**Data Science for Social Scientists**

Psyc 546, Spring 2023

Homework Assignment 9

**Due Date**: April 5th (by 8:15 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 homework assignment below. **Once completed, you should submit a completed version of this document and your final R script file to the Homework Assignment 9 – Submission Portal on Canvas**.

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

1. Q1 uses the **bankloan.sav** data file from Canvas. Perform a binary logistic regression model that predicts whether the individual has ever defaulted on a loan or not (0 = No, 1 = Yes). The regression model should include the following four predictors: the individual’s age, the years with their current employer, household income, and credit card debt. In your code, also make sure to estimate the odds ratios (with their 95% confidence intervals) after running the model. Below, report the directionality of the four predictors (e.g., which variables were associated with greater or lesser odds of defaulting) and whether each predictor was statistically significant or not. Finally, choose one of the four predictors (regardless of if it was statistically significant or not) and describe the effect in sentence format. That is, describe one of the effects in a sentence or two in a way that you would see reported in an APA research article. [2 points]

* The result showed that years with their current employer (B = -0.24, *p* < .001), household income (B = -0.14, *p* = .01), and credit card debt (B = 0.81, *p* < .001) were statistically significant predictors of defaulting on a loan while age (B = -0.01, *p* = .64) wasn’t a significant predictor. Both years with their current employer and household income were negatively related, while credit card debt was positively related to criterion variable defaulting. Credit card debt was associated with greater odds of defaulting, while years with their current employer and household income were associated with lesser odds of defaulting.
* Higher credit-card debt was associated with an increase in the likelihood of defaulting on a loan (OR=2.25, 95%CI [1.89, 2.69]).

1. This question uses the **titanic.csv** data file on Canvas. This data set contains information about the passengers on the Titanic and whether they survived or not. Perform a binary logistic regression with survival as the DV and the following three predictors: the passenger’s sex, age, and class. Usually, you would want to treat the categorical class predictor as a factor, but for this question just keep it as a numeric predictor. After running the model, use the predict() function to estimate the probability of survival for a 40 year-old passenger with the characteristics below. That is, in all predictions, imagine the passenger is 40 years-old but vary whether the passenger was male or female and their class status. This tutorial should be helpful guidance: <https://www.statology.org/r-glm-predict/>. [2 points]

|  |  |  |
| --- | --- | --- |
|  | Female | Male |
| First Class | 2.27 | -0.32 |
| Second Class | 1.04 | -1.55 |
| Third Class | -0.19 | -2.78 |

1. In Q3, use the **survey.csv** data file. Imagine you have a research question about whether the relationship between recent negative affect levels (Mnegaff; IV) and optimism towards the future (Moptim; DV) differs for males vs. females. Perform a moderated linear regression model that would test this research question. Make sure to mean-center the continuous IV. For the biological sex variable, 0 = Males and 1 = Females.

First, assess the multicollinearity of the model and report whether there is any problematic multicollinearity or not. Second, report the conclusions of the interaction effect in APA format. That is, state the conclusion of whether moderation is present or not and report the regression statistics of the interaction effect to support your conclusion. Finally, explore any potential moderation effect further by performing a simple slopes analysis. How does the relationship between negative affect and optimism differ for males compared to females in this sample? Include a figure below and the estimates/results of a simple slope analysis to back up your conclusions. [2 points]

* VIF tests indicate that the model has met the assumption of collinearity, as multicollinearity was not a concern (Negative affect, VIF = 2.49; sex, VIF = 1.01; Negative affect \* sex, VIF = 2.48).
* The interaction between negative affect and sex was found to be statistically significant (*B* = -0.20, *p* < .05). Thus, the linear regression model indicates that sex is a moderator of the relationship between negative affect and optimism toward the future.
* The simple slope analysis reveals that the relationship between negative affect and optimism toward the future is significant for both males (sex = 0) and females (sex = 1). There is a negative relationship between negative affect and optimism for males (b = -0.23, t = -3.09, *p* < .001) and for females (b = -0.42, t = -7.01, *p* < .001). This implies that as the negative affect increases, the optimism for the future decreases for both males and females. But this negative relationship is stronger for females than for males.

Chart, line chart

Description automatically generated

1. This question uses the **popular2.sav** data file from Canvas. It includes popularity ratings of students from 100 different classrooms. The popularity scores are derived from a sociometric measure based on the ratings of other students in the classroom. Because these ratings are nested within class, you know that you will need to perform a multilevel model with classroom acting as the group variable. First, perform an intercept-only multilevel linear regression model to assess the degree of clustering in the popularity 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 the students’ extraversion scores as a predictor of popularity. Is extraversion a significant predictor of popularity in this sample of students? If so, what is the directionality of the effect? [2 points]

* ICC = 0.7021 / (0.7021 + 1.2218)

= 0.36

* As ICC is greater than 0.05, the degree of clustering warrants a multilevel model.
* Yes, extraversion is a statistically significant predictor of popularity in this sample of students (*p* <.01). Extraversion (B = 0.49) is positively related to the criterion variable popularity, suggesting that students with higher ratings of extraversion are predicted to have higher ratings of popularity.

1. Q5 uses the **gpa2long.sav** data file from Canvas. This data set contains longitudinal data for a sample of high school students. Specifically, semester GPA was measured in these students over six semesters. Other information was collected as well, including the students’ job status at the time. Because multiple measurements are occurring for each participant, the data violates the independence assumption of the traditional linear regression model. As a result, perform a multilevel model that uses the individual student as the group/cluster variable. In the model, include the student’s job status as a predictor as well as the student’s biological sex to adjust for any sex differences in HS GPA. Job status is a multi-categorical variable, so make sure you treat it is a factor in your model. Based on the results, how does job status relate to GPA in this sample of high school students? [2 points]

* The results of the multilevel linear regression model revealed that there was a significant negative relationship between job status and GPA in this sample of high school students (*p* < .001). This implies that students with a job had a lower GPA than students without a job. Specifically, students who worked 10 or more hours per week (B = -0.60, *p* < .001) are predicted to have a lower GPA than students who worked 1-10 hours per week (B = -0.42, *p* < .001). So, students who work more hours are predicted to have lower GPAs.