Module 6 R Practice

Angel Waters

2022-07-01

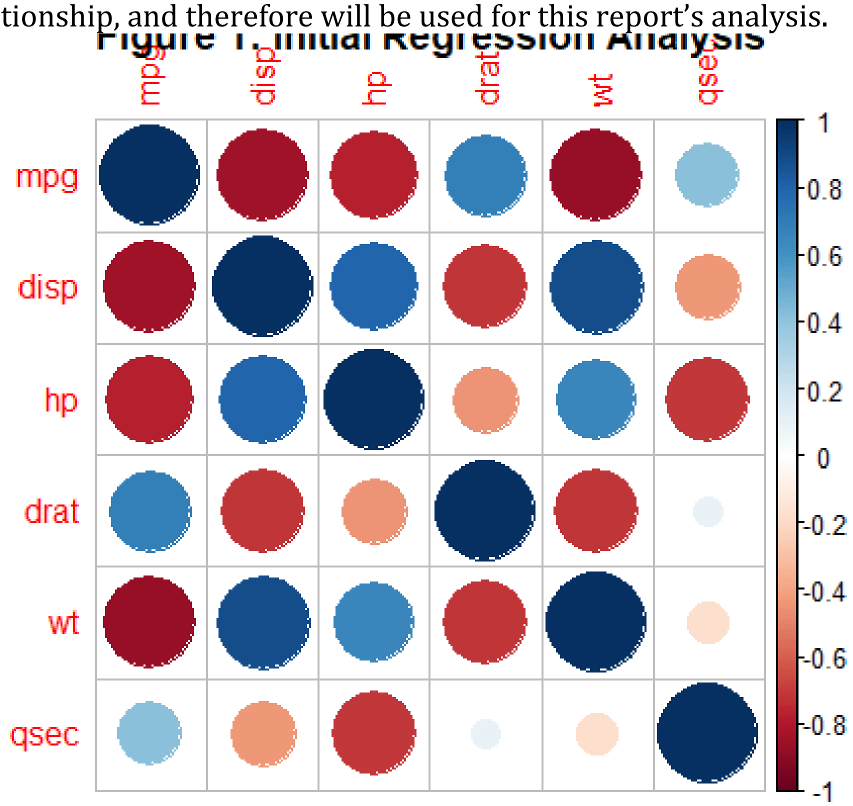
Data Cleaning and Setup

mtcars from core R was selected as the data set to be analyzed. It was loaded into the script and put in a format that was familiar for this analysis. The dummy variable identified was the number of cylinders each vehicle had, therefore an additional column was added to the data for a character version of the cylinders columns.

```
## 'data.frame':    32 obs. of  12 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 ...
##  $ Cyl.Dummy: chr  "6" "6" "4" "6" ...
```

Regression

An initial look at the relationships between the continuous variables was warranted to understand if there were any strong positive or negative linear relationships. **Figure 1** shows that analysis, where strong positives are represented by dark blue circles and strong negatives represented by dark red circles. These were the relationships that were considered for this analysis. Displacement vs weight shows a strong positive linear relationship, and therefore will be used for this report's analysis.



Correlation

After plotting the data (see **Fig. 2**), it is clear there is a positive linear relationship with some groupings in the data. The significance was tested to ensure the analysis could continue at confidence of 0.95. The correlation coefficient was calculated at 0.888 with a p-value < 0.05. This means there is enough evidence to support that the correlation for displacement by weight is not equal to 0.

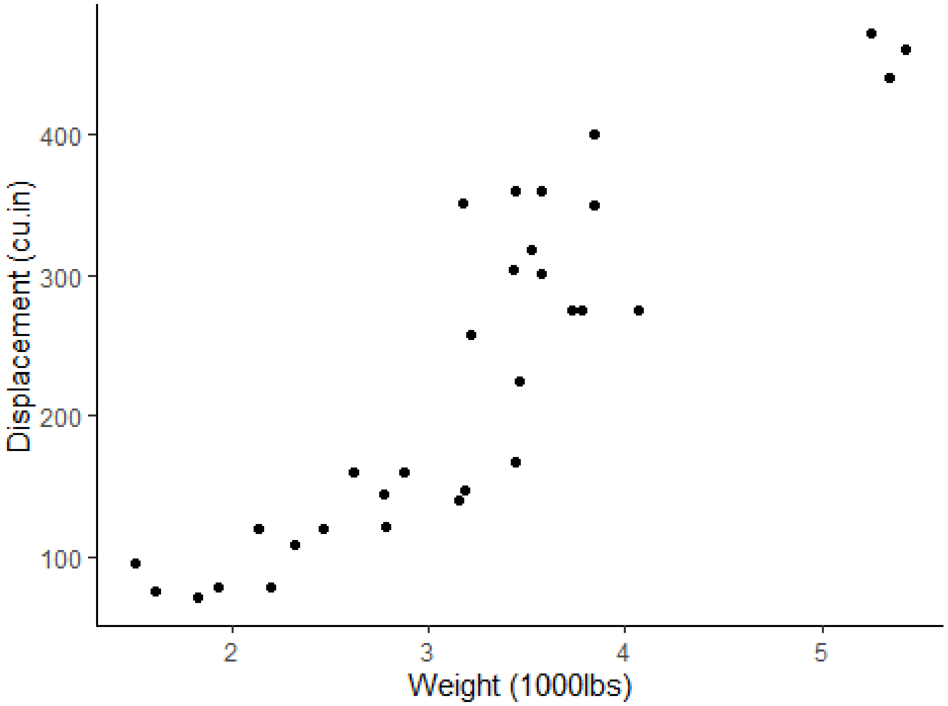


Figure 2: Displacement by Weight

```
##
##  Pearson's product-moment correlation
##
## data:  data$wt and data$disp
## t = 10.576, df = 30, p-value = 1.222e-11
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7811586 0.9442902
## sample estimates:
##       cor
## 0.8879799
```

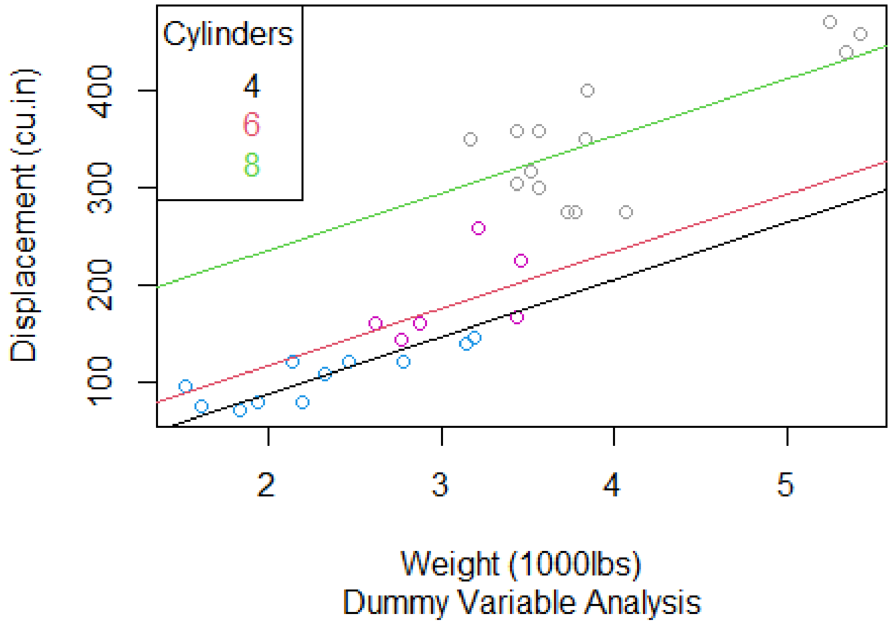
Dummy Variable Analysis

Because there are visible buckets, the data was analyzed to optimize the regression equations. The first look was using a dummy variable in which the regression was normalized by cylinder and offsets were calculated to shift the regression curve to the appropriate y intercept. The calculated regression lines were plotted and shown on **Figure 3**. Distinct groupings become more apparent once the legend was applied to the graph.

```
##
## Call:
## lm(formula = disp ~ wt + Cyl.Dummy, data = data)
##
## Coefficients:
```

