Regression Assignment

Jean-Paul Courneya

11/20/2017

The classic question confronted. Is an Automatic or Manual transmission better for MPG?

Executive summary

In this report will address the question : "Is automatic or manual transmission better for mpg ?". To answer this question we used a dataset from the 1974 Motor Trend US magazine, and ran some statistical tests and a regression analysis. On one hand the statistical tests show (without controlling for other car design features) a difference in mean of about 7 MPG more for the cars with a manual transmission. On the other hand, the regression analysis indicates that by taking into account other variables like horsepower and weight, the story can become different. In this case manual transmission cars are only 4.9 miles better MPG than cars with an automatic transmission. Turns out that the difference in MPG is heavily contingent on other regressors. We see that if we are relying on auto or manual transmission as the defining variable for best MPG we may want to consider whats "under the hood" also.

Cleaning data

The first step of our analysis is simply to make a copy of the data and take a look.

```
mtcars2 <- mtcars
str(mtcars2)
```

Next the Auto and Manual factors for "am" variable were coerced as factors and levels were created "Auto"" and "Manual".

```
mtcars2$am_fac1 <- factor(mtcars2$am, labels = c("auto", "manual"))
mtcars2$am_fac2 <- factor(mtcars2$am)
```

Graphics

We begin by plotting boxplots of the variable "mpg" when "am" is "Auto" or "Manual" (see Figure 1 in the appendix). This plot hints at an increase in mpg with manual transmission but this data may have other variables which may play a bigger role in determination of mpg.

T-test

Assuming the a normal distribution of mtcars2 a t-test will be used to see if the mean MPG of auto VS manual transmission is significantly different .

```
t.test(mpg ~ am_fac1, data = mtcars2)
```

The p-value of 0.0013736 allows us to reject the null hypothesis that there is NO difference in the mean MPG of auto VS manual transmission.

Regression analysis

Getting a birds eye view, we plotted the relationships between all the variables of the dataset (see Figure 2 in the appendix). We may note that variables like "wt", "cyl", "disp" and "hp" seem highly correlated together.

```
model.all <- lm(mpg ~ ., data = mtcars2)
```

This model shows the relationship between transmission and MPG alone and matches the results of the t-test

```
a0 <- lm(mpg~am_fac1, data = mtcars2)
summary(a0)$coef
```

As we can see the Intercept coincides with the mean for automatic transmission shown in the table (up to rounding error), and the coefficient am_fac1manual is the difference between the means of each transmission type. The very low p-values suggest that the coefficients are significantly different from zero.

More accurately, we are 95% confident that the difference in miles per gallon between manual and automatic transmission cars lies somewhere in the interval [3.6415096, 10.848369].

Horsepower is a measure of work. In terms of cars it means how much work the engine can do in a time inerval and get results. So it stands to reason that more HP = more work. So cars with a high HP can do more work over a similar interval of time as those with lower HP. This is definitely something that should play an effect on MPG

```
a1 <- lm(mpg~am_fac1 + hp, data = mtcars2)
summary(a1)$coef
```

Even though the coefficient of hp is small, it's statistically significant, which means that there is a small but definitely non-zero effect of horsepower on miles per gallon. This coefficient is the variation of mpg due to the increase of one unit of hp keeping auto fixed. Also of note, the coefficient of am_fac1manual is positive, which reinforces the conclusion that manual transmissions give more mpg.

Now, we are going to add weight wt:

```
a3 <- lm(mpg~am_fac1 + hp + wt, data = mtcars2)
summary(a3)$coef
```

Again, the positive coefficient of automanual says that a manual transmission gives more miles per gallon. The coefficient of wt is the change in mpg per unit of weight keeping the other variables constant.

Let's see whether the addition of variables to our model gives us more detail using ANOVA.

```
anova(a0, a1, a3, test = "Chisq")
```

The low p-values show that the two models built on top of a0 are significantly different from a0 and, hence, they make a good selection.

Residual analysis

The residual plots allow us to check our data (Figure 3 in the appendix). These plots allow us to verify some assumptions made before.

1. The Residuals vs Fitted plot seem to verify the independance assumption as the points are randomly scattered on the plot
2. The Normal Q-Q plot seem to indicate that the residuals are normally distributed as the points hug the line closely
3. The Scale-Location plot seem to verify the constant variance assumption as the points fall in a constant band.

Conclusions

Each one of the three nested models that we used reduced the residual sum of squares with respect to its predecessor, which means that each variable that we chose added more detail to our overall understanding of the variations of mpg. In each one of the models, we saw that a manual transmission gives more miles per gallon, a conclusion drawn from the sign of the coefficient of auto.

