Regression analysis of fuel efficiency using R dataset mtcars

Alex Istrate

Contents

The Rmd file for this document is available online at https://github.com/calzzone/coursera_regression_final

Summary

The purpose of this analysis is to find if fuel efficiency (miles / gallorn) is affected by transmission type (automatic vs. manual) and, if yes, by how much. The R dataset, 'mtcars' contains data from 32 cars (each row has 10 columns).

I found that a manual transmission significantly increased fuel efficiency by approximatibely 7.24 miles / gallon compared to automatic. However, adjusting for the ohter covariates resulted in significant fuel usage differences with a particular selection of covariates. Adjusted R-squared values increased from 34% for the unadjusted model to 80-84% for the adjusted models.

These results suggest that fuel efficiency is better in cars with manual transmission than in cars with automatic transmission, but other factors are able to compensate for this difference.

Data preparation

Starting from the 'mtcars' dataset, I prepared two custom datasets for analysis. `mtcars2` kept the numerical type of discrete variables with numerical levels: `cyl` (Number of cylinders: 4, 6 or 8), `gear` (Number of forward gears: 3, 4 or 5) and `carb` (Number of carburetors: 1-4, 6 or 8). Therefore, regression coefficients for these variables show the expected difference of fuel efficiency for evey unit increase in their numeric value and the intercept uses values of 0, even when these values do not appear in the dataset. `mtcars3` coded these variables as factors resulting in dummy variables compared by the regression algorithm to their first level.

In both datasets, variables `am` (Transmission: manual compared to automatic) and `vs` (Engine: straight compared to V-shaped) were coded as factors.

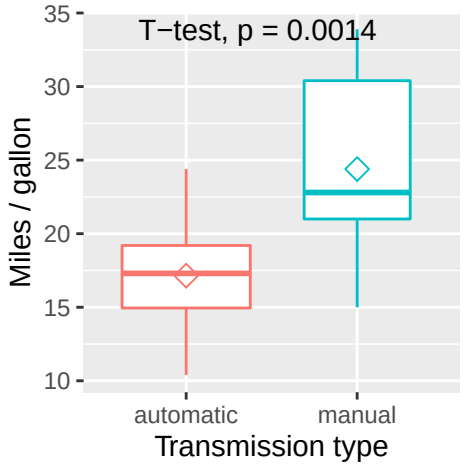
Variables `disp` (Displacement) was rescaled from cubic inches to litters and `hp` (Gross horsepower) was downscaled by a factor of 100 in order to make their coefficients easier to visualize graphically since their unscaled coefficients and confidence intervals were very close to 0, which made them appear as dots rather than ranges. Rescaling only affects the numerical values of their coefficients and their visual aspect but does not affect any related statisitcal inference.

Exploratory analysis

```
##       mpg        cyl         disp             hp             drat
##  Min.   :10.40   4:11   Min.   :1.165   Min.   :0.520   Min.   :2.760
##  1st Qu.:15.43   6: 7   1st Qu.:1.980   1st Qu.:0.965   1st Qu.:3.080
##  Median :19.20   8:14   Median :3.217   Median :1.230   Median :3.695
##  Mean   :20.09          Mean   :3.781   Mean   :1.467   Mean   :3.597
##  3rd Qu.:22.80          3rd Qu.:5.342   3rd Qu.:1.800   3rd Qu.:3.920
##  Max.   :33.90          Max.   :7.735   Max.   :3.350   Max.   :4.930
##        wt             qsec              vs              am     gear   carb
##  Min.   :1.513   Min.   :14.50   V-shaped:18   automatic:19   3:15   1: 7
##  1st Qu.:2.581   1st Qu.:16.89   straight:14   manual   :13   4:12   2:10
##  Median :3.325   Median :17.71                               5: 5   3: 3
##  Mean   :3.217   Mean   :17.85                                      4:10
##  3rd Qu.:3.610   3rd Qu.:18.90                                      6: 1
##  Max.   :5.424   Max.   :22.90                                      8: 1
```

MPG by Transmission

A boxplot of fuel efficiency by type of transmission shows that manual cars use less fuel than automatic cars. This difference is statisticaly signifficant accoding to a T-test.



