

Regression Models: Assignment

ND

24/10/2014

Executive Summary

When considering fuel efficiency (mpg) and transmission type (automatic or manual) alone, i.e. without accounting for other variables, the manual transmission cars in the `mtcars` data set are, on average, 7 mpg more efficient than the automatics. This does not necessarily mean manual transmission cars are more efficient as there are other variables such as weight and power that must be accounted for.

It's worth bearing in mind that the cars in the data set are 1973-1974 models. Since automatic transmission technology will have improved since then, a similar analysis on modern cars may show less of a difference between transmission types.

Introduction

Much of the R code in this document has been hidden to aid readability. The full R markdown source for this document can be found in my [github repository](#). The first step in the analysis is to load the data, convert binary or categorical variables to factor variables, and then assign meaningful factor names.

Exploratory Analysis

Once the data has been loaded the data can be inspected for patterns. The pairwise relationship between the data variables is shown in the scatterplot matrix (appendix 1, fig. 2). The relationship of interest is that between fuel efficiency (mpg) and transmission type (am). The boxplot (fig. 1) shows a clear difference between the median mpg for automatic and manual transmission types.

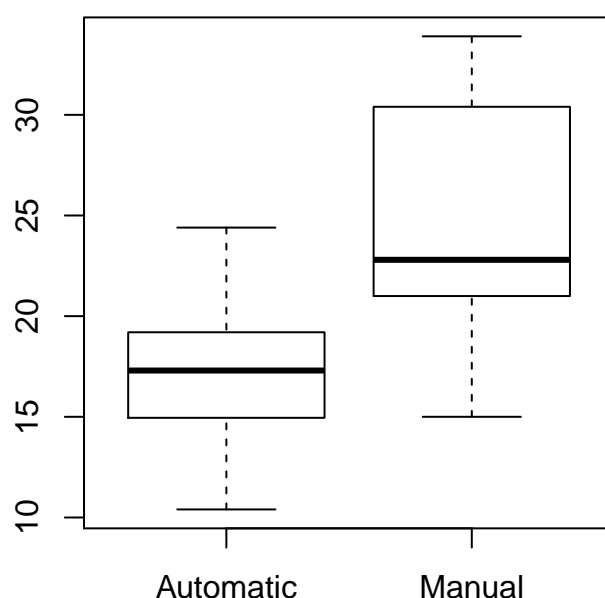


Figure 1: Boxplot for MPG by transmission type

The normal q-q plots (appendix 1, fig. 3) show that the mpg of automatic and manual transmission cars in the data set is approximately normally distributed. The Welch Two Sample t-test suggests a significant difference between the means of the two groups ($P = 0.0014$, two sided).

Simple Model

The relationship shown above can be expressed as a simple linear model $\text{mpg} \sim \text{am}$. This model results in an intercept of 17.147, which corresponds to the mean mpg of the reference group (automatic). The second coefficient, 7.245 corresponds to the increase in mean of the manual group compared to the reference group. This result is not surprising considering the obvious difference between the boxplots and the statistically significant difference between the two means. The simple model however only accounts for 36.0% of the variation in mpg, and has a residual sum of squares of 720.90. This is not particularly surprising as the scatterplot matrix (appendix 1, fig. 2) shows mpg to be influenced by a range of variables besides transmission type.

Full Model

A linear model of the form $\text{mpg} \sim .$ adds terms for every variable in the data set. After throwing every predictor variable into the model, it now accounts for 86.9% of the variation in mpg. When considering multiple variate regression the adjusted R^2 can be used to provide a better indication of explanatory power (as R^2 will increase for every predictor added to the model, regardless of the model's explanatory power). The adjusted R^2 for the full model is 0.807. The values of the beta coefficients represent the change in response variable to a unit change in the relevant predictor variable, given the other variables are held constant. The magnitude of the values themselves depend on the measurement unit and therefore are not a reliable indication of correlation strength. The t-values and associated p-values indicate the significance of the effect. The most significant coefficient is weight (wt), with a p-value of 0.0633, followed by transmission type (am), acceleration (qsec) and engine power (hp).

Stepwise Regression

Stepwise regression uses an automatic procedure for choosing predictor variables on the basis of AIC, which evaluates the trade-off between model fit and complexity. The forwards stepwise regression starts with the simple model and results in the model $\text{mpg} \sim \text{am} + \text{hp} + \text{wt} + \text{qsec}$. The adjusted R^2 for the forward stepwise model is 0.837. The backwards stepwise regression starts with the full model and results in the model $\text{mpg} \sim \text{wt} + \text{qsec} + \text{am}$. The adjusted R^2 for the backward stepwise model is 0.834. The model $\text{mpg} \sim \text{am} + \text{hp} + \text{wt} + \text{qsec}$, from the forward stepwise has the highest adjusted R^2 of the models considered so far. The four predictor variables in this model correspond to the four most significant coefficients in the full model.

