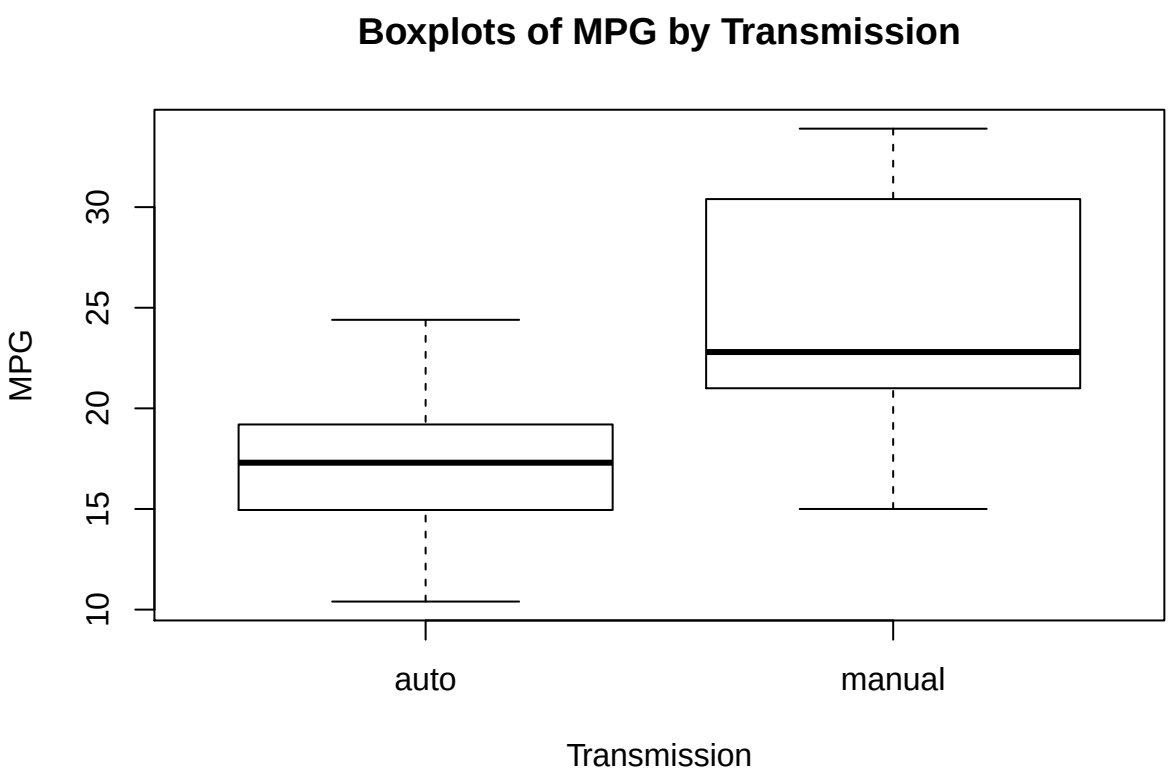


Figure 2: Relationship between `mpg` and `disp` separated by `am`:

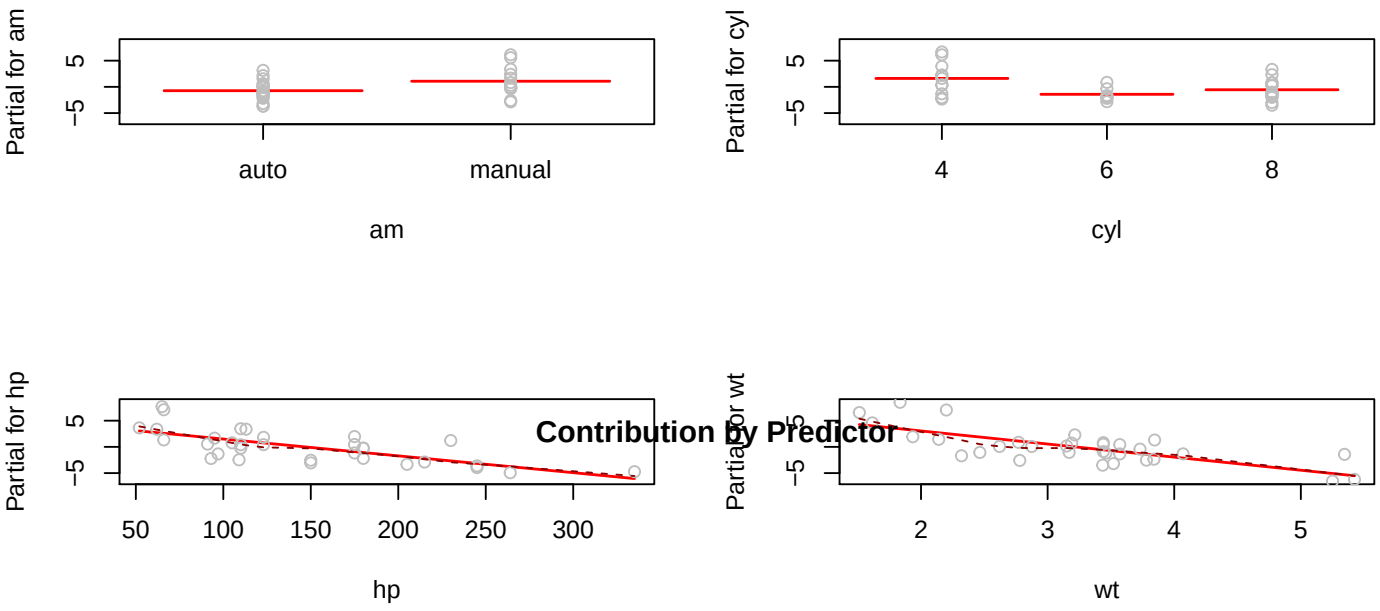
