Regression Project - mtcars dataset

Willem Abrie

2022-12-07

Executive Summary

Study questions: 1. "Is an automatic or manual transmission better for MPG" 2. "Quantify the MPG difference between automatic and manual transmissions"

We found that, based on the sample that was studied, there is a trend for manuals to on average have a higher MPG, but we cannot infer with much certainty that the transmission type (manual or auto) necessarily makes a difference to fuel mileage. The differences borne out by the data is most likely due to other factors such as vehicle weight and engine displacement. Without knowing how the sample was collected we cannot assume that it is representative of the population and we strongly suspect that it is not.

The average difference in MPG performance in the mtcars sample is that manual cars have a 7.245 (+-1.764422 SE) higher miles per gallon performance compared to autos. But the transmission type only explains about 34% of the variation in the data.

When modeled along with vehicle weight as a predictor, it can be shown that manuals have a gas mileage advantage at at the low end of the max spectrum, but that the MPG deteriorates faster (an extra -5.30 +-1.44 SE mpg per 1000 lbs) as mass increases compared to autos. This model explains 81% of the variation, yet the data is too limited to able to draw firm conclusions. No data is available for autos in the low mass range and the opposite for manuals. The entire region of data overlap (in terms of mass) has overlapping confidence intervals.

Loading Initial Packages

```
package_names <- c("tidyverse", "GGally", "rstatix", "ggpubr", "car")

check_install_load_packages <- function(package_names){
    for (i in package_names) {
            if(!i %in% installed.packages()){
                install.packages(i)
                }
            if(!i %in% (.packages())){
                library(i, character.only = TRUE)
                }
    }
}

check_install_load_packages(package_names)
#(.packages())
```

Load the data - mtcars.

```
data_raw <- mtcars

#save data into new data frame and convert categorical variables to factors
data_cars <- data_raw
data_cars$cyl <- factor(data_cars$cyl, levels = c("4", "6", "8"),
                                       labels = c("four_cyl", "six_cyl", "eigth_cyl"))
data_cars$vs <- factor(data_cars$vs, levels = c("0", "1"),
                                       labels = c("V_cyls", "Straight_cyls"))
data_cars$am <- factor(data_cars$am, levels = c("0", "1"),
                                       labels = c("auto", "manual"))
```

Data Exploration

Plot of the data

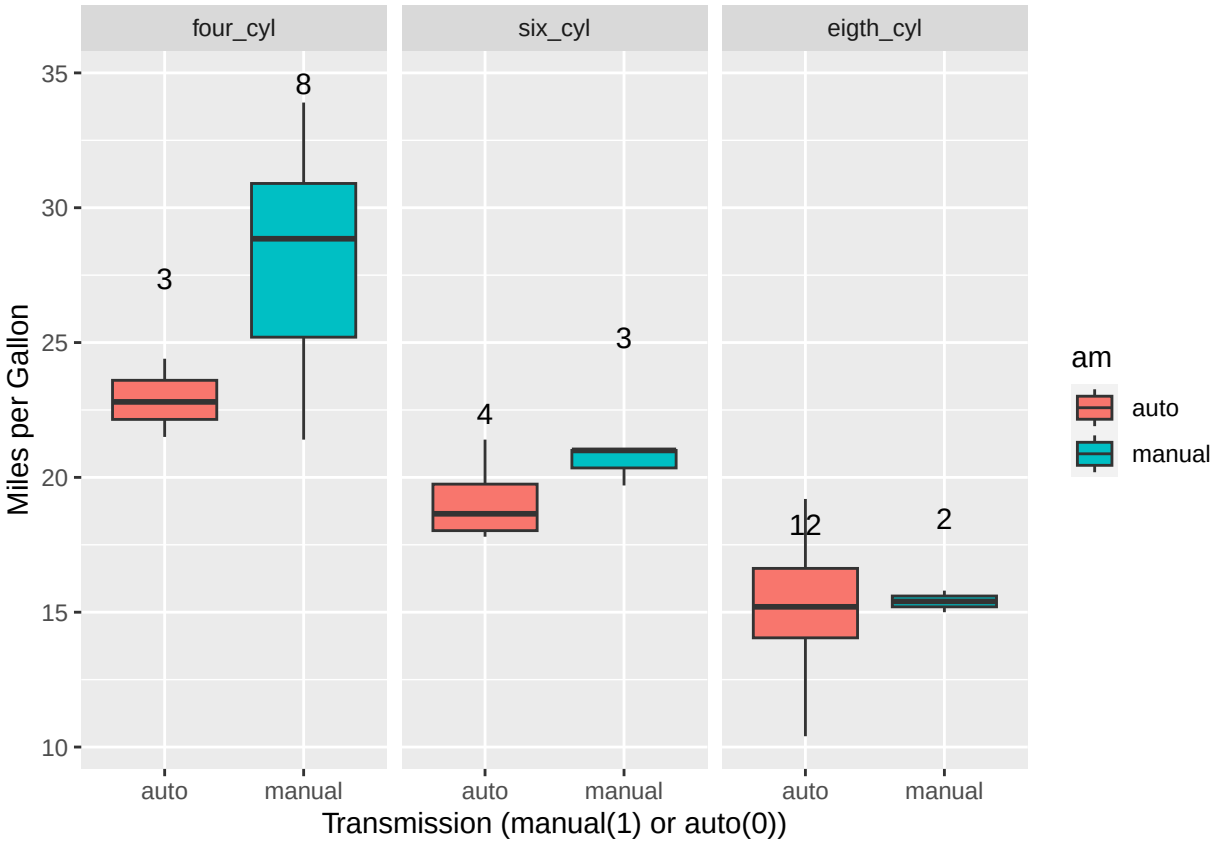
Let's see if there is more to the data than this simple relationship by looking at how the performance varies by cylinders.

```
give.n <- function(x){
  return(c(y = median(x)*1.2, label = length(x)))
  # experiment with the multiplier to find the perfect position
}

boxPlot_mpg_am = ggplot(data = data_cars, aes(x = am, y = mpg)) +
        xlab("Transmission (manual(1) or auto(0))") +
        ylab("Miles per Gallon") +
        geom_boxplot(aes(fill = am)) +
        facet_wrap("cyl") +
        stat_summary(fun.data = give.n, geom = "text", fun.y = median,
                     position = position_dodge(width = 0.75))
```

```
## Warning: The ‘fun.y‘ argument of ‘stat_summary()‘ is deprecated as of ggplot2 3.3.0.
## i Please use the ‘fun‘ argument instead.
```

```
boxPlot_mpg_am
```



