**PUBHBIO 6270: Introduction to SAS for Public Health Students**

**Final Project**

**Due Tuesday, 12/08/2020, 11:59pm via Carmen Dropbox**

**Submission instructions:**

This is an open book/notes project. You **should NOT** consult anyone except the instructor for any clarification. You must submit a single Word document that includes your name, your responses to the questions, and your SAS codes. Please be sure to include all SAS codes as an appendix, do NOT separate them under each question.

**Datasets:**

The National Health and Nutrition Examination Survey (NHANES) 4-Year (2011-2014) Data

* nhanes\_4yr.sas7bdat

2012-2014 National Health Interview Survey (NHIS) person files

* personsx12.sas7bdat
* personsx13.sas7bdat
* personsx14.sas7bdat

2012-2014 National Health Interview Survey (NHIS) family files

* familyxx12.sas7bdat
* familyxx13.sas7bdat
* familyxx14.sas7bdat

2012-2014 National Health Interview Survey (NHIS) injury/poisoning files

* injpoiep12.sas7bdat
* injpoiep13.sas7bdat
* injpoiep14.sas7bdat

**Question 1 is based on the NHANES 4-year data. You can find the description of variables in its codebook (attached with final project in Carmen).**

1. (20 points)
2. (i) Read the dataset ‘nhanes\_4yr.sas7bdat into SAS.

(ii) Create the variables BMICAT and HYPTS using the following conditions.

For **non-missing values**, BMICAT is “Underweight” if body mass index (BMI) is less than 18.5, BMICAT is “Healthy” if BMI is 18.5 but less than 25, BMICAT is “Overweight” if BMI is 25 but less than 30, and BMICAT is “Obese” if BMI is 30 and above. For **non-missing values**, HYPTS is equal to 1 if systolic blood pressure is greater than or equal to 120 **or** if diastolic blood pressure is greater than or equal to 80. And HYPTS is equal to 0 if systolic blood pressure is less than 120 **and** if diastolic blood pressure is less than 80.

Also, create labels for the variables BMICAT and HYPTS as follows:

BMICAT ~ BMI Categories

HYPTS ~ Hypertension Status

Print out the **first 20 observations** for **only** the variables SEQN, BMICAT, and HYPTS. To obtain full credit, ensure that your output contains the **labels** for each variable (and **not the variable names**).

Copy and paste your output (**with an appropriate title**) here.

1. Create formats for the age variable RIDAGEYR according to the following

scheme.

RIDAGEYR (in years) 13-18 ~ Teen

19-64 ~ Adults

65 and above ~ Elderly

For each of the age category above (Teen, Adults, Elderly), perform appropriate

procedures to generate the following summary statistics listed below (**rounded to 3 decimal places**) for the variables ‘RIDAGEYR’, ‘BMXWT’, ‘BMXHT’, ‘BMXBMI’:

*Number of non-missing values*, *number of missing values*, *mean*, *median*, *standard deviation*, *minimum value*, *maximum value, and 97% confidence*

*interval for the population mean*.

**Note**: Here, **exclude respondents 12 years and below** in the analysis.

Copy and paste your output (**with an appropriate title**) here.

1. (i) Create formats for the variable HYPTS you created in part-(a)-(ii) according to

the following scheme.

HYPTS 0 ~ No hypertension

1 ~ Pre hypertension/Stage 1 or 2 HTN

Perform appropriate procedures to answer the following questions.

1. Using the created formats in parts -(b) and -(c)-(i), generate two-way

contingency tables (**without the total percentages**) for the variables:

RIDAGEYR and BMICAT

RIDAGEYR and HYPTS

**Note**: Here, **exclude respondents 12 years and below** in the analysis.

Copy and paste your output (**with an appropriate title**) here.

1. Based on your output in part-(c)-(ii), determine and report the percentage of

Teens considered to be *overweight* or *obese* in this dataset.

1. Based on your output in part-(c)-(ii), which *age group* has the largest proportion among people with pre hypertension or stage 1 or 2 HTN in this dataset?

**Question 2 is based on your modified NHANES 4-year data (from question 1 part-a). You can find the description of variables in its codebook (attached with final project in Carmen).**

1. (a) (8 points)
2. Use PROC UNIVARIATE to perform a 2-sided one sample t-test (and using ***alpha* = 0.01**) to test the null hypothesis: the *mean weight* = *65* kg.

Make sure to produce the following graphs/statistics:

*Histogram* with KERNEL density plot, *mean*, *median*, *standard deviation*, *25%*

*quantile*, *75% quantile*, *90th percentile* (**Note**: place the summary statistics with an appropriate header at the middle-right portion of the plot window).

Copy and paste your output (**with an appropriate title**) here. Only include results for the t-test and the histogram with the KERNEL density plot and summary statistics.