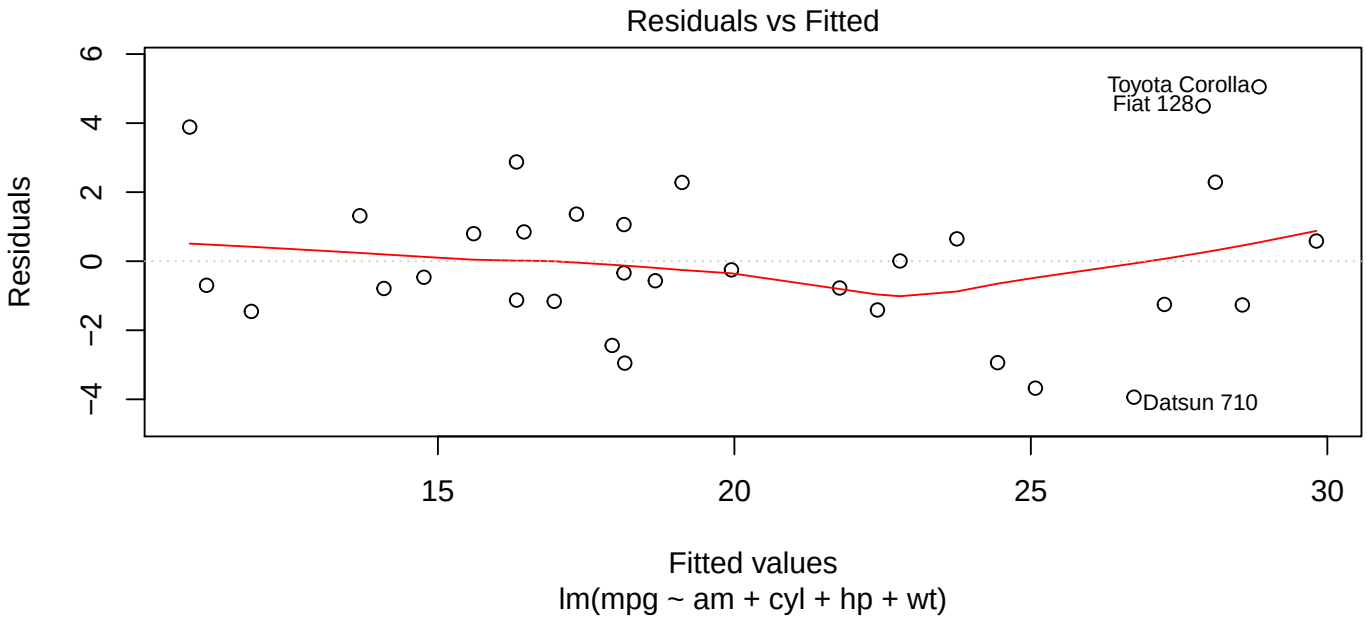
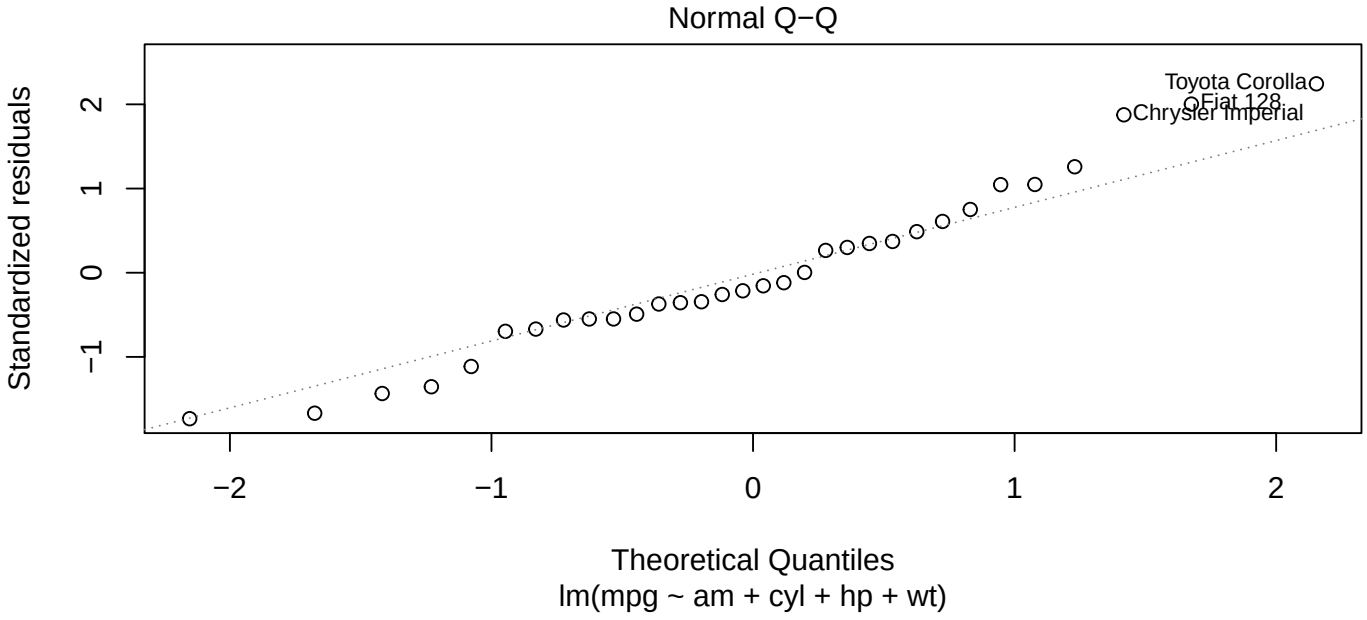