We note: 1. The mgp performance is clearly better for smaller numbers of cylinders 2. The difference in mpg performance between autos and manuals decreases for more cylinders 3. There are very few observations in each group so all of this is pretty sketchy

We did a linear model predicting mpg by number of cylinders and found that it would have a greater R squared - i.e. 0.72 compared to about 0.34 for modelling on transmission type.

For a bit more detail on some of the other variables that we intuitively know are important we run the ggpairs() function from the GGally package. Refer to Plot 1 in the Appendix.

We note the following from the pairs plots: 1. The auto cars in the dataset were generally heavier 2. The auto cars in the dataset generally had higher displacement engines 3. The auto cars in the dataset tended to have more cylinders

Hence the auto cars will probably seem to be worse performers on mpg but probably because of confounders.

Linear Model

**Fist model: mpg_hat = beta0 + beta1*am_manual **

```
linmod <- lm(formula = mpg ~ am, data_cars)
summary(linmod)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## ammanual     7.244939   1.764422  4.106127 2.850207e-04
```

We can see from the plot and coefficients that mean MPG of the auto cars is 17.147 miles per gallon, which is the intercept. The difference in mean when going to the manual group is 7.245 miles per gallon (the slope or estimated coefficient "am"). I.e. the mean mpg for the manuals is 17.147 + 7.245 = 24.392 miles per gallon. This linear relationship only explains about 34% of the variation (R squared).

We are not surprised at this outcome, given what we noted in our data exploration.

Second model: mpg_hat = beta0 + beta1*wt* + *beta2*am_manual - i.e. we include weight as a predictor, which we reason should be the main real predictor. We add interaction so we can see the change in slope of considering manuals relative to autos. Weight is prefered as predictor above cyl and disp because it is more highly correlated with mpg, as can be shown with *cor(x = mtcars[,c("wt", "cyl", "disp")], y = mtcars$mpg)*.

```
linmod2 <- lm(formula = mpg ~ wt * am , data_cars)
summary(linmod2)$coef
```

