Name:

**Multivariate Statistics**

PSYC 739, Fall 2021

***Homework #3, Part A: Logistic Regression***

For Part A, you will be using data from the 2018 General Social Survey. Make sure to use the data file on Canvas named **GSS2018\_HW3A** (Files 🡪 Data Sets 🡪 SPSS files 🡪 GSS2018\_HW3A.sav). This modified GSS data file only contains the variables included in the assignment, which have been recoded to be more interpretable:

* **Sex:** Biological sex (0 = Female, 1 = Male)
* **Marital**: Marital status (0 = Not married, 1 = Married)
* **Health**: Respondent’s health in general (1 = Poor, 5 = Excellent)
* **Unhappy:** Happy or unhappy with life today (1 = Very Happy, 4 = Not at all Happy)
* **Depression:** Respondent told have depression (0 = No, 1 = Yes)

You will be completing the following questions, which involve conducting analyses in SPSS. You will hand in the answers to the questions; enter them into this Word document. You will also hand in your *complete* **SPSS syntax** (i.e., the syntax for all the analyses that comprise this assignment; this is just good practice!).

For the logistic regression analysis, you will be employing a **list-wise deletion** approach to handling missing data (which is also what logistic regression does by default). You will be using the following five variables in your analysis: Sex, Marital, Health, Unhappy, and Depression.

1. How many participants have complete data on these five variables? What percentage is this of the full sample?
2. Now, I want you to see if there are any differences between complete and incomplete cases on the five variables in the study. You can treat all five study variables as categorical variables. As a result, run five chi-square tests between the binary Complete cases variable you created for Question 1 and each study variable. No need to paste any of the output here. Just simply list which study variables, if any, were significantly related to the Complete cases variable.
3. Paste the output of the frequency descriptive statistics for the study variables. Make sure that you provide statistics for the complete cases only. That is, the sample size should be consistent throughout.
4. Now, run your binary logistic regression predicting Depression by the Sex, Marital, Health, and Unhappy variables. Remember to save your syntax. Fill out the following chart.

|  |  |  |
| --- | --- | --- |
| *Description of Statistic* | *Name of Statistic* | *Result (including df and p if applicable)* |
| This statistic compares the baseline model to the model with the predictors (i.e., Block 0 vs. Block 1). | Intercept | -0.4229 |
| This is ONE of the measures of effect size for the overall model, and approximates *R2* in linear regression. |  |  |
| Correct classification rate (%) for the baseline model (Block 0) | N/A |  |
| Correct classification rate (%) for the Block 1 model | N/A |  |

1. Which of the individual predictors were significantly associated with the odds of depression?
2. For each of the significant predictors, write a brief statement interpreting the result, including the odds ratio and 95% confidence interval. If a significant odds ratio is less than one, feel free to invert it to better aid your interpretation.

***Homework #3, Part B: ANCOVA***

You will use the survey.sav data set for this assignment. You can find the file on Canvas under Files 🡪 Data Sets 🡪 SPSS files 🡪 survey.sav. The following variables will be used in the HW #3B assignment:

* **age\_group:** a categorical age-group variable (0 = Younger Adults, 1 = Middle-aged Adults, 2 = Older Adults)
* **Mslfest**: mean self-esteem
* **Mlifesat**: mean life satisfaction

You will answer the following questions with this dataset. **Please remember to save your SPSS syntax and include it with your submission**. Do not worry about any missing data for this part of the assignment.

I think that, adjusting for life satisfaction, there will be differences in self-esteem among younger adults, middle-aged adults, and older adults.

1. In the above example, what is the:
   1. IV:
   2. DV:
   3. Covariate:
2. Before running the ANCOVA, perform a one-way ANOVA to test if the IV has a significant relationship with the Covariate. Is the effect of the IV significant or not significant?
3. Carry out the ANCOVA. Paste the “Between-Subjects Factors” and “Descriptive Statistics” tables you get from SPSS here.
4. Did the model pass the Levene’s Test of Equality/Homogeneity of Error Variances? Provide the *F* and *p* values of the test to back up your conclusion.
5. Paste the “Tests of Between-Subjects Effects” table from the SPSS output here. Was the covariate and the IV effects significant? What were their effect sizes? Provide the *F*, *p*, and partial eta squared values to support your conclusions.
6. If the IV effect was significant, which levels were different from one another? Use the *p* values from the Estimated Marginal Means Pairwise Comparisons table to support your conclusions.
7. Test the homogeneity of regression slopes assumption. This assumption ensures that the relationship between the covariate and the DV is similar for all levels of the IV.

***Homework #3, Part C: Logistic Regression in R***

You will use the titanic.csv data set for this assignment. You can find the file on Canvas under Files 🡪 Data Sets 🡪 csv files 🡪 titanic.csv. The following variables will be used in the HW #3C assignment:

* **Survived:** whether a passenger survived or not (0 = died, 1 = survived)
* **Pclass:** passenger’s class (1 = First, 2 = Second, 3 = Third)
* **Sex:** passenger’s sex (“male”, “female”)
* **Age:** passenger’s age (in years)

Part C is all about logistic regression in R (see the PowerPoint slides from Session 7 for guidance). Specifically, you will explore whether a passenger’s class (Pclass), sex (Sex), and age (Age) significantly predicted the odds of survival on the Titanic.

Import the titanic.csv file into RStudio to complete the following questions. **You need to upload your R script file along with your Word document**.

1. Run a frequency analysis to see how many passengers in the data set survived or not. Include code below.
2. Use the glm() function to perform a logistic regression with Survived as the DV and Pclass, Sex, and Age as the predictors. Pclass is a multicategorical variable, so make sure to use the as.factor() function so that Pclass is properly dummy coded in the regression model. Include code below and the output of the summary() function.
3. Compute the odds ratios and 95% confidence interval for the predictor coefficients. Include code and output below.
4. Which (if any) predictors were statistically significant? Provide an interpretation of the odds ratios for all significant predictors.