What is the impact of an automatic or manual transmission on MPG?

Executive Summary

The results of this study were inconclusive. The weight of the vehicle and the displacement of the vehicle have for more impact on the MPG than does the transmission. To that end, light cars with small engines are being outfitted with manual transmissions far more often than a heavy cars with large engines, most likely as a cost-saving measure, and so a direct comparison could not be made using this dataset.

Exploratory Analysis

"Is an automatic or manual transmission better for MPG"

"Quantify the MPG difference between automatic and manual transmissions"

Data loading and manipulation First, we load the data.

```
cars <- mtcars
str(cars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

The variables are described as:

- mpg Miles/(US) gallon
- cyl Number of cylinders
- disp Displacement (cu.in.)
- hp Gross horsepower
- drat Rear axle ratio
- wt Weight (1000 lbs)
- qsec 1/4 mile time
- vs Engine (0 = V-shaped, 1 = straight)
- am Transmission (0 = automatic, 1 = manual)
- gear Number of forward gears

- carb Number of carburetors

We need to recode vs and am as factor variables, and convert weight to it's appropriate value

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
cars <- cars %>%
        mutate(vs = factor(vs, levels = c(0,1), labels = c("V-Shaped", "Straight")),
               am = factor(am, levels = c(0,1), labels = c("Automatic", "Manual")),
               wt = wt * 1000)
str(cars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2620 2875 2320 3215 3440 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : Factor w/ 2 levels "V-Shaped","Straight": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am  : Factor w/ 2 levels "Automatic","Manual": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

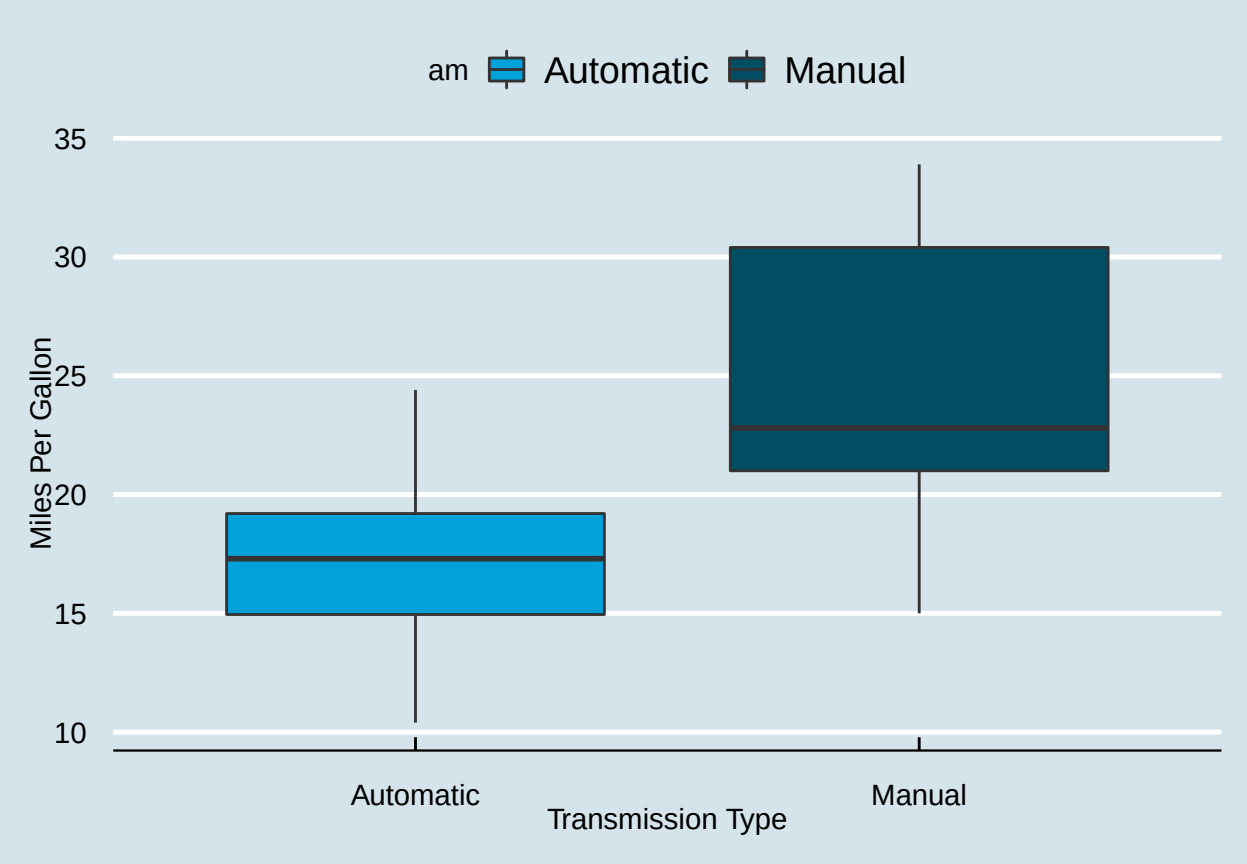
Exploratory Visualizations Now that variables are properly coded, we can begin exploring our variables that will be used in a model.

First, a boxplot of our outcome variable, MPG, by the kind of transmission.

