**Data Science for Social Scientists**

Psyc 546, Spring 2023

Week 10 – In-Class Assignment

**Due Date**: March 30th (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 10 – 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. For Q1, imagine the following contingency table of results is found in your study that included two conditions (experimental vs. control group) and classified participants as passing or failing some cognitive test:

|  |  |  |
| --- | --- | --- |
| Condition | Fail | Pass |
| Experimental | 25 | 55 |
| Control | 36 | 45 |

In decimal format, calculate [2 points overall]:

1. The odds of passing for the experimental group: 2.2
2. The odds of passing for the control group: 1.25
3. The odds-ratio comparing the experimental vs. control group: 1.76
4. In a sentence, describe how you would report the odds-ratio found in 1c: The odds of passing is 1.76 times higher for the experimental group than the control group.
5. Q2 uses the **patient.sav** data file from Canvas. The data set consists of patient health information. Imagine you are tasked with exploring potential factors associated with patients arriving to the hospital dead on arrival (doa: 0 = No, 1 = Yes). Perform a binary logistic regression model with dead on arrival as the criterion variable and the following three predictors: the patient’s age in years, the patient’s gender, and the patient’s cholesterol level classification. Make sure to include code that outputs the summary of the model as well as code to extract the odds-ratios for the three predictors. Below, state which predictors (if any) are statistically significant and the directionality of these effects. [2 points]

* In this binary regression model, patient’s age and cholesterol are statistically significant predictors of the criterion variable dead on arrival (*p* < .01). Both patient’s age (B = 0.03) and cholesterol (B = 0.47) are positively related to the criterion variable dead on arrival.

1. This question uses the **survey.csv** data file from Canvas and focuses on the impact of centering on the multicollinearity of moderation models. Imagine you are interested in if positive affect (Mposaff) and negative affect (Mnegaff) interact together in predicting life satisfaction (Mlifesat).

Run two moderated linear regression models. In the first model, do not center the two predictors. In the second model, center the two predictors. Then, assess the multicollinearity (e.g., variance inflation factors) of the two models and report the multicollinearity results below (and state whether centering reduced multicollinearity to more acceptable levels or not). Finally, compare the two main effects and the interaction effect across the two models. Which effects are the same and which are different? [2 points overall]

* VIF tests indicate that Model 1 (without centering the predictors) has not met the assumption of collinearity as multicollinearity was a concern (VIF > 5). (Positive affect, VIF = 6.98; Negative affect, VIF = 14.46; Positive affect \* Negative affect, VIF = 14.98)
* VIF tests indicate that Model 2 (mean centering the predictors) has met the assumption of collinearity as multicollinearity was not a concern. (Positive affect, VIF = 1.13; Negative affect, VIF = 1.14; Positive affect \* Negative affect, VIF = 1.1)
* The main effect of positive affect in Model 2 (0.65) was larger than that of Model 1 (0.59), whereas the main effect of negative affect in Model 1 (-0.50) was smaller than that of Model 2 (-0.40). Additionally, the interaction effect was similar in both Model 1 (0.03) and Model 2 (0.03).

1. Q4 also uses the survey.csv data file. Imagine you have a research question about whether age moderates the relationship between stress (Mpstress; IV) and self-esteem (Mslfest; DV). Usually you would want to mean-center the continuous age and stress variables, but for this question you do not need to. Perform a moderated linear regression model. Report below whether there is significant moderation present? What are you basing this conclusion on? Finally, explore any potential moderation effect further by performing a simple slopes analysis. How does the relationship between stress and self-esteem change as one gets older in this sample? Include a figure below and the estimates/results of a simple slope analysis to back up your conclusions. [2 points]

* A moderation linear regression model tells us that there is a significant moderation effect present in the model. This is because the interaction term is statistically significant (*p* <.01) according to the model.
* As one gets older, the negative relationship between stress and self-esteem decreases. This means that as one gets older, their stress level has comparatively smaller effects on their self-esteem. At age -1SD (24.21), the slope of the relationship was -0.62 (t = -12.20, p < 0.001); at the mean age (37.34), the slope was -0.53 (t = -14.52, p < 0.001); and at age +1 SD (50.46), the slope was -0.44 (t = -8.89, p < 0.001).

Chart

Description automatically generated

1. This question uses the **pupil.sav** data file from Canvas. It includes achievement scores for students (pupils) nested within primary schools. First, perform an intercept-only multilevel linear regression model to assess the degree of clustering within the achievement scores. Specifically, estimate the intraclass correlation coefficient (ICC), report it below, and state your conclusion about whether the degree of clustering warrants a multilevel model. Regardless of your conclusion to the previous question, perform a multilevel model that still includes the random intercept from the intercept-only model but adds Level-1 predictors of pupil gender and pupil SES (for this question, just treat SES as a continuous variable with higher scores implying higher SES). Are these two predictors statistically significant? If so, what is the directionality of the effects? [2 points]

* ICC = 0.1756 / (0.5771 + 0.1756)

= 0.2332953

= 0.23

* As ICC is greater than 0.05, the degree of clustering warrants a multilevel model.
* Yes, the two predictors, pupil gender and pupil SES are statistically significant predictors of student achievement scores (*p* < .01). Both pupil gender (B = 0.26) and Pupil SES (0.12) are positively related to the criterion variable student achievement scores.