*Correlation and regression*

**R**

**Practice**

Rp

ALY6010 Probability Theory and Introductory Statistics

Module 5 R Practice

**PREPERATION:**

By: John DiSessa

For: Professor Goulding

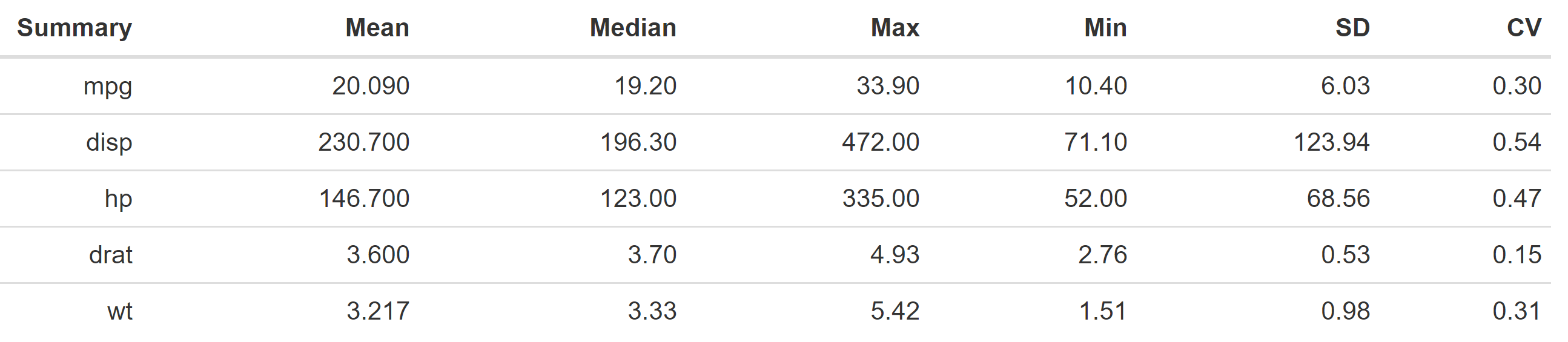
On: April 4th,2021

Introduction

For this assignment, I wanted to go back to my roots and revisit the mtcars dataset which was the first dataset I experimented with when learning R. The goal of my analysis was to find which variables most influenced the cars’ miles per gallon (mpg). I created a correlation table and a multivariate regression table to find the best predictors of mpg.

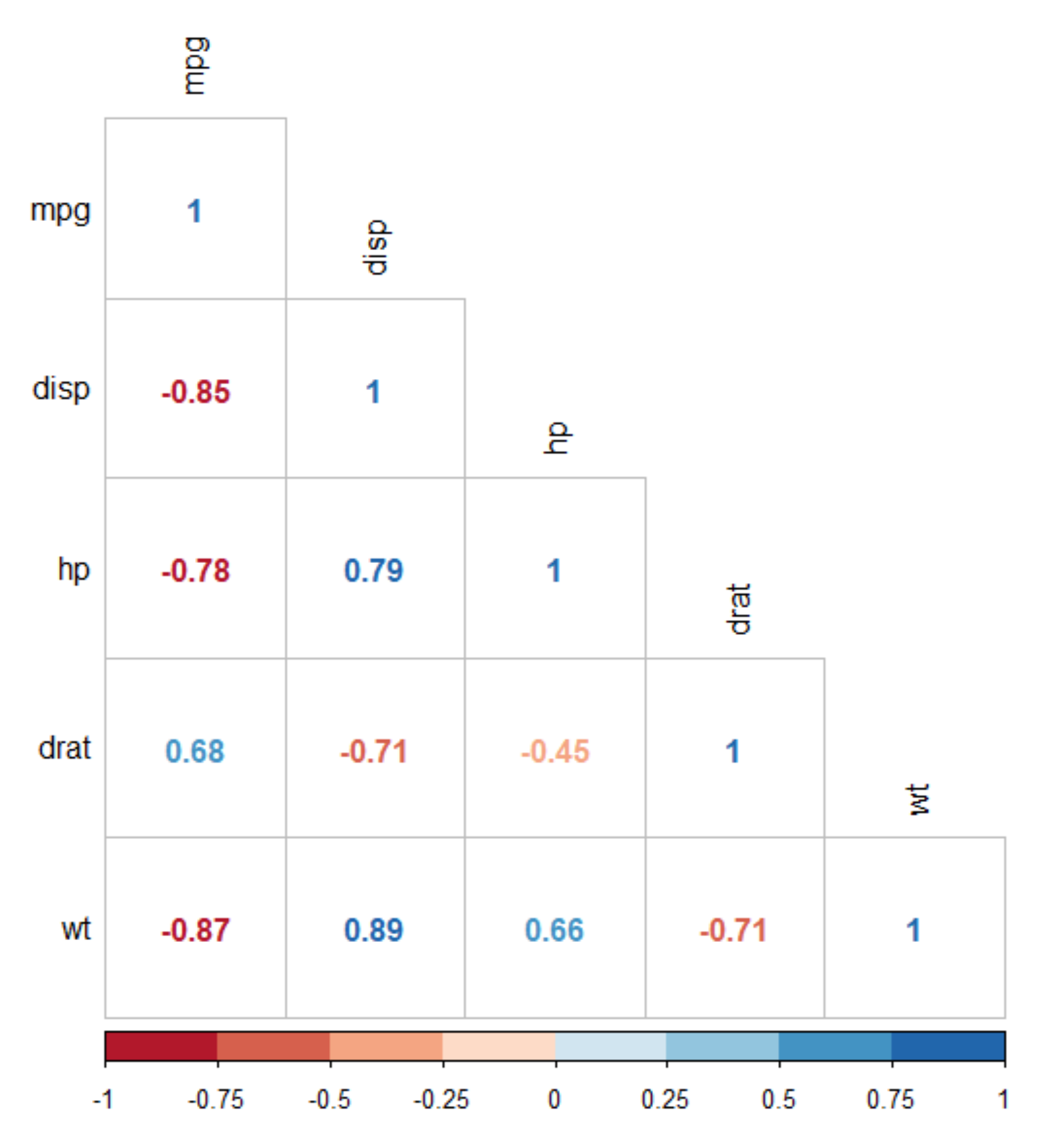
Analysis

I initially started my analysis by calculating correlation and regression for all 11 variables in mtcars, however I quickly realized that running analysis with both numerical and categorical variables did not work well. I removed the categorical variables (cyl, vs, am, gear, and carb) in order to only analyze continuous data. In my initial findings, I also discovered that the ¼ mile time variable (qsec) was insignificant and strongly correlated with other predictors, so I removed it as well in order to limit multicollinearity. I then modified the mtcars dataset to only include miles per gallon (mpg), displacement (disp – cubic inches), horsepower (hp), rear axle ratio (drat), and weight (wt – per 1000lbs). With this condensed dataset, I conducted the rest of my analysis starting with a descriptive statistics summary table.



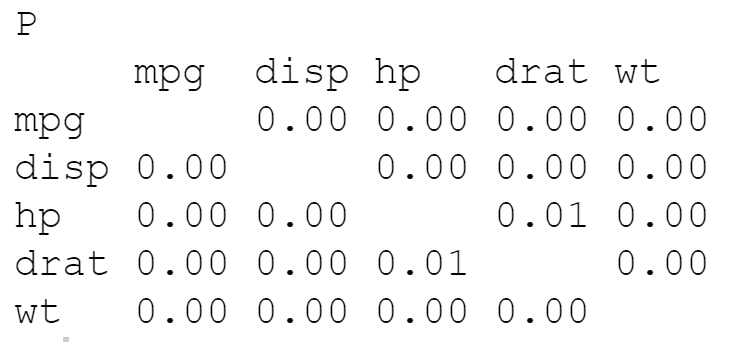
Even though mtcars is a standard dataset built into R, it is still important to note the Coefficient of Variations (CV) in the last column. CV standardizes the standard deviations in order to determine if variation is high or low. All of them are well below 1 which is a good indicator that variation within each variable is low.