Appendix

Figure 1 : Boxplots of "mpg" vs. "am"

```
boxplot(mpg ~ am_fac1, data = mtcars2, main = "MPG by transmission type", xlab = "Transmission type", yl
```

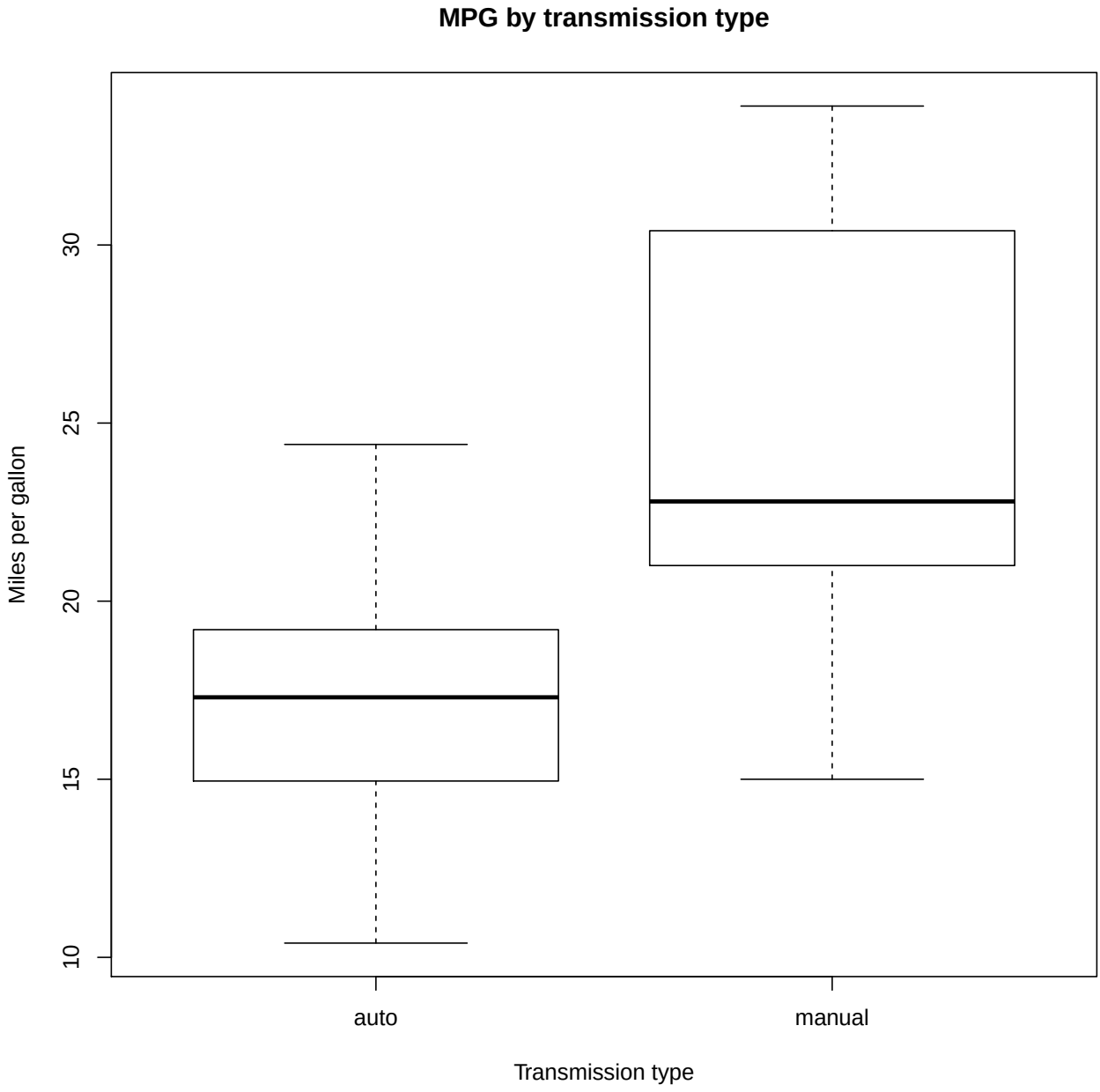

Figure 2 : Pairs graph

```
panel.cor <- function(x, y, digits=2, cex.cor)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- abs(cor(x, y))
  txt <- format(c(r, 0.123456789), digits=digits)[1]
  test <- cor.test(x,y)
  Signif <- ifelse(round(test$p.value,3)<0.001,"p<0.001",paste("p=",round(test$p.value,3)))
  text(0.5, 0.25, paste("r=",txt))
  text(.5, .75, Signif)
}
pairs(mtcars2, upper.panel = panel.smooth, lower.panel = panel.cor, main = "Pairs graph for MTCars")
```