```
library(ggplot2)
library(ggthemes)
```

```
## Warning: package 'ggthemes' was built under R version 4.1.1
```

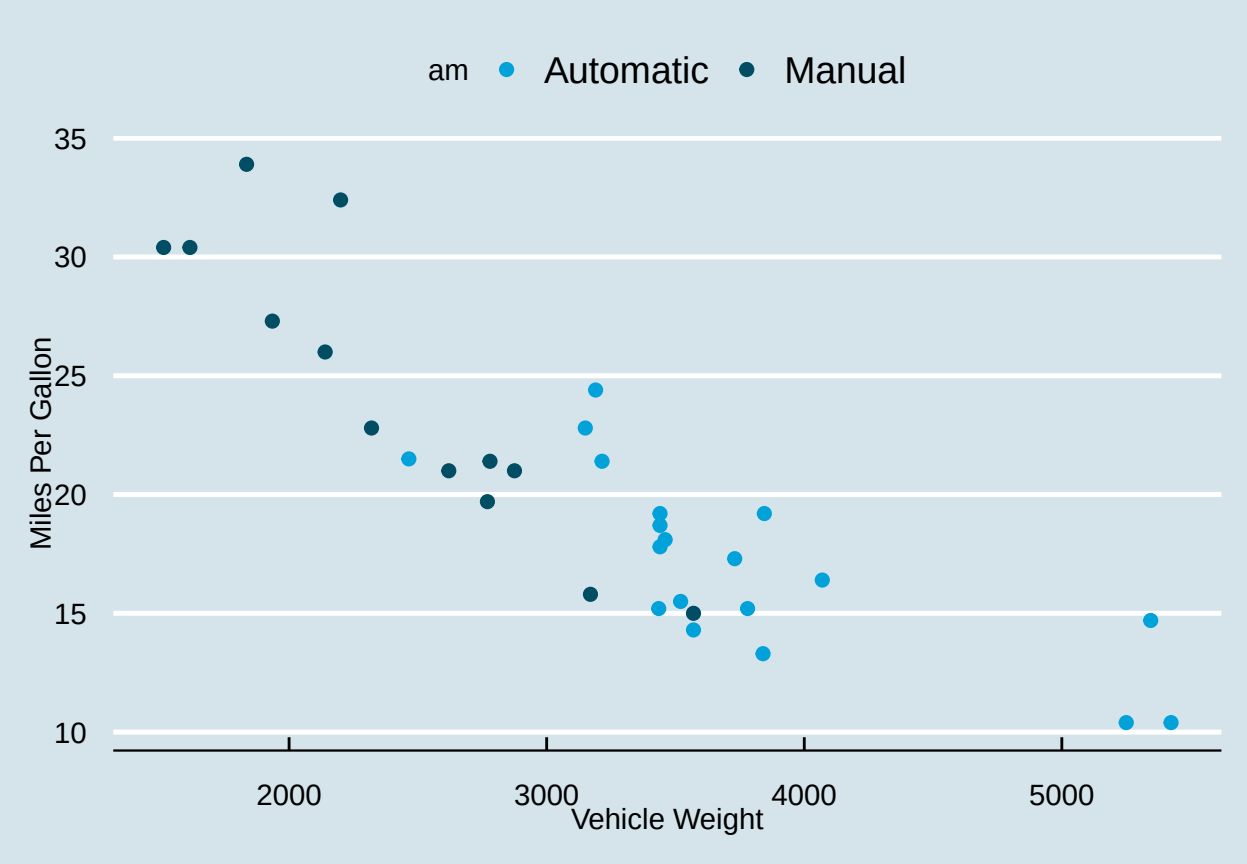
```
ggplot(data = cars, mapping = aes(x = am, y = mpg, fill = am)) +
        geom_boxplot() +
        xlab("Transmission Type")+
        ylab("Miles Per Gallon") +
        theme(legend.position = "none") +
        theme_economist() +
        scale_fill_economist()
```



Generally speaking, it appears manual transmission vehicles get higher MPG.