I then created the following correlation table in order to see if any of the variables are strongly correlated with mpg. Dark red r values indicate strong negative correlation (as one variable increased the other variable decreased) and dark blue r values indicate strong positive correlation (both variables increase or decrease at similar rates). Mpg is strongly correlated with all 4 of my dependent variables. The correlation table is also great for detecting multicollinearity. The issue with multivariate regression is that it assumes all of the dependent variables are independent of each other. In reality, these variables are often related. For example, disp is correlated with all of the other variables. This is not surprising because cars with lots of horsepower often have bigger engines for more weight and displacement. I considered removing this variable but decided to keep it in case it is statistically significant in the regression.

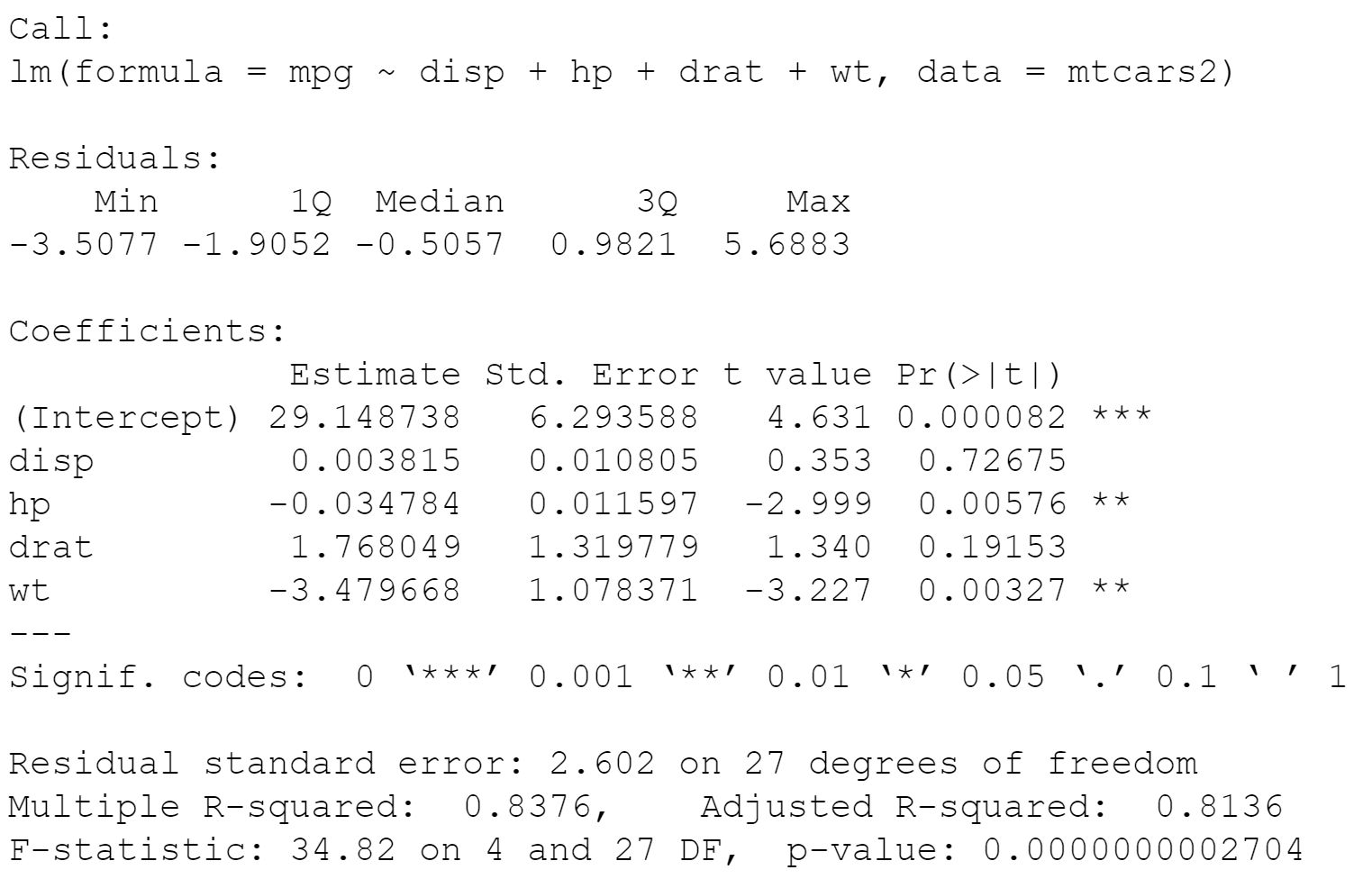


It is important that the correlation table only has 5 variables to keep it visually appealing. Every variable added creates a new row, a new column, and more r values to interpret. I do not want my readers spending too much time trying to dissect this table. I try to keep my graphics simple and visually appealing so that anyone can glance at my charts and understand the information I am trying to convey.

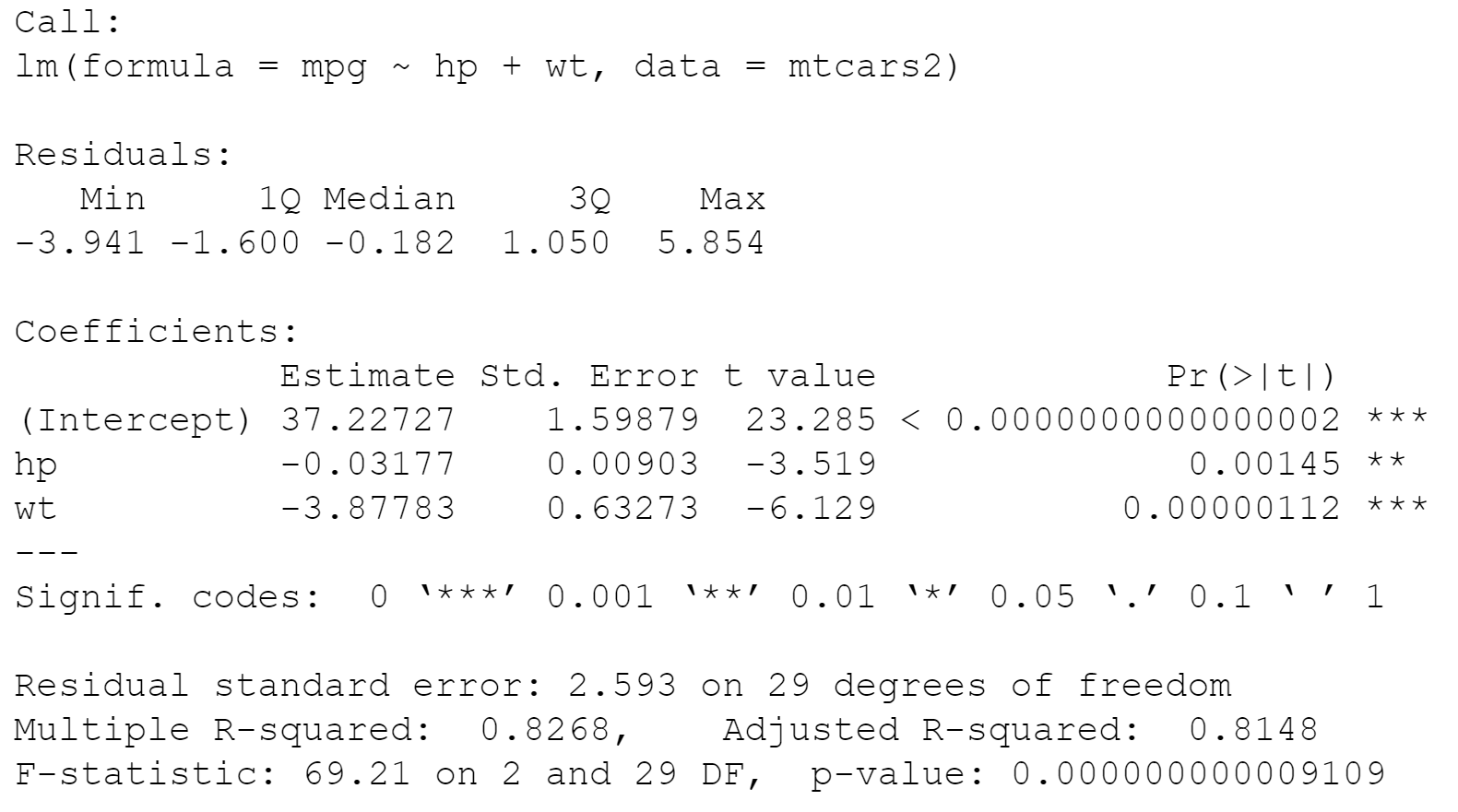
Before moving to multivariate regression, I created a correlation table of p-values to make sure the correlations were significant. When I created my initial correlation table, many of the correlations had p-values greater than .1 so it would have been a waste of time to produce a correlation table with insignificant r values. As you can see below, all of the correlations are significant.



My goals of the following multivariate regression were to determine which variables are the best predictors of mpg and given certain dependent variable values, estimate an expected mpg. I used a confidence level of 90% (α = .1) because there are no consequences of a false positive. I am not pitching this data to a car company or changing a business model. This is strictly for research purposes. I created the following regression output using a summary of the lm function.



First, we had to determine if the regression results were significant. The p-value of almost 0 indicates there is very strong evidence this output is significant. Next, the adjusted R2 value of .8136 means that 81.36% of variation in mpg is explained by the other variables. This is a very high value and indicates this is a strong fitting model. I ignored the R2 value because that does not take into account the number of predictors. Simply adding more predictors (even of insignificant) would increase R2, thus I prefer adjusted R2. Now that we know the model as a whole is significant, we had look at each individual predictor. Considering the high p-values of disp and drat, we can see these are not significant predictors of mpg. Hp and wt have p-values lower than our alpha of .1 so we can confidently say these impact mpg. Before I created my regression equation, however, I wanted a more accurate model so I removed disp and drat and ran a multivariate regression of hp and wt and seen below.



Since we removed 2 variables, the R2 decreased, but the adjusted R2 increased slightly so we technically can say this is a better fitting model than before. Using this regression output, I created the following regression equation to estimate mpg with any given hp and wt values.

Mpg = 37.23 - .03(hp) – 3.88(wt) ± 2.24

If we had a car that weighed 2,600 lbs with 110 horsepower, we would expect that car to get 23.84 mpg on average (between 21.6 and 26.1 mpg when factoring standard error). Lastly, I calculated the residual standard error (rse) to determine the error rate of my model. With an rse of .13, on average any calculated mpg value will deviate from the true regression line by about 13%. This is a very low error rate so coupled with the adjusted R2 value, my regression equation fits the data extremely well.

Summary

The initial correlation table indicated that our variables were all correlated with mpg. However, the multivariate regression calculations showed us that only 2 of the variables (horsepower and weight) are accurate predictors of mpg. Correlation calculations produced only 1 value indicating the relationship between one variable and another, whereas the regression helped produce an entire equation showing how the predictors influence mpg. The regression equation allows us to estimate mpg given any hp and wt values with great accuracy. Our high adjusted R2 value and low residual standard error means the model above is very accurate.

Citations

“Visualize Correlation Matrix Using Correlogram.” STHDA, www.sthda.com/english/wiki/visualize-correlation-matrix-using-correlogram.

“Elegant Correlation Table Using Xtable R Package.” STHDA, www.sthda.com/english/wiki/elegant-correlation-table-using-xtable-r-package.

Kassambara, et al. “Multiple Linear Regression in R.” STHDA, 10 Mar. 2018, www.sthda.com/english/articles/40-regression-analysis/168-multiple-linear-regression-in-r/.

Menon, Pradeep. “Data Science Simplified Part 5: Multivariate Regression Models.” Medium, Towards Data Science, 5 Aug. 2017, towardsdatascience.com/data-science-simplified-part-5-multivariate-regression-models-7684b0489015.

Heitmann, Soren. A Discussion of R's Mtcar Dataset Variables, rstudio-pubs-static.s3.amazonaws.com/61800\_faea93548c6b49cc91cd0c5ef5059894.html.

“Correlation Matrix : A Quick Start Guide to Analyze, Format and Visualize a Correlation Matrix Using R Software.” STHDA, www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software.

“Multicollinearity in Multiple Regression.” GraphPad, www.graphpad.com/support/faq/multicollinearity-in-multiple-regression/.

Stephanie. “Multicollinearity: Definition, Causes, Examples.” Statistics How To, 16 Sept. 2020, www.statisticshowto.com/multicollinearity/.