***Statistical Methods for Health Professionals II***

**Spring 2021**

**Name:**

**Course Project**

***Please read through this entire document before starting to work on the project. Doing so will give you an idea of the scope of the project and the steps involved, and allow you to gauge the amount of time it will take you to complete the project, so you can plan accordingly. It is highly recommended that you work on this project incrementally throughout the semester so you can ask for help, if needed. It is highly not recommended that you wait until the end of the semester to begin work on this project. You may work together with your classmates on this project, provided everyone submits their own work.***

***Good luck!***

**Objectives**

1. Perform a regression with Stata predicting systolic blood pressure (bpsyst) (for people who are 21 to 79 years old and not pregnant), from BMI level (bmilevel), age in years (ridageyr), sex (riagendr), race (ridreth3), being currently married (married), and the interaction term of sex by being currently married (married). Make sure to consider the complex survey nature of the data.
2. Perform a regression with Stata predicting having hypertension (hyperten) (for people who are 21 to 79 years old and not pregnant), from BMI level (bmilevel), age in years (ridageyr), sex (riagendr), race (ridreth3), being currently married (married), and the interaction term of sex by being currently married (married). Make sure to consider the complex survey nature of the data.
3. Perform the same model as for objective 2, but include age group rather than continuous age in the model treating the youngest age group as the reference group. Compare how the findings are similar or different from the model you performed for objective 2. What do we learn about the relationship between age level and blood pressure from this model that we did not learn from the model you performed for objective 2?

Note: UCLA has a good page on complex survey data analysis ([link](https://stats.idre.ucla.edu/stata/seminars/applied-svy-stata13/)). It will be very useful for you to review this prior to starting work on this project. Additionally, pay close attention to the “Analysis of subpopulations” section in this page, which covers the proper use of the “subpop” command. You should use this command when running any analyses (including descriptive statistics) on subpopulations in your dataset (i.e. specific age groups, etc.) instead of deleting observations to obtain that subset (the reason for this is explained in the link).

**Data source**

The National Health and Nutrition Examination Survey (NHANES) combines interviews and physical examinations to assess the health and nutritional status of adults and children in the United States. This annual survey examines a nationally representative sample of about 5,000 persons. Individuals of all ages are interviewed in their homes and complete the health examination component of the survey, which is conducted in a mobile examination center (MEC).[[1]](#footnote-1)

For this project, we will use data from the 2015-2016 survey, and will focus on the variables shown in Table 1. We are interested in looking only at participants who were 21 to 79 years old and not pregnant at the time of the survey.

**Reference Table 1. Variables required for this analysis.**

|  |  |  |
| --- | --- | --- |
| **NHANES survey part** | **Variable name** | **Variable description** |
| Demographics Data | riagendr | Sex |
| ridageyr | Age (years) |
| ridreth3 | Race |
| dmdmartl | Marital status |
| ridexprg | Pregnancy status at exam |
| wtmec2yr | Full sample 2 year MEC exam weight |
| sdmvpsu | Masked variance pseudo-PSU |
| sdmvstra | Masked variance pseudo-stratum |
| Examination Data | bmxbmi | Body Mass Index (kg/m^2) |
| bpxsy1 | Systolic blood pressure first reading (mm Hg) |
| bpxsy2 | Systolic blood pressure second reading (mm Hg) |
| bpxsy3 | Systolic blood pressure (third reading) mm Hg |
| bpxsy4 | Systolic blood pressure fourth reading if necessary (mm Hg) |
| bpxdi1 | Diastolic blood pressure first reading (mm Hg) |
| bpxdi2 | Diastolic blood pressure second reading (mm Hg) |
| bpxdi3 | Diastolic blood pressure third reading (mm Hg) |
| bpxdi4 | Diastolic blood pressure fourth reading if necessary (mm Hg) |
| Questionnaire Data | bpq020 | Told to have high blood pressure |
| diq010 | Told to have diabetes |

