**Data Science for Social Scientists**

Psyc 546, Spring 2023

Week 9 – In-Class Assignment

**Due Date**: March 23rd (by 11:59 PM)

**Reminder**: See the assigned readings, resources on Canvas, and the lecture slides for a tutorial on how to use R to perform the various functions included in the in-class assignment below. **Once completed, you should submit a completed version of this document and your final R script file to the Week 9 – In-Class Assignment – Submission Portal on Canvas**.

Your submitted R script file should contain code to answer the questions below (when relevant). Please use comments (e.g., #Question 1) to label the code for each question.

1. This question uses the **survey.csv** data file from Canvas. It can be helpful to get a sense of how even basic descriptive statistics are calculated. For this one, let’s imagine there are not functions in R that can calculate the standard deviation for you, which means you need to calculate it in steps. Say you want to do this for the variable **age**. First, create a new column in the data frame called **age\_squared\_deviation,** which contains the squared deviation from the mean for each participant on the age variable (e.g., this calculation is part of the numerator of the variance formula).

Then, go about calculating the standard deviation utilizing this age\_squared\_devaiation column. It probably will be helpful to break this final part down in steps (e.g., calculating sum of squares, variance, and then standard deviation). Finally, you should be able to confirm that your calculation is correct by utilizing a standard deviation function in R on the age variable. [2 points]

1. This question does not utilize any data file in R, but feel free to create code to calculate the manual calculations the question requires. Say that it is known that the scores on an IQ test are normally distributed with a population μ = 100 and population σ = 15. A researcher hypothesizes that special training will produce a change in the scores for the individuals in the population.

A sample of *n* = 36 individuals is selected, and the special training is given to this sample. Following the training, the average IQ score for this sample is *M* = 103.5. Is this enough evidence to conclude that the training has a significant effect on IQ scores using a two-tailed sample *z* test at α = .05? [2 points]

* 1. What is the calculated sample *z* for this problem? 1.4
  2. The conclusion of the hypothesis test for this study would be to:

Reject the Null Hypothesis OR Fail to Reject the Null Hypothesis

1. Using the survey.csv data file, perform an independent samples *t* test with positive affect over the past two weeks (Mposaff) as the dependent variable and whether the participants have a child or not (child) as the independent variable. You can assume equal variances or not for this question. Report below the *t* value, *p* value, whether the effect would be considered statistically significant, and which group had the higher mean. [2 points]

t = -2.3645, df = 396.85, p-value = 0.01854, as the p-value is less than 0.05, the effect is statistically significant.

The child group/group 1 had a higher mean (m = 3.464674) than the non-child group/group 0 (m = 3.299203)

1. In the survey.csv, the variable Mmast includes the mean scale score for a mastery scale. Perform a one-way ANOVA on this with education (educ) as the grouping variable. Was there a statistically significant relationship? If so, perform a Tukey’s HSD post-hoc test to explore which educational groups differed from one another. For reference, categories on the educ variable are: 1 = Primary, 2 = Some secondary, 3 = Completed high school, 4 = some additional training, 5 = completed undergraduate, and 6 = postgraduate completed. Finally, calculate an effect size for the ANOVA (e.g., eta-squared) and assess the magnitude of the effect (e.g., small, medium, or large). Report the conclusions below and your assessment if the relationship between education and mastery in this sample appears to be large and robust or relatively small in magnitude. [2 points]

* The ANOVA test suggests that there is a statically significant relationship between education and mastery (p = 0.0167)
* The Post Hoc test suggests that only those who completed an undergraduate degree (group 5) differed significantly from those who completed some secondary education (group 2). The rest of the groups didn’t differ significantly from each other.
* The relationship between education and mastery in this sample appears to be relatively small in magnitude, as the Eta2 = 0.03

1. Using the survey.csv data file, perform a multiple linear regression with optimism (Moptim) as the criterion variable and the following predictors included in the model: biological sex (sex), age in years (age), education (educ: treated as a factor dummy coded variable), positive affect (Mposaff), and negative affect (Mnegaff).

Report below which coefficients were statistically significant and which directions they were in. Also, assess if there was any problematic multicollinearity in the model and inspect the residuals using a histogram to assess the normality of the residuals of the regression model. Include the histogram below and state your conclusions regarding the multicollinearity and normality of the residuals assumptions. [2 points]

* age in years (p = 0.000629), positive affect (p = 1.75e-15), and negative affect (p = 5.22e-06) are the coefficients that are statistically significant in the multiple linear regression model. Age (0.008674) and positive affect (0.368406) are positively related, while negative affect (-0.211385) is negatively related to the criterion variable optimism.
* The model doesn’t have any issues with multicollinearity, as the variance inflation factor (VIF) shows that all the predictor variables have VIF values closer to 1.
* The histogram of residuals is normally distributed, so it meets the underlying assumptions of multiple regression.

Chart, histogram

Description automatically generated