Manual vs automatic car transmission. Which one has better fuel economy?

An Inferential Data Analysis of the Mtcars Data. Regression Models

Darwin Reynell Nava

Jun 09, 2021

- **Background**: Motor Trend is a magazine about the automobile industry. They are interested in exploring the relationship between car's transmission type and miles per gallon (MPG).
- **Objectives**: Determine association between car's transmission type and miles per gallon (mpg) using regression models. Determine which one has better fuel economy. Quantify the MPG difference between automatic and manual transmissions.
- **Methods**: An statistical inference analysis in R. Regresión models shoulb be used. The mtcars data available at The R Datasets Package are used.
- **Results**: 1. It was found a significant association between car's transmission type and miles per gallon (mpg) in our Fit1 model (unadjusted) where manual transmission has better fuel economy than automatic one. The fit4 model (adjusted and with best performance) show than holding wt and cyl constant, the transmission types appears to have almost the same impact on mpg. 2. Related to the MPG difference between automatic and manual transmission, we see at the Fit1 model that the mean (y-value) of the level "Manual" is 7.245 units higher (24.392) than the mean (y-value) of "Automatic" (which is listed as the intercept, 17.147), pvalue:0.000285.
- **Conclusions**: From a mechanical design point of view, manual transmission engines are less complex, weigh less, and have more gears than automatics,thus favoring higher mpg for manual transmissions.The hypothesis tests showed that the mean wt (weight) for automatic transmission is higher than for manual transmission. Analogously, the average mpg is lower for automatic transmission.

Github link (all project and code can be observed here)

Data processing. An exploratory statistical analysis. Summary of the data.

Figures 1 to 3 in Appendix let's see some aspects of automobile design and performance.

Observations: Grouping by engine, number of cylinders, number of forward gears, and number of carburetors, manual transmission shows better fuel efficiency (mpg) than the automatic one. Continuous variables vs mpg plots show similar trends for both types of transmissions (for example: the lower the weight, the higher the mpg or the higher the Rear axle ratio, the lower the mpg, for both transmissions).

Mean of continuous variables by transmission type and T.test for wt and mpg

Assumptions: independent samples and randomly selected from the population. Outliers were not observed.
Unequal variances. 95% confidence level. Ha: mpg|automatic < mpg|manual and Ha: wt|automatic > wt|manual

```
## # A tibble: 2 x 7
##   am        mpg_mean wt_mean disp_mean hp_mean drat_mean qsec_mean
##   <fct>        <dbl>   <dbl>     <dbl>   <dbl>     <dbl>     <dbl>
## 1 automatic     17.1    3.77      290.    160.      3.29      18.2
## 2 manual        24.4    2.41      144.    127.      4.05      17.4
```

**Results of T-Test: 1. mpg|automatic < mpg|manual,. 2. wt|automatic > wt|manual.

An Inferential Data Analysis of the "mtcars" data

Adjusted models

```
fit1 <- lm(mpg ~ am, data=by_am_vs_cyl_gear_carb)
fit2 <- lm(mpg ~ am + wt, data=by_am_vs_cyl_gear_carb)
fit3 <- lm(mpg ~ am + wt + disp , data=by_am_vs_cyl_gear_carb)
fit4 <- lm(mpg ~ am + wt + cyl , data=by_am_vs_cyl_gear_carb)
fit5 <- lm(mpg ~ am + wt + disp + cyl , data=by_am_vs_cyl_gear_carb)
fit6 <- lm(mpg ~ am + wt + cyl + carb, data=by_am_vs_cyl_gear_carb)
fit7 <- lm(mpg ~ am + wt + cyl + gear, data=by_am_vs_cyl_gear_carb)
fit8 <- lm(mpg ~ am + wt + cyl + hp, data=by_am_vs_cyl_gear_carb)
fit9 <- lm(mpg ~ am + wt + cyl + vs, data=by_am_vs_cyl_gear_carb)
fit10 <- lm(mpg ~ am + wt + cyl + drat, data=by_am_vs_cyl_gear_carb)
fit11 <- lm(mpg ~ am + wt + cyl + qsec, data=by_am_vs_cyl_gear_carb)
fit12 <- lm(mpg ~ ., data=by_am_vs_cyl_gear_carb)
```

Observations: Some significant improvements in some model comparisons. Fit4 and Fit8 show better performance (Anova analysis). Note: if the resulting p-value is sufficiently low (usually less than 0.05), we conclude that the more complex model is significantly better than the simpler model, and thus favor the more complex model. If the p-value is not sufficiently low (usually greater than 0.05), we should favor the simpler model.

