

The Relations between Automobile's Gas Mileage and Transmissions

Taedong Yun

Overview and Explorative Analysis

In this article we analyze the relationship between various functional aspects of automobiles. In particular we study the relationship between a set of variables and miles per gallon (MPG) of a car. We will try to answer the following two questions in this report.

1. Is an automatic or manual transmission better for MPG?
2. If so, what is the MPG difference between automatic and manual transmissions?

Throughout this report we will use the `mtcars` dataset, about which you can find more details here <https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html>. Let us first explore the data set.

```
data(mtcars)
names(mtcars) # see variables in the data set
n <- nrow(mtcars); n # total number of observations
c(mean(mtcars$mpg), sd(mtcars$mpg)) # mean and st.dev of MPG
c(sum(mtcars$am), n - sum(mtcars$am)) # number of cars with manual vs automatic trans.
```

In our dataset we have 11 variables including MPG, transmission type, horsepower, etc. and there are 32 observations in total. The mean MPG is 20.091 and st. dev is 6.027. Among the 32 observations 13 cars were manual and 19 were automatic.

Model Selection

Let us first consider the simplest model where we have only the transmission type as a factor variable and the MPG as the outcome.

```
mtcars$am <- factor(mtcars$am)
fit1 <- lm(mpg ~ am, data = mtcars)
summary(fit1)
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125   15.247 1.13e-15 ***
## am1           7.245      1.764    4.106 0.000285 ***
```

From this model we obtain the slope coefficient 7.245 with an extremely low p-value of 0.000285. Recall that when linear regression is applied to a factor variable, the slope coefficient is the expected difference of the outcomes in the two groups. Hence, without considering other variables, the expected MPG of the cars with manual transmission is 7.245 higher than the automatic cars, with significantly low p-value.

However, there are many other possible confounders in the data set we are considering. We know (from physics) that the weight and the horsepower of a car will affect the MPG of the car. Hence, let us consider the linear models containing these two (continuous) variables.

```
fit2 <- update(fit1, mpg ~ am + hp)
fit3 <- update(fit1, mpg ~ am + hp + wt)
anova(fit1, fit2, fit3)
```

The above R code compares the three linear models: (MPG vs Transmission), (MPG vs Transmission and Horsepower), and (MPG vs Transmission, Horsepower, and Weight), by the *nested likelihood ratio tests*. From this we obtain the p-value of $2.445e-09$ when comparing the first and the second model, and the p-value of 0.003574 when comparing the second and third model, both of which are less than 0.01 . From this we choose the third linear model where we have all three variables.

Figure 1 in appendix shows the *residual plots* for the linear model with three variables we chose. We can confirm that there are no strong patterns in the figures.

Results and Summary

Since we have chosen our model, let us fit our data to the model and attempt to answer the two questions we asked.

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

##	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	34.002875	2.642659	12.867	2.82e-13
## am1	2.083710	1.376420	1.514	0.141268
## hp	-0.037479	0.009605	-3.902	0.000546
## wt	-2.878575	0.904971	-3.181	0.003574

Appendix

```
par(mfrow=c(2, 2))
res <- resid(fit3)
plot(as.vector(mtcars$am), res, main='Residual Plot for Transmission Type',
     xlab='Transmission Type', ylab='Residuals')
legend("top", legend=c("0 Automatic", "1 Manual"))
plot(mtcars$hp, res, main='Residual Plot for Horsepower',
     xlab='Horsepower', ylab='Residuals')
plot(mtcars$wt, res, main='Residual Plot for Weight',
     xlab='Weight (1000 lb)', ylab='Residuals')
```

