**BS730 – Spring 2020**

**Project 1**

**Due Date: Tuesday 31st March by 10am (beginning of class)**

**Details**

This project must be completed individually – no collaboration at all is allowed between students on all aspects of the project. The project is worth 15% of your final grade. We encourage you to work as independently as possible on the project although you may ask the teaching team for assistance with programming issues when absolutely necessary. Your project submission should include completed tables, write-up and any graphs for each part. Your R code needs to be easily readable (neat and tidy), clearly labelled and included as an Appendix. Please submit on Blackboard by the beginning of class on 31st March 2020 by 10am before class begins. Points are assigned for presentation and organization. Projects must be a maximum of 4 single-sided A4 pages (which includes the R Code appendix). Late submissions are penalized 5 points per day unless prior arrangements have been made with your course instructor. No late projects will be accepted after three days past the due date and time.

**Background**

The Jackson Heart Study (JHS), which initiated in 1998, is a single-site, prospective, epidemiologic investigation of cardiovascular disease among African Americans. The JHS is funded by the National Heart, Lung and Blood Institute (NHLBI) and the National Institute on Minority Health and Health Disparities (NIMHD). It is a population-based longitudinal study. The JHS recruited 5306 African American residents living in the Jackson, Mississippi, metropolitan area of Hinds, Madison, and Rankin Counties. Participants were enrolled from four recruitment pools: random, 17%; volunteer, 30%; currently enrolled in the Atherosclerosis Risk in Communities Study (ARIC), 31% and secondary family members, 22%. Recruitment was limited to non-institutionalized adult African American men and women, 35-84 years old, except in a nested family cohort where those 21 to 34 years of age were also eligible. The final cohort of participants enrolled during the baseline exam included 6.6% of all African American men and women residents of the Jackson Mississippi Metropolitan Statistical Area aged 35-84 (N=76,426, US Census 2000). Because there is a greater prevalence of cardiovascular disease among African Americans compared to other ethnic groups, the purpose of the Jackson Heart Study is to explore the reasons for this disparity and to uncover new approaches to reduce it. (1)

The JHS TRANS-Data Package consists of a de-identified build of JHS data that has been modified by steps such as: creating a set of anonymized IDs, including a ~50% random sample of participants consenting to all study data usage, truncating low frequency classifications for categorical variables (<5), and setting all dates to the 15th day of the month. The JHS TRANS-Data Package is intended to be TRANSformative for research and research education. (2)

Life's Simple 7 is defined by the American Heart Association as the 7 risk factors that people can improve through lifestyle changes to help achieve ideal cardiovascular health. (3) These include managing blood pressure, controlling cholesterol, reducing blood sugar, being active, eating better, losing weight and stopping smoking. There are multiple ways of evaluating these 7 risk factors. One is awarding an individual one point for each category for which they have “ideal health” resulting in a score that would range from 0 to 7. While numerous studies have examined Life’s Simple 7 in terms of cardiovascular disease (see (4) for a meta analysis) there is also interest in how 6 (“Simple 6”) of these factors relate to diabetes. (For studies of diabetes blood sugar is not included as it is a clinical measure used in the diagnosis of diabetes.) The relationship between the Simple 6 and incident diabetes has been examined in the Jackson Heart Study (5) longitudinal data. In this project the main focus will be on the association between the Simple 6 and diabetes status at baseline (visit 1).

The current project will use a subset of the JHS TRANS-Data package. The aims of the project are:

1. To determine the prevalence of diabetes among JHS participants included in the TRANS-Data package at baseline (visit 1).

2. To examine the association between various potential risk factors at baseline (including the Simple 6) and diabetes among JHS participants included in the TRANS-Data package. **Please treat Simple 6 as a continuous variable in this analysis.**

3. Finally to look at longitudinal data for fasting plasma glucose, which is one of the clinical factors used in determining diabetes status.

A description of the variables in the dataset “**jhst\_proj1.csv**” can be found on pages 3-4. Any missing data is denoted by a blank space (“”) for character variables unless otherwise specified in the data description.

**References**

1. Wyatt SB, Diekelmann N, Henderson F, Andrew ME, Billingsley G, Felder SH et al. A community-driven model of research participation: the Jackson Heart Study Participant Recruitment and Retention Study. Ethn Dis 2003; 13(4):438-455.
2. <https://www.jacksonheartstudy.org/Research/Data-Science>
3. <https://www.heart.org/en/healthy-living/healthy-lifestyle/my-life-check--lifes-simple-7>
4. *Guo L, Zhang S. Association between ideal cardiovascular health metrics and risk of cardiovascular events or mortality: A meta-analysis of prospective studies. Clin Cardiol. 2017;40(12):1339–1346. doi:10.1002/clc.22836*
5. *Effoe VS, Carnethon MR, Echouffo-Tcheugui JB, et al. The American Heart Association Ideal Cardiovascular Health and Incident Type 2 Diabetes Mellitus Among Blacks: The Jackson Heart Study. J Am Heart Assoc. 2017;6(6):e005008. Published 2017 Jun 21. doi:10.1161/JAHA.116.005008*

**Data Dictionary**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Variable Type** | **Description** | **Coding** |
| SUBJID | Numeric | Participant ID | Integer value between 1 and 2750 |
| AGEV1 | Numeric | Participant age at visit 1 | Range 22.9 to 83.2 years |
| AGEV2 | Numeric | Participant age at visit 2 | Range 28.2 to 87.8 years |
| AGEV3 | Numeric | Participant age at visit 3 | Range 31.9 to 89.8 years |
| SexV1 | Character | Participant sex | Female  Male |
| BMIV1 | Numeric | Body mass index (kg/m2) at visit 1 | |  | | --- | | Range 18.5 to 59.57 kg/m2 | |  | |  | |
| BMIV2 | Numeric | Body mass index (kg/m2) at visit 2 | Range 18.4 to 58.80 kg/m2 |
| BMIV3 | Numeric | Body mass index (kg/m2) at visit 3 | Range 18.23 to 59.55 kg/m2 |
| DMMEDSV1 | Character | Takes medication for diabetes at visit 1 | Yes  No |
| DMMEDSV2 | Character | Takes medication for diabetes at visit 2 | Yes  No |
| DMMEDSV3 | Character | Takes medication for diabetes at visit 3 | Yes  No |
| SBPV1 | Numeric | Systolic blood pressure (mmHg) at visit 1 | 93.57 to 188.93 mmHg |
| SBPV2 | Numeric | Systolic blood pressure (mmHg) at visit 2 | 90.82 to 200.00 mmHg |
| SBPV3 | Numeric | Systolic blood pressure (mmHg) at visit 3 | 90.00 to 194.00 mmHg |
| DBPV1 | Numeric | Diastolic blood pressure (mmHg) at visit 1 | 52.64 to 99.95 mmHg |
| DBPV2 | Numeric | Diastolic blood pressure (mmHg) at visit 2 | 49.00 to 106.00 mmHg |
| DBPV3 | Numeric | Diastolic blood pressure (mmHg) at visit 3 | 50.00 to 108.00 mmHg |
| HTNV1 | Character | Hypertension at visit 1 | Yes  No |
| HTNV2 | Character | Hypertension at visit 2 | Yes  No |
| HTNV3 | Character | Hypertension at visit 3 | Yes  No |
| FPGV1 | Numeric | Fasting plasma glucose (mg/dL) at visit 1 | 70.00 to 302.00 mg/dL |
| FPGV2 | Numeric | Fasting plasma glucose (mg/dL) at visit 2 | 66.00 to 305.00 mg/dL |
| FPGV3 | Numeric | Fasting plasma glucose (mg/dL) at visit 3 | 69.00 to 302.00 t mg/dL |
| DIABETESV1 | Character | Diabetes at visit 1 | Yes  No |
| DIABETESV2 | Character | Diabetes at visit 2 | Yes  No |
| DIABETESV3 | Character | Diabetes at visit 3 | Yes  No |
| TOTCHOLV1 | Numeric | Fasting total cholesterol (mg/dL) at visit 1 | 111.00 to 327.00 mg/dL |
| TOTCHOLV2 | Numeric | Fasting total cholesterol (mg/dL) at visit 2 | 106.00 to 330.00 mg/dL |
| TOTCHOLV3 | Numeric | Fasting total cholesterol (mg/dL) at visit 3 | 106.00 to 328.00 mg/dL |
| INCOMEV1 | Character | Income category at visit 1 | Affluent  Upper-middle  Lower-middle  Poor  Unknown |
| SMK3CATV1 | Character | Smoking category at visit 1 | Ideal Health  Intermediate Health  Poor Health |
| PA3CATV1 | Character | Physical activity category at visit 1 | Ideal Health  Intermediate Health  Poor Health |
| NUTRITION3CATV1 | Character | Nutrition category at visit 1 | Ideal Health  Intermediate Health  Poor Health |
| PRIVATEPUBLICINSV1 | Character | Public or Private Insurance at visit 1 | Uninsured  Public Only  Private Only  Public & Private  Unknown |

**Part 1. Data Manipulation**

Create an R dataset called “JHS” using the csv file **jhst\_proj1.csv**. Recode any missing values for any variable denoted as blank spaces as NA. Make the following changes:

* 1. For character variables, recode any values of “Unknown” to missing (i.e. NA).
  2. Using the coding in the table below, create the following new numeric variables: IdealHealthBMIV1, AGEgrp, IdealHealthBPV1, IdealHealthCholV1, IdealHealthSMKV1, IdealHealthPAV1, IdealHealthNutritionV1, and Simple6.

|  |  |  |
| --- | --- | --- |
| **Old Variable** | **New Variable** | **Coding for New Variable** |
| BMIV1 | IdealHealthBMIV1 | 1 = BMIV1 <25 kg/m2  0 = BMIV1 >= 25 |
| AGEV1 | AGEgrp | 1 = 20 to <=40 years  2 = >40 to <=50 years  3 = >50 to <=60 years  4 = >60 to <=70 years  5 = >70 years |
| SBPV1  DBPV1 | IdealHealthBPV1 | 1= if SBP < 120 and DBP < 80  0= if SBP >=120 or DBP >=80 |
| TOTCHOLV1 | IdealHealthCholV1 | 1= if TOTCHOLV1 < 240  0= if TOTCHOLV1 >=240 |
| SMK3CATV1 | IdealHealthSMKV1 | 1= “Ideal Health”  0= “Poor Health” or “Intermediate Health” |
| PA3CATV1 | IdealHealthPAV1 | 1= “Ideal Health”  0= “Poor Health” or “Intermediate Health” |
| NUTRITION3CATV1 | IdealHealthNutritionV1 | 1= “Ideal Health”  0= “Poor Health” or “Intermediate Health” |
| IdealHealthBMIV1  IdealHealthBPV1  IdealHealthCholV1  IdealHealthSMKV1  IdealHealthPAV1  IdealHealthNutritionV1 | Simple6 | Value from 0 to 6 that is the sum over the following variables IdealHealthBMIV1, IdealHealthBPV1, IdealHealthCholV1, IdealHealthSMKV1, IdealHealthPAV1  IdealHealthNutritionV1 |

C. Exclude participants who have missing values for any of the following variables: DIABETESV1, IdealHealthBMIV1, IdealHealthBPV1, IdealHealthCholV1, IdealHealthSMKV1, IdealHealthPAV1 or IdealHealthNutritionV1.

**Part 2. Statistical Analysis**

1. Using your final dataset created in Part (1C) above, fill in the table below with the appropriate summary statistics and p-values and add a title.

**Table 1.** Distribution of risk factors stratified by diabetes status at visit 1 among Jackson Heart Study participants in the TRANS data set.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Visit 1 | **Total Participants**  **(N=xxx)** | **Diabetes** | | **P-value** |
| **No**  **(N=xxx)** | **Yes**  **(N=xxx)** |
| **DEMOGRAPHICS** |  |  |  |  |
| Age (years) |  |  |  |  |
| Age category  20-40 years  >40-50 years  >50-60 years  >60-70 years  >70 years |  |  |  |  |
| Male sex |  |  |  |  |
| **LIFE’s Simple 6** |  |  |  |  |
| Body mass index (kg/m2) |  |  |  |  |
| Ideal Health BMI |  |  |  |  |
| SBP |  |  |  |  |
| DBP |  |  |  |  |
| Ideal Health BP |  |  |  |  |
| Total Cholesterol |  |  |  |  |
| Ideal Health total cholesterol |  |  |  |  |
| Ideal Health Nutrition |  |  |  |  |
| Ideal Health Smoking |  |  |  |  |
| Ideal Health Physical Activity |  |  |  |  |
| Simple 6 total score |  |  |  |  |
| **SOCIOECONOMIC FACTORS** |  |  |  |  |
| Income  Poor  Lower Middle  Upper Middle  Affluent |  |  |  |  |
| Type of Insurance  Uninsured  Public Only  Private Only  Public and Private |  |  |  |  |

1. Based on your results in Table 1, which risk factors showed a statistically significant association with diabetes?
2. Is there an association between type of insurance and diabetes status at baseline (visit 1)? Report the name of the test, the null and alternative hypotheses, test statistic, degrees of freedom, p-value, the appropriate measure of effect (and its interpretation), and your conclusion.
3. So far in this analysis we have only examined cross sectional data from visit 1. Now we will consider longitudinal data. Is there a change in mean fasting plasma glucose between visit 1 and visit 2 among individuals who are not diabetic? What about between visit 2 and visit 3? Report the name of the test, the null and alternative hypotheses, test statistic, degrees of freedom, p-value, the appropriate measure of effect (and its interpretation), and your conclusion for visit 1 versus visit 2 and for visit 2 versus visit 3.

**PASTE YOUR R CODE HERE AS AN APPENDIX.**