The effect of transmission type on miles per gallon (MPG)

Regression Models course project – Hedley Stirrat – 27 May 2020

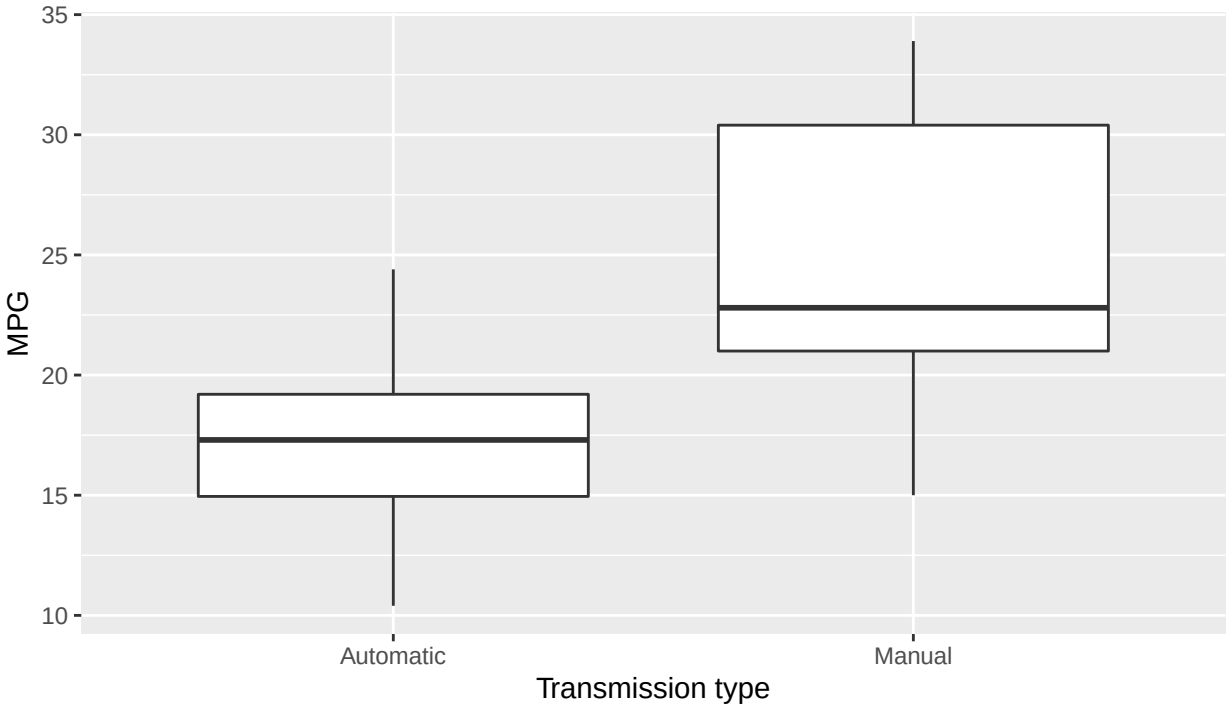
Executive summary

Miles per gallon (MPG) is an important metric in fuel efficiency. This report explores the effect of transmission type (manual or automatic) on MPG, using the `mtcars` dataset. Using a two-sample t-test, it was found that automatic cars have significantly lower MPG than manual cars. This conclusion was supported by the fitting of a simple linear regression model. Finally, a more robust multivariate model was constructed that took into account other covariates in the dataset, and the initial hypothesis—that cars with automatic transmission have lower fuel efficiency than cars with manual transmission—was found to be well supported.

Exploratory analysis

Figure 1 shows that, for the cars in this dataset, a manual transmission is generally associated with a higher MPG than an automatic transmission.

Figure 1: Effect of transmission type on MPG



To formally test this hypothesis—that cars with manual transmission are more fuel-efficient than those with an automatic transmission—we will conduct a two-sample t-test.

Hypothesis testing

t	p	LowerConfidence	UpperConfidence	Mean.Automatic	Mean.Manual
-3.767123	0.0013736	-11.28019	-3.209684	17.14737	24.39231

The results of the *t*-test suggest that there is a statistically significant difference in the mean MPG of cars with automatic versus manual transmission. The p-value is ~0.001, with a 95% confidence interval placing the mean difference between 3–11 MPG.

To further explore this relationship, we will next fit a linear regression model to the data.

Linear regression

Univariate linear regression

The first linear regression model that we fit is using only the `am` variable (transmission type) to predict MPG. `am` is a binary factor variable where automatic cars are coded `0` and manual cars are coded `1`.

```
##              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
```

The intercept of this linear model is simply the mean MPG of automatic cars. The `am` coefficient tells us the estimated increase in MPG when we switch from an automatic to a manual car, and the coefficient value of ~7 is consistent with our *t*-test that estimated a mean difference of between 3–11 MPG. Additionally, R^2 = `0.36`, so only a moderate amount of variance in MPG is able to be explained by transmission type with this model.

The p-value of <0.001 suggests that the likelihood of the observed trend being due to chance is very low. However, it is possible that other covariates in the `mtcars` dataset are contributing to this effect. Indeed, there are several other variables that probably have an effect on MPG, including `wt` (weight), `cyl` (number of cylinders) and so on. To explore these effects and see whether the relationship between transmission type and MPG holds true when other variables are accounted for, we will explore multivariate linear regression.

Multivariate linear regression

The process to determine which other variables to include in the regression model is an important one, as extra regressors increase the standard errors of the other regressors. Conversely, we can introduce bias into the model if we omit variables correlated with the outcome variable.

One way to examine the effect of adding variables to the model is to calculate the variance inflation factor (VIF), which is the increase in variance when a regressor is added to a model. A high VIF suggests that a variable is strongly correlated with another included regressor.

We fit a model predicting MPG using all available variables as regressors, and calculate the associated square-roots of the VIFs:

```
##      cyl     disp       hp     drat       wt     qsec       vs       am
## 3.920948 4.649757 3.135608 1.837014 3.894212 2.743712 2.228424 2.156035
##     gear     carb
## 2.314617 2.812249
```

The high VIFs for `cyl` (number of cylinders), `disp` (engine displacement) and `wt` (weight) suggest high collinearity with other regressors. Indeed, all three of those variables are likely to be correlated with one another (think about large engines!).

However, purposefully excluding variables like those above that are intuitively correlated with MPG would introduce bias into our model. Therefore, our final model will be fitted with transmission type and `wt`, in order to take that correlation into account.

```
##                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
```

From the coeffients of the model fitted above, we can see that the inclusion of `wt` significantly increases the prediction ability of the model ($p = 1.87 \times 10^{-7}$). On the other hand, using `wt` and `am` to predict MPG is not significantly better than using `wt` alone ($p = 0.99$). Car weight is clearly a strong predictor of fuel efficiency. Indeed, for this model, $R^2 =$ `0.75`, representing a significant increase in explanation of variance in MPG over the simple transmission-only linear model.

Residuals and diagnostics

Finally, we can plot the residuals to check for non-normality. In Figure 2 (see Appendix), the Residuals vs Fitted plot doesn't appear to suggest any obvious non-normality or heteroskedasticity that would impact on the model's underlying assumptions.

Conclusion

It has been found that automatic cars have lower fuel efficiency, as measured by MPG, than manual cars. The effect of transmission type on MPG was supported by a *t*-test and regression models. The quantified difference is that cars with manual transmission tend to get about 7 MPG higher than cars with automatic transmission. It was also found that other variables, such as car weight, were good predictors of MPG, and thus should be taken into account when evaluating fuel efficiency affected by transmission type.

Appendix

Exploratory analysis

The first plot was generated with the below code, using the `ggplot2` and `dplyr` packages:

```
library(ggplot2)
library(dplyr)
mtcars %>%
    mutate(am = case_when(am==1 ~ "Manual", am==0 ~ "Automatic")) %>%
    ggplot(aes(x = factor(am), y = mpg)) +
        geom_boxplot() +
        ggtitle("Figure 1: Effect of transmission type on MPG") +
        labs(x = 'Transmission type', y = 'MPG') +
        theme(plot.title = element_text(hjust=0.5))
```

Hypothesis testing

The t-test and summary table were produced with the following code:

```
test <- with(mtcars, t.test(mpg ~ am))
knitr::kable(data.frame(
    t = test$statistic,
    p = test$p.value,
    LowerConfidence = test$conf.int[1],
    UpperConfidence = test$conf.int[2],
    Mean.Automatic = test$estimate[1],
    Mean.Manual = test$estimate[2], row.names=""))
```

Linear regression

The first regression model was fitted using the following code:

```
summary(with(mtcars, lm(mpg ~ am)))$coefficients
```

After loading the `car` package, the multivariate regression model taking all variables as regressors was fit, and the VIFs were calculated, using the following code:

```
library(car)
fit <- lm(mpg ~ ., data = mtcars)
sqrt(vif(fit))
```

The final regression model taking `wt` into account was produced with the following code:

```
summary(lm(mpg ~ am + wt, data = mtcars))$coefficients
```

The residuals and diagnostics plot is shown below, and was produced with the following code:

```
par(mfrow = c(2,2))
plot(fit)
mtext("Figure 2: Residuals and Diagnostics", side = 3, line = -2, outer = TRUE, cex=1.5)
```

Figure 2: Residuals and Diagnostics