**Reference Table 2. Variables that need to be created for this analysis.**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Variable name** | **Details** |
| BMI level | bmilevel | 1 = Underweight (bmxbmi < 18.5)  2 = Healthy weight (bmxbmi ≥ 18.5 & < 25)  3 = Overweight (bmxbmi ≥ 25 & < 30)  4 = Obese (bmxbmi ≥ 30) |
| Currently married | married | 1 = Married (dmdmartl = 1)  0 = Not married (dmdmartl = 2, 3, 4, 5, or 6)  . = Missing (dmdmartl = 77, 99, or .) |
| Age category | agecat | 1 = 21 to 34 (ridageyr ≥ 21 & < 35)  2 = 35 to 44 (ridageyr ≥ 35 & < 45)  3 = 45 to 54 (ridageyr ≥ 45 & < 55)  4 = 55 to 64 (ridageyr ≥ 55 & < 65)  5 = 65 to 74 (ridageyr ≥ 65 & < 75)  6 = 75 to 79 (ridageyr ≥ 75 & < 80) |
| Systolic blood pressure | bpsyst | rowmean(bpxsy1, bpxsy2, bpxsy3, bpxsy4) |
| Diastolic blood pressure | bpdias | rowmean(bpxdi1, bpxdi2, bpxdi3, bpxdi4) |
| Hypertension | hyperten | 1 = Yes (bpq020 = 1 OR bpsyst ≥ 140 OR bpdias ≥ 90)  0 = No (bpq020 = 2 AND bpsyst < 140 AND bpdias < 90) |

**Data processing**

As in most statistical projects, you must first process the data before it can be analyzed. First, open the code book and get familiar with the variables that we will use in this study. Notice how missing/don’t know/refused variables are coded, notice the range of possible values, Conduct the following data processing before proceeding:

1. Keep only the variables that we will use in this analysis (see Reference Table 1).
2. Create a BMI level (bmilevel) variable from existing BMI data (bmxbmi). See Reference Table 2.
3. Create an age category (agecat) variable from existing age data (ridageyr). See Reference Table 2.
4. Create a currently married (married) variable from existing marital status data (dmdmartl). See Reference Table 2.
5. Notice that systolic and diastolic blood pressure were measured three to four times. Average those readings into a single value for systolic blood pressure (bpsyst) and a single reading for diastolic blood pressure (bpdias). See Reference Table 2. The code to calculate the mean by row is:

*egen newvar = rowmean (varlist)*

Note: rowmean(varlist) creates the (row) means of the variables in varlist, ignoring missing values; for example, if three variables are specified and, in some observations, one of the variables is missing, in those observations newvar will contain the mean of the two variables that do exist. Other observations will contain the mean of all three variables. Where none of the variables exist, newvar is set to missing.

1. Create a new hypertension variable based on participants being told they have hypertension (bpq020) or having a high systolic or diastolic blood pressure reading (bpsyst ≥ 140 or bpdias ≥ 90). People who have high blood pressure readings but have not been told they have hypertension could have recently developed hypertension and it has not yet been diagnosed by a physician. Thus, we want to take into consideration all three variables (bpq020, bpsyst, and bpdias) when creating the new hypertension variable. See Reference Table 2.
2. To take into account the complex survey nature of the data, you will need the following svyset command:

*svyset sdmvpsu [pw = wtmec2yr], strata(sdmvstra) singleunit(centered)*

**Objective 1**

Before running any regressions, you should always calculate descriptive statistics. For this objective, fill in Table 4. Make sure that you calculate the descriptive statistics only for non-pregnant adults between 21 and 79 years old. As mentioned previously, do not drop observations to obtain the population subset, instead, use the *subpop* command as described in the UCLA tutorial.

**Table 4. Weighted sample demographic and health characteristics for non-pregnant adults between 21 and 79 years old.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **n** | **mean (sd)** | **%** |
| Systolic blood pressure in mmHg |  |  |  |
| Age in years |  |  |  |
| BMI Level  Underweight  Healthy Weight  Overweight but not obese  Obese |  |  |  |
| Sex  Male  Female |  |  |  |
| Race  Mexican American  Other Hispanic  Non-Hispanic White  Non-Hispanic Black  Non-Hispanic Asian  Other Race - Including Multi-Racial |  |  |  |
| Currently married  No  Yes |  |  |  |

* 1. Perform a regression with Stata predicting systolic blood pressure (bpsyst) for people who are 21 to 79 years old and not pregnant from BMI level (bmilevel), age in years (ridageyr), sex (riagendr), race (ridreth3), being currently married (married), and the interaction term of sex and being currently married. Given what you know about the outcome variable, make sure to run the correct type of regression. Make sure to consider the complex survey nature of the data. The reference categories for each variable should be:
* Sex: female
* BMI level: “healthy weight”
* Race: “non-hispanic white”
* Married: no
  1. Write the equation of the regression from part 1.1 in terms of beta coefficients (i.e. ) and the variable names. Do not plug in the values from the regression.
  2. Present the results of the regression in the form of a table (Table 5). Make sure to include a descriptive title for your table. Then, describe your results in 1-2 paragraphs of text comparable to what would be found in the results section of a journal article—make sure you mention the type of regression you ran, what the variables in the regression were, which results were statistically significant (main effects and interaction term), the magnitude of the main effects and interaction term (point estimate and CI), and the interpretation of the main effects and interaction term. Make sure to add a descriptive title to your table.