Regression analysis

Models

I defined 3 regression model to study the relation between the type of transmission (binary variable `am`) and fuel consumption (in miles / gallon, variable `mpg`).

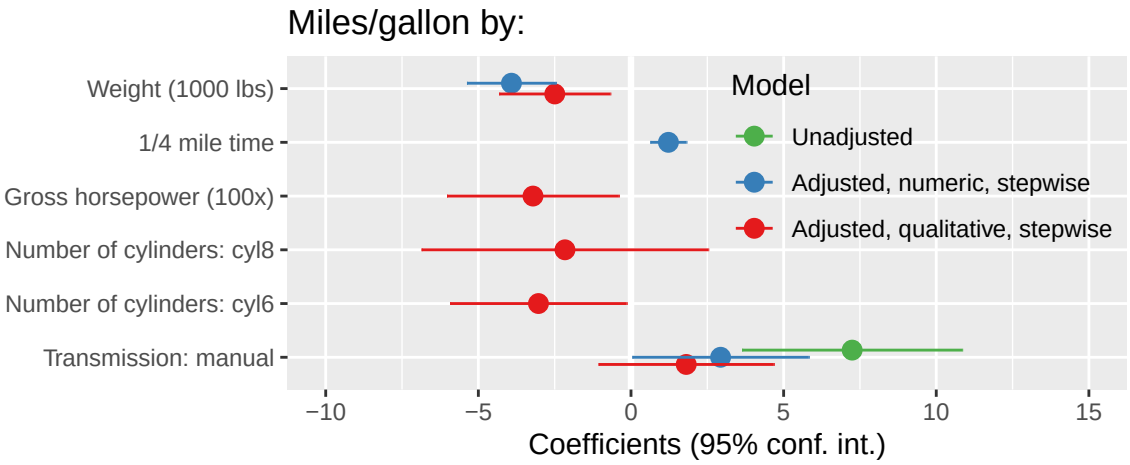
The unadjusted model is just another way to perform the same T-test as above. It is used as refference point for the other models.

I created 2 other models using all other variables as covariates. One of them uses numerical coding for numerical dicrete variables and the other codes them as factors. In order to reduce the number of coeficients, I used a stepwise selection algorithm to select the most important covariates from both adjusted models.

I used the `sjPlot::plot_models` function to create a chart of the coefficients for the two adjusted models and the unadjusted model. This chart shows each coeficient and its 95% confidence interval. There are 3 models, each with a different color (green: unadjusted; blue: adjusted, numeric coding; red: adjusted, qualitative coding).

Using numeric coding, I found that a manual transmission significantly increases fuel economy with only aproximativey 3 miles / gallon compared to automatic, when adjusted for car weight and 1/4 mile time. Also, each 1000lbs car weight reduces reduces fuel economy with approximatively 4 miles / gallon and a slower acceleration (each second added to the 1/4 mile time) increases fuel economy by about 1.2 miles / gallon.

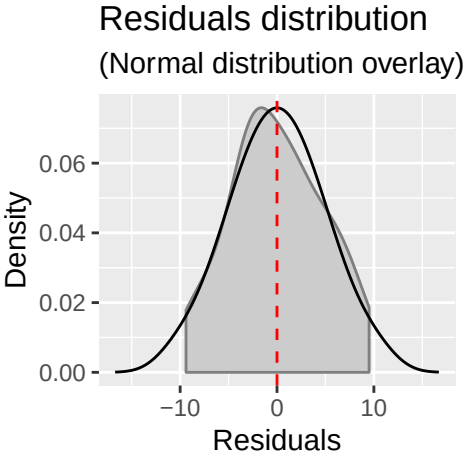
Using factor coding, I found that a manual transmission did significantly increase fuel economy, adjusted for weight, horsepower and number of cylinders. However, larger weight reduced fuel economy by about 2.5 miles / gallon for every 1000 lbs, a more powerful motor reduced fuel economy by about 3.2 miles / gallon for every 100 hp and a 6 cylinder motor reduced fuel economy by about 3 miles / gallon compared to 4 cylinders.