Interaction Terms

The scatterplot matrix (appendix 1, fig. 2) shows that many of the data set variables are correlated. Appendix 1, fig. 4 illustrates the individual relationships between the three most significant additional variables. The weight plot shows that the automatic cars tend to be heavier than the manual cars. This may be a reason why the difference between automatic and manual varies according to weight, i.e. there is an interaction between weight and transmission type. The quarter mile time plot shows that automatic cars have a tendency towards lower mpg than the manual cars. The difference increases as quarter mile time increases, i.e. there is also an interaction between quarter mile time and transmission type. The horsepower plot shows that across the range of horsepower, the automatic cars have a consistent tendency towards lower mpg than the manual cars, i.e. there is no interaction between transmission type and horsepower.

Model Selection

The final model was selected on the basis of the most significant variables in the full model, as confirmed by the stepwise regression. Since there is a clear interaction between weight and transmission type an

interaction term `am:wt` was added to the model. This reduced the significance of the `hp` variable, which was removed from the model. The final model `mpg ~ am + wt + qsec + am:wt` has an adjusted R^2 of 0.880 and a residual sum of squares of 117.28.

Interpretation

Diagnostics

Diagnostic plots (appendix 1, fig. 5)

Appendix 1: Plots

Scatterplot matrix

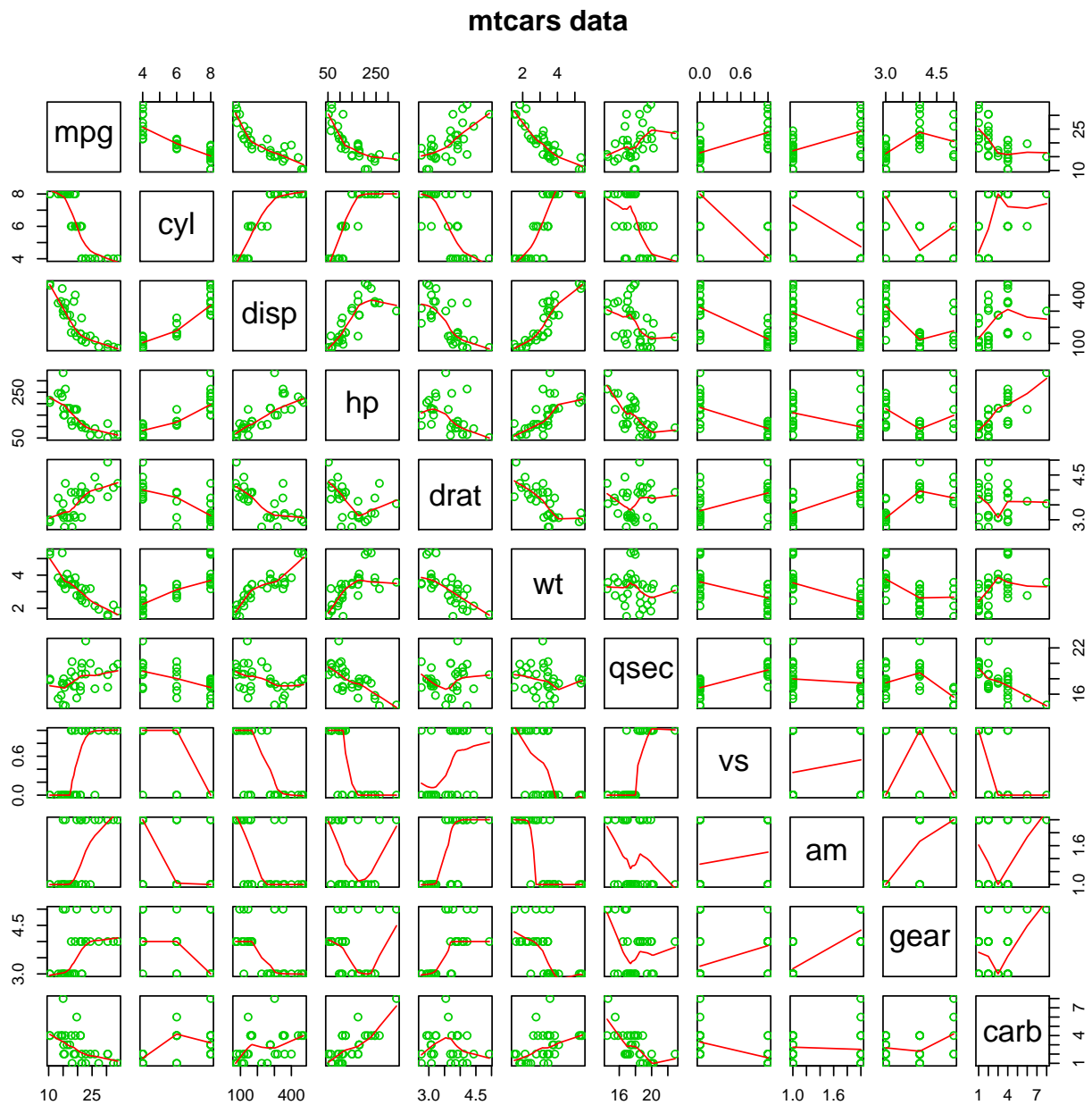


Figure 2: Scatterplot matrix for the mtcars data set

Normal q-q plot

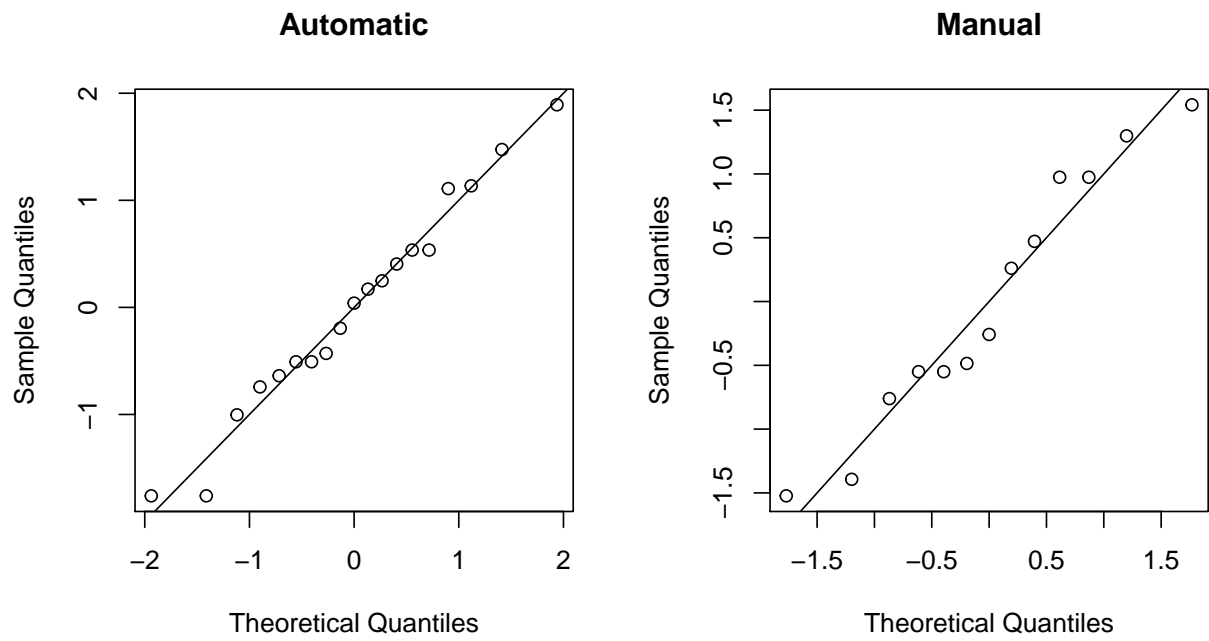


Figure 3: Normal q-q plot for MPG by transmission type

Individual interactions

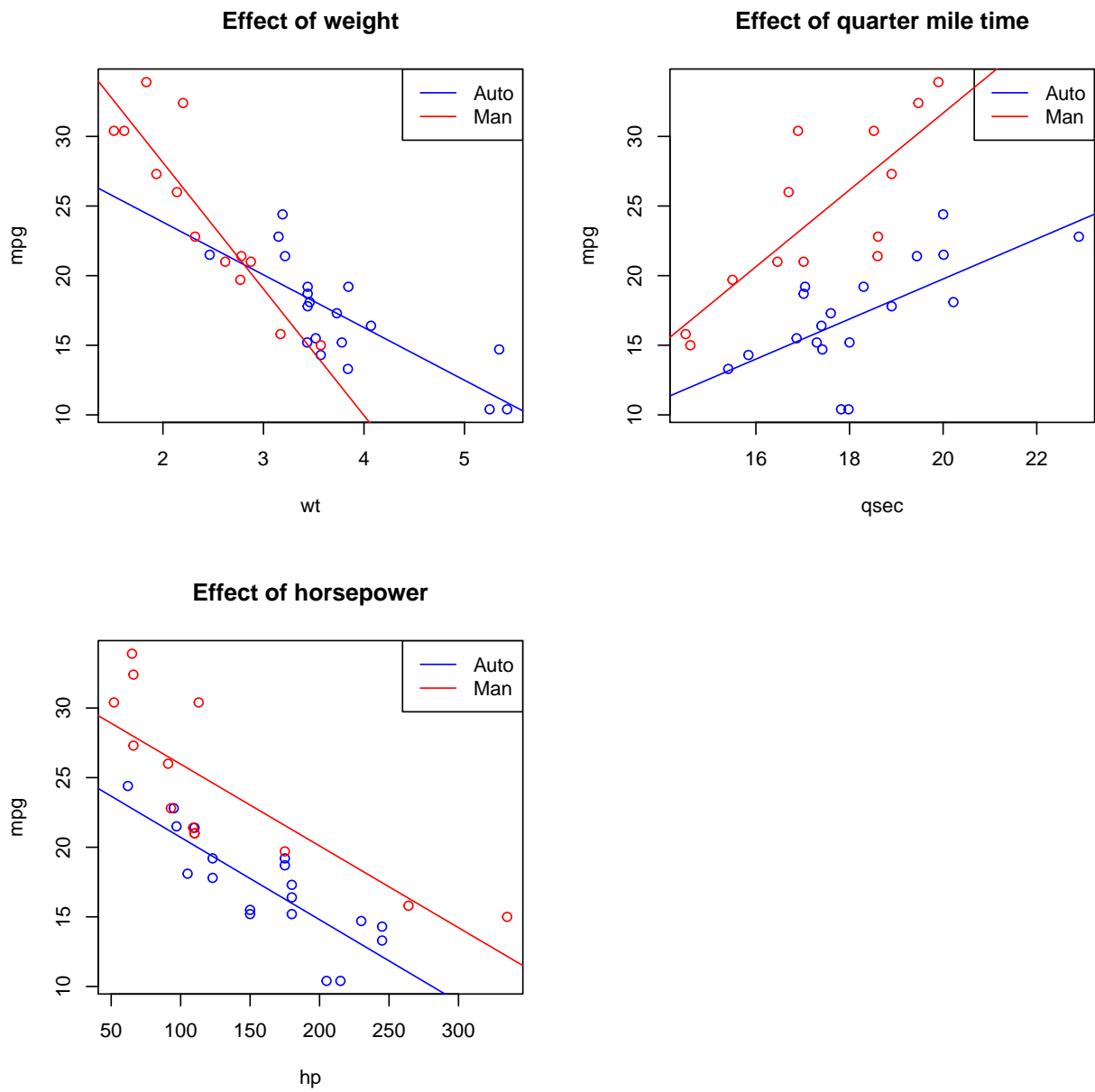


Figure 4: Individual interactions

Diagnostic plots

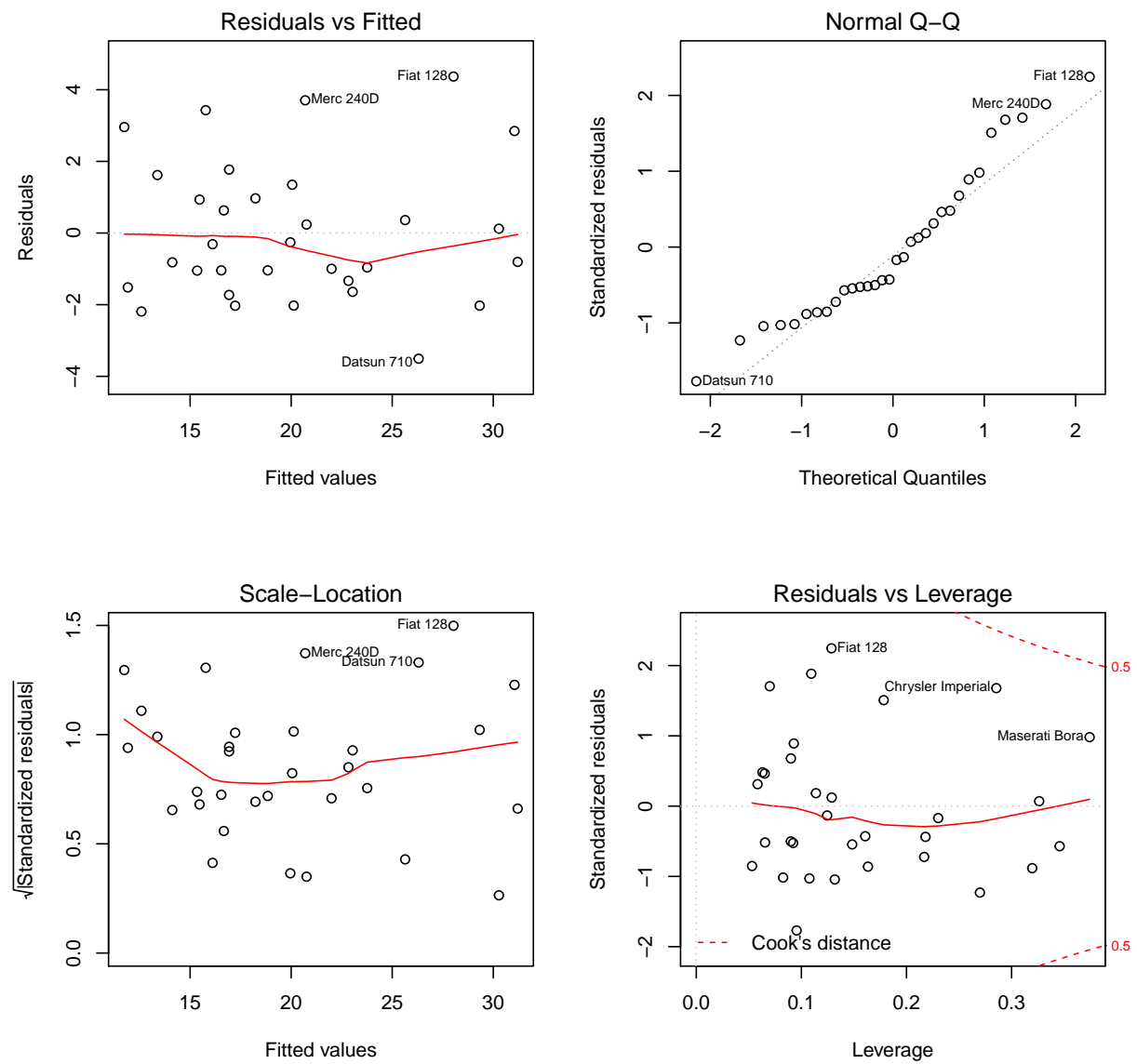


Figure 5: Model diagnostics plots

Appendix 2: Model Summaries

```
summary(mSimple)
```

```
##
## Call:
## lm(formula = mpg ~ 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 ***
## amManual       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
```

```
summary(mFull)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs          0.31776    2.10451   0.151   0.8814
## amManual     2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869, Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF, p-value: 3.793e-07
```

```
summary(mStepF)
```



```
##
## Call:
## lm(formula = mpg ~ am + hp + wt + qsec, data = mtcars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.4975 -1.5902 -0.1122  1.1795  4.5404
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.44019     9.31887   1.871  0.07215 .
## amManual      2.92550     1.39715   2.094  0.04579 *
## hp           -0.01765     0.01415  -1.247  0.22309
## wt           -3.23810     0.88990  -3.639  0.00114 **
## qsec          0.81060     0.43887   1.847  0.07573 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.435 on 27 degrees of freedom
## Multiple R-squared:  0.8579, Adjusted R-squared:  0.8368
## F-statistic: 40.74 on 4 and 27 DF,  p-value: 4.589e-11
```

```
summary(mStepB)
```

```
##
## 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 ***
## amManual      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
```

```
summary(mInt)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + qsec + am:wt, data = mtcars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.5076 -1.3801 -0.5588  1.0630  4.3684
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)    9.723      5.899    1.648 0.110893
## amManual      14.079      3.435    4.099 0.000341 ***
## wt           -2.937      0.666   -4.409 0.000149 ***
## qsec          1.017      0.252    4.035 0.000403 ***
## amManual:wt   -4.141      1.197   -3.460 0.001809 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.084 on 27 degrees of freedom
## Multiple R-squared:  0.8959, Adjusted R-squared:  0.8804
## F-statistic: 58.06 on 4 and 27 DF,  p-value: 7.168e-13
```

```
anova(mInt)
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value    Pr(>F)
## am          1 405.15  405.15   93.276 2.980e-10 ***
## wt          1 442.58  442.58  101.892 1.161e-10 ***
## qsec        1 109.03  109.03   25.102 2.963e-05 ***
## am:wt        1  52.01   52.01   11.974 0.001809 **
## Residuals  27 117.28    4.34
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```