Regression Models Assignment

Peder Lewenhaupt

Executive Summary

The following is an analysis of how fuel efficiency is affected by transmission type, and possibly other variables, in the mtcars data set. The analysis was undertaken using multivariable regression and associated statistical methods. The final model chosen explained 85% of the total variance, and included transmission, weight and number of cylinders as regressors. It showed that manual transmissions are more efficient in lighter cars. However, that efficiency is reduced as the cars get heavier. All other variables being constant, a manual car gets 11.5 more mpg than an automatic. However, due to an interaction effect between transmission and weight, this difference is reduced with increasing weight. At 2840 lbs, the difference in efficiency is 0.

Loading and exploring the data

The data used came from the mtcars data set, available in R. It was loaded using the following code:

```
data(mtcars)
```

After the data was loaded, the next step consisted of exploring it.The more basic aspects of this (using the head, tail, summary and str functions), will not be included here for the sake of brevity. Neither will the code for the data transformations, where factor variables were converted to the correct format.

A boxplot was drawn to explore the difference in mpg between transmission types (see Appendix, figure 1). The boxplot showed a difference in mpg between manual and automatic. To further explore this relationship, a regression model with only transmission type (am) and mpg was built.

```
fit <- lm(mtcars$mpg~mtcars$am)
summary(fit)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## mtcars$am1   7.244939   1.764422  4.106127 2.850207e-04
```

The model shows a significant effect of transmission type on mpg. However, the model also shows that only 36 percent (R-squared = 0.3598) of the variance is explained by transmission type.

Development of multivariable model

The development of a more accurate model started with thinking about the different variables that intuitively would be influential on fuel efficiency, based on physics and vehicle mechanics. The weight of the vehicle should influence fuel efficiency (more weight needs more energy to move it). Also, the number of cylinders and the rear axle ratio should have an effect on fuel efficiency (based on vehicle mechanics). However, the degree of influence is more uncertain, particularly for axle ratio. The number of carburetors and forward gears might also have an impact, but this would be dependent on driving styles and other factors, so these variables were regarded as less certain to influence fuel efficiency and not included in the model.

The remaining variables (horsepower, 1/4 time, V or straight cylinder arrangement) were also considered. All of these should be dependent on a host of mechanical factors and were thus not included in the model.

Based on the above reasoning, the following variables were included in the first model: transmission type, weight, number of cylinders and rear axle ratio.

```
fit <- lm(mpg~am+wt+cyl+drat, data=mtcars)
summary(fit)$coef
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 34.7974227  6.5735684  5.2935363 1.551227e-05
## am1          0.2757821  1.5036306  0.1834108 8.558985e-01
## wt          -3.1655913  0.9292258 -3.4066976 2.148190e-03
## cyl6        -4.3374275  1.5072743 -2.8776631 7.903955e-03
## cyl8        -6.2054068  1.8581533 -3.3395559 2.543597e-03
## drat        -0.2698849  1.5296489 -0.1764358 8.613187e-01
```

The summary shows that both transmission type and rear axle ratio show non-significant p-values, with rear axle ratio having the highest value. To see if the model could be further improved, the least significant variable, rear axle ratio, was removed. The new model looks as follows:

```
fit <- lm(mpg~am+wt+cyl, data=mtcars)
summary(fit)$coef
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 33.7535920  2.8134831 11.9970836 2.495549e-12
## am1          0.1501031  1.3002231  0.1154441 9.089474e-01
## wt          -3.1495978  0.9080495 -3.4685309 1.770987e-03
## cyl6        -4.2573185  1.4112394 -3.0167231 5.514697e-03
## cyl8        -6.0791189  1.6837131 -3.6105432 1.227964e-03
```

Interestingly, the two expanded models show non-significant p-values for transmission type. The adding of weight and number of cylinders to the model seems to have removed the effect of transmission type.

To explore this further, three plots were drawn, showing the distribution of weight in relation to mpg, and number of cylinders in relation to mpg, respectively (see Appendix, figure 2).

The plots show that: 1. A majority of the heavier cars (+3000 lbs) have automatic transmission, with the opposite of course being true for manual. 2. Cars with more cylinders tend to be heavier, which is perfectly logical, considering engine size, etc. 3. As hypothesized above, there is a linear relationship between weight and mpg.