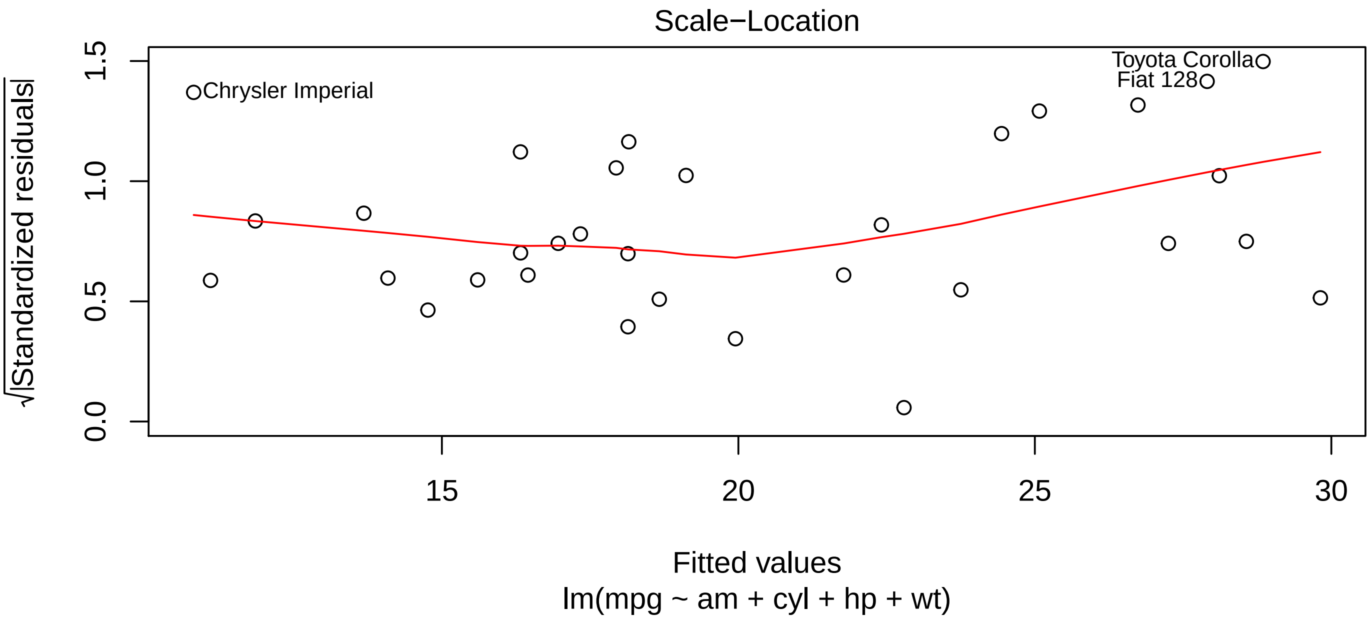