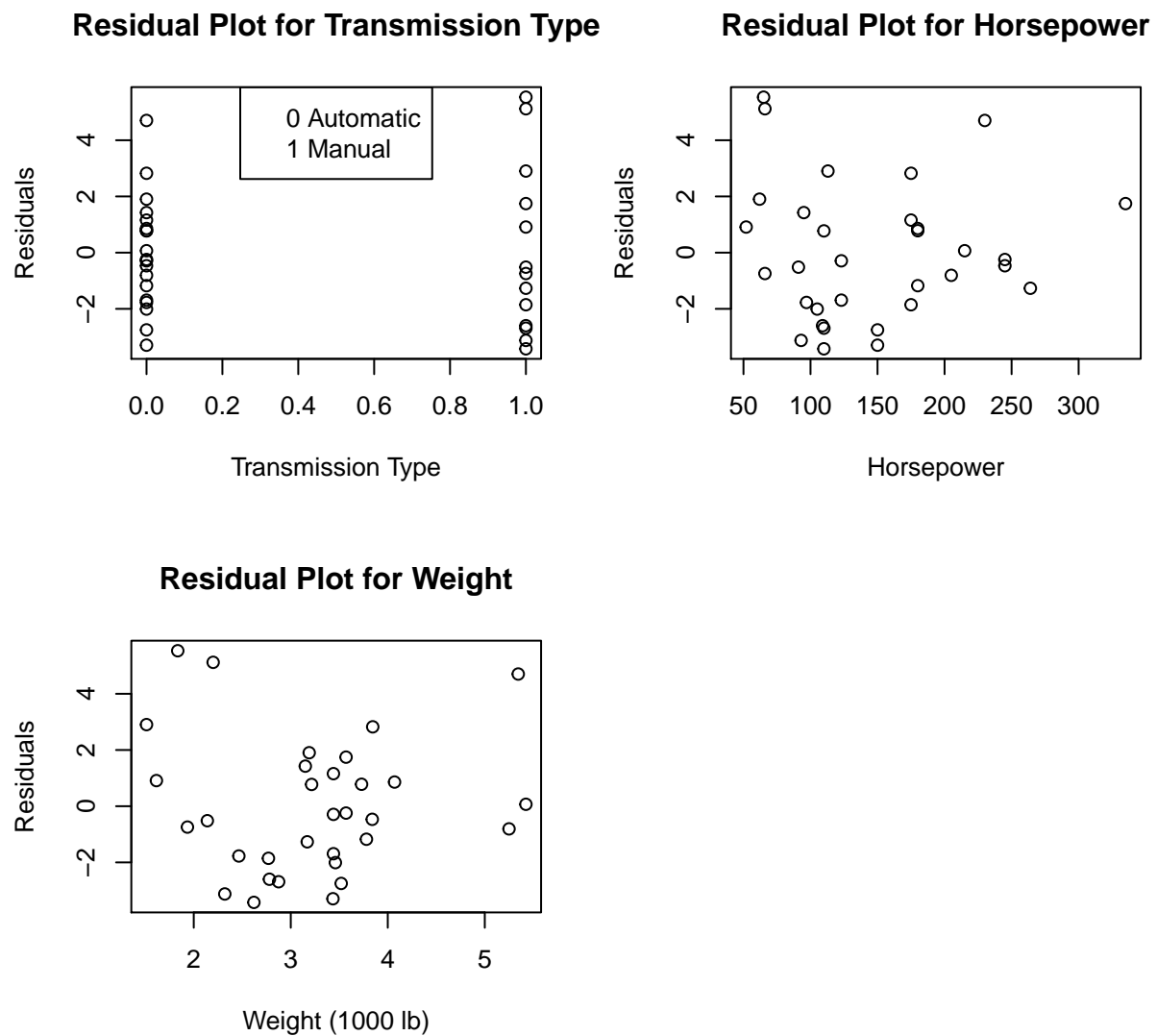


FIGURE 1. Residual Plots

- interpret coefficients

- exploratory data analysis
- fit multiple models and detail their strategy for model selection
- answer the questions of interest (or detail why the question is not answerable)
- residual plot and some diagnostics
- quantify the uncertainty in the conclusions and/or perform an inference
- report < 2 page, including appendix of figures < 5 page
- executive summary