Let's explore MPG as a fuction of other variables, starting with weight.

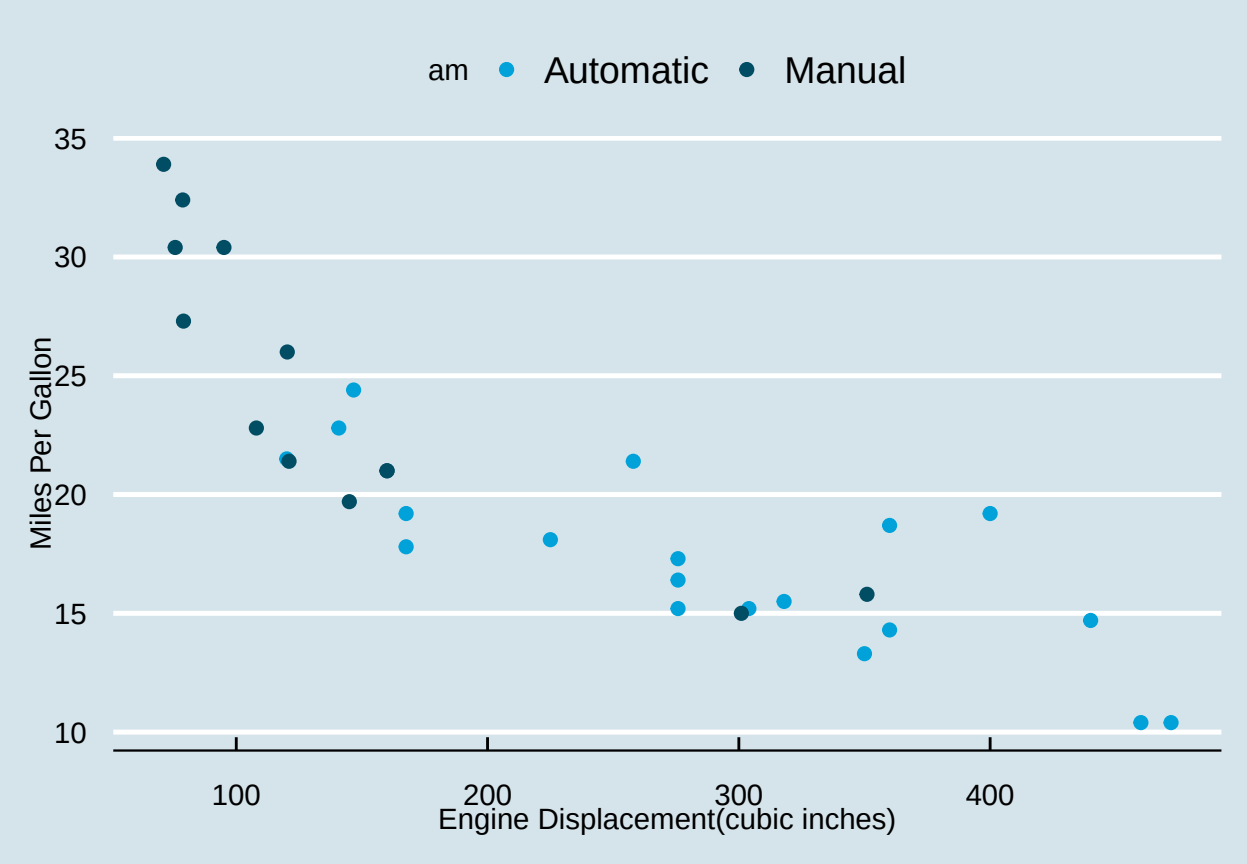
```
ggplot(data = cars, mapping = aes(x = wt, y = mpg, color = am)) +
        geom_point(size = 2) +
        xlab("Vehicle Weight")+
        ylab("Miles Per Gallon") +
        theme(legend.position = "none") +
        theme_economist() +
        scale_color_economist()
```



Here we can see that lighter vehicles tend to be manual transmissions, and such have higher MPG.

Finally, let's explore MPG as a function of displacement.

