Motor Trend: Does auto transmission improve mpg performance?

Paolo Saracco (with knitr, Rmd to pdf)

2024-05-10

Executive summary

We examine the effect of the transmission type (manual vs automatic) on the miles per gallon (mpg) in the mtcars dataset. Regression analysis using transmission type, weight, and quarter mile time as explanatory variables leads to conclude that manual cars get on average 14 more mpg than automatic cars, holding the other effects constant.

Packages: We will use the `dplyr` package for working with data frames and the `ggplot2` package for graphs.

```
library(dplyr); library(ggplot2)
```

Data: The data set comes from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

```
library(datasets); data("mtcars")
```

Exploratory data analysis

We are dealing with a 32 x 11 data set: we have 11 observations for 32 cars. Let us have a look at the data (we refer the reader to the `mtcars` help page for the explanation of the variable names).

```
head(mtcars); tail(mtcars); str(mtcars); summary(mtcars)
```

To begin with, it is useful to look at how the different `mpg` values vary across the two groups: see **Figure 1**. It seems that cars with manual transmission cover more `mpg`, on average. To quantify this, we perform a linear regression with `mpg` as outcome and the factor variable `am` as predictor, without intercept (which is the same as a T-test)

```
naive_fit <- lm(mpg ~ as.factor(am)-1, data = mtcars)
```

	mean	std error	95% conf int
automatic	17.1474	1.1246	[14.8506, 19.4441]
manual	24.3923	1.3596	[21.6157, 27.1689]

However, we reasonably expect that other factors play a role. For example, `mpg` is also influenced by `cyl`, `hp`, `wt`, `qsec`, and not all of them are uncorrelated, as we can see in **Figure 2**.

Model selection

It may happen that the effect of transmission on miles per gallon is biased by the fact that we are not considering other factors. Therefore we start nesting models in order to see if adding regressors have a significant effect above and beyond the change in degrees of freedom. First we convert factor variables into factors:

```
data <- mtcars %>% mutate(cyl = as.factor(cyl), vs = as.factor(vs), am = as.factor(am),
                          gear = as.factor(gear), carb = as.factor(carb))
```

Since it is clear that the weight of the car influences its miles per gallon, we start adding `wt` to our model and then we ask whether adding another variable has a significant effect, by relying on the ANOVA test.

```
seek <- function(var, params = c()) {
        formula <- reformulate(c(params, var), response = 'mpg'); fit_new <- lm(formula, data)
        cat("p-value for adding", var, "is", anova(fit, fit_new)[2,6], "\n")}
fit <- lm(mpg ~ am + wt, data)
invisible(sapply(names(select(data, !c(mpg,am,wt))), seek, params = c('am','wt')))
```

```
## p-value for adding cyl is 0.003473216
## p-value for adding disp is 0.0678774
## p-value for adding hp is 0.0005464023
## p-value for adding drat is 0.2849371
## p-value for adding qsec is 0.0002161737
## p-value for adding vs is 0.008454158
## p-value for adding gear is 0.1200295
## p-value for adding carb is 0.2470563
```

Since we will perform 20 tests, to take into account the multiple hypothesis testing issue and to control the false positive rate at level $\alpha = 0.05$, we call significant the p-values below $0.05/20 = 0.0025$. It turns out that `hp` and `qsec` have a significant effect. Since `qsec` is the one with weaker correlation with `wt` (-0.17 vs 0.66 of `hp`), then we add `qsec` to our model and we ask ourselves the same question.

```
fit <- lm(mpg ~ am + wt + qsec, data)
invisible(sapply(names(select(data, !c(mpg,am,wt,qsec))), seek, params = c('am','wt','qsec') ))
```

We conclude that no other variable has a significant effect, so we search for potential cross-effects:

ANOVA w.r.t. adding	`am:wt`	`am:qsec`	`wt:qsec`
p-value	0.0018086	0.0384079	0.2823973

Linear regression

We can fit our linear model with `mpg` as output and `am`, `wt`, `am:wt`, `qsec` as regressors and view the inference results:

```
fit <- lm(mpg ~ am*wt + qsec, data); summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ am * wt + qsec, data = data)
##
## 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
## am1           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 ***
## am1: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
```

Therefore, holding constant `qsec` and `wt`, cars with manual transmissions get about 14 more miles per gallon:

	estimate	std error	95% conf int
manual	14.0794	3.4353	[7.0309, 21.128]

Moreover, we are satisfied because our model significantly explains about 90% of the variance (R-squared is 0.8959).

We can check if regression assumptions are met, namely the normality assumption for the residuals, with a Shapiro-Wilk test and diagnostic plotting (see **Figure 3**):

```
shapiro.test(fit$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit$residuals
## W = 0.94444, p-value = 0.1001
```

Normality assumptions don't seem far off (approximately), and heteroskedasticity doesn't seem to be an issue.

Conclusions

Regression analysis using transmission type, weight, and quarter mile time as explanatory variables leads to the conclusion that manual cars get on average 14.08 ± 3.44 more mpg than automatic cars, holding the other effects constant.

However, the number of observations is relatively small, whence it is difficult to draw trustful conclusions. **Figure 4** offers a glimpse of the problem. It is advisable to obtain a new and larger data set to challenge these results.

