Week 7 Overview

Last week we took a bit of a breather and focused on R Markdown as part of our Data Science workflow for report generation. Markdown is a weird language, honestly, but it translates to so many other aspects of what the computer folks do that it is necessary to learn. You will also find the stakeholders you are interacting with in Data Science are a cross between business/technical and will want to find a good method for presenting your results. R Markdown provides that capability.

This week, we will analyze how we can express the relationships between variables statistically by looking at two measures: Covariance and the correlation coefficient. We will also discover and practice how to carry out and interpret correlations in R. Recall that both textbooks for this course include many examples that are similar to your weekly assignments. If you haven't been working along with the examples in your textbook, I would highly recommend doing so to build your confidence and proficiency.

Contents of the Week

Objectives

After completing this week, you should be able to:

- Describe different types of distribution
- Calculate and explain covariance given a set of variables
- Choose an appropriate correlation and partial correlation test to perform and justify their choice
- Make predictions about positive and negative correlation results
- Calculate the correlation coefficient and the coefficient of determination from a set of variables

Weekly Resources

- Comprehensive R Archive Network
- R Studio
- Sage Publications. (2021). Discovering Statistics Using R
- RStudio, PBC. (2021). RStudio Documentation
- RStudio, PBC. (2021). R Studio Cheatsheets

Week 7 Readings, Assignments, and Tasks

Here are your tasks for this week:

Read the following:

- *R for Everyone:* Chapters 17-18

- *Discovering Statistics Using R*: Chapter 6

Complete the following:

- 7.1 Discussion/Participation
- 7.2 Exercise

Helpful Sources

Magnusson, Kristoffer. Interpreting Correlations.

Wagih, Omar. Guess the Correlation.

7.1 Discussion/Participation

Here are optional topics for discussion via Teams this week.

Remember, these topics aren't required, but if you are struggling to know what to post about, these can be used to initiate discussion!

1. What are options for producing distributions in R?
2. What is binomial distribution?
3. What is poisson distribution?
4. We've covered this before, but now that we are further in the course – what are summary or descriptive statistics? Why do they matter and when do you use them?
5. What are correlation and covariance and what methods do you have in R to perform? How should you visualize correlation and covariance?
6. What is a T-Test?
7. What is ANOVA? How is it used?
8. What is correlation? What is positive vs negative correlation?
9. What is standardization and correlation coefficient?
10. What is Pearson's correlation coefficient?
11. What is the significance of the correlation coefficient?
12. What are confidence intervals?
13. What does it mean when there is statistical significance?
14. What are confidence intervals?
15. What is causality and what do we need to be cautious of?
16. What are the two types of correlation?
17. How do we handle missing data?
18. What is Pearson's correlation coefficient?
19. What are P-Values?
20. How is r^2 used? What is it?
21. What is Spearman's correlation coefficient and when is it used?
22. What is Kendall's tau? When is it used?
23. What is bootstrapping? When do you do this?
24. What is biserial and point-biserial correlation?
25. What is partial correlation?
26. How do you compare correlations? Why would you do this?
27. How do you calculate the effect size?
28. How do you report correlation coefficients?

7.2 Exercise

For the remainder of the course, you should complete your exercises in an RMarkdown file.

Complete the following exercises

1. Complete assignment05 - this is loaded into GitHub as a R script - it will need to be exported to RMarkdown/PDF
2. Student Survey
    a. As a data science intern with newly learned knowledge in skills in statistical correlation and R programming, you will analyze the results of a survey recently given to college students. You learn that the research question being investigated is: “Is there a significant relationship between the amount of time spent reading and the time spent watching television?” You are also interested if there are other significant relationships that can be discovered? The survey data is located in this StudentSurvey.csv file.
        i. Use R to calculate the covariance of the Survey variables and provide an explanation of why you would use this calculation and what the results indicate.
        ii. Examine the Survey data variables. What measurement is being used for the variables? Explain what effect changing the measurement being used for the variables would have on the covariance calculation. Would this be a problem? Explain and provide a better alternative if needed.
        iii. Choose the type of correlation test to perform, explain why you chose this test, and make a prediction if the test yields a positive or negative correlation?
        iv. Perform a correlation analysis of:
            1. All variables
            2. A single correlation between two a pair of the variables
            3. Repeat your correlation test in step 2 but set the confidence interval at 99%
            4. Describe what the calculations in the correlation matrix suggest about the relationship between the variables. Be specific with your explanation.
        v. Calculate the correlation coefficient and the coefficient of determination, describe what you conclude about the results.
        vi. Based on your analysis can you say that watching more TV caused students to read less? Explain.
        vii. Pick three variables and perform a partial correlation, documenting which variable you are “controlling”. Explain

Submission Instructions

For all assignments in this course, you must export the script or Markdown file to PDF. You are welcome to submit your URL to GitHub in addition, but all submissions must include a PDF (no zip files will be accepted either).

The assignment is due by Sunday, 11:59 p.m. CT.

how this changes your interpretation and explanation of the results.

Include all of your answers in an R Markdown report. Refer to the example template presented as a guide.