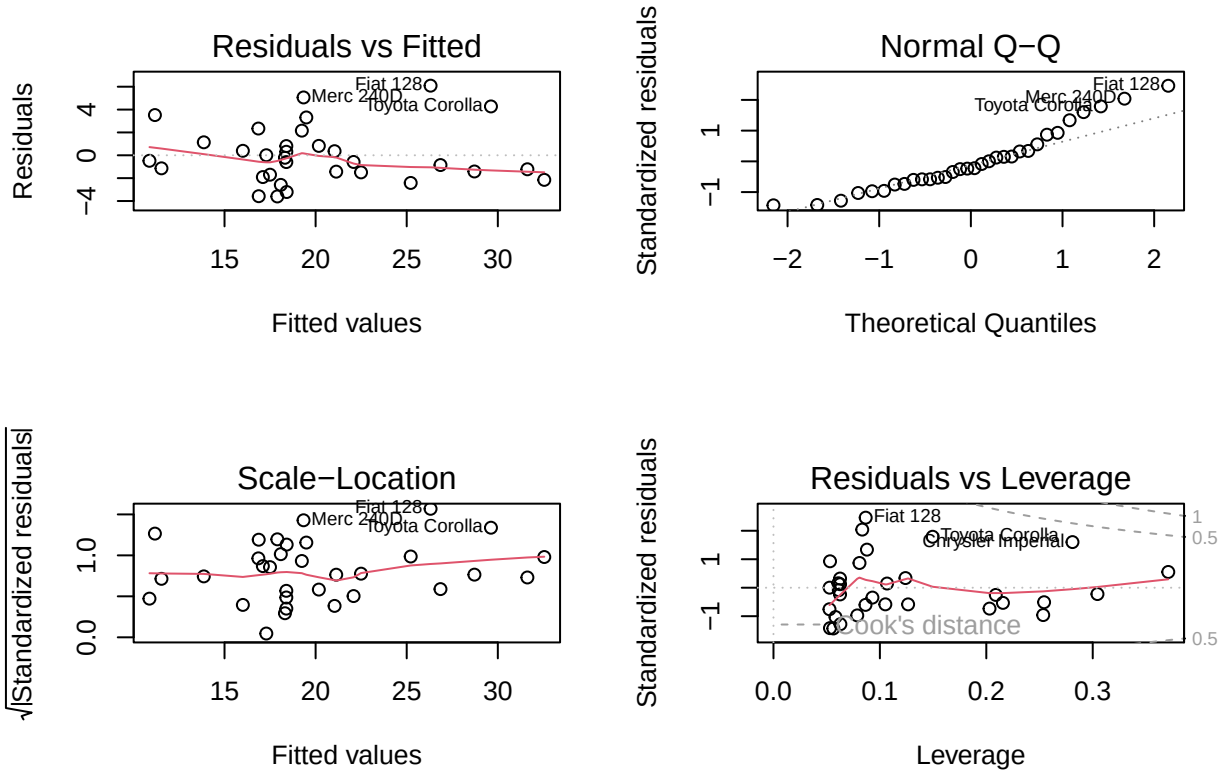
```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 31.416055  3.0201093 10.402291 4.001043e-11
## wt          -3.785908  0.7856478 -4.818836 4.551182e-05
## ammanual    14.878423  4.2640422  3.489277 1.621034e-03
## wt:ammanual -5.298360  1.4446993 -3.667449 1.017148e-03
```

See plot 2 in the appendix. Manuals are more fuel efficient under roughly the 2800 pound mass. Thereafter, autos perform better.

Residuals

But let's check the residuals to see if our linear model is reliable. Residuals Diagnostic plots:

```
par(mfrow = c(2,2))
plot(linmod2)
```