Figure 3: Residual plots for the chosen model:


Residuals vs Fitted
Residuals
Fitted values
lm(mpg ~ am + cyl + hp + wt)
Toyota Corolla
Fiat 128
Datsun 710


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(mpg ~ am + cyl + hp + wt)
Toyota Corolla
Fiat 128
Chrysler Imperial



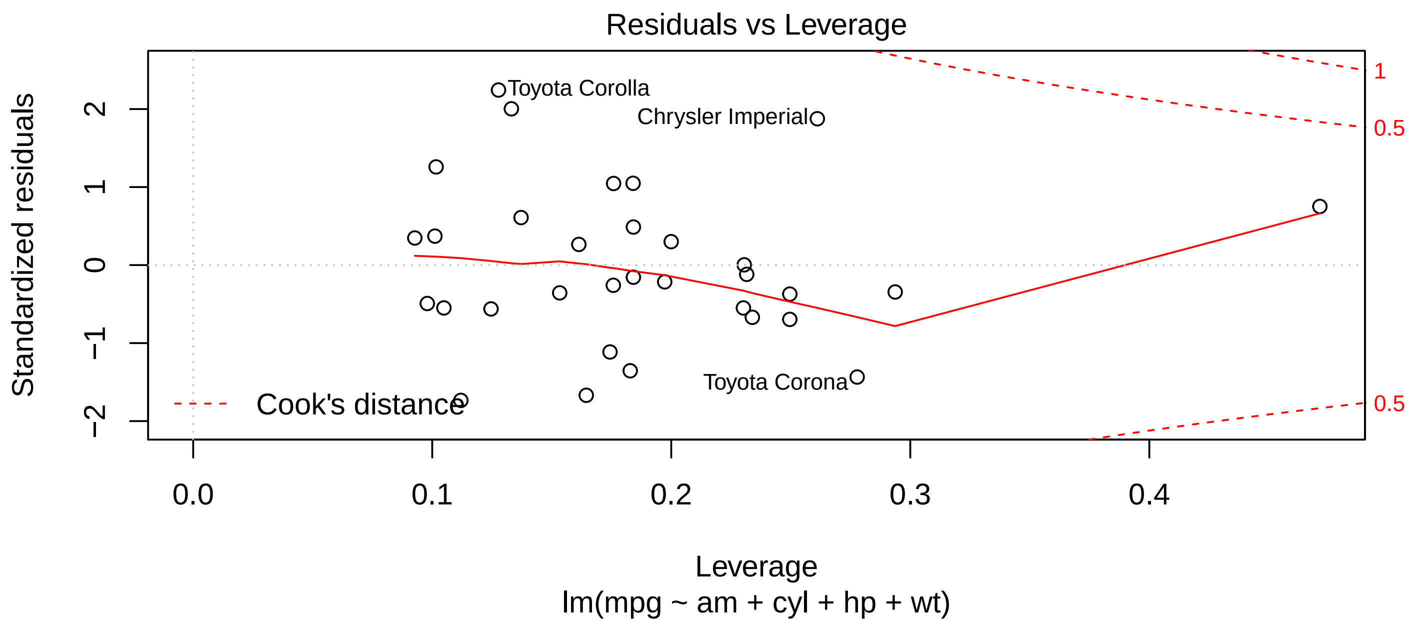


Figure 4: Relationship between `mpg` and `wt` separated by `am`:

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

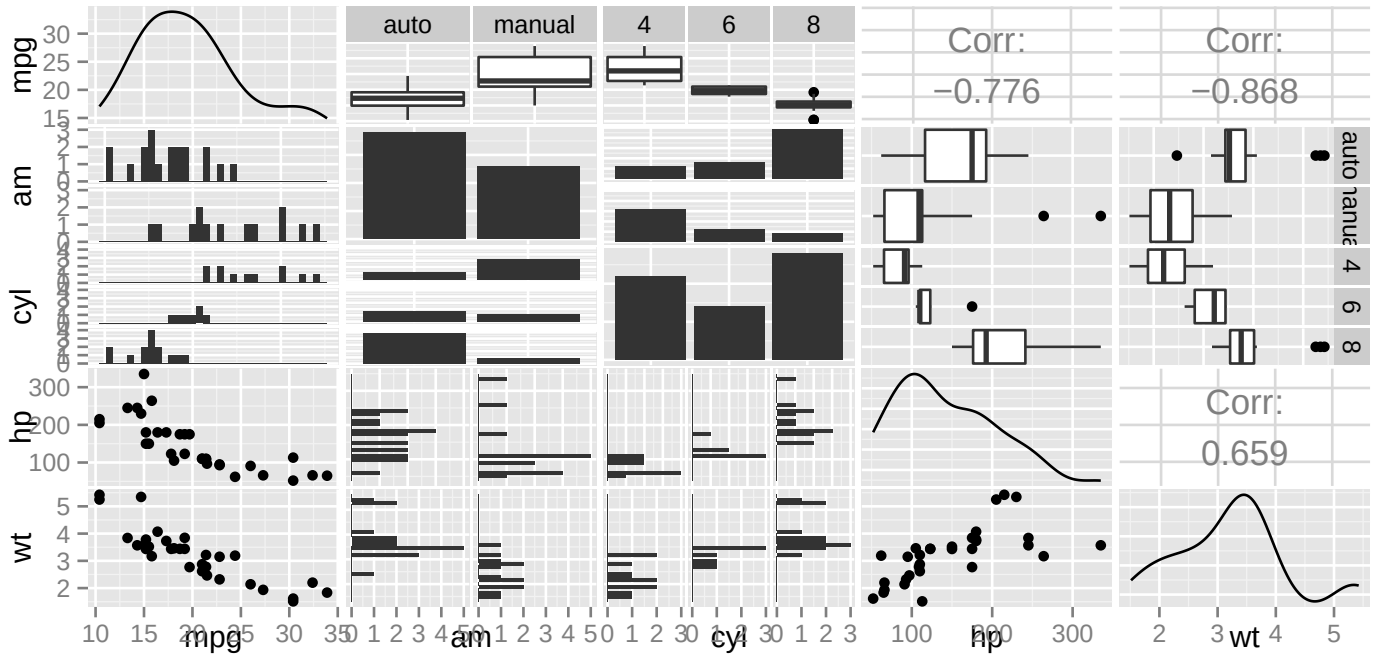


Figure 5: Residual plot and plot of distribution for model `mpg` ~ `am`: