Weeks 8 & 9 Overview

Last week, you analyzed the relationship between variables by looking at two measures: covariance and the correlation coefficient. You also discovered and practiced how to carry out and interpret correlations in R, including how to control and make interpretations about a third variable. These next two weeks are all about regression – and correlation and regression are the bread and butter of frequentist statistics. We combined them because there is a lot of content to be covered and this should give you the necessary time to focus on understanding a topic fundamental to data science. Don't forget that when two weeks are combined, you have to do 10 discussion/participation posts **each** week, for a total of 20 in Weeks 8 & 9. As always, there are topics to aide you in discussion with your peers. Your projects are a great topic of discussion, as well as providing some feedback about the course and what you have learned.

We will take the methods from last week and move a step forward. We will begin to predict outcomes using what is known as regression. We will look at how influential cases and outliers can affect the accuracy of a regression model and methods used to identify those factors. You will perform regression using R, interpret the output, create regression models and test those models' reliability and generalizability. Influential cases and outliers can influence the accuracy of a regression model. In that light, methods that are commonly used to identify those factors are being highlighted.

Some imporant statistics tips!

> An important reminder before starting this week's assignments: One thing not to do is select hundreds of random predictors, bring them all into a regression analysis and hope for the best.
>
> Also remember that R is a tool and it will perform calculations on the data you 'feed' it; garbage in garbage out may result if you are not careful with the quality of the data inputs.
>
> In addition to the problem of selecting predictors, there are several ways in which variables can be entered into a model. When predictors are all completely uncorrelated, the order of variable entry has very little effect on the parameters calculated; however, we rarely have uncorrelated predictors and so the method of predictor selection is crucial. Keep all of this in mind as you work through this week's activities.

Contents of the Week

Objectives

After completing this week, you should be able to:

- Identify differences between Simple and Multiple Regression
- Use regression and multiple regression to determine the strength of relationships between variables
- Combine scatterplots and best-fit linear regression line to make inferences
- Create a correlation matrix and describe what the calculations suggest
- Calculate and explain the coefficient of determination and the correlation coefficient
- Use regression models to make predictions

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

Weeks 8 & 9: Readings, Assignments, and Tasks

Here are your tasks for this week:

Read the following:

- *R for Everyone:* Chapter 19
- *Discovering Statistics Using R*:

Chapter 7

Complete the following:

- 8.1 Discussion/Participation
- 9.1 Discussion/Participation
- 8.2 Exercise
- 8.3 Final Project Step 1

Helpful Sources

Gallo, Amy. (2015). A Refresher on Regression Analysis.

Martin, Rose. (2018). Using Linear Regression for Predictive Modeling in R.

Statistics How To. Regression Analysis: Step by Step Articles, Videos, Simples Definitions.

Berkeley. Multiple Regression

8.1 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 is simple linear regression? What is a linear model?
2. What is the difference between the outcome variable and the input variable?
3. What options in R do you have for regression?
4. What is multiple regression?
5. What are regression coefficients?
6. What is the method of least squares?
7. What is goodness of fit? What does this mean and how do you apply it?
8. How do you interpret simple regression results?
9. How do we use the model we created? What is the t-statistic?
10. What does the sum of squares, r, and r^2 tell us?

9.1 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. How do we decide which predictors to use?
2. What are different regression methods? How do you choose?
3. How can you assess your model? What do we need to test for?
4. What is generalization related to assessing a regression model?
5. What does cross-validation do? How do we do it?
6. How much data is enough for a model?
7. What is multi-collinearity?
8. How do you determine model parameters?
9. Why should you compare models and how do you in R?
10. How do you test for accuracy?
11. How should you report multiple regression results?
12. How do you do multiple regression with categorical predictors?
13. When do you use dummy variables?

8.2 Exercise

This week we are completing the last 2 exercises from GitHub - again, these are both saved as R Scripts and you will want to complete the work in RMarkdown, exported to PDF for final submission.

1. Complete assignment06
2. Complete assignment07
3. Housing Data
    a. Work individually on this assignment. You are encouraged to collaborate on ideas and strategies pertinent to this assignment. Data for this assignment is focused on real estate transactions recorded from 1964 to 2016 and can be found in Housing.xlsx. Using your skills in statistical correlation, multiple regression, and R programming, you are interested in the following variables: Sale Price and several other possible predictors.
        i. If you worked with the Housing dataset in previous week – you are in luck, you likely have already found any issues in the dataset and made the necessary transformations. If not, you will want to take some time looking at the data with all your new skills and identifying if you have any clean up that needs to happen.

    b. Complete the following:
        i. Explain any transformations or modifications you made to the dataset
        ii. Create two variables; one that will contain the variables Sale Price and Square Foot of Lot (same variables used from previous assignment on simple regression) and one that will contain Sale Price and several additional predictors of your choice. Explain the basis for your additional predictor selections.
        iii. Execute a summary() function on two variables defined in the previous step to compare the model results. What are the R2 and Adjusted R2 statistics? Explain what these results tell you about the overall model. Did the inclusion of the additional predictors help explain any large variations found in Sale Price?
        iv. Considering the parameters of the multiple regression model you have created. What are the standardized betas for each parameter and what do the values indicate?
        v. Calculate the confidence intervals for the parameters in your model and explain what the results indicate.
        vi. Assess the improvement of the new model compared to your original model (simple regression model) by testing whether this change is significant by performing an analysis of variance.
        vii. Perform casewise diagnostics to identify outliers and/or influential cases, storing each function's output in a dataframe assigned to a unique variable name.
        viii. Calculate the standardized residuals using the appropriate command, specifying those that are +-2, storing the results of large residuals in a variable you create.
        ix. Use the appropriate function to show the sum of large residuals.
        x. Which specific variables have large residuals (only cases that evaluate as TRUE)?
        xi. Investigate further by calculating the leverage, cooks distance, and covariance rations. Comment on all cases that are problematics.
        xii. Perform the necessary calculations to assess the assumption of independence and state if the condition is met or not.
        xiii. Perform the necessary calculations to assess the assumption of no multicollinearity and state if the condition is met or not.
        xiv. Visually check the assumptions related to the residuals using the plot() and hist() functions.

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 of Week 9, 11:59 p.m. CT.

Summarize what each graph is informing you of and if any anomalies are present.

xv. Overall, is this regression model unbiased? If an unbiased regression model, what does this tell us about the sample vs. the entire population model?

8.3 Final Project Step 1

You will be working on a research paper for your final project. This project will include identifying a topic/problem that you want to solve using data science. While the final solution to the problem does not need to be provided via programming – you will be doing some exploratory data analysis, transformations, and summary statistics on the data via R. You are welcome to create a model based on what you have learned in this course to solve the problem, but this is not required. Instead, a recommendation is required for a model or method you would implement to solve the problem. There are 3 steps to this project, with the final deliverable being a formal paper (completed in R Markdown) that outlines the problem, shows the analysis done with the data, and concludes with your recommendation for next steps. Each step provides detailed information that you must include in each phase of the project.

- Identify a topic or a problem that you want to research. Provide an introduction that explains the problem statement or topic you are addressing. Why would someone be interested in this? How is it a data science problem?
- Draft 5-10 Research questions that focus on the problem statement/topic.
- Provide a concise explanation of how you plan to address this problem statement.
- Discuss how your proposed approach will address (fully or partially) this problem.
- Do some digging and find at least 3 datasets that you can use to address the issue. (There is not a required number of fields or rows for these datasets)
    - Original source where the data was obtained is cited and, if possible, hyperlinked.
    - Source data is thoroughly explained (i.e. what was the original purpose of the data, when was it collected, how many variables did the original have, explain any peculiarities of the source data such as how missing values are recorded, or how data was imputed, etc.).
- Identify the packages that are needed for your project.
- What types of plots and tables will help you to illustrate the findings to your research questions?

- What do you not know how to do right now that you need to learn to answer your research questions?

You can use the following template for Step 1:

- Introduction
- Research questions
- Approach
- How your approach addresses (fully or partially) the problem.
- Data (Minimum of 3 Datasets - but no requirement on number of fields or rows)
- Required Packages
- Plots and Table Needs
- Questions for future steps

Submission Instructions

Submit an initial draft of your proposed project via PDF (of R Markdown file) to the assignment link - as a heads-up, you should add your next steps (Step 2 & 3) to the same file in upcoming weeks.

The assignment is due by Sunday of Week 9, 11:59 p.m. CT.