Variance inflation factor, VIF for fit4 and fit 8

```
## [1] "VIF - FIT4"
##         GVIF Df GVIF^(1/(2*Df))
## am  1.925620  1        1.387667
## wt  3.611208  1        1.900318
## cyl 2.585745  2        1.268079
## [1] "VIF - FIT8"
##         GVIF Df GVIF^(1/(2*Df))
## am  2.590777  1        1.609589
## wt  4.007113  1        2.001778
## cyl 5.824545  2        1.553515
## hp  4.703625  1        2.168784
```

Observations: Fit4 has VIFs less than 5. This model is the selected. VIF measures the correlation and strength of correlation between the predictor variables in a regression model: VIF=1, no correlation / 1<VIF<=5,moderate correlation but his is often not severe / VIF>=5 severe correlation in this case, the coefficient estimates and p-values in the regression output are likely unreliable**.**

Model 1 Review (fit1). The 95% confidence intervals for both coefficients

```
## lm(formula = mpg ~ am, data = by_am_vs_cyl_gear_carb)
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## ammanual       7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                2.5 %   97.5 %
## (Intercept) 14.85062 19.44411
## ammanual     3.64151 10.84837
```

Observations: There is a significant association between college major category and income.(p-value: 0.000285) The intercept is the estimated mean for the reference level "Automatic". The intercept t-test tests for whether or not the mean for the reference level is 0 (p-value. 1.13e-15, Ho: rejected). The other t-tests is for comparison of the other level versus the reference level "Automatic". We see that the mean (y-value) of the level "Manual" is 7.245 units higher than the mean (y-value) of "Automatic" (which is listed as the intercept). The t-test is simply testing if the difference between, say the 'Manual' coefficient and the reference category, "Automatic" is different than zero: 24.392 - 17.147 = 7.245; is that absolute difference greater than zero? Yes, pvalue: 0.000285.

Model 4 Review (fit4). The 95% confidence intervals for both coefficients

```
## lm(formula = mpg ~ am + wt + cyl, data = by_am_vs_cyl_gear_carb)
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.7536     2.8135  11.997  2.5e-12 ***
## ammanual      0.1501     1.3002   0.115  0.90895
## wt           -3.1496     0.9080  -3.469  0.00177 **
## cyl6         -4.2573     1.4112  -3.017  0.00551 **
## cyl8         -6.0791     1.6837  -3.611  0.00123 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.603 on 27 degrees of freedom
## Multiple R-squared:  0.8375, Adjusted R-squared:  0.8134
## F-statistic: 34.79 on 4 and 27 DF,  p-value: 2.73e-10

##                 2.5 %    97.5 %
## (Intercept) 27.980802 39.526382
## ammanual    -2.517734  2.817941
## wt          -5.012761 -1.286434
## cyl6        -7.152943 -1.361694
## cyl8        -9.533813 -2.624425
```

Observations: Holding wt and cyl constant, the transmission types appears to have almost the same impact on mpg.