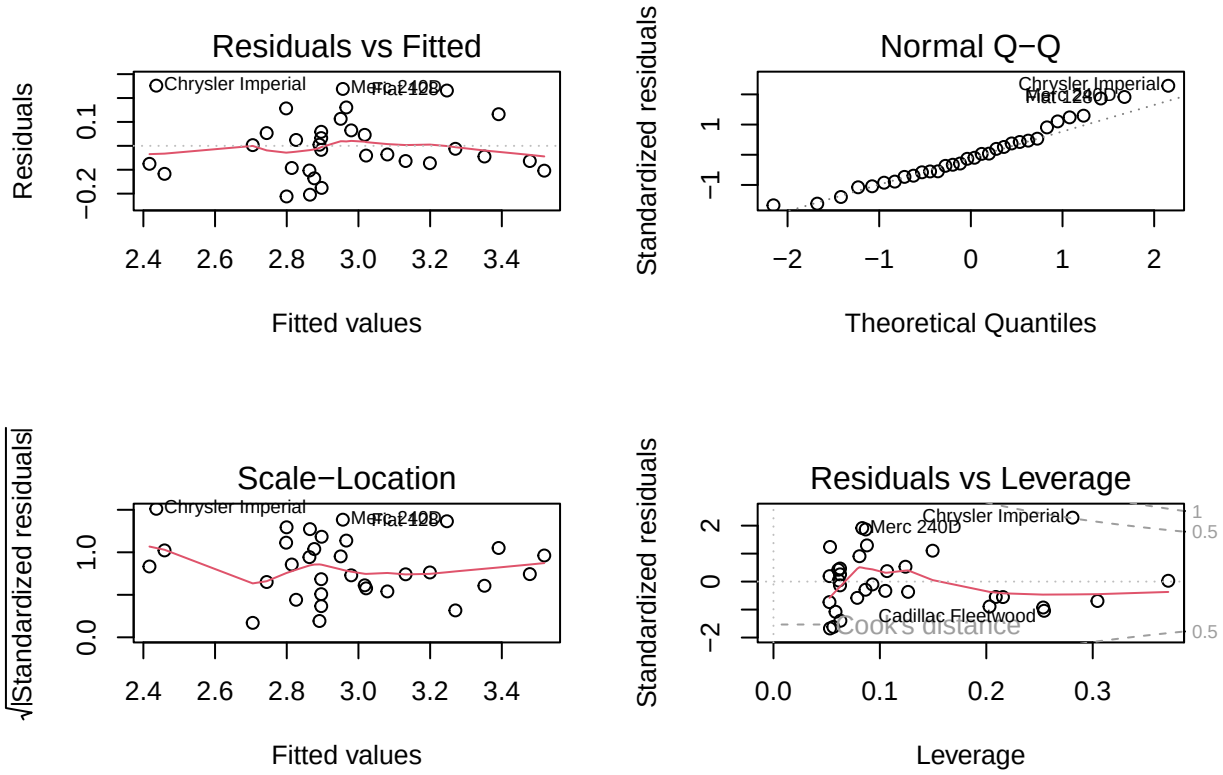


The data looks roughly normal on the QQ plot but shows some departure at the higher residuals. The top left hand plot shows some heteroskedasticity, which violates our linear regression assumption of equal variances in the residuals. We'll try to remove the heteroskedasticity by doing a log transform of the response.

```
linmod3 <- lm(formula = log(mpg) ~ wt * am , data_cars)
summary(linmod3)$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  3.7267980 0.15145179 24.607157 1.656822e-20
## wt          -0.2414783 0.03939850 -6.129124 1.293878e-06
## ammanual     0.3898126 0.21383227  1.822983 7.900069e-02
## wt:ammanual -0.1539089 0.07244847 -2.124391 4.260222e-02
```

```
par(mfrow = c(2,2))
plot(linmod3)
```



We note now the residuals look a bit better but still not great - indicating there is still a problem of unequal variance, which casts doubt over the accuracy of our linear model. A square root and a cube root transformation has a similar result.

After the slightly helpful log transformation, the influence of the transmission type has lost much of its statistical significance with the t test probability going from 0.16% to 8%. At a 95% confidence requirement we'll fail to reject the null hypothesis that transmission type has no effect. Some additional residual plots are shown in the Appendix demonstrating that there isn't a very strong relationship between mpg and transmission type after having removed the effect of weight.

Conclusions and Discussion of Assumptions

We were not able, by linear regression, to determine a convincing relationship between transmission type and MPG performance. Not enough direct evidence exists in this data set to perform a meaningful comparison of the variables in question. Unequal variances result in inaccurate linear regression modeling outcomes.