1. Is there evidence to reject the null hypothesis? Briefly give your reasons and clearly state your conclusions.

(b) (10 points)

1. Perform an ANOVA test (and using *alpha* = 0.05) to examine whether the mean

body mass index differs by race and Hispanic origin. Create format for the race/hispanic variable ‘RIDRETH1’ as below and apply the created format in your analysis so that your output will be easy to read/understand.

RIDRETH1 1, 2 ~ Hispanic

3 ~ Non-Hispanic White

4 ~ Non-Hispanic Black

5 ~ Other Race

Copy and paste your output (**with an appropriate title**) here. Only include results for the ANOVA test and the graph of boxplots.

1. Report the F-statistic and the *p*-value for the F-test?
2. Is there evidence that the mean body mass index differs by race of Hispanic origin? Briefly give your reasons.
3. If your answer to part-(b)-(iii) is YES, then use **TUKEY’s method** to determine where there are pairwise differences.

Copy and paste your output here. Include **ONLY** the table for the comparisons.

1. (14 points) Investigators want to examine the linear relationship between body mass index (BMI) and HDL-Cholesterol. They decided to regress BMI on HDL-Cholesterol.

**Note**: Here, **exclude respondents 12 years and below** in the analysis.

1. Use the appropriate procedure to estimate the intercept (β0) and slope (β1) of the least squares (or regression) line. Also, produce a scatter plot (with the regression line, and confidence and prediction bands) and a diagnostic plot.

Copy and paste your output (**with an appropriate title**) here.

1. From your diagnostic plot, what can say about the normality assumption for the error term?
2. Give practical interpretations to the estimates of the intercept (β0) and the slope (β1).

Is the interpretation of the intercept (β0) practically meaningful? Briefly explain.

1. Write down the equation of the least squares (regression) line.
2. What is the expected BMI for a participant with an HDL-Cholesterol of 85 mg/dL in the study? (Round your answer to 2 decimal places.)
3. Is there a significant (linear) relationship between BMI and HDL-Cholesterol? Use the appropriate results in the output to briefly explain.
4. (18 points) Another researcher is also interested in how variables, such as RIDAGEYR, BMICAT, RIDRETH1, LBDLDL, and RIAGENDR, affect hypertension status (HYPTS).

**Note**: Here, **exclude** **respondents** **12 years and below** in the analysis.

Create format for the variable ‘RIAGENDR’ as below.

RIAGENDR 1 ~ Male

2 ~ Female

1. Using the *DESCENDING* and *PARAM=REF* options, run a binary logistic regression model with HYPTS (hypertension status) as the outcome/response variable and RIDAGEYR, BMICAT, RIDRETH1, LBDLDL, and RIAGENDR as the predictor variables. Apply the formats for RIDRETH1and RIAGENDR [created in parts -(b)-(i) and -(d)] for easy reading/understanding of your output.

Copy and paste your output (**with an appropriate title**) for ONLY the maximum likelihood estimates and odds ratio estimates here.

1. Which variable(s) is/are significant in predicting whether the participant has hypertension? Briefly explain.
2. Interpret the *odds ratio estimates* for **all** the significant variable(s) from part-(d)-

(ii) in your output.

**Question 3 is based on the NHIS data.**

1. (a) (8 points)

Generate a dataset ‘person\_family12\_14’ that contains both person and family information of people interviewed by NHIS from 2012 to 2014. **Hint**: merge datasets by HHX (household number) and FMX (family number). Ensure to sort the datasets by the variables (in the order listed) HHX and FMX.

How many observations and variables are contained in the dataset ‘person\_family12\_14’?

(b) (6 points)

Generate a dataset ‘person\_injury\_13’ that contains both person and injury information of people interviewed by NHIS in 2013. Keep **only** the following variables:

HHX FMX FPX AGE\_P SEX HISCODI3 EDUC1 R\_MARITL NOTCOV ICD9\_1 ICD9\_2 ICD9\_3 ICD9\_4 ICD9\_5 ICD9\_6 ICD9\_7 ICD9\_8 ICAUS

**Hint**: merge datasets by HHX (household number), and FMX (family number), and FPX

(person number). Ensure to sort the datasets by the variables (in the order listed) HHX, FMX, and FPX.

How many observations and variables are contained in the dataset ‘person\_injury\_13’?

1. (16 points)

Based on the dataset ‘person\_injury\_13’, generate frequency tables for the variables SEX,

HISCODI3, EDUC1, R\_MARITL, and NOTCOV. Create formats for the variables listed

below, and apply these formats in your analysis so that your output will be more readable.