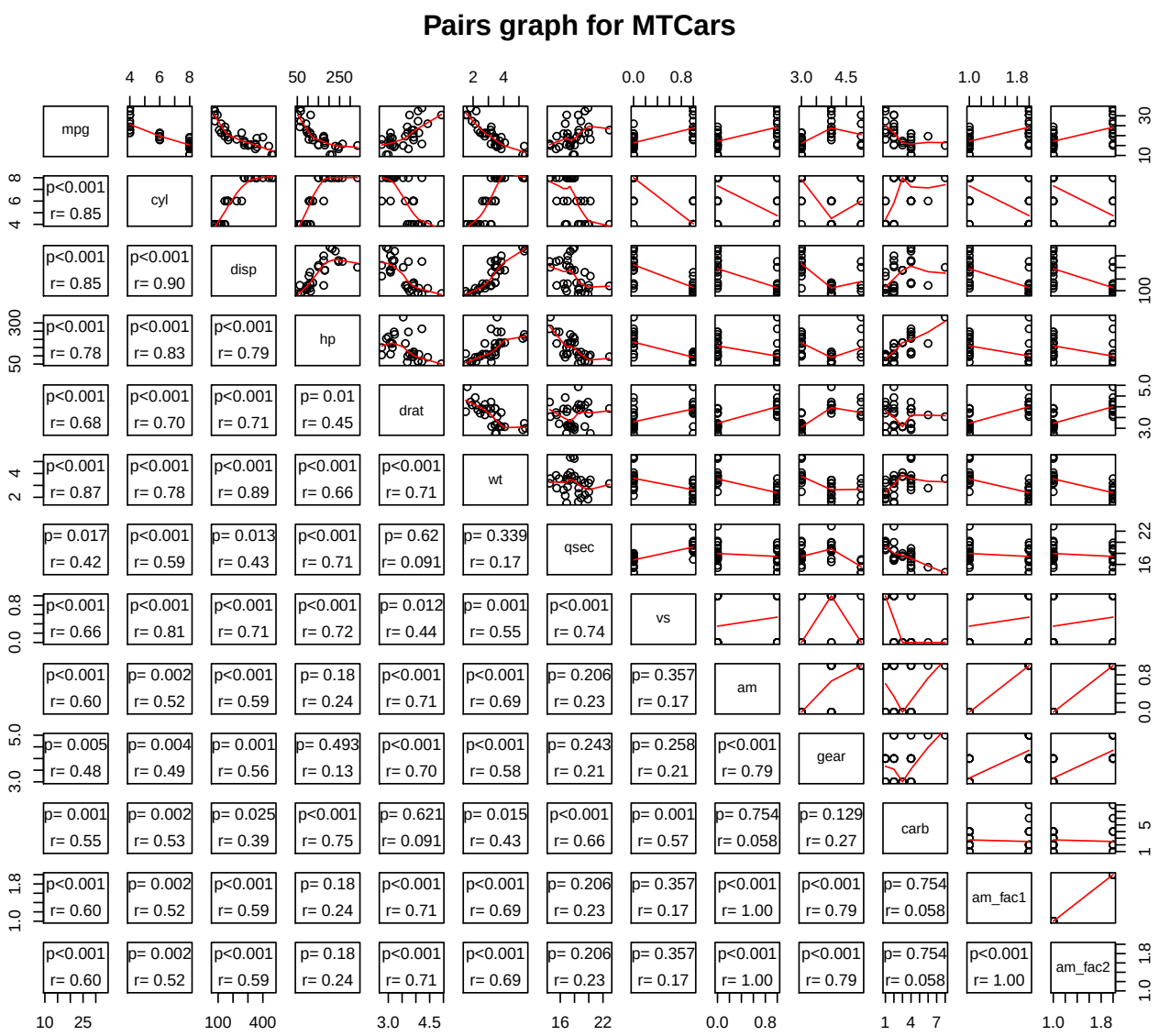

Figure 3 : Residual plots

```
par(mfrow = c(2, 2))
plot(fitted(a3), residuals(a3), xlab = "Fitted values", ylab = "Residuals", main = "Residuals vs Fitted
abline(h = 0, col = "red")
qqnorm(residuals(a3))
qqline(residuals(a3), col = "red")
plot(fitted(a3), sqrt(abs(rstandard(a3))), xlab = "Fitted values", ylab = "Square Root of Standardized
```