Appendix: Model diagnostics

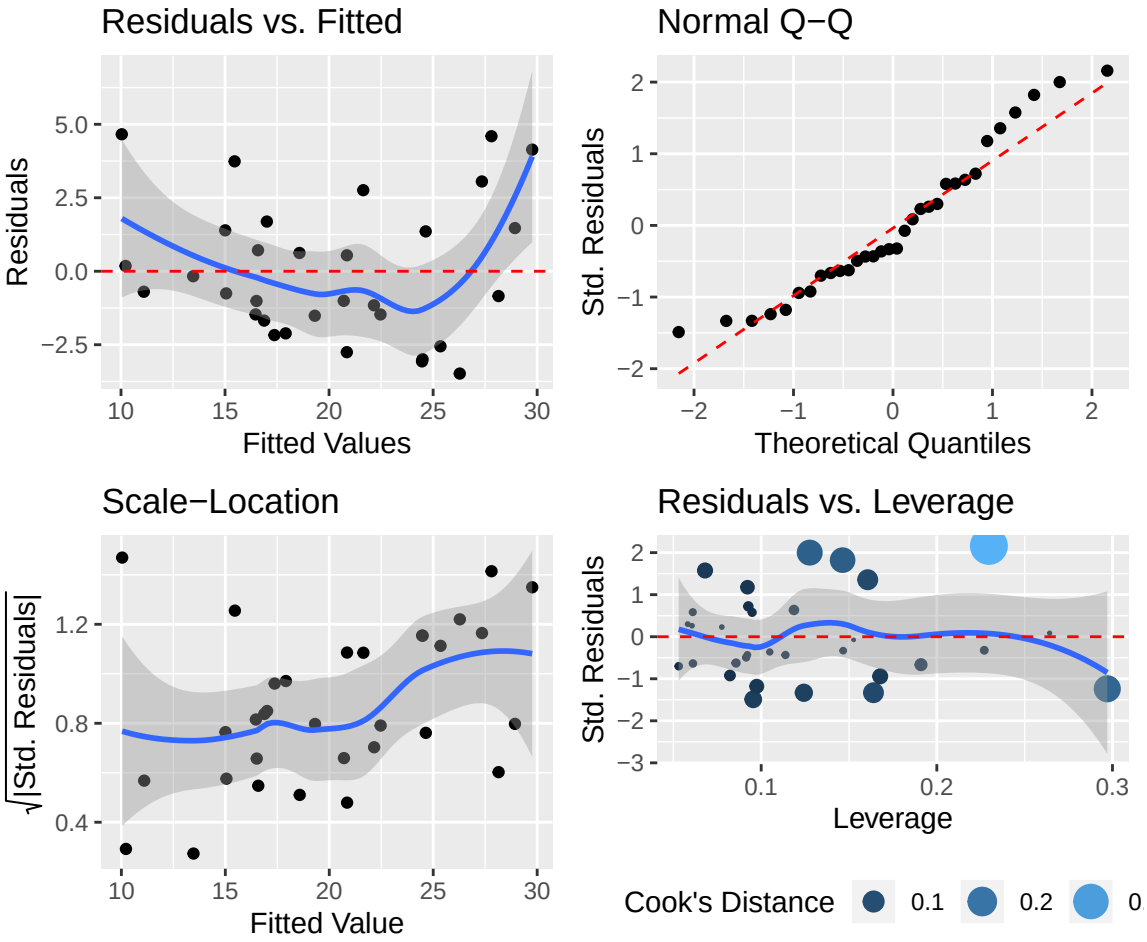
Unadjusted model

The full diagnostics chart-grid is less usefull for the unadjusted model. The chart below shows the density distribution of the residuals, which are approximatively normal and centerd at 0.