SEX (gender) 1 ~ Male

2 ~ Female

HISCODI3 (race) 1 ~ Hispanic

2 ~ Non-Hispanic White

3 ~ Non-Hispanic Black

4, 5 ~ Others

EDUC1 (education level) 0-14 ~ High school or below

15-18 ~ Some college/Associate degree/Bachelors degree

19-21 ~ Master’s degree or above

96-99 ~ Children/Refused/Not ascertained/Dont know

R\_MARITL (marital status) 0 ~ Under 14 years

1-3 ~ Married

4-9 ~ Not married/Separated/Unknown marital status

NOTCOV (medical insurance coverage) 1 ~ Not covered

2 ~ Covered

9 ~ Unknown

Copy and paste your outputs (**with an appropriate title**) here and use them to answer the following questions.

1. What is the percentage of females?
2. What is the percentage of Non-Hispanic White people?
3. What is the percentage of people who had Some college/Associate/Bachelor’s

degree or above (i.e., values 15 through 21)?

1. What is the percentage of married people?
2. What is the percentage of people who are NOT covered by medical insurance?
3. (6.5 points)

Based on the dataset ‘person\_injury\_13’, create a new dataset ‘injury2013’.

In the ‘injury2013’ dataset, using Array-Do loops, search through all 8 diagnosis codes (ICD9\_1 – ICD9\_8) to generate a new variable ‘injury\_mul’ such that injury\_mul = 1 if ***at*** ***least three*** of the ICD-9-CM diagnosis codes falls into the range of 800-959, otherwise injury\_mul = 0. Create a format for the variables ‘injury\_mul’ as indicated below and apply the format in your analysis so that your output will be more readable.

Injury\_mul (injury status) 0 ~ Had 0-2 injuries

1 ~ Had 3 or more injuries

1. Generate a frequency table of the created variable ‘injury\_mul’.

Copy and paste your output (**with an appropriate title**) here.

1. What is the percentage of people who sustained ***at*** ***least three*** injuries in the 2013 NHIS?
2. (8 points)

Create formats for the variable ‘ICAUS’ according to the following scheme.

ICAUS (cause of injury) 01 ~ In a motor vehicle

02 ~ On a bike/scooter/skateboard/horse

03 ~ Pedestrian who was struck by a vehicle

04 ~ In a boat, train, or plane

05 ~ Fall

06 ~ Burn

07 ~ Other

97,98,99 ~ Refused/Not ascertained/Dont know

1. Based on the dataset ‘injury2013’, draw a **horizontal** bar chart for the variable ‘ICAUS’. Apply the formats for ‘ICAUS’ in your analysis so that your output will be more readable. Make sure to ***exclude*** ***only*** those who had missing cause of injury episode (i.e., ICAUS = . ) **and** those who did not sustain injuries in 2013. (**Hint**: you may have to create a new binary variable to determine if someone sustained injury or not.)

Copy and paste your bar chart (**with an appropriate title**) here.

1. In addition, draw a **horizontal** boxplot for ‘AGE\_P’ by ‘ICAUS’. Apply the formats for ‘ICAUS’ in your analysis so that your output will be more readable. Make sure to

***exclude*** ***only*** those who had missing cause of injury episode (i.e., ICAUS = . ) and those who did not sustain injuries in 2013. (**Hint**: you may have to create a new binary variable to determine if someone sustained injury or not.)

Copy and paste your boxplot (**with an appropriate title**) here.

1. (8.5 points)

Based on the dataset ‘injury2013’ (from part d), perform a chi-square test (and using *alpha* = 0.05) to test null hypothesis: injury status is NOT associated with insurance coverage status. In addition to the format you created for the variable ‘injury\_mul’ in part-(d), create a format for the variable ‘NOTCOV’ as indicated below, and apply the formats for these variables in your analysis so that your output will be more readable. Make sure to ***exclude*** those who had unknown coverage status (i.e., NOTCOV = 9).

NOTCOV (medical insurance coverage) 1 ~ Not covered

2 ~ Covered

Copy and paste all your output (**with an appropriate title**) here.

1. Report the value of *chi-square statistic* and the corresponding *p-*value?
2. Is there evidence to reject the null hypothesis? Briefly give your reasons and clearly state your conclusions.
3. (9 points)

Based on the dataset ‘injury2013’ (from part d), perform a two-sample t-test (and using *alpha* = 0.05) to test null hypothesis: mean age among people who had 0-2 injuries = mean age among people who had 3 or more injuries. Apply the formats for the variable ‘injury\_mul’ created in part-(d) in your analysis so that your output will be more readable.

Copy and paste your output (**with an appropriate title**) here. Include ONLY the tables for the summary statistics and test results here. (Do not include any graph.)

1. What are the sample mean ages among people who had 0-2 injuries and those

who had 3 or more injuries?

1. Is there evidence to reject the null hypothesis? Briefly give your reasons and clearly state your conclusions.

**Appendix**: SAS codes (10 points)