These three statements led to the exploration of interaction effects between the three regressor variables in the model. Four different models were tested:

Interaction between transmission type and number of cylinder, plus weight alone.

```
intAmCyl <- lm(mtcars$mpg~mtcars$am*mtcars$cyl+mtcars$wt)
```

Interaction between transmission type and weight, plus number of cylinders alone.

```
intAmWeight <- lm(mtcars$mpg~mtcars$am*mtcars$wt+mtcars$cyl)
```

Interaction between weight and cylinders, plus transmission type alone.

```
intWeightCyl <- lm(mtcars$mpg~mtcars$wt*mtcars$cyl+mtcars$am)
```

Interaction between all terms.

```
intWeightCylAm <- lm(mtcars$mpg~mtcars$wt*mtcars$cyl*mtcars$am)
```

```
summary(intAmWeight)$coef
```

```
##                       Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)          29.774836  2.8403415 10.482836 7.870715e-11
## mtcars$am1           11.568790  4.0877912  2.830083 8.853842e-03
## mtcars$wt            -2.398713  0.8439884 -2.842116 8.603904e-03
```

```
## mtcars$cyl6          -2.709777  1.3573517 -1.996370 5.646509e-02
## mtcars$cyl8          -4.776110  1.5558306 -3.069814 4.964603e-03
## mtcars$am1:mtcars$wt -4.067981  1.3974151 -2.911075 7.295503e-03
```

```
summary(intAmWeight)$adj.r.squared
```

```
## [1] 0.8538884
```

The best model, both in terms of fit and significance of coefficients, is the model that includes transmission type, weight, number of cylinders, and the interaction between transmission type and weight, with adjusted R-squared at 85%. In this model, all coefficients are significant at a 5% or less (except the six cylinder-coefficient that narrowly misses significance).

Residuals

Looking at the residuals, the values seem to be normally distributed, showing no systematic anomalies in any of the plots (see Appendix, figure 3).

Interpretation for transmission type coefficients

The interpretation of the model for coefficients related to transmission type is the following: 1. The difference in mpg between auto and manual cars is that manual cars get 11.6 more mpg, holding the other variables constant. 2. For a car with automatic transmission, a 1000 lbs increase in weight results in a decrease of 2.4 mpg. 3. If the car has manual transmission, a 1000 lbs increase in weight leads to a decrease in mpg by 6.467. Thus, although a manual transmission is more efficient than an automatic with the other variables constant, increasingly heavy cars with manual transmissions would in the end be more inefficient, possibly explaining the choice of the automakers of heavier cars to put automatics in them. The intercept between the two slopes is at weight=2.84.

Confidence intervals

Because the central questions of this here writ concerns the relationship between transmission type and mpg, these are the main variables for which confidence intervals will be calculated.

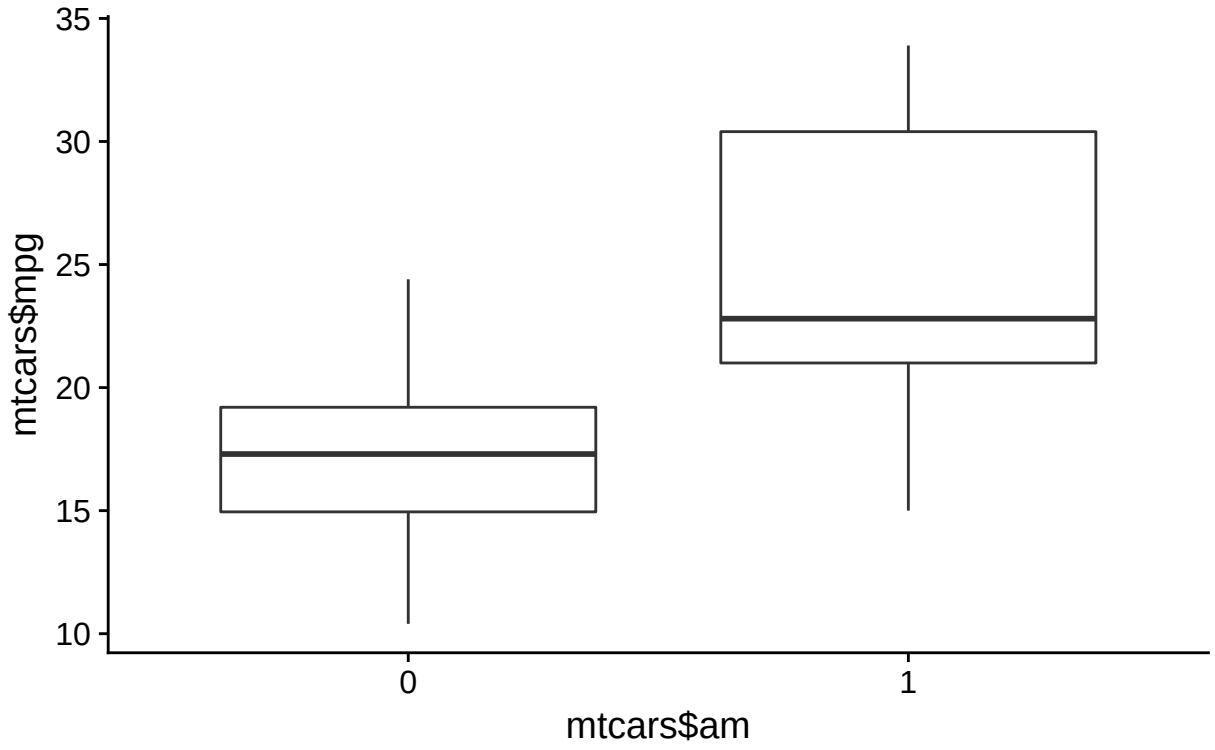
```
confint(intAmWeight, c(2, 3, 6), level=.95)
```

```
##                          2.5 %     97.5 %
## mtcars$am1            3.166215 19.9713651
## mtcars$wt            -4.133556 -0.6638702
## mtcars$am1:mtcars$wt -6.940409 -1.1955528
```

