Stat 311 Autumn 2020 Final Group Assignment
Due Uploaded to Canvas by 11:30 PM on Thursday December 17th

Introduction

For this project, your group will analyze some aspects of the data that we collected as a class. All necessary files are posted on Canvas under Modules > Group Project.

The purpose of this project is to determine if some of the independent variables we collected have associations with our response variables (ClimateScore, SmartPhoneUse, SIFStudy). ClimateScore and SmartPhoneUse are quantitative responses, whereas SIFStudy is a categorical response. In addition to our three response variables, we have seven potential quantitative explanatory variables (Age, NumSocialMedia, AgeFirstPhone, AgeSmartPhone, HoursGV, PoliticalLean, and FollowNews), and 11 potential qualitative explanatory variables (Ethnicity, Gender, Religion, Study, Origin, SES, PoliticalAff, Education, DataPlan, SubUse, and Diet). The project data set was preprocessed to create climate sentiment scores and number of social media apps, and sleep scores from the raw data. The sleep scores (SleepC) were developed from the raw data using a modified version of the Pittsburg Sleep Quality Index (PSQI).

The preprocessed data set is called Stat311Au20GroupProjectSurveyDataProcessed.csv. The file Stat311Au20GroupProjectDataDictionary.pdf is the data dictionary that provides definitions for each variable in the survey data file. The file Stat311Au20GroupProjectCatalystSurvey.pdf contains a print version of the Catalyst survey. You will also need the file Stat311Au20Project.Rmd, which is an R Markdown template to get you started.

We reviewed the survey responses and have developed a set of questions that will allow you to demonstrate what you have learned this quarter. You will need to think about the data, how it was collected and to which population the data may apply. You will do some exploratory analyses as well as some inference. Calculations and graphing will be done in R. All of your interpretations will be added to the R Markdown file as well. Your job will be to choose the correct analyses and then focus on the written communication of results, making sure to include context.

Not all variables have been used in the problems below as not all variables produced interesting relationships for the types of problems we wanted to assign, and we needed to vary the types of problems while keeping the project a reasonable length. Complete the problems listed below.

Project Problems

1. **Data set Summary (5 points):** In your own words, write one or two paragraphs describing the survey data used for the project. Your description should briefly address how the data were collected. Then in more detail, you should specify the sample size and describe the types of variables. You should give details regarding the preprocessing data reductions that were done, as in the climate sentiment scores, number of social media apps, and sleep scores.
2. **Sampling (1+5 points)**
   A. What type of sampling was used to obtain this data set? Briefly explain.
   B. For the purposes of this project, we are treating the data as a representative random sample from some population. What population(s) do you think may be represented by this sample and why?
3. **Exploratory Analysis of the Climate Sentiment Scores (6+6 points)**
   A. Explore the ClimateScore variable using appropriate tables and graphs. Describe the distribution of ClimateScore.
   B. Repeat Part A for ClimateScore after subsetting the data by PoliticalAff, looking only at Democrat and No Affiliation subsets.

4. **PoliticalAff and SIFStudy (3+4+6 points)**

   A. We will focus on PoliticalAff values for "Democrat" and "No affiliation" since they have the highest counts. Also, drop the SIFStudy "I'm not sure" and "I prefer not to answer" options and only use total responses for either "Yes" or "No." Create a contingency table for these two variables with the specified rows and columns deleted. Just include the R code to display the table; no interpretation is necessary.

   B. Construct a 95% confidence interval for the difference in the population proportion of democrats vs. no political affiliation that favors a three-year SIF study in Seattle. Use "Democrats" minus "No affiliation." Be sure to check large sample conditions. Report and interpret the interval in the context of the problem.

   C. Conduct a test of the hypothesis to determine if the proportion of students in favor of a three-year SIF study in Seattle is greater for students that identify as democrats versus no political affiliation. Use "Democrats" minus "No affiliation." Be sure to state your statistical hypotheses, report the *p*-value, report the decision for your test, and provide a conclusion in the context of the problem. Use a 5% significance level to make your decision.

5. **Inference for Climate Sentiment (2+2+2+7+2 points)**

   A. Construct a 90% confidence interval for mean climate sentiment. Report and interpret the interval in the context of the problem.

   B. Construct a 99% confidence interval for mean climate sentiment. Report and interpret the interval in the context of the problem.

   C. Explain why the lower and upper bounds for the interval from part A are different than those from part B.

   D. Conduct a test of the hypothesis that the mean climate sentiment score is different than 32.4 (the climate sentiment score from spring quarter's Stat 311). Be sure to state your statistical hypotheses, report the test statistic, degrees of freedom, *p*-value, your decision, and provide a conclusion in the context of the problem. Use $\alpha = 0.05$ for this hypothesis test.

   E. For the hypothesis test in part D, do you risk making a Type I or Type II error? Describe the possible error in the context of the problem.

6. **Complete one of the following two problems. Only do one. No extra points for doing both.**

Climate Sentiment and Political Affiliation/Origin (5+7+7 points)

Explore climate sentiment scores (ClimateScore) by political affiliation (PoliticalAff). Subset the climate sentiment scores into two groups: one for students identifying as democrats and one for everyone else (not democrat). For all inference, assume equal population variances.

A. Construct a 99% confidence interval for the difference in mean climate sentiment scores for students identifying as democrats versus everyone else (democrat minus other). Report and interpret the interval in the context of the problem**.**

B. Test the claim that the mean climate sentiment score for students that identify as democrats is greater than the mean score for students that identify as something other than democrat (democrat minus other). Use a 1% significance level. Be sure to state your statistical hypotheses, report the test statistic, degrees of freedom, *p*-value, your decision, and provide a conclusion in the context of the problem.

C. Test the claim that the mean climate sentiment score for students that indicated they were raised in an urban location is different than the mean score for students that indicated they were raised in a suburban location. Use a 1% significance level. Be sure to state your statistical hypotheses, report the test statistic, degrees of freedom, *p*-value, your decision, and provide a conclusion in the context of the problem.

Climate Sentiment and Political Leaning (3+1+2+3+3+5+2 points)

Consider the two quantitative variables climate sentiment score (ClimateScore) and political leaning (PoliticalLean). Before starting this problem, make a subset that only includes records with a PoliticalLean score > 0. 0 was used for students that identified as apolitical or that preferred not to answer the question.

A. Using the subset of data, make a scatterplot of ClimateScore "on" PoliticalLean. Describe any pattern you see in the context of the variables.

B. Using the subset of data, run a simple linear regression model for ClimateScore (y) on PoliticalLean (x). Use `summary(object name)` to display the regression summary. No summary explanation for this part.

C. Use the output from part B to write out the least-squares regression equation.

D. Interpret the estimated slope parameter in the context of the problem.

E. What are R^2 and s_e for this regression? What do they suggest about the utility of this model?

F. Explore the model residuals to determine if the required model assumptions for inference are being met. Do you believe model assumptions are being met? If so why, if not, what assumptions might be violated?

G. Assume that model assumptions are met (regardless of what you found in part F). What is the 95% prediction interval for mean climate sentiment when political leaning is 75? Interpret this interval in the context of the problem.

Extra Credit (5 points)

Often when we are looking at categorical variables, we will find that there are too few people in some of the categories to feel that we have a representative sample. One way to deal with this is to combine categories. Pick **<u>one</u>** variable from among Ethnicity, SES, and Study, and think about how it might make sense to combine students so that you have only two groups. Show the original tabled values and then describe and summarize your new binary variable and your rationale for how you decided to combine responses.

Project Checklist

This project is both an analysis exercise and a writing project that allows you to use and demonstrate your understanding of many of the ideas and methods we have covered this quarter. R will be used for plotting, key summary statistics, confidence intervals, and test statistics/*p*-values for hypothesis tests, and regression. While choosing the correct summaries and methods is important, the presentation and interpretation of the results are just as or more important than producing graphs and numeric output. Please refer to the checklist below when doing this project and creating the final knitted R Markdown product.

- ☐ We did the writeup by problem number in order, with headers for each problem and parts of a problem added to the markdown file.
- ☐ We used reasonable rounding for numeric summaries when referring to the numbers in any written discussions/interpretations.
- ☐ We made use of `par(mfrow=c(rows, cols))` to put multiple related plots into a single figure where appropriate.
- ☐ We fully labeled all axes on graphs, including units if applicable.
- ☐ For confidence intervals we only show the R output for the interval, not all the other output that applies to hypothesis testing.
- ☐ For all hypothesis tests, we made sure to include the null and alternative hypotheses, the test statistic and degrees of freedom for t-tests, the *p*-value, our decision (reject the null or fail to reject the null) and an interpretation in the context of the problem, including units as appropriate.
- ☐ We used meaningful subscripts, such as mu[D] for mean climate sentiment for students that identify as democrats, or we used mu[1] and mu[2] but defined what group belongs to 1 and 2.
- ☐ We made sure to use the correct=FALSE argument for any calls to `prop.test`.
- ☐ For the longer free response writing and for interpretations, we developed thoughtful responses by focusing on what we considered to be the important features (the TAs and I do not want to read "brain dump")
- ☐ We included the name of all group participants on the first page of the final write-up (no need for a cover page). **The TAs and I are assuming that if your name is on the paper, you read/approved of the final product.**
- ☐ We uploaded a pdf of our knitted R Markdown file to Canvas by the 12/17 11:30 PM PST deadline.

Other Guidelines

- You are welcome to split up the work in any way that works for your group. You can divide and conquer, assign a couple of people to work on each problem, work on all problems together, etc. Try to make the best use of each group member's strengths.
- RStudio has a spellchecker. On the main R Studio ribbon, go to Edit > Spell Checker. Spelling and grammar do matter!
- We recommend that each group comes together for a final review of analysis methods/outputs and final edit of the writing. You should produce a project report that sounds like it was written by one person/one voice **(there should not be any sentences starting with I)**.

- Part of completing the project will be the completion of a self-evaluation and peer-review. A form for this will be provided at the beginning of finals week. Failure to upload this self-evaluation/peer-review will result in individual deductions on the final project score.
- Each student in a group initially receives the same project score. Evaluations may be used to make changes to individual scores.
- The project is worth a total of 70 points. The extra credit problem may be used to cancel up to five points of deductions from the required problems. Final scores, however, cannot exceed the total of 70 points for Problems 1 – 6.