```
ggplot(data = cars, mapping = aes(x = disp, y = mpg, color = am)) +
        geom_point(size = 2) +
        xlab("Engine Displacement(cubic inches)")+
        ylab("Miles Per Gallon") +
        theme(legend.position = "none") +
        theme_economist() +
        scale_color_economist()
```



Vehicles with manual transmissions also appear to be built with smaller engines, allowing for more MPG.

Predictive Modelling

With intial analysis, let's begin to model our assumptions. MPG is a continuous variable, and so our model will be an ordinary least squares regression. Let's begin by modelling MPG as the outcome, and transmission type as the predictor.

```
library(broom)

linear1 <- lm(mpg~am, data = cars)

tidy(linear1)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    17.1       1.12     15.2  1.13e-15
## 2 amManual        7.24      1.76      4.11 2.85e- 4
```

According to our simplest model, having a manual transmission put in your car will, on average, increase MPG by 7.2.

Let's explore further, using weight and displacement as predictors with transmission.

```
linear2 <- lm(mpg~am + disp + wt, data = cars)

tidy(linear2)
```

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept) 34.7       3.24       10.7   2.12e-11
## 2 amManual     0.178     1.48        0.120 9.06e- 1
## 3 disp        -0.0178    0.00937    -1.90  6.79e- 2
## 4 wt          -0.00328   0.00133    -2.47  1.99e- 2
```

With this model, having a manual transmission is no longer a signficant predictor of any change in MPG!

Let's take a look the correlations between our predictors.

```
library(corrr)
```

```
## Warning: package 'corrr' was built under R version 4.1.1
```

```
library(corrplot)
```

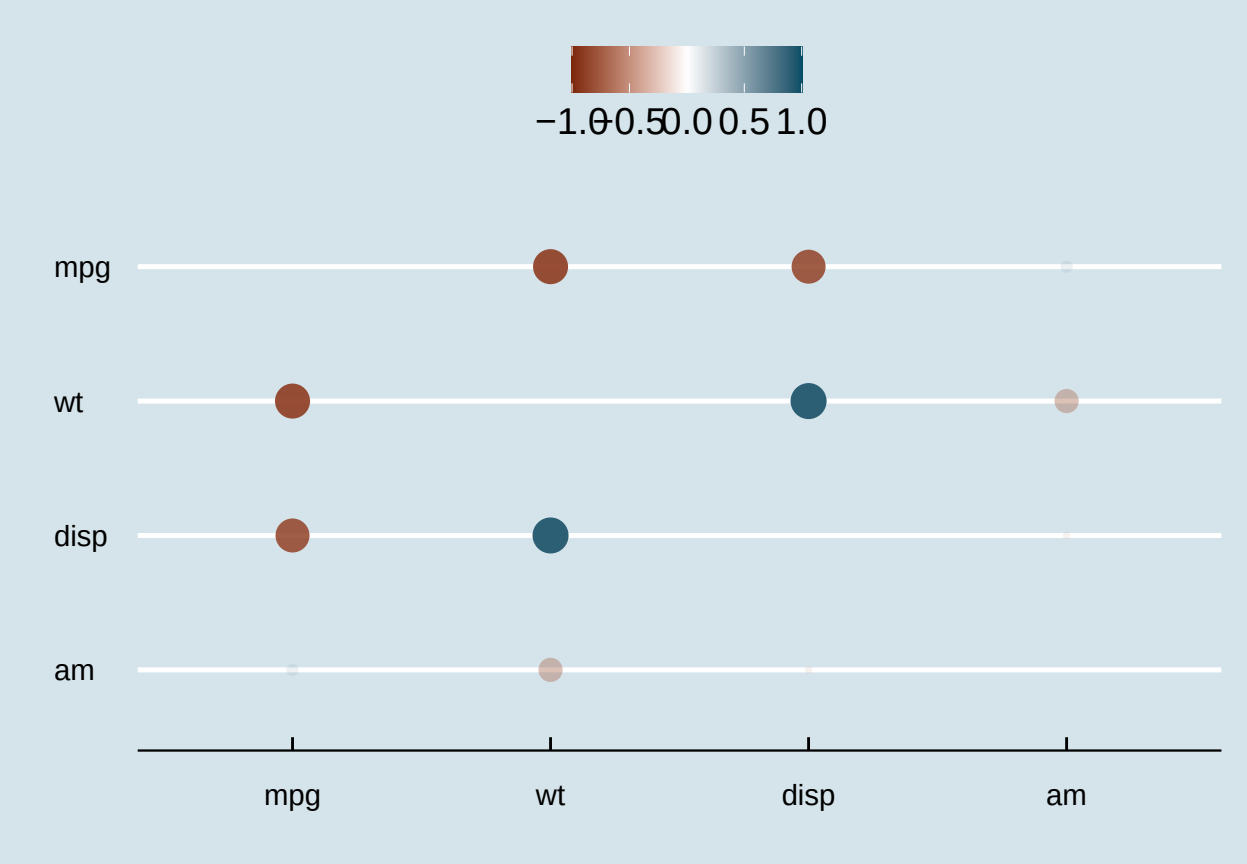
```
## corrplot 0.90 loaded
```

```
corr <- cars %>%
        select(mpg, wt, disp, am) %>%
        mutate(am = as.numeric(am)) %>%
        correlate()
```

```
##
## Correlation method: 'pearson'
## Missing treated using: 'pairwise.complete.obs'
```

```
rplot(corr, colours = c("#7C260B", "white", "#014D64")) +
        theme_economist()
```

```
## Don't know how to automatically pick scale for object of type noquote. Defaulting to continuous.
```



MPG is heavily correlated to displacement and weight, and transmission type is correlated to weight. Displacement and weight are also heavily correlated. Let's do some more modelling to see what a good combination of these may be.