The confidence intervals show that the original coefficients are quite uncertain at a 95% level.

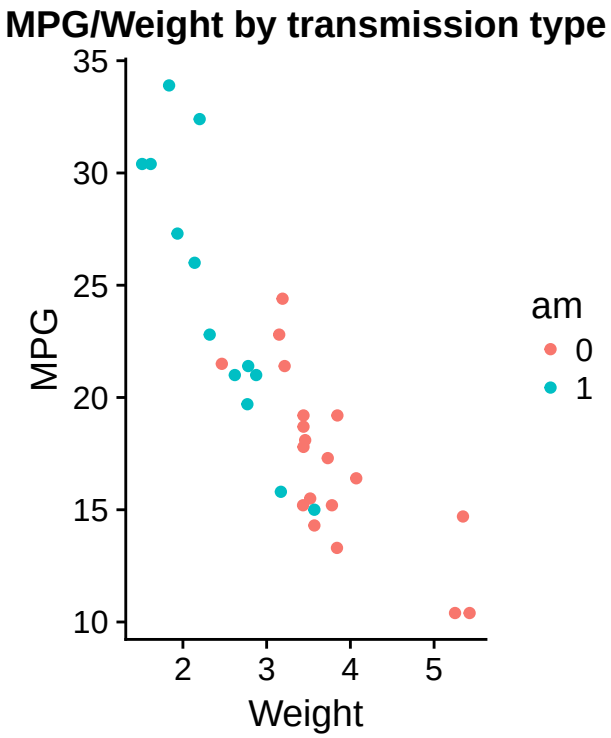
Appendix

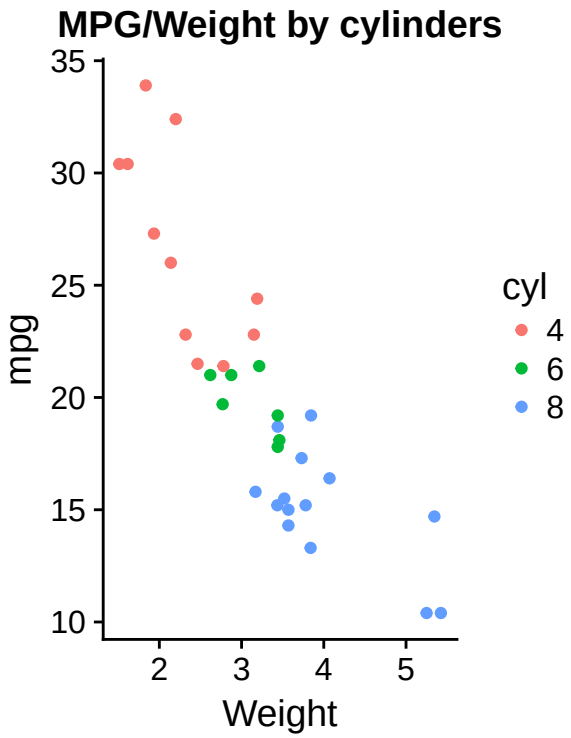
```
b <- ggplot(mtcars, aes(x=mtcars$am))
b <- b + geom_boxplot(aes(y=mtcars$mpg))
b
```




```
mpgwtpl <- ggplot(mtcars, aes(x=wt, y=mpg, group=am, colour=am, height=3, width=3))
mpgwtpl <- mpgwtpl + geom_point()
mpgwtpl <- mpgwtpl + xlab("Weight") + ylab("MPG") + ggtitle("MPG/Weight by transmission type")

mpgwtcyl <- ggplot(mtcars, aes(x=wt, y=mpg, group=cyl, colour=cyl, height=3, width=3))
mpgwtcyl <- mpgwtcyl + geom_point()
mpgwtcyl <- mpgwtcyl + xlab("Weight") + ggtitle("MPG/Weight by cylinders")
plot_grid(mpgwtpl, mpgwtcyl)
```






```
par(mfrow = c(2, 2))
plot(intAmWeight)
```