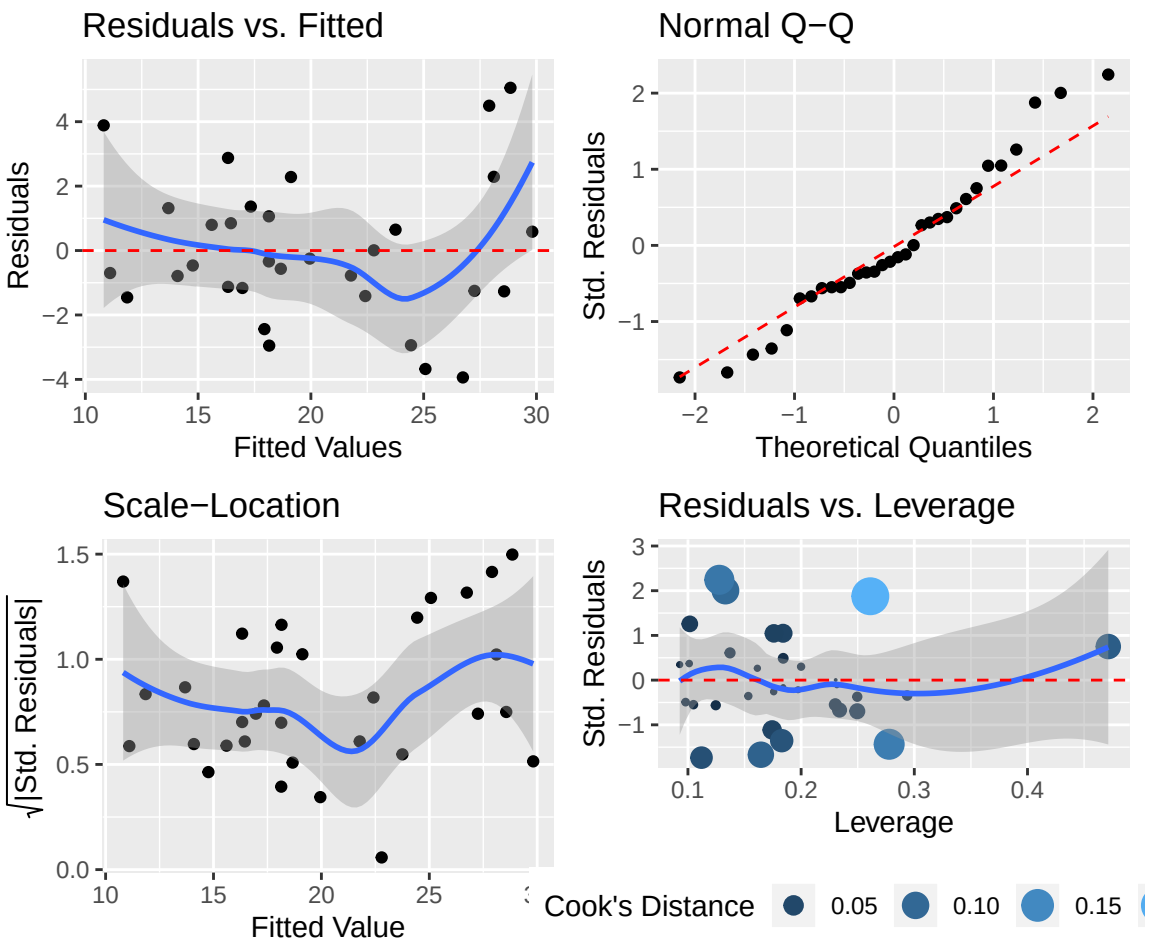
Adjusted, numeric

The adjusted model, using numeric coding, does not show many allarming issues. The residuals are normally distributed and scattered around 0 but show a "U"-shaped trend as more cars are added. They show no pattern with regards to fitted values. As expected, larger residuals show the highest leverage.



Adjusted, qualitative

The adjusted model, using factor coding, does not show many allarming issues. The residuals are less normally distributed, scattered around 0 and do not show any particular pattern as more cars are added. There may be some heteroskedacity. As expected, larger residuals show the highest leverage.


Residuals vs. Fitted
Residuals
Fitted Values
Normal Q−Q
Std. Residuals
Theoretical Quantiles
Scale−Location
√|Std. Residuals|
Fitted Value
Residuals vs. Leverage
Std. Residuals
Leverage
Cook's Distance
0.05
0.10
0.15