Let's start by removing displacement and seeing if that improves our model using ANOVA to evaluate.

```
linear3 <- lm(mpg~am + wt, data = cars)


tidy(anova(linear3,linear2))
```

```
## # A tibble: 2 x 6
##   res.df   rss    df sumsq statistic p.value
##    <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl>
## 1     29  278.    NA  NA       NA     NA
## 2     28  247.     1  31.8      3.61  0.0679
```

H_0 = Adding Displacement to the model did not change the model.

H_a = Adding Displacement to the model did change the model.

As we can see, our P-value is greater than the standard of .05, and so we fail to reject the null. However, a strong argument can be made about any of the variables. Let's try the same process, instead of adding displacement, let's add the interaction between weight and displacement.

```
linear4 <- lm(mpg ~ am + wt + disp + (wt*disp), data = cars)

tidy(linear4)
```

```
## # A tibble: 5 x 5
```

```
##   term          estimate  std.error statistic  p.value
##   <chr>            <dbl>      <dbl>     <dbl>    <dbl>
## 1 (Intercept) 48.4       4.40           11.0  1.77e-11
## 2 amManual    -1.81      1.31           -1.38 1.78e- 1
## 3 wt          -0.00773   0.00157        -4.92 3.77e- 5
## 4 disp        -0.0617    0.0136         -4.54 1.05e- 4
## 5 wt:disp      0.0000136 0.00000348      3.90 5.70e- 4
```

```
tidy(anova(linear2,linear4))
```

```
## # A tibble: 2 x 6
##   res.df   rss    df sumsq statistic   p.value
##    <dbl> <dbl> <dbl> <dbl>     <dbl>     <dbl>
## 1     28  247.    NA  NA        NA   NA
## 2     27  158.     1  89.0      15.2  0.000570
```

H_0 = Adding the interaction between weight and displacement to the model did not change the model.

H_a = Adding the interaction between weight and displacement to the model did change the model.

Our P-Value is far below the standard of .05, and so we reject the null that the model did not change.

Finally, let's work through interactions between the transmission type and the weight and displacement.

```
linear5 <- lm(mpg ~ am + wt + disp + (wt*disp) + (am*wt), data = cars)
linear6 <- lm(mpg ~ am + wt + disp + (wt*disp) + (am*disp), data = cars)
linear7 <- lm(mpg ~ am + wt + disp + (wt*disp) + (am*wt) + (am*disp), data = cars)
linear8 <- lm(mpg ~ wt + disp + (wt*disp), data = cars)


comparison1 <- tidy(anova(linear4, linear5))
comparison2 <- tidy(anova(linear4, linear6))
comparison3 <- tidy(anova(linear4, linear7))
comparison4 <- tidy(anova(linear8, linear4))

comparison1$p.value
```

```
## [1]        NA 0.2898137
```

```
comparison2$p.value
```

```
## [1]       NA 0.474685
```

```
comparison3$p.value
```

```
## [1]        NA 0.5751443
```

```
comparison4$p.value
```

```
## [1]        NA 0.1782267
```

Adding the interaction term between transmission and any of the other predictors did not add anything to the model.

However, there was another interesting model at the end of the comparison list. Comparison4 compared a model with no Transmission variable against one that did, and found that adding transmission to the model was not useful!

Let's compare the model with transmission to that without.

```
tidy(linear4)
```

```
## # A tibble: 5 x 5
##   term          estimate  std.error statistic  p.value
##   <chr>            <dbl>      <dbl>     <dbl>    <dbl>
## 1 (Intercept) 48.4       4.40           11.0  1.77e-11
## 2 amManual    -1.81      1.31           -1.38 1.78e- 1
## 3 wt          -0.00773   0.00157        -4.92 3.77e- 5
## 4 disp        -0.0617    0.0136         -4.54 1.05e- 4
## 5 wt:disp      0.0000136 0.00000348      3.90 5.70e- 4
```

```
tidy(linear8)
```

