**Final Exam**

100 points

**Due date: 5/2/2021 midnight**

Please refer to the **General Written Homework Assignment Submission Guidelines** to prepare and submit this Final Exam.

Please **do not copy and paste the entire software output as** an answer but only the requested statistics organized in a meaningful way.

**Include the software CODE and any relevant output as a reference in an Appendix at the end of the document.**

**Introduction**

The data analysis is just one of the steps of conducting a study. In general a study involves generation and formulation of a study question, identifying the type of the study needed to be conducted to address this question, identifying the population of interest about which to draw the conclusion; sampling strategy to select the study population from where the data will be collected; deciding what data to be collected and data collection tools to be used; developing data collection tools; deciding the data management and checking data integrity. Statistics is more than just analyzing numbers and statistics is involved in all of these aforementioned activities.

In this project you will need to find a dataset based on you interest or data availability. Based on the collected variables in this dataset you will need to identify and formulate a study question and address this question using appropriate method(s) that you learned in this class. The goal of this project is to learn how to formulate a study question based on the available data and current literature; how to answer the question using appropriate statistical methods, how to carry out the statistical analysis; and how to report and interpret the findings of the study to the general reader.

**Guidelines**

Take some time to select a dataset that is interesting and it will be suitable for analysis (Linear, Logistic, regression analysis, PCA, Discriminant analysis, Classification/Regression trees, etc.).

The project output should be a report that includes the following components:

**I. Introduction**

1. What is the problem or the motivation for this study?

*Example:*

*Depression (major depressive disorder or clinical depression) is a serious mood disorder and it is one of the most common mental disorders in the U.S. It causes severe symptoms that affect how you feel, think, and handle daily activities, such as sleeping, eating, or working......*

2. What is/are the objective(s) or hypotheses of the study?

*Example:*

*The purpose of this analysis is to report the prevalence of depression and to examine potential factors associated with depression in Mexican Americans adults aged 18 years or older.*

**II. Methods**

**II.1 Study population and Measures**

3. What is the study design? (cross-sectional, longitudinal, ...)

*Example:*

*This is a cross-sectional study.*

4. What is the study population?

*Example:*

*The study population of interest is Mexican-American population in Cameron County Hispanic Cohort (CCHC), in Brownsville, Texas located on the U.S./Mexico border.*

5. What is the sampling design and the sample size of the study sample (if this information is available)?

*Example:*

*CCHC participants were recruited using two-stage stratified random sampling of U.S. census tracks and blocks, and households in the city of Brownsville.*

6. What is the outcome variable of interest and how it is measured ( if this information is available)?

*Example:*

*The outcome variable of interest is the continuous variable Depression score measured using ..........*

7. What are the explanatory variables of interest ( and how they are defined or measured if this information is available)?

*Example:*

*The potential factors of interest are: age, gender, language preferences, country of birth, education, marital status, employment status, BMI, BMI groups: normal, overweight and obese; metabolic syndrome and metabolic syndrome components: systolic and diastolic blood pressure, HDL, and triglycerides. Age was categorized into three groups: 18-39 years, 40-64 years, and 65 years or more. Body mass index (BMI) was categorized into normal weight (BMI<25 kg/m2), overweight (25≤BMI <30 kg/m2), and obesity (BMI≥30 kg/m2).....*

