Does Transmission Type Affect Miles Per Galon?

Executive summary

We analize data from the R mtcars dataset (data for 32 1973-1974 models). We show that Transmission Type (manual or automatic) is interacting with Weigth as a reasonable predictor of Miles per Galon (MPG). Charting the predictor lines for the model shows that for low weigth vehicles manual transmission gives higher MPG while for high weigth it is automatic transmission that gives higher MPG. Both types of transmission are equivalent near the 3000 lb weigth.

Exploratory Analysis

Our first step is to compare the means for both Automatic and Manual cars and create two figures in the Apendix. Figure 1 is a set of pair-wise plots comparing pairs of variables from the dataset. Figure 2 is a box plot for these two means.

```
means <- aggregate(mtcars$mpg, by = list(mtcars$am), FUN = mean);
names(means) <- c("Transmission Type", "Mean MPG")
means
```

```
##   Transmission Type Mean MPG
## 1                 0 17.14737
## 2                 1 24.39231
```

We will fit a linear model trying to use Transmission Type as a predictor of Miles Per Galon, to see whether there is a significant relation worth exploring:

```
fit0 <- lm(mpg ~ am, mtcars)
summary(fit0)$coef
```

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

We can see that the Intercept (the base MPG value for Automatic Transmission) is the same value as the mean for Automatic Transmission we previously calculated. We can also verify that the mean for Manual Transmission is the same as the mean we previously calculated ($24.3923077 = 17.1473684 + 7.2449393$) which corresponds to the Intercept plus the coeficient for the binary factor Transmission Type when the value is *Manual*. Also, note that the p-values for both coefficients are small, so we can say with confidence that the coefficients are correct.

Linear Model Selection and MPG Quantification

We began considering two additional models by including factors likely to affect MPG. The first factor was Weight (column *wt* measured in *lb / 1000*). The second factor was Horsepower (column *hp*). We fit then the first factor:

```
fit1 <- lm(mpg ~ am + wt, mtcars)
summary(fit1)$coef
```

```
##                Estimate Std. Error     t value     Pr(>|t|)
## (Intercept) 37.32155131  3.0546385 12.21799285 5.843477e-13
## am          -0.02361522  1.5456453 -0.01527855 9.879146e-01
## wt          -5.35281145  0.7882438 -6.79080719 1.867415e-07
```

What we realized is that the p-value for the factor Transmission Type (column *am*) was likely to be rejected (with near 99% probability), so we decided to actually explore the interaction between Transmission Type and Weight in our third model instead of the planed Horsepower, as follows:

```
fit2 <- lm(mpg ~ am * wt, mtcars)
summary(fit2)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 31.416055  3.0201093 10.402291 4.001043e-11
## am          14.878423  4.2640422  3.489277 1.621034e-03
## wt          -3.785908  0.7856478 -4.818836 4.551182e-05
## am:wt       -5.298360  1.4446993 -3.667449 1.017148e-03
```

Here, we see that all p-values are very low, so this second model seems better than the model without the interaction.

The model with interaction uses the following equation for MPG prediction:

$$mpg_i = \beta_0 + \beta_1 * Transmission_i + \beta_2 * Weight_i + \beta_3 * Transmission_i * Weight_i + \epsilon_i$$

And the expeected values for **automatic** (am = 0) versus **manual** (am = 1) are:

$$E_a[mpg] = \beta_0 + \beta_2 * Weight = 31.4160554 + -3.7859075 * Weight$$
$$E_m[mpg] = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) * Weight = (31.4160554 + 14.8784225) + (-3.7859075 + -5.2983605) * Weight$$

Figure 3 in the Apendix shows the different values and the predictor lines for this last model. We can see that for weigths lower than 2800 lb manual transmission is better and for weigths of over 3000 lb automatic transmission is better, with a zone where there is no better choice between 2800 and 3000 lb (where both predictor lines cross).

Residuals and Variances

Our first step will be to analize the Residuals from the Model that considers Transmission Type is interacting with Weigth. We must remeber that Residuals must have 0 (zero) mean and that we should observe no pattern.

```
mean(resid(fit2))
```

```
## [1] 1.283695e-16
```

For any practical effect, we can consider that this value is zero. Figure 4 shows a scatter plot of the Residuals versus the Fitted values, where we see no obvious pattern.