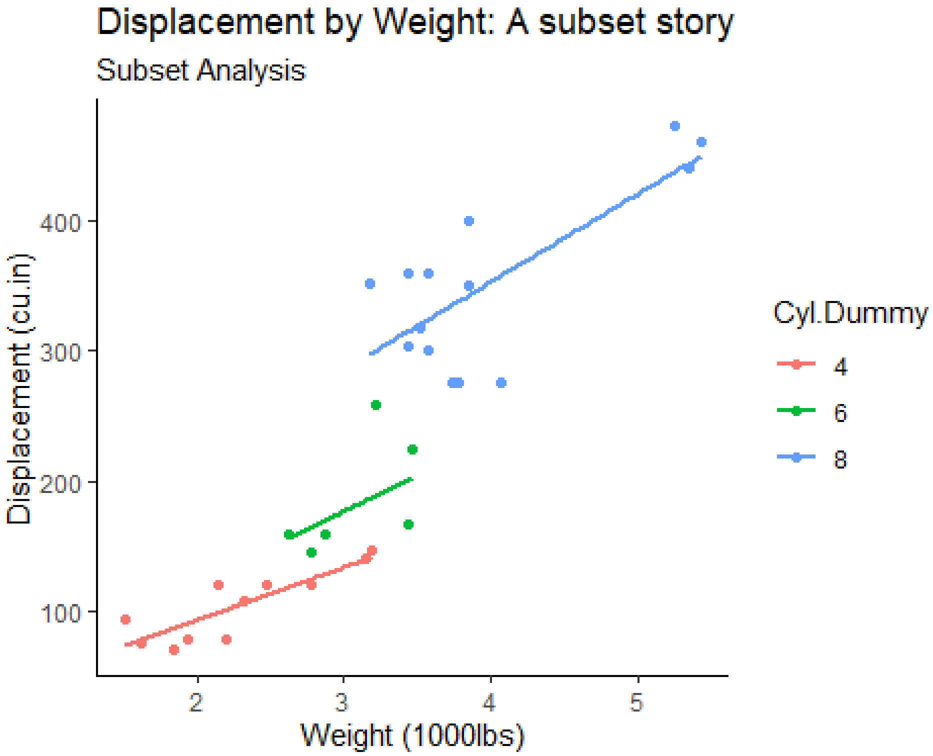
```
## (Intercept)          wt    Cyl.Dummy6    Cyl.Dummy8
##      -29.67       58.98         29.14        146.91
```

Figure 3: Displacement by Weight: A Dummy's Ta



Subset Analysis

An alternative method to looking at data subgroups is to subset the subgroups and calculate each group’s regression individually. This shows offsets in the data as well as the change in slope by subgroup. **Figure 4** shows the subset analysis for the displacement by weight relationship. 8 Cylinder vehicles have a higher slope than the other two as shown in the figure with the steeper line. This could show that 8 cylinder vehicles pull the data’s total regression in a direction more so than the other two cylinders.



Conclusion

Dummy variables help to look at how each subgroup offsets the y intercept. It was a quick way to show the data is effected by subgroups, specifically the Cylinders within a vehicle. Further analysis could be completed because there were clear subgroups. The subset analysis shows how individual regressions between displacement and weight look by subgroups. By using the subset method, it can offer a three dimensional analysis of the relationship between two variables by subgroups. It gives more insights into how the variables under analysis relate to each other by their individual groups. In this analysis, there is a higher slope and intercept for 8 cylinder vehicles. This may pull the regression slightly in favor of that subgroup, which is a good to know because the regression calculation is sensitive to data being pulled in one direction or another.