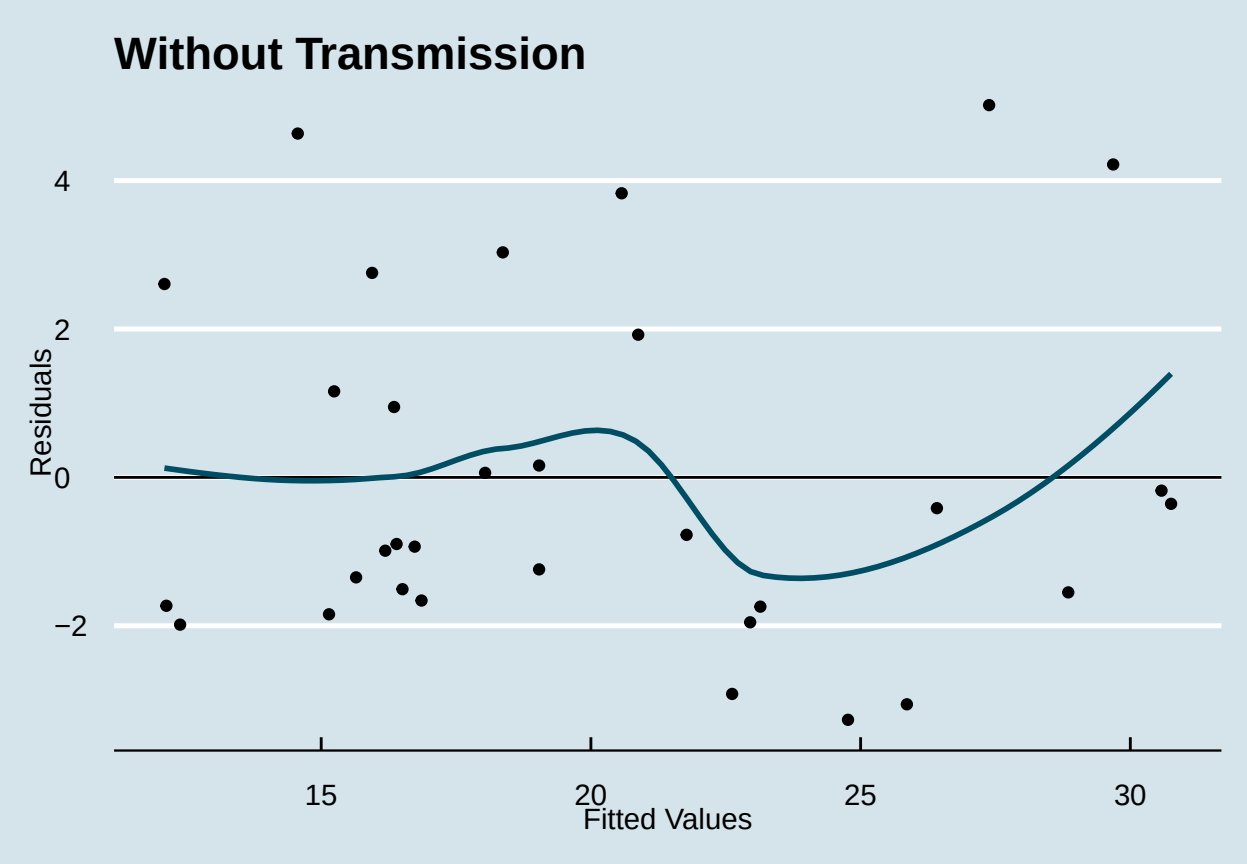
```
## # A tibble: 4 x 5
##   term          estimate  std.error statistic  p.value
##   <chr>            <dbl>      <dbl>     <dbl>    <dbl>
## 1 (Intercept) 44.1       3.12           14.1  2.96e-14
## 2 wt          -0.00650   0.00131        -4.95 3.22e- 5
## 3 disp        -0.0564    0.0132         -4.26 2.10e- 4
## 4 wt:disp      0.0000117 0.00000326      3.60 1.23e- 3
```

Finally, let's plot the residuals of our final model to check for any strangeness.

```
plt1 <- ggplot(mapping = aes(x = linear8$fitted.values, y = linear8$residuals)) +
        geom_point() +
        geom_hline(mapping = aes(yintercept = 0)) +
        geom_smooth(se = FALSE, color = "#014d64") +
        labs(title = "Without Transmission") +
        ylab("Residuals")+
        xlab("Fitted Values")+
        theme_economist()

plt1
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Not fantastic, or terrible.

Let's plot the residuals of the model with a transmission variable

