**Your Name:** Hendi Kushta **Partner’s Name:** Jiachen Li

**IST 772 Week 8 Class Exercise: Correlation**

**Instructions: Paste all code, results, and interpretations of results in this document for posting in Blackboard.**

In Blackboard there is a CSV file for this exercise entitled Hybrid. The data contain prices for midsize hybrid models of vehicles, acceleration rate and 2 measures of fuel economy.

Variables:

**msrp:** Manufacturer's suggested retail price in 2013

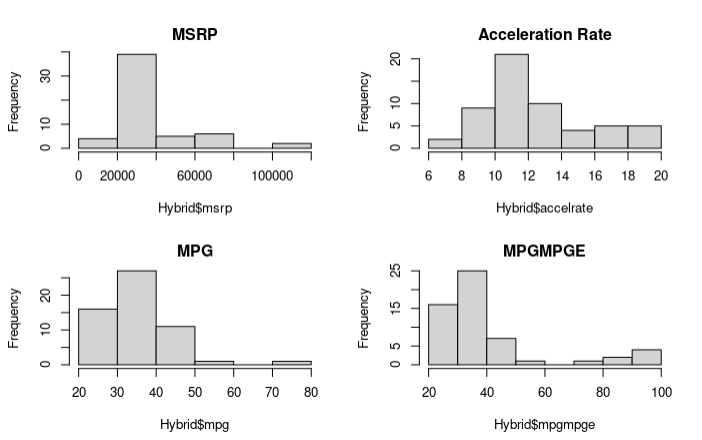
**accelerate:** acceleration rate in km/hour/second

**mpg:** Fuel economy in miles/gallon

**mpgmpge:** Max of Mpg and Mpge

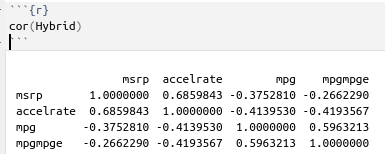
(for fully electric plug-ins, MPGe = 33.7\*driverange/batterycapacity)

1. Read in the CSV file. Display histograms of the 4 variables. **Describe what you see.**



Based on the histograms, we can infer that the distributions of MSRP, MPG, and MPGMPGE are likely right-skewed, meaning that they have longer right tails than left tails. On the other hand, the distribution of Acceleration Rate appears to be approximately symmetric, although it has a slight right skewness.

1. Generate a correlation matrix of these variables and **write about the pairwise correlations.**

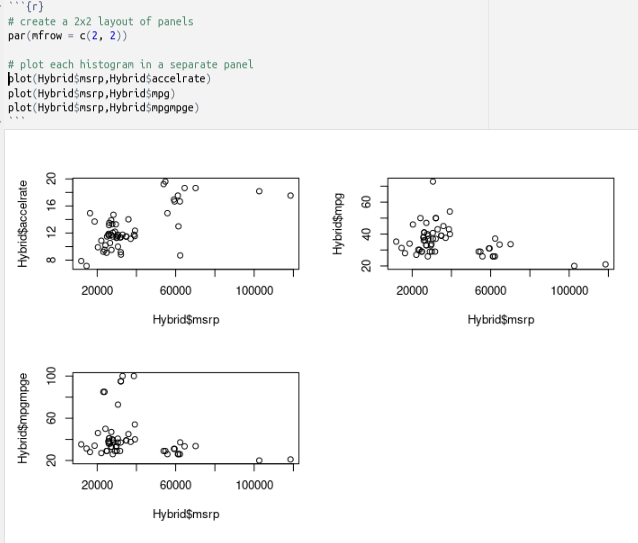


MSRP has a negative correlation with MPG and MPGMPGE, but the correlations are moderate to weak, ranging from -0.26 to -0.37. MSRP has a strong positive correlation with accelerate, with a correlation coefficient of 0.69.

Acceleration Rate has a negative correlation with MPG and MPGMPGE, with similar strength to the correlations (-0.41).

MPG has a negative correlation with MSRP and Acceleration, but the correlations are moderate to weak, ranging from -0.26 to -0.41. MPG has a strong positive correlation with MPGMPGE, with a correlation coefficient of 0.59.

1. Create 3 scatterplots, each with msrp as the independent variable and the other 3 variables as the dependent variable (one at a time). **Compare each scatterplot to the corresponding correlation above. Does it look as you expected?**

****

The scatter plot do not look as we were expecting, but still removing some of the outliers, we can confirm that MSRP and Acceleration have a positive correlation, and MSRP and MPG or MPGMPGE have negative relationships.

1. You can test the significance of correlations with cor.test(). The cor.test() function can take an extra argument called method= which can be set to one of three correlation techniques listed below.

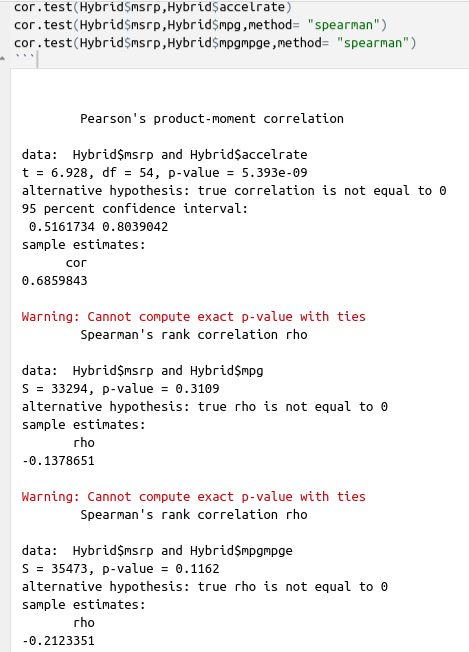
* method= "pearson" – This produces the Pearson product moment correlation, suitable for normally distributed metric variables (also the default if method is not specified)
* method= "kendall" – This produces Kendall’s Tau, suitable for arbitrary distributions with or without outliers, suitable for ordinal data
* method= "spearman" – This produces Spearman’s Rho, also suitable for distributions with or without outliers, more widely used

**a. Choose a method to run the cor.test() for each of the 3 pairs of variables from the scatterplots and explain your choice.**

For the pair MSRP and Accelerate, we said that accelerate is normally distributed. Hence, we can use the Pearson correlation method with the cor.test() function.

For the pair MSRP and MPG (City), both variables are continuous but their distributions may not be normal. Therefore, we can use either Kendall's Tau or Spearman's Rho correlation methods, which are more suitable for arbitrary distributions with or without outliers. We will choose the Spearman correlation method because it is more widely used and has been shown to perform better than Kendall's Tau in many situations.

**b. What are the results of the tests in terms of significant correlations?**



For the pair MSRP and Accelerate, the Pearson correlation coefficient is 0.685, which is positive and highly significant with a p-value < 5.393e-09.

For the pair MSRP and MPG, the Pearson correlation coefficient is -0.137, which is negative and not significant with a p-value = 0.3109.

For the pair MSRP and MPGMPGE, the Pearson correlation coefficient is -0.212, which is negative and not significant with a p-value = 0.11

1. Now run the Bayesian correlation test on these pairs with the bfCorTest function.

bfCorTest<-function(x,y)

{

zx<-scale(x)

zy<-scale(y)

zData<-data.frame(x=zx,rhoNot0=zy)

bfOut<-generalTestBF(x~rhoNot0,data = zData)

mcmcOut<-posterior(bfOut,iterations = 10000)

print(summary(mcmcOut[,"rhoNot0"]))

return(bfOut)

}

**Compare the Bayesian correlation results to the frequentist results in #4.**

1-   
The first set of output represents the results of a Pearson's product-moment correlation analysis between two variables, "msrp" and "accelrate", in a dataset called "Hybrid". The correlation coefficient is 0.686, indicating a moderate to strong positive relationship between the two variables. The t-value is 6.928, with 54 degrees of freedom, and the p-value is less than 0.001, indicating that the correlation is statistically significant at the 0.05 level. The 95% confidence interval for the correlation coefficient is between 0.516 and 0.804.

In this case, the analysis compares the evidence for the alternative hypothesis (that the correlation coefficient between the "msrp" and "accelrate" variables is not equal to 0) to the evidence for the null hypothesis (that the correlation coefficient is 0). The output indicates that the Bayes factor is for the hypothesis that the correlation coefficient is not equal to 0, denoted as "rhoNot0". The value of 1605547 indicates the strength of evidence in favor of the alternative hypothesis relative to the null hypothesis.

2-

The first set of output represents the results of a Spearman's rank correlation analysis between two variables, "msrp" and "mpg", in a dataset called "Hybrid". The correlation coefficient is -0.138, indicating a weak negative relationship between the two variables. The S statistic is 33294 and the p-value is 0.3109, indicating that the correlation is not statistically significant at the 0.05 level. The alternative hypothesis is that the true correlation coefficient is not equal to 0.

The Bayes factor for the hypothesis that the correlation coefficient is not equal to 0, denoted as "rhoNot0", is 9.215305, indicating moderate evidence in favor of the alternative hypothesis relative to the null hypothesis. The ±0% represents the uncertainty in the Bayes factor estimate, which is zero, indicating that the estimate is very precise and has no uncertainty.

3-

The first set of output represents the results of a Spearman's rank correlation analysis between two variables, "msrp" and "mpgmpge", in a dataset called "Hybrid". The correlation coefficient is -0.212, indicating a weak negative relationship between the two variables. The S statistic is 35473 and the p-value is 0.1162, indicating that the correlation is not statistically significant at the 0.05 level. The alternative hypothesis is that the true correlation coefficient is not equal to 0.

The Bayes factor for the hypothesis that the correlation coefficient is not equal to 0, denoted as "rhoNot0", is 1.456079, indicating only anecdotal evidence in favor of the alternative hypothesis relative to the null hypothesis. The ±0% represents the uncertainty in the Bayes factor estimate, which is very small, indicating that the estimate is very precise and has almost no uncertainty.