Appendix

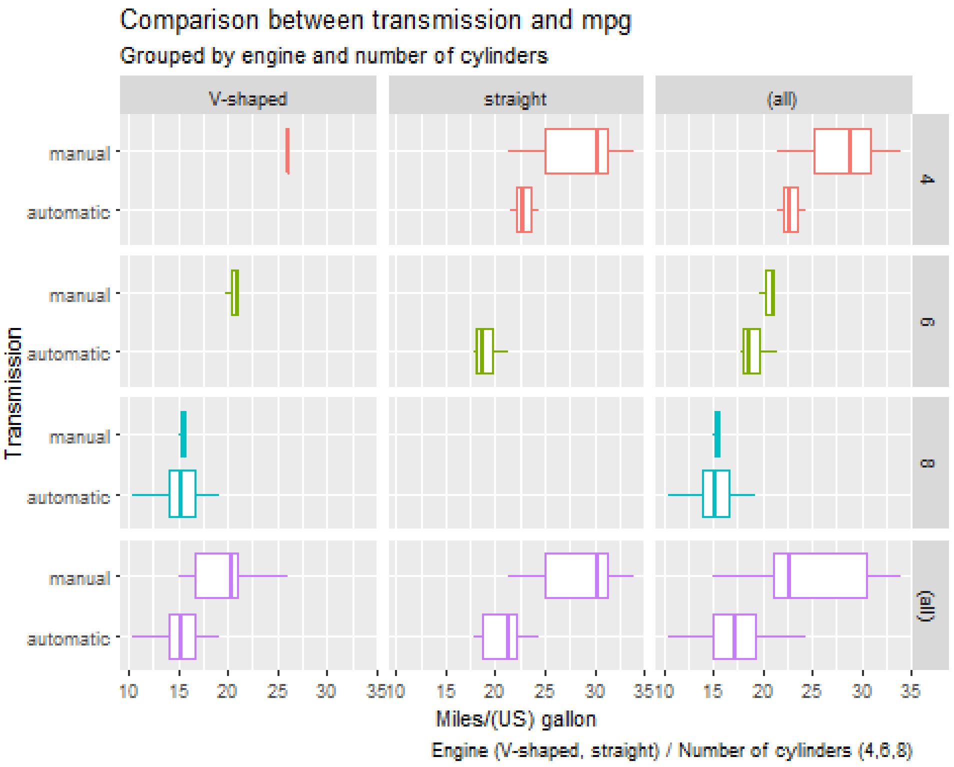


Figure 1

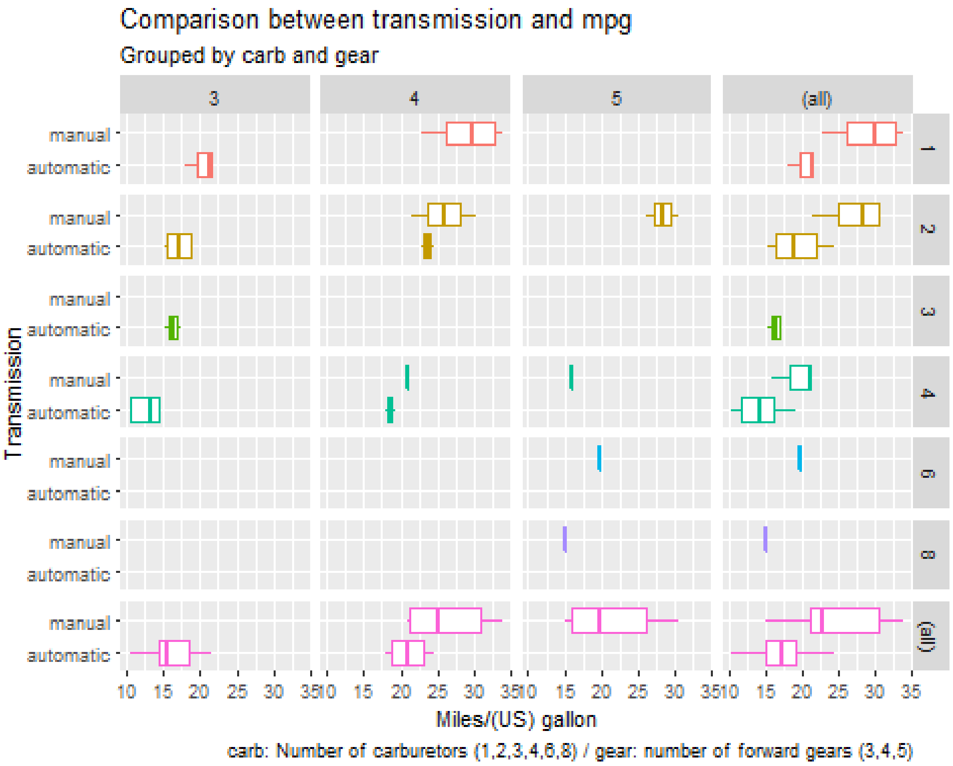
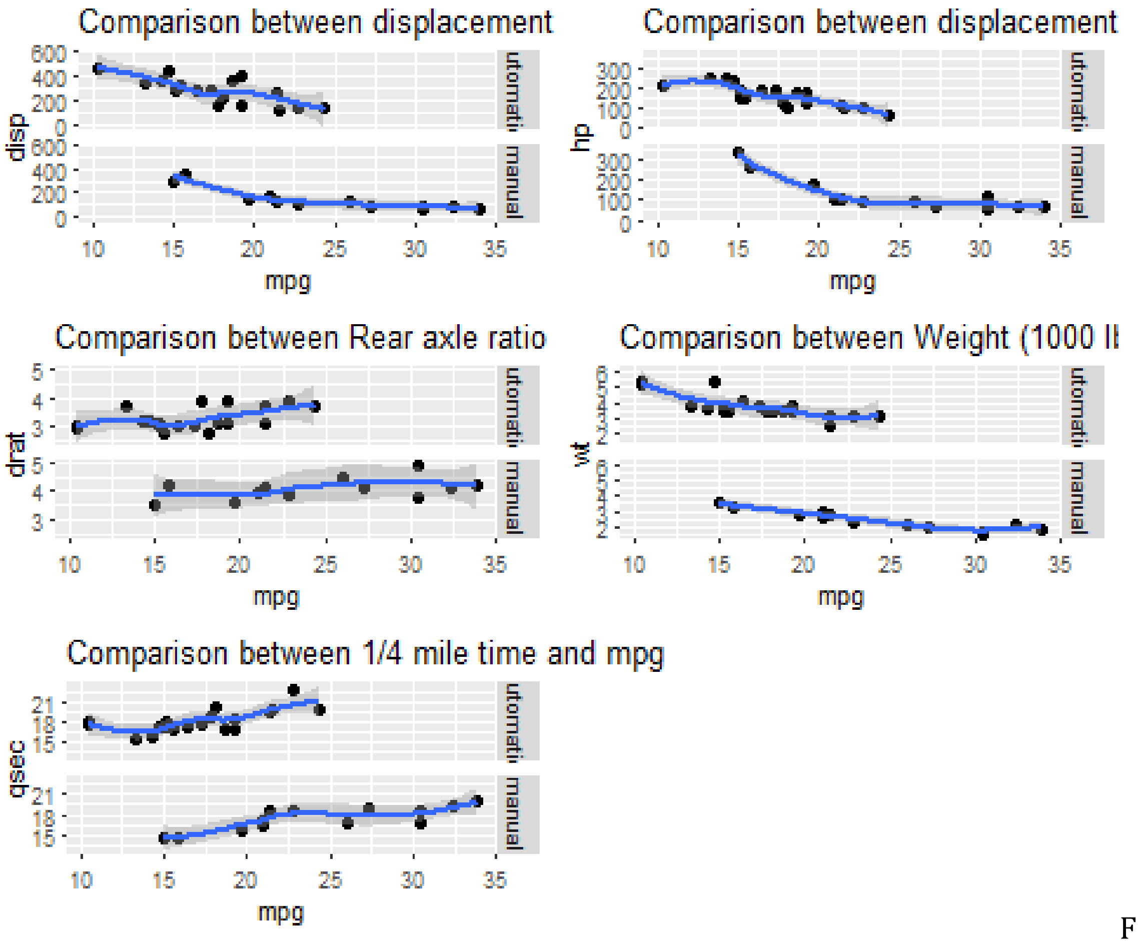


Figure2



Comparison between displacement

Comparison between displacement

Comparison between Rear axle ratio

Comparison between Weight (1000 lb

Comparison between 1/4 mile time and mpg

Figure3

Fit4: Residuals, variation, diagnostics. Influential, high leverage and outlying points