**II.2 Statistical Methods**

8. What are the statistical methods used in the study?

9. What is/are the type of the regression analysis/analyses used in the study?

10. How the regression analysis was carried out?

*Example:*

*We conducted survey-weighted analysis with age- and sex- adjusted probability weights, taking into account the sampling stratification by socio-economic status and clustering by census tracks and blocks and the clustering effects due to multiple participation from the same household. The prevalence of depression was calculated using all cohort participants. Data were summarized and reported as means and standard errors for continuous variables, and unweighted frequencies and weighted percentages for categorical variables. Differences in sociodemographic and clinical characteristics between depressed and non-depressed was defined groups were evaluated using Student’s t test for continuous variables and Rao-Scott F adjusted chi-square statistic for categorical variables. To identify independent factors associated with the presence of depression various multivariable weighted logistic regression models were fitted with variables selected based on the univariable analysis and controlling for potential confounders. Assessment for multicollinearity and potential two-interactions between the variables included in the models were performed as previously suggested. BMI and WC were highly correlated (r=0.89, p<0.0001) and were evaluated in separate regression models. Regression model assumptions and goodness-of-fit were checked and the models with best fit were reported. All statistical tests were two-sided and were performed using significance (alpha) level of 0.05. All statistical analyses were conducted using SAS 9.4 (SAS Institute, Inc*

**III. Results**

11. Report basic descriptive statistics of the study population including summary statistics organized in tables and/or graphs and interpretation.

*Example:*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *Table 1.[TITLE] Characteristic* | *All (n=2220)* | | *Depressed (n=971)* | | *Not depressed*  *(n=1249)* | | *P-value* |
| *Categorical variables* | | *n (%)* | | *n (%)* | | *n (%)* | |
| *Age groups* | | | | | | | |
| *18-39 years* | *1066 (45.9)* | | *340 (37.1)* | | *726 (53)* | | *<0.0001* |
| *40-64 years* | | *978 (41.1)* | | *528 (44.3)* | | *450 (38.5)* | |
| *≥65 years* | | *176 (13)* | | *103 (18.6)* | | *73 (8.5)* | |
| *Sex* | | | | | | | |
| *Male* | *770 (43.6)* | | *378 (51.1)* | | *392 (37.6)* | | *<0.0001* |
| *Female* | | *1450 (56.4)* | | *593 (48.9)* | | *857 (62.4)* | |
| *Employment status* | | | | | | | |
| *Employed* | *1148 (52.4)* | | *517 (55.8)* | | *631 (49.7)* | | *0.0692* |
| *Unemployed* | | *1070 (47.6)* | | *452 (44.2)* | | *618 (50.3)* | |
| *Country of birth* | | | | | | | |
| *Marital status* | | | | | | | |
| *Married* | *1349 (59.5)* | | *624 (63.9)* | | *725 (56)* | | *0.0177* |
| *Single or divorced* | | *867 (40.5)* | | *346 (36.1)* | | *521 (44)* | |
| *Continuous Variables* | *mean (SE)* | | *mean (SE)* | | *mean (SE)* | | *p-value* |
| *Age (years)* | *43.3 (0.65)* | | *47.24 (1.06)* | | *40.15 (0.71)* | | *<0.0001* |
| *Years of education* | *10.9 (0.15)* | | *10.5 (0.25)* | | *11.3 (0.15)* | | *0.0073* |
| *Waist circumference(cm)* | *99.9 (0.44)* | | *103.5 (0.62)* | | *96.9 (0.55)* | | *<0.0001* |
| *Hip circumference(cm)* | *108.7 (0.35)* | | *110.6 (0.53)* | | *107.1 (0.42)* | | *<0.0001* |
| *Diastolic blood pressure (mmHg)* | *71.3 (0.33)* | | *72.9 (0.57)* | | *70.1 (0.39)* | | *<0.0001* |
| *Systolic blood pressure (mmHg)* | *115.5 (0.58)* | | *119.2 (0.94)* | | *112.4 (0.64)* | | *<0.0001* |
| *BMI* | *29.9 (0.19)* | | *31.2 (0.29)* | | *28.8 (0.24)* | | *<0.0001* |

**IV. Discussion**

13. What are the main findings in this study? Are they supported by other studies conducted in similar or different populations (if you have literature reviewed)?

14. What are the strengths and the weaknesses of the study and the analysis?

15. Do you think that there is any bias?

16. Are the results generalizable to the general population?

**V. Conclusions**

17. What are the conclusions of this study?

**Link to dataset**

https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset

