Computational Statistics - Seminar 2

Andrew Struthers

February 2023

Honor code: I pledge that I have neither given nor received help from anyone other than the instructor or the TAs for all work components included here. – Andrew

1 Introduction

The purpose of this report is to analyze the given data file containing six columns of data. The data will be analyzed for correlation using the Pearson's R and R^2 and Spearman's rho methods. The following analysis will be reported for each group: covariance, significance, and standard error.

R Code

Throughout this lab, the calculations for Pearson's R, Spearman's Rho, and many other statistical values are calculated by the following two functions:

```r
#Define function for Pearson's R and R^2
calc_pearson <- function(g)
{
  x <- g[, 1]
  y <- g[, 2]
  r <- cor(x, y, method = "pearson")
  r2 <- r^2
  c <- cov(x, y)
  se <- sqrt(1/(length(x)-2)) * sqrt((1-r^2)/r^2)
  sig <- cor.test(x, y)$p.value

  return(c(r = r, r2 = r2, covariance = c, significance = sig, standard_error = se))
}
```

Figure 1: R Function for Calculating Pearson's R and Other Values

```r
#Define a function to calculate Spearman's rho
calc_spearman <- function(g) {
  x <- g[, 1]
  y <- g[, 2]
  rho <- cor(x, y, method = "spearman")
  c <- cov(x, y)
  se <- 1/sqrt(length(x)-3)
  sig <- cor.test(x, y, method = "spearman")$p.value
  return(c(rho = rho, covariance = c, significance = sig, standard_error = se))
}
```

Figure 2: R Function for Calculating Spearman's Rho and Other Values

Additionally, the bootstrapping to make absolutely sure our correlation conclusions were correct were done using the following functions:

```r
# Define a function to calculate Pearson's R and R^2 using bootstrapping
calc_pearson_boot <- function(x, y, R) {
  boot_result <- boot(data = cbind(x, y),
                      statistic = function(d, i) cor(d[i, 1], d[i, 2], method = "pearson"),
                      R = R)
  r <- boot_result$t[1]
  sig <- boot.ci(boot_result, type = "bca")$bca[4]
  return(c(r = r, significance = sig))
}
```

Figure 3: R Function for Calculating Pearson's R Using Bootstrapping

```r
# Define a function to calculate Spearman's rho using bootstrapping
calc_spearman_boot <- function(x, y, R) {
  boot_result <- boot(data = cbind(x, y),
                      statistic = function(d, i) cor(d[i, 1], d[i, 2], method = "spearman"),
                      R = R)
  rho <- boot_result$t[1]
  sig <- boot.ci(boot_result, type = "bca")$bca[4]
  return(c(rho = rho, significance = sig))
}
```

Figure 4: R Function for Calculating Spearman's Rho Using Bootstrapping

Analysis of Group 1 Data

Below, we can see the scatter plot of the first two columns of data. Doing a quick visual analysis of the picture, we can see that there doesn't seem to be really any correlation between the two variables being plotted. They seem to be pretty evenly spread out throughout the entire plot.

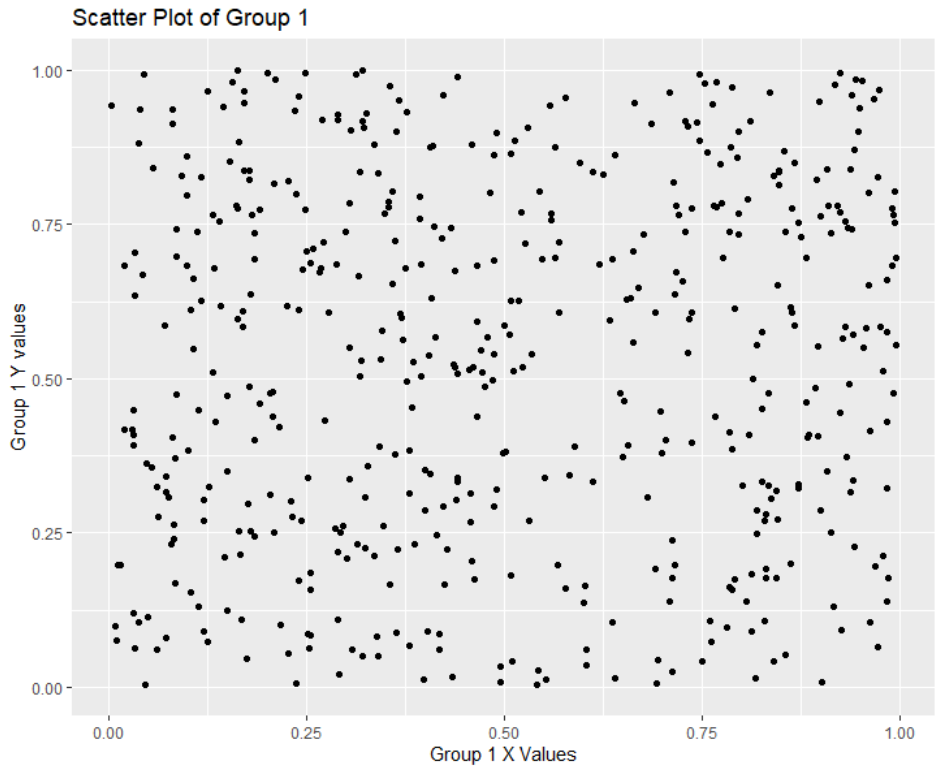

Figure 5: Separated visualization with three classes in DV

For group 1, the Pearson's correlation coefficient (r) was calculated to be 0.0716, with a R-squared value of 0.0051. The significance of the Pearson's correlation calculation was calculated to be 0.1097. Similarly, the Spearman's correlation coefficient (rho) for group 1 was found to be 0.0702, with a significance of 0.117.

Group 1	R	R^2	Rho	Significance
Pearson	0.071619357	0.005129332	N/A	0.109708710
Spearman	N/A	N/A	0.070174169	0.117057063

The covariance between the two columns was calculated to be 0.006216265, and the standard errors were found to be 0.624077064 for Pearson's R and 0.044856130 for Spearman's Rho.

Analysis of Group 2 Data

Below, we can see the scatter plot of the second two columns of data. Like before, we can see from visual analysis of the picture, there doesn't seem to be really any correlation between the two variables being plotted. These points are more tightly clustered around $(-1, -1)$ than the points in figure 5, but they still don't appear to be very correlated.

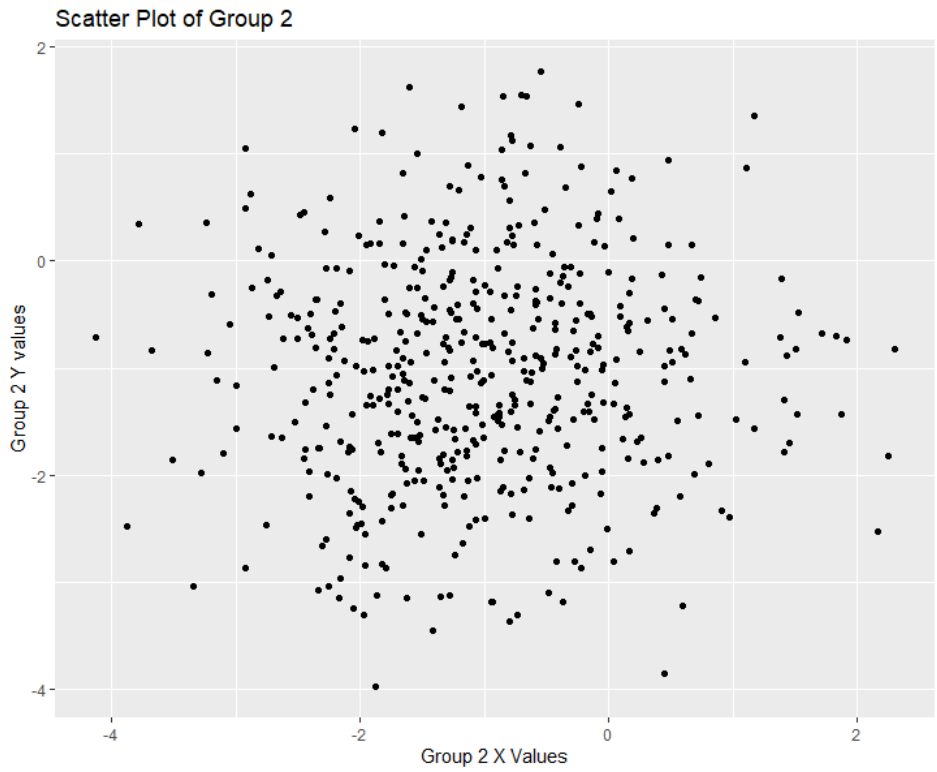

Figure 6: Separated visualization with three classes in DV

For group 2, the Pearson's r was calculated to be 0.0492, with a R-squared value of 0.0024. The significance of the Pearson's correlation calculation was calculated to be 0.2726. Likewise, the Spearman's rho for group 2 was to be 0.0625, with a significance of 0.163.

Group 2	R	R^2	Rho	Significance
Pearson	0.049154334	0.002416149	N/A	0.272625403
Spearman	N/A	N/A	0.06247916	0.16298042

The covariance between the two columns was calculated to be 0.056064082, and the standard errors were found to be 0.910538296 for Pearson's R and 0.044856130 for Spearman's Rho.

Analysis of Group 3 Data

Below, we can see the scatter plot of the second two columns of data. Like before, we can see from visual analysis of the picture, there doesn't seem to be really any correlation between the two variables being plotted. These points are more tightly clustered around $(1, 1)$ than the points in figure 5, and they are more spread out than the points in figure 6, but they still don't appear to be very correlated.

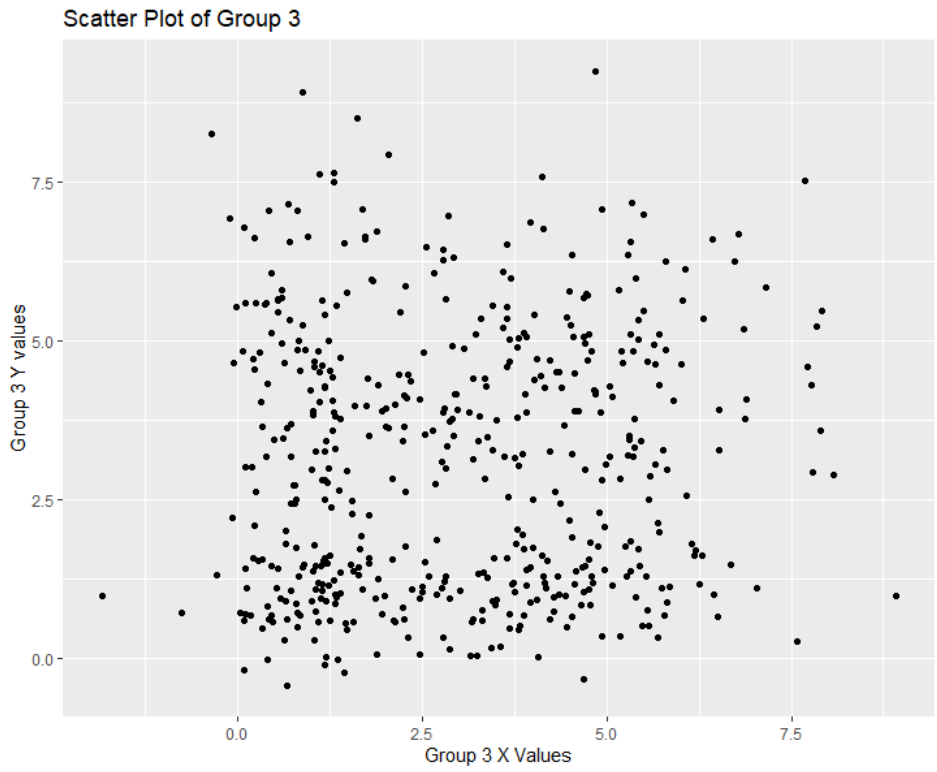

Figure 7: Separated visualization with three classes in DV

For group 3, the Pearson's r was calculated to be 0.0492, with a R-squared value of 0.0024. The significance of the Pearson's correlation calculation was calculated to be 0.2726. Likewise, the Spearman's rho for group 2 was to be 0.0625, with a significance of 0.163.