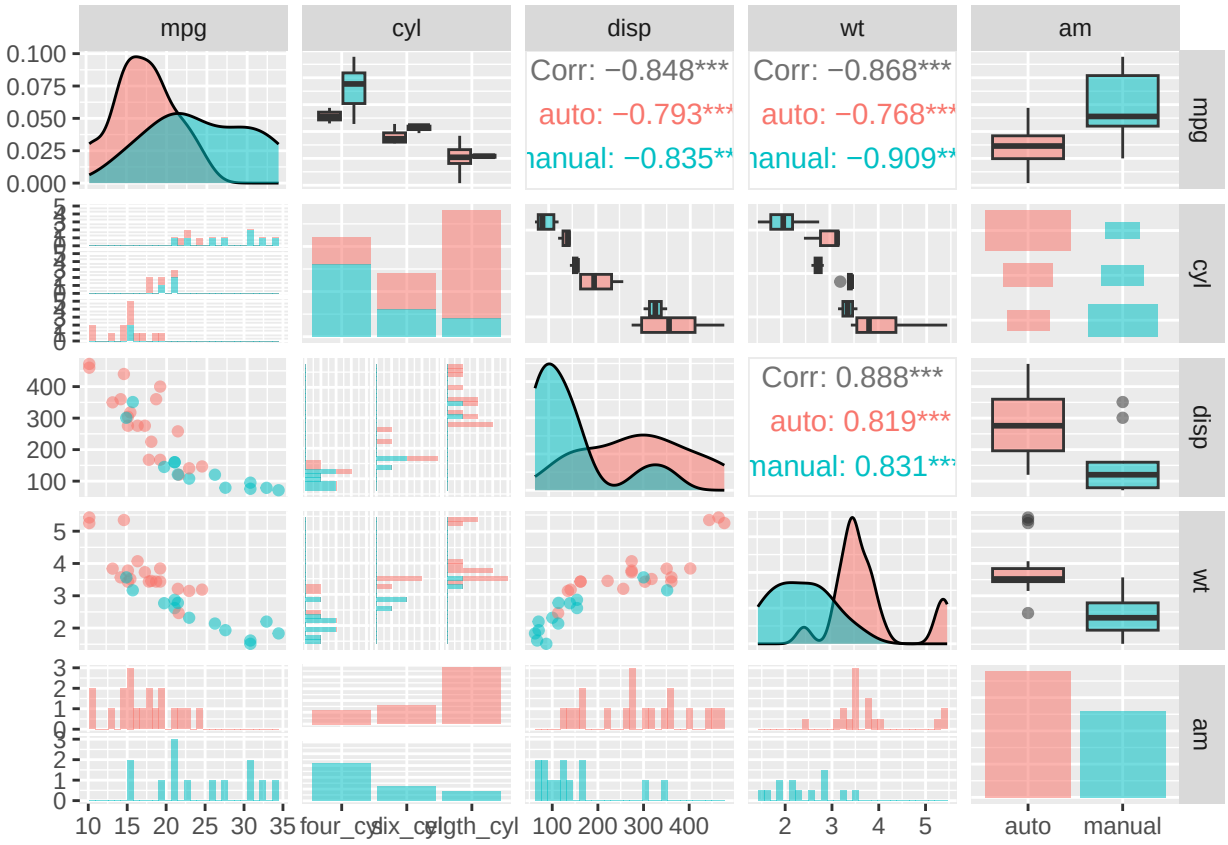
Appendix

Plot 1 - ggpairs exploration

```
#based on the questions, we'll limit our exploration to mpg and transmission, and sneak in displacement
ggpairs(subset(data_cars, select = c(mpg, cyl, disp, wt, am) )
        ,ggplot2::aes(colour=am, alpha = 0.8)#,
```

```
        #upper = list(continuous = "density", combo = "box_no_facet")
        )
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Plot 2 - linear model of mpg predicted on weight and transmission type, with interaction

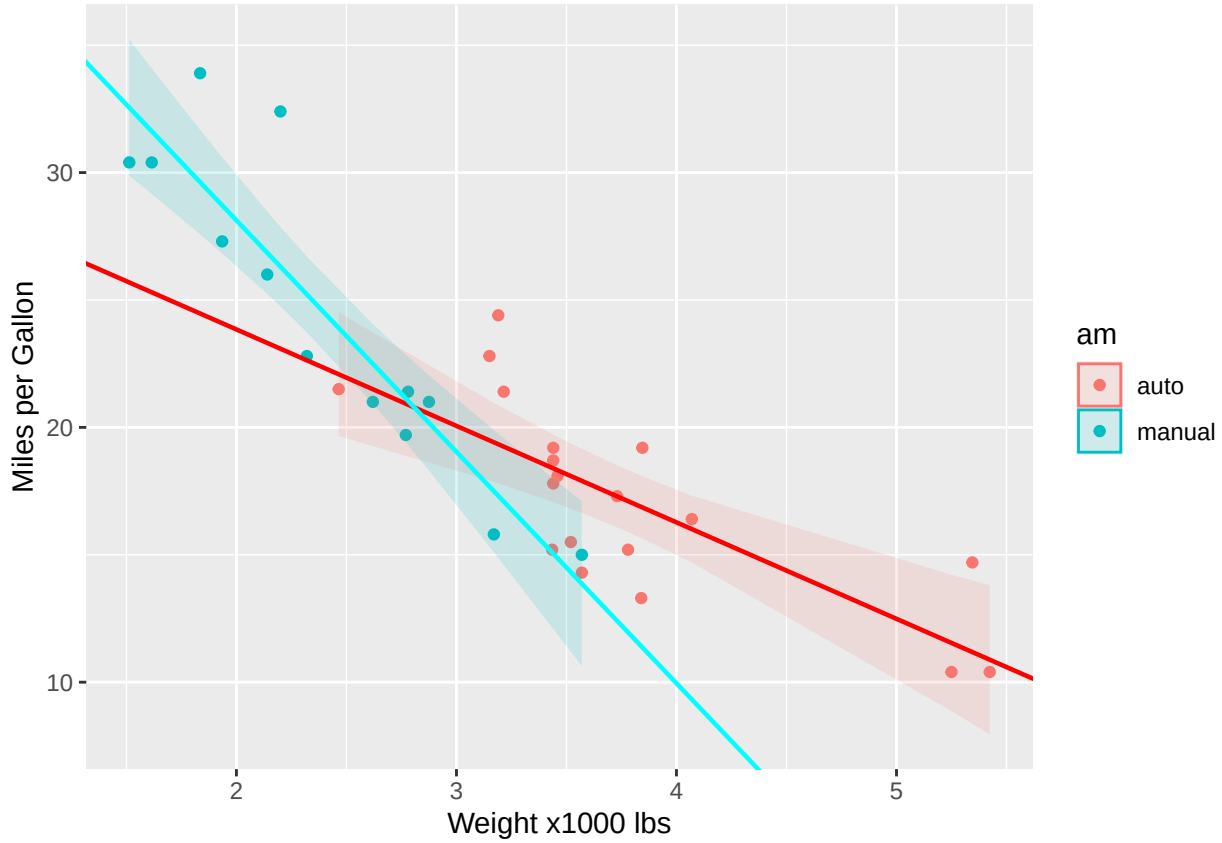
```
predlm = predict(linmod2, interval = "confidence")
data_cars_pred <- cbind(data_cars, predlm)