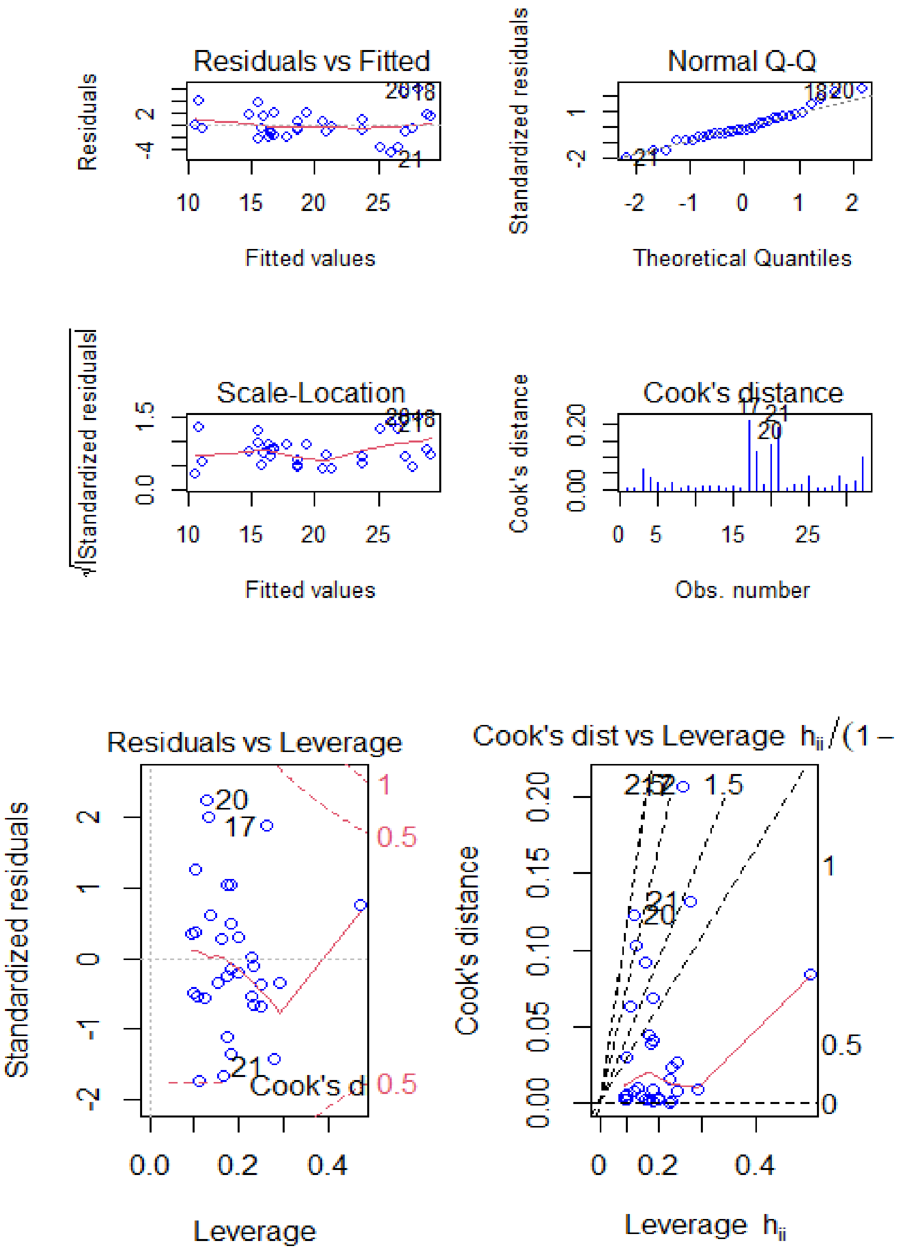


Figure 4

Observations:

The model is good! 1. Residuals vs Fitted plot shows the residuals on the vertical axis and the independent variable on the horizontal axis, the points are randomly dispersed around the horizontal axis (no pattern observed). 2. Normal Q-Q. This plot shows if residuals are normally distributed. QQ plot evaluates the fit of a linear regression model, many points lie approximately on the line so the residuals are Gaussian and thus the errors too. 18 y 20 might be a potential problem. 3. Scale-Location. This plot shows if residuals are spread equally along the ranges of predictors. This is how you can check the assumption of equal variance (homoscedasticity). The residuals are randomly spread and the red smooth line is not horizontal and shows a steep angle. 4. Residuals vs Leverage. This plot helps us to find influential cases if any. Some outliers can be influential in linear regression analysis. Look for cases outside of a dashed line, Cook's distance. When cases are outside of the Cook's distance (meaning they have high Cook's distance scores), the cases are influential to the regression results. The regression results will be altered if we exclude those cases.In our graph there is no influential case, or cases. You can barely see Cook's distance lines (a red dashed line) because all cases are well inside of the Cook's distance lines.