Group 2	R	R^2	Rho	Significance
Pearson	0.052276196	0.002732801	N/A	0.243289732
Spearman	N/A	N/A	0.05795764	0.19565826

The covariance between the two columns was calculated to be 0.217900577, and the standard errors were found to be 0.856026328 for Pearson's R and 0.044856130 for Spearman's Rho.

Conclusion

Based on the data, it appears the correlation between the two columns in group 1 is not strong, as the Pearson's R and Spearman's rho values are both close to 0 and the significance values are high (0.109708710 and 0.117057063, respectively). This means there is a weak relationship between the two columns, and this relationship is not statistically significant.

In group 2, the correlation between the two columns is also not strong, as the Pearson's R and Spearman's rho values are both close to 0 and the significance values are relatively high (0.272625403 and 0.16298042, respectively). Therefore, there is also a weak relationship between the two columns, and like in group 1, this relationship is not statistically significant.

The data for group 3 also has a weak correlation between the two columns since the Pearson's R and Spearman's rho values are both close to 0 once again and the significance values are still relatively high (0.243289732 and 0.19565826, respectively). This once again means that there is not a strong relationship between these two columns, and the relationship is not statistically significant.

Analyzing the covariance of each of the groups, we can see that there were covariance values of 0.0062 for group 1, 0.0561 for group 2, and 0.2179 for group 3. For group 1, the very close to zero covariance value implies that there is almost no effect between the variables. There is a very weak positive linear relationship, meaning that one column tends to move in the same direction as the other, but since the covariance is so close to zero, the effect is very weak. Group 2 is a similar case. Because the covariance is close to zero, there is a slightly positive linear relationship, but again, the effect is very small. Group 3 is a bit of an outlier from the other two groups. Since the covariance is larger than in the other two groups, there is more of a positive relationship here. However, due to the Pearson's R and Spearman's Rho values being very close to zero, with large significance values in both cases, we can conclude that while there might be a slight positive relationship, it is almost purely coincidental, since the two variables are very insignificantly related.

These conclusions about each group of data all relate to the scatter plots of the columns. It is visually clear that there is little correlation between each column in the three groups, and now comparing the Pearson's R and Spearman's Rho values as well as the covariance of each group,

we can see that there is indeed statistically insignificant correlation between columns in the dataset.

We can also compare the standard errors for each of the groups to the correlation coefficient and the level of significance. In group 1, there was a standard error of 0.6241, group 2 had a standard error of 0.9105, and group 3's standard error was 0.856. Comparing these values to the correlation coefficients for each group, we can see that the standard error in all three cases is much larger. This implies that there are low levels of precision in the data. We can then use bootstrap resampling to attempt to provide ourselves with more accurate estimates of the data. Using bootstrapping on each of these three groups with 1000 samples each gave the following results:

Group 1	R	R^2	Rho	Significance
Pearson	0.070960515	0.005035395	N/A	−0.007883696
Spearman	N/A	N/A	0.00795256	−0.01321442
Group 2	-	-	-	-
Pearson	−0.044382614	0.001969816	N/A	−0.094233532
Spearman	N/A	N/A	−0.01624779	−0.10023914
Group 3	-	-	-	-
Pearson	0.04856635	0.00235869	N/A	−0.04413416
Spearman	N/A	N/A	0.13508054	−0.03496485

Even with resampling, we still have values very close to zero for Pearson's R and Spearman's Rho. Therefore, we can determine that the the results from both Pearson's R and Spearman's Rho as well as the bootstrapping analysis suggest a weak correlation between the two columns of data.

Overall, the results from the Pearson's R and Spearman's rho tests indicate that the relationships between the columns in each group are weak and not statistically significant.

References

[1] Dr. Donald Davendra. Ch. 6 correlation, 2023. Lecture on Pearson's correlation coefficient, Spearman's Rho, and Kendall's tau.

[2] Dr. Donald Davendra. Ch. 6 bootstrapping, 2023. Lecture on Bootstrap resampling in R.

[3] Zoe FIELD, Jeremy MILES, and Andy FIELD. *Discovering statistics using R*. Sage Publications, 2012.

[4] PhD Leihua Ye. A practical guide to bootstrap in r, Dec 2021.