lmPlot_mpg_wt_am = ggplot(data_cars_pred, aes(x = wt, y = mpg, colour = am)) +
        xlab("Weight x1000 lbs") +
        ylab("Miles per Gallon") +
        geom_point() +
        geom_ribbon( aes(ymin = lwr, ymax = upr, fill = am, color = NULL), alpha = .15) +
        geom_abline(intercept = coef(linmod2)[1], slope = coef(linmod2)[2], size = 0.8, colour = "red")
```

```
        geom_abline(intercept = coef(linmod2)[1] + coef(linmod2)[3],
                    slope = coef(linmod2)[2] + coef(linmod2)[4], size = 0.8, colour = "cyan")
```

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
```

```
lmPlot_mpg_wt_am
```

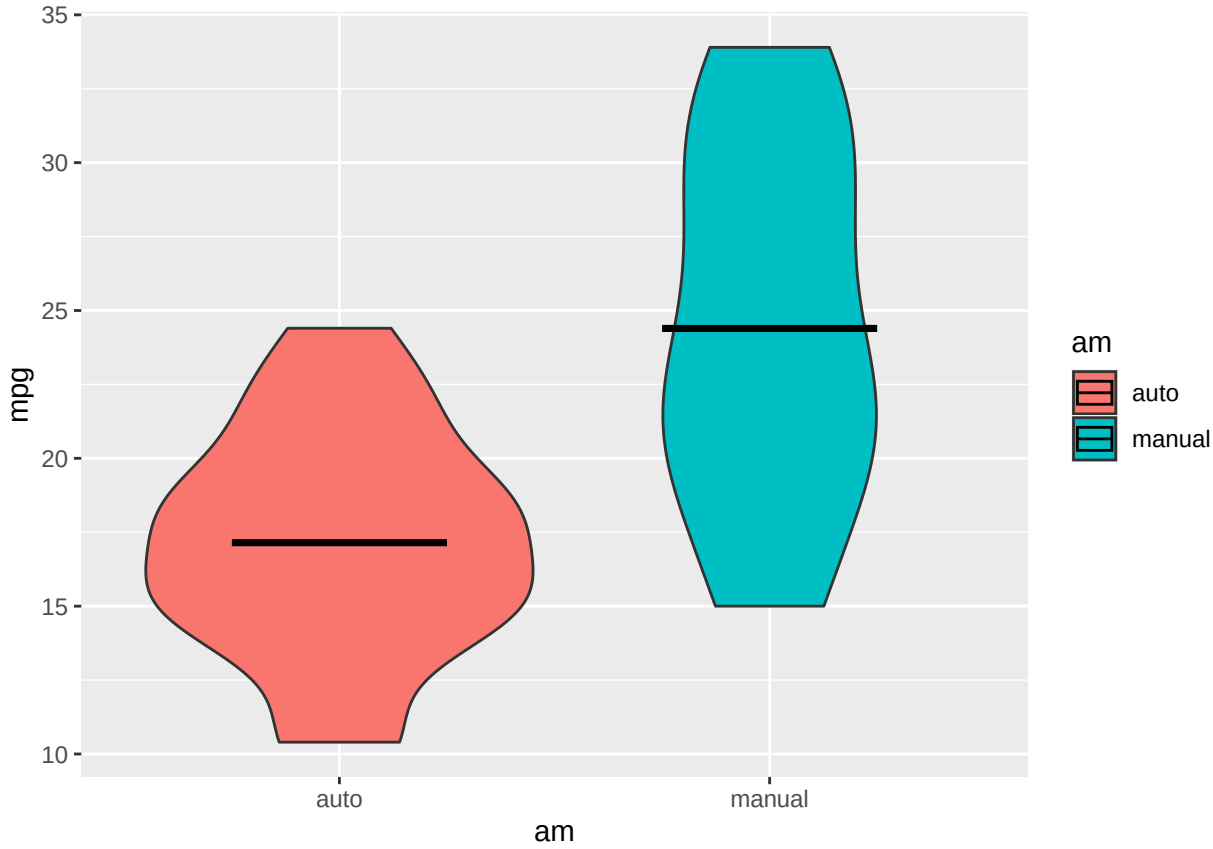


Once again this shows that the data is too limited to able to draw many conclusions. No data is available for autos in the low mass range and the opposite for manuals. The entire region of data overlap (in terms of mass) has overlapping confidence intervals.