**Table 5. Add the title here**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **N= \_\_\_\_\_\_\_\_** |  | **Coefficient** | **95% Confidence Interval** | **p-value** |
| **Body Mass Index level**  **Underweight**  **Healthy weight**  **Overweight but not obese**  **Obese** | **Reference** |  |  |  |
| **Age (years)** |  |  |  |  |
| **Sex**  **Male**  **Female** | **Reference** |  |  |  |
| **Race**  **Mexican American**  **Other Hispanic**  **Non-Hispanic White**  **Non-Hispanic Black**  **Non-Hispanic Asian**  **Other Race - Including Multi-Racial** | **Reference** |  |  |  |
| **Currently married**  **No**  **Yes** | **Reference** |  |  |  |
| **Male & Currently married** |  |  |  |  |

**Text describing your results:**

* 1. Create a graph of predicted systolic blood pressure based on the interaction of sex and being currently married for non-pregnant adults between 21 and 79 years old. Include the graph (Figure 1) in this document and describe what you observe in the graph. Make sure to add a title to your figure.

**Objectives 2 and 3**

Calculate descriptive statistics for objectives 2 and 3 - fill in Table 6. Just as for objective 1, make sure that you calculate the descriptive statistics only for non-pregnant adults between 21 and 79 years old. For this table, calculate descriptive statistics only for those participants who had no missing data. You can do this by adding this additional condition to the *subpop* command: *!missing(hyperten) & !missing(married)*.

**Table 6. Weighted sample demographic and health characteristics of non-pregnant adults between 21 and 79 years old with and without hypertension.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Total sample**  **(N=)** | | **With hypertension**  **(n=)** | | **Without hypertension (n=)** | |
| **Variable** | **mean (sd)** | **%** | **mean (sd)** | **%** | **mean (sd)** | **%** |
| Age in years |  |  |  |  |  |  |
| Age category  21 to 34  35 to 44  45 to 54  55 to 64  65 to 74  75 to 79 |  |  |  |  |  |  |
| BMI Level  Underweight  Healthy Weight  Overweight but not obese  Obese |  |  |  |  |  |  |
| Sex  Male  Female |  |  |  |  |  |  |
| Race  Mexican American  Other Hispanic  Non-Hispanic White  Non-Hispanic Black  Non-Hispanic Asian  Other Race - Including Multi-Racial |  |  |  |  |  |  |
| Currently married  No  Yes |  |  |  |  |  |  |

* 1. Perform a regression with Stata predicting hypertension (hyperten) for people who are 21 to 79 years old and not pregnant from BMI level (bmilevel), age in years (ridageyr), sex (riagendr), race (ridreth3), being currently married (married), and the interaction term of sex and being currently married. Given what you know about the outcome variable, make sure to run the correct type of regression. Make sure to consider the complex survey nature of the data. The reference categories for each variable should be the same as for objective 1.
  2. Is the interaction term of sex and being currently married significant? If it is not, re-run the regression without the interaction term.
  3. Write the equation of the regression from part 2.2 in terms of beta coefficients. Do not plug the values from the regression.
  4. Present the results of the regression from part 2.2 in the form of a table (Table 7) and text that would be found in the results section of a journal article. Make sure to add a title to your table.
  5. Perform the same model as for part 2.2, but include age group rather than continuous age in the model treating the youngest age group as the reference group.
  6. Present the results of the regression from part 2.2 in the form of a table (Table 8). Make sure to add a title for the table. Compare how the findings are similar or different from the model you performed for part 2.2. What do we learn about the relationship between age level and blood pressure from this model that we did not learn from the model you performed for part 2.2?

1. Adapted from: https://www.cdc.gov/nchs/nhanes/about\_nhanes.htm [↑](#footnote-ref-1)