Our second step is to analize the changes in Variance between the different models. To do this we will use the ANOVA method, as follows:

```
anova(fit0, fit1, fit2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am * wt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 65.913 7.717e-09 ***
## 3     28 188.01  1     90.31 13.450  0.001017 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here, we can see that the second model (predictors Transmission Type and Weigth) is better than the first model (Transmission Type) because of the very low p-value (7.71708e-09) and that the third model, that has Transmission Type and Weigth interacting is even better, because it's p-value is low (0.00102) and because the sum of squared errors is lower than the second model (442.58 versus 90.31).

Apendix

Figure 1: Pairwise mtcars dataset variable display

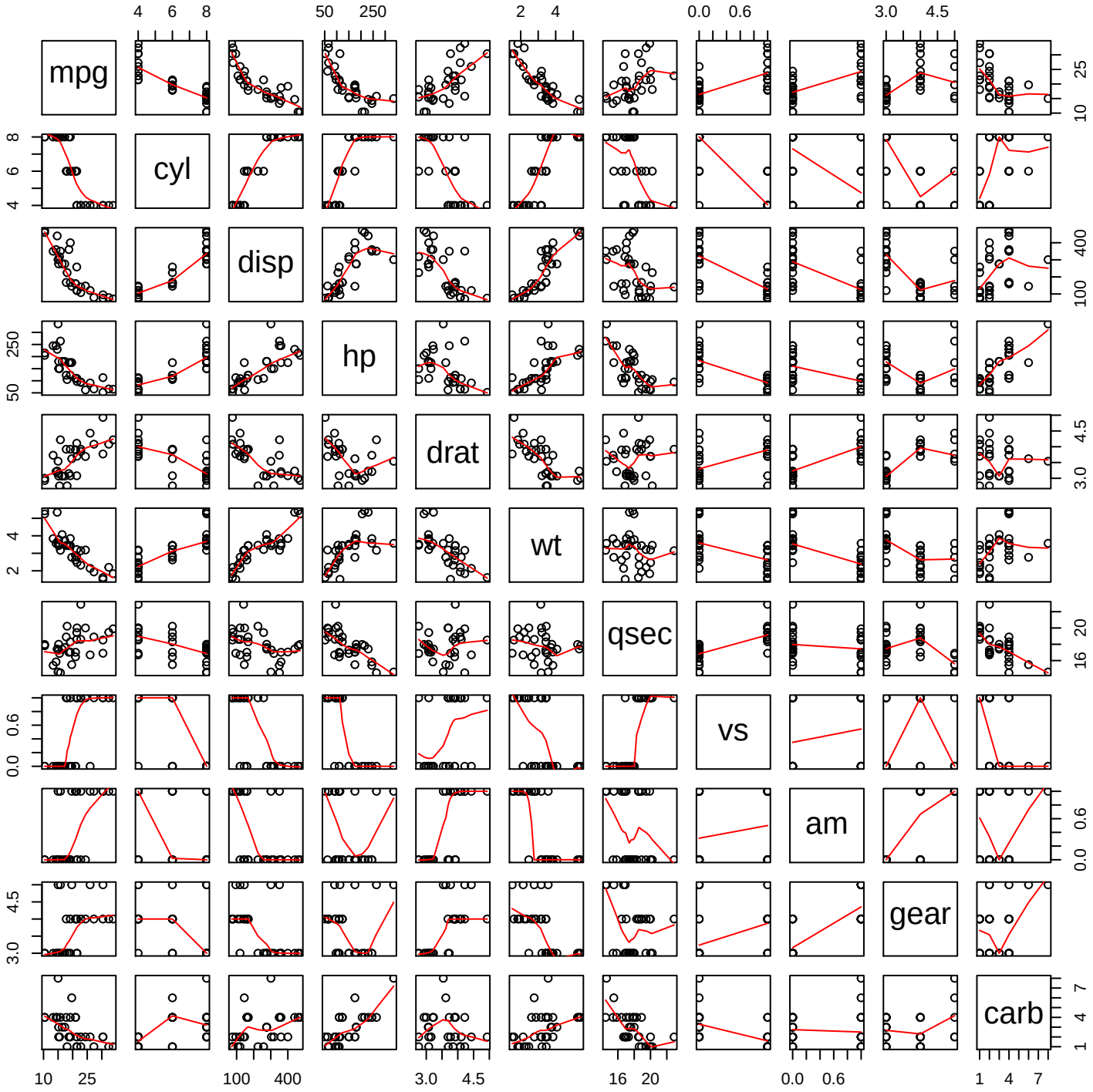

Figure 2: MPG depending on Transmission Type

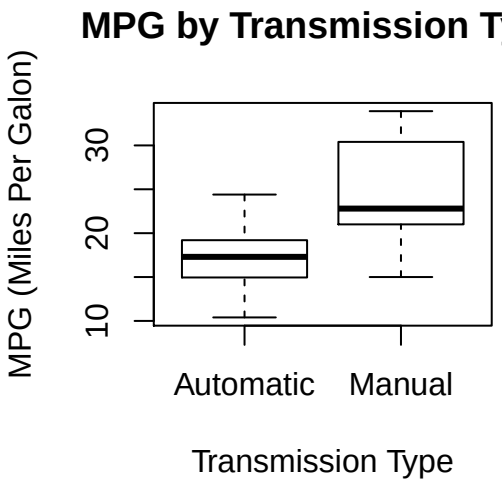

Figure 3: Weigth and Trasnmission Type as predictors of MPG

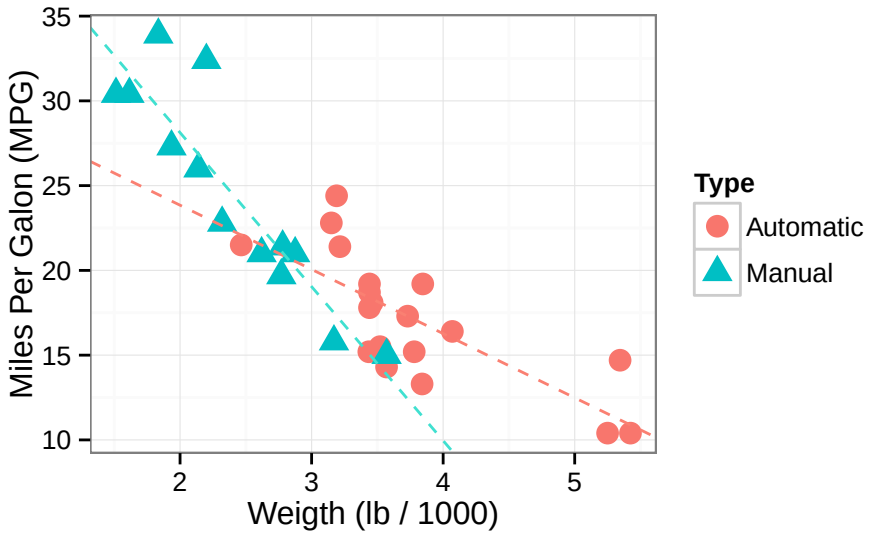

Figure 4: Residuals versus Fitted