Plot 3 - violin plot.

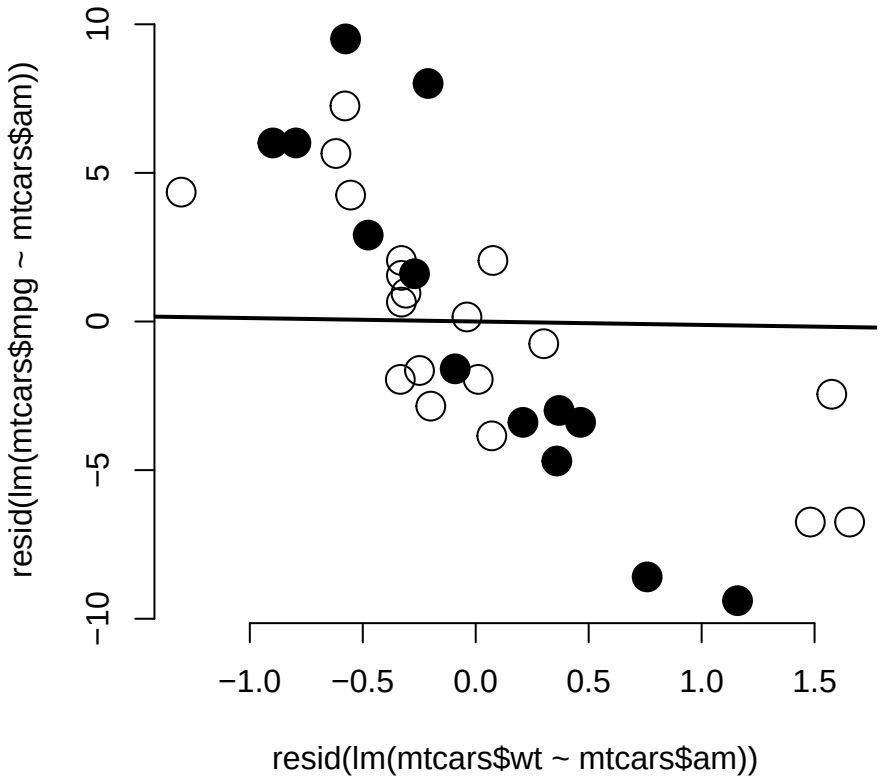
This shows that the mpg values are not identically distributed around the mean when comparing between transmission type groups.

```
v_plot <- ggplot(data_cars, aes(x = am, y = mpg, fill = am)) +
        geom_violin() +
        stat_summary(fun = "mean",
               geom = "crossbar",
               width = 0.5,
               color = "black")
v_plot
```



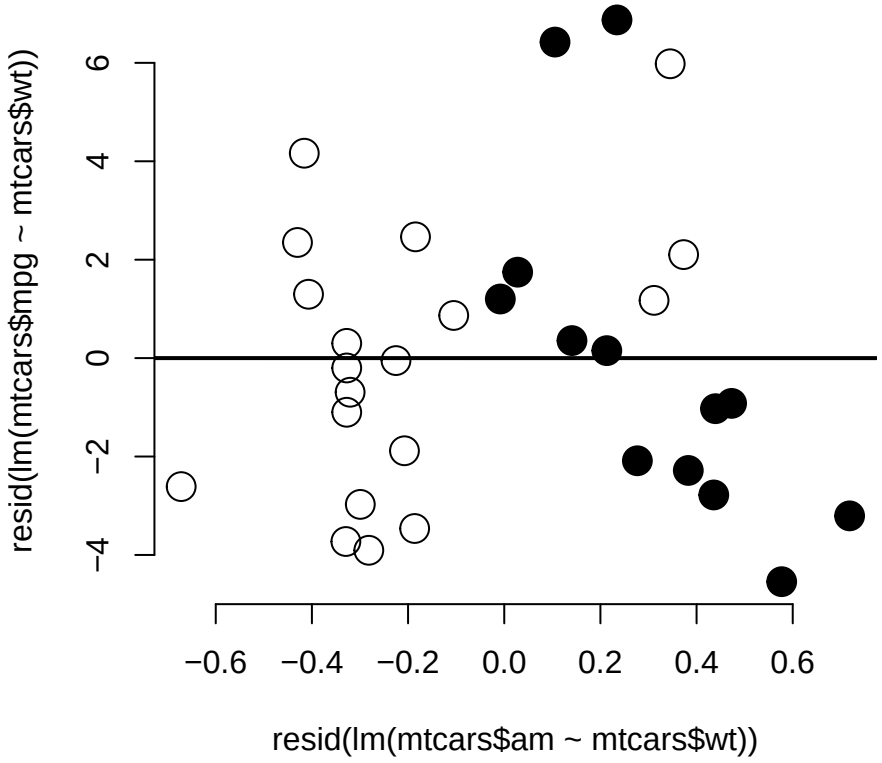
Plot6 - residuals for adjusted mpg

First we plot the residuals of mpg against weight after have removed the affect of transmission type.



There is a clear relationship between mpg and weight (wt) after having removed the effect of transmission type (am).

Secondly we plot the residuals of mpg against transmission type after have removed the affect of weight.

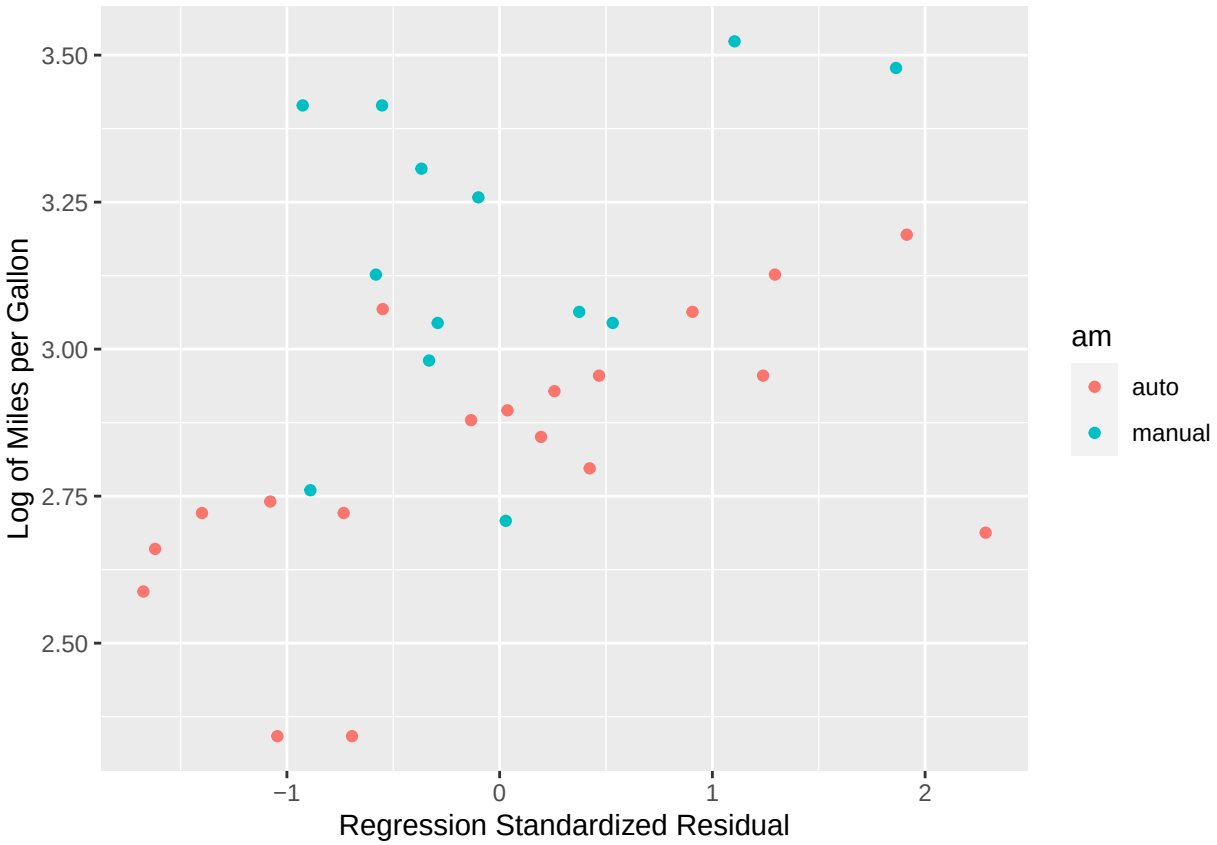


There is not a clear relationship between mpg and transmission type (am) after having removed the effect of weight (wt).

Standardized residual plot showing fanning out pattern - indicative of unequal variance:

```
standard_res <- rstandard(linmod3)

resplot1 = ggplot(data_cars_pred, aes(x = standard_res, y = log(mpg), colour = am)) +
        xlab("Regression Standardized Residual") +
        ylab("Log of Miles per Gallon") +
        geom_point()
resplot1
```


3.50
3.25
3.00
2.75
2.50
Log of Miles per Gallon
−1
0
1
2
Regression Standardized Residual
am
auto
manual