```
library(cowplot)
```

```
## Warning: package 'cowplot' was built under R version 4.1.1
```

```
##
## Attaching package: 'cowplot'
```

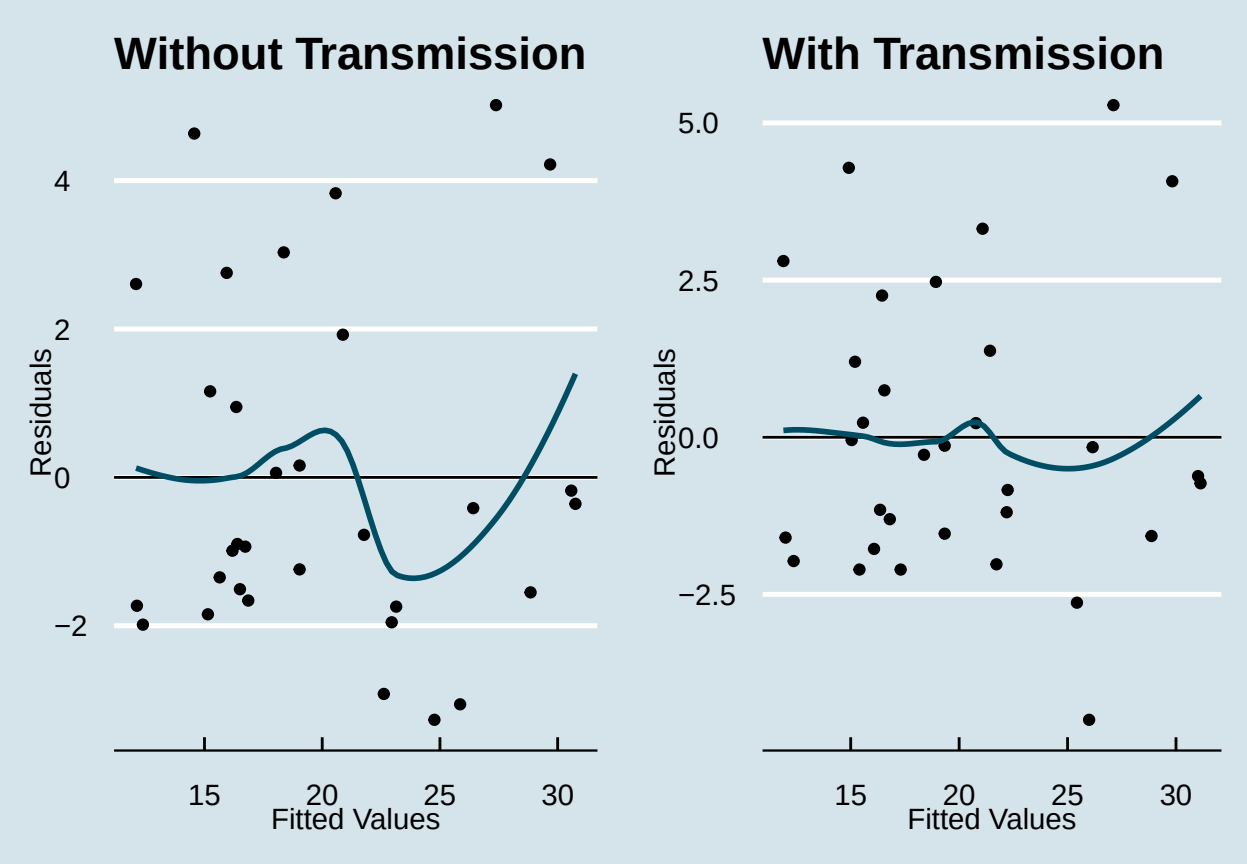
```
## The following object is masked from 'package:ggthemes':
##
##     theme_map
```

```
plt2 <- ggplot(mapping = aes(x = linear4$fitted.values, y = linear4$residuals)) +
        geom_point() +
        geom_hline(mapping = aes(yintercept = 0)) +
        geom_smooth(se = FALSE, color = "#014d64") +
        labs(title = "With Transmission") +
        ylab("Residuals") +
        xlab("Fitted Values") +
        theme_economist()

cowplot::plot_grid(plt1, plt2)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