Appendices

Box plot mpg vs transmission

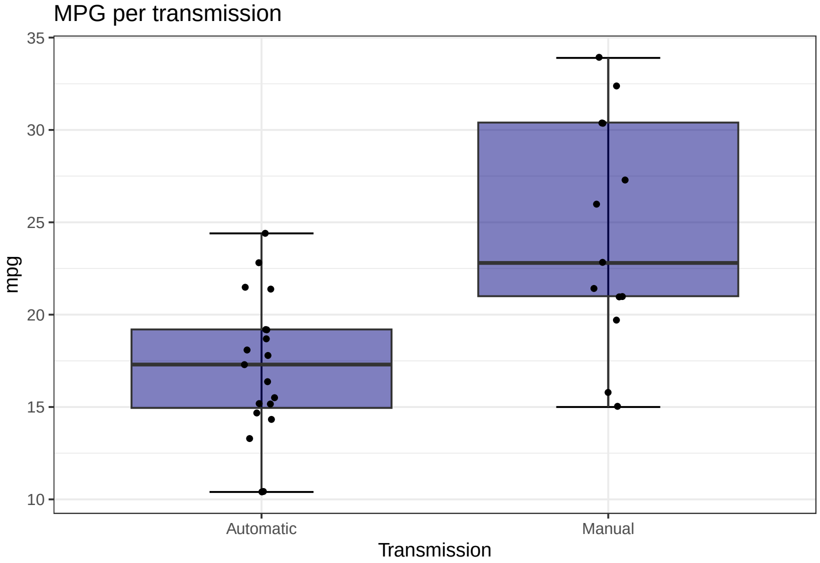

Figure 1: Box plot

Pair plot of mtcars

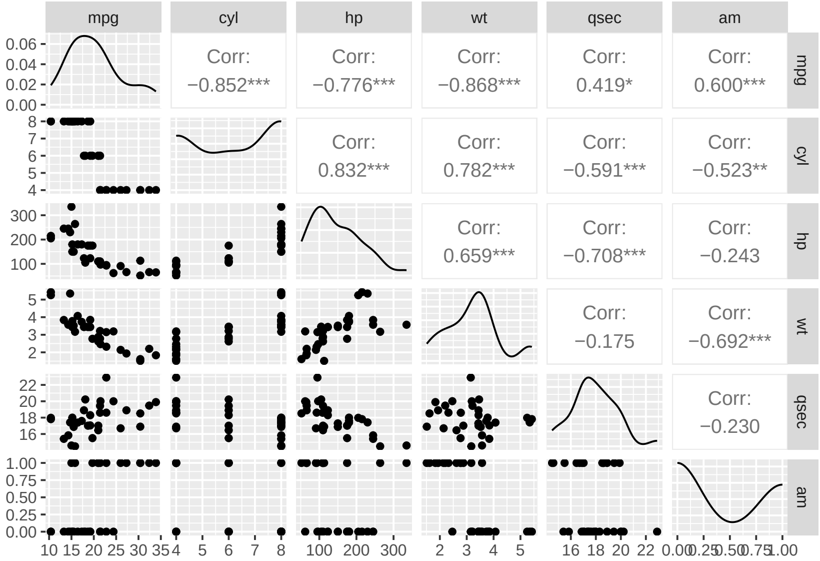

Figure 2: Pair plot

Diagnostic plotting for residuals

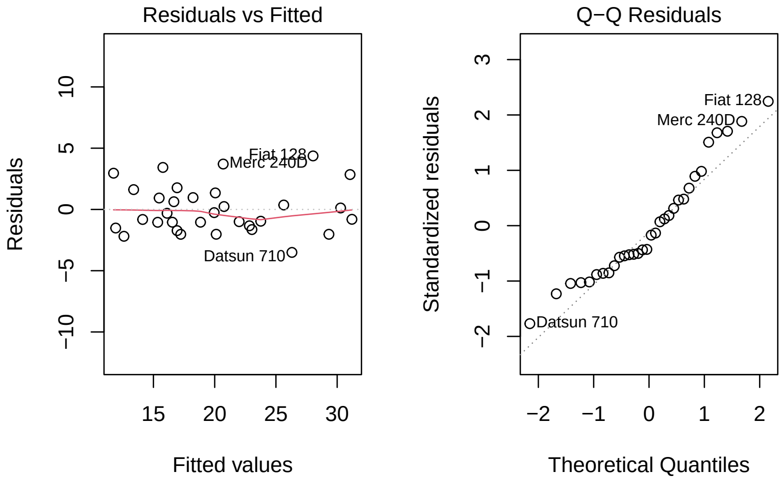

Figure 3: Diagnostic plotting

Scatter plot of the effect of 1/4 mile time, transmission and weight on mpg

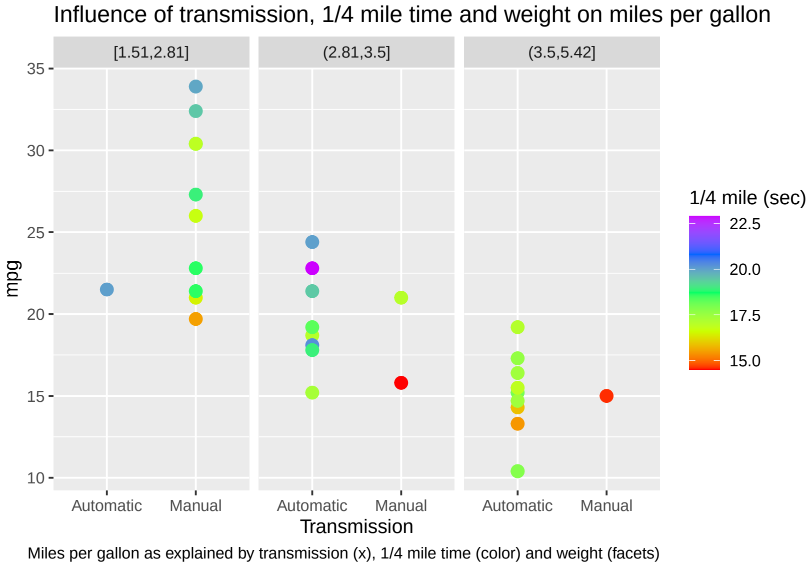

Figure 4: 4-dim scatterplot