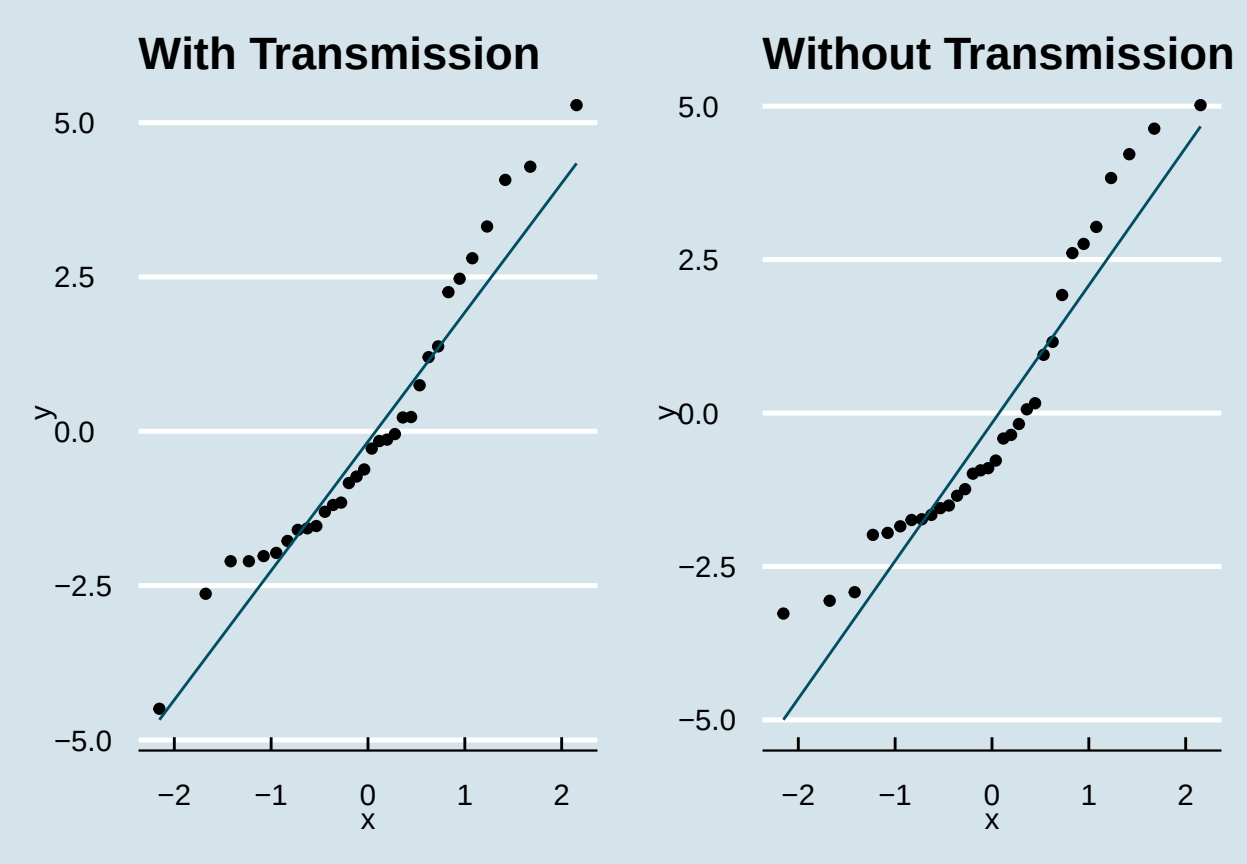


These residuals are much better, so maybe having the transmission type adds something the ANOVA can't see. Let's do a qq plot and find out if there are any other patterns

```
plt1 <- ggplot(cars, aes(sample = linear4$residuals)) +
        geom_qq() +
        geom_qq_line(color = "#014d64") +
        labs(title = "With Transmission")+
        theme_economist()

plt2 <- ggplot(cars, aes(sample = linear8$residuals)) +
        geom_qq() +
        geom_qq_line(color = "#014d64") +
        labs(title = "Without Transmission")+
        theme_economist()

cowplot::plot_grid(plt1, plt2)
```



It looks like adding transmission to the model disrupts some of the issues with the residuals.

Interpretation

Let's look at our coefficients again

```
model <- linear8
tidy(model)
```

```
## # A tibble: 4 x 5
##   term          estimate  std.error statistic  p.value
##   <chr>            <dbl>      <dbl>     <dbl>    <dbl>
## 1 (Intercept) 44.1       3.12           14.1  2.96e-14
## 2 wt          -0.00650   0.00131        -4.95 3.22e- 5
## 3 disp        -0.0564    0.0132         -4.26 2.10e- 4
## 4 wt:disp      0.0000117 0.00000326      3.60 1.23e- 3
```

- All else being equal, a car starts with 44 MPG.
- All else being equal, for every one pound increase in weight, the MPG drops by -.006 MPG, or for
- All else being equal, for every cubic inch increase in displacement, the MPG drops by -.05 MPG

The original questions of

- Is an automatic or manual transmission better for MPG
- Quantify the MPG difference between automatic and manual transmissions

are not answerable in this dataset. All indications are that transmission has much less to do with the MPG of car as opposed to weight and the displacement of the engine.

To conduct further analysis, a dataset with the same car, but with two different kinds of transmission, and their MPG, would be necessary for a direct apples to apples comparison.

As it stands, it is far more likely for a car to be made light, with small engines, and with a manual transmission (most likely to save money) than for a car to be made heavy, with a large engine, and a manual transmission for that comparison.