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

**Final Project**

**Due Friday, 12/10/2021, 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 in Carmen by the due date. 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) 6-Year (2011-2016) Data

* nhanes\_6yr.sas7bdat

2011-2017 National Health Interview Survey (NHIS) person files

* personsx11.sas7bdat
* personsx12.sas7bdat
* personsx13.sas7bdat
* personsx14.sas7bdat
* personsx15.sas7bdat
* personsx16.sas7bdat
* personsx17.sas7bdat

2011-2017 National Health Interview Survey (NHIS) family files

* familyxx11.sas7bdat
* familyxx12.sas7bdat
* familyxx13.sas7bdat
* familyxx14.sas7bdat
* familyxx15.sas7bdat
* familyxx16.sas7bdat
* familyxx17.sas7bdat

2011-2017 National Health Interview Survey (NHIS) injury/poisoning files

* injpoiep11.sas7bdat
* injpoiep12.sas7bdat
* injpoiep13.sas7bdat
* injpoiep14.sas7bdat
* injpoiep15.sas7bdat
* injpoiep16.sas7bdat
* injpoiep17.sas7bdat

**Question 1 is based on the NHANES 6-year data. You can find the description of variables in its codebook (attached with final project in Carmen).**

1. (20 points)
2. (i) Read/load the dataset ‘nhanes\_6yr.sas7bdat into SAS.

(ii) Update the loaded data in SAS by creating the variables BMICAT and HYPTS using the following conditions.

**Hint**: consider the length of the characters for the new variables when writing your code

**Note**: Here, **do not** **exclude** respondents with missing values for any of the

variables.

If the body mass index (BMI) is missing, then BMICAT is equal to “**Missing**”. 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.

If **both** systolic blood pressure and diastolic blood pressure are missing, then HYPTS is equal to “**Missing**”. For **non-missing values**, HYPTS is equal to “**Prehypertension/Stage 1 or 2 HTN**” 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 “**No hypertension**” 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

(iii) From your updated data (in part-a-ii), use PROC MEANS (with appropriate title) to check whether the variable ‘BMICAT’ is correctly coded.

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

According to the output, do you think the variable ‘BMICAT’ is correctly coded? Explain why or why not.

(iv) Using your updated data (part-a-ii), print out the **third 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

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

For each of the age category above (i.e., Teen, Adults, Elderly), perform appropriate procedures to generate the following summary statistics listed below (**rounded to 2 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 two-sided 85% confidence interval for the population mean*.

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

1. (i) Perform appropriate procedures to answer the following questions.

**Note**: Here, **exclude** respondents 12 years and below **and** with “Missing” value for BMICAT in the analysis.

Using the format you created in part-(b), generate a two-way contingency table (**without the total percentages**) for the variables:

RIDAGEYR and BMICAT

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

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

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

1. Perform appropriate procedures to answer the following questions

**Note**: Here, **exclude** respondents 12 years and below **and** with “Missing” value for HYPTS in the analysis.

Using the format you created in part-(b), generate a two-way contingency table (**without the total percentages**) for the variables:

RIDAGEYR and HYPTS

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

Based on your output in part-(c)-(ii), which *age group* had the smallest proportion

among people with “Prehypertension/Stage 1 or 2 HTN” in this dataset?

**Question 2 is based on your updated NHANES 6-year data (from question 1 part-a-ii). 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* = *66* kg.

**Note**: Here, **exclude** respondents who weigh below 40 kg in the analysis.

Make sure to produce the following graphs/statistics:

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

*quantile*, *75% quantile*, *99th percentile* (**Note**: place the summary statistics with an appropriate header at the middle-right portion (inside) of the plot window).

Copy and paste your output (**with an appropriate title**) here. Only include output of 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 indicated below and apply the created format in your analysis so that your output will be easy to read/understand.

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

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 178

mg/dL in the study? (Round your answer to 2 decimal places.)

1. Is there a significant (linear) relationship between BMI and HDL-Cholesterol? Use the appropriate results in the output to briefly explain.
2. (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 **and** with “Missing” values for both HYPTS and BMICAT in the analysis.

Create format for the variable ‘RIAGENDR’ as below.

RIAGENDR 1 ~ Male

2 ~ Female

1. Run a 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), respectively] for easy reading/understanding of your output. Also, use the “PARAM=REF” option, and set EVENT=“Prehypertension/Stage 1 or 2 HTN” for HYPTS; and REF=“Male” for RIAGENDR; REF=“Healthy” for BMICAT; and REF=“Hispanic” for RIDRETH1.

Copy and paste your outputs (**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\_17’ that contains both person and family information of people interviewed by NHIS from 2012 to 2017. Here, sort and merge datasets by (in the order listed) HHX (household number) and FMX (family number)

How many observations and variables are contained in the dataset ‘person\_family12\_17’? 603794 observations and 814 variables

(b) (6 points)

Generate a dataset ‘person\_injury\_11’ that contains both person and injury information of people interviewed by NHIS in 2011. Here, your dataset must include **ONLY** observations/people from the “person file” and their (additional) information from the “injury/poisoning file.” Also, keep **only** the following variables:

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

**Hint:** sort and merge datasets by (in the order listed) HHX (household number), FMX (family number), and FPX (person number)

How many observations and variables are contained in the dataset ‘person\_injury\_11’?

1. (16 points)

Based on the dataset ‘person\_injury\_11’, generate frequency tables for the variables SEX, HISCODI3, EDUC1, R\_MARITL, and NOTCOV. Create formats for the variables as indicated 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 males?
2. What is the percentage of Non-Hispanic Black people?
3. What is the percentage of people who had Some college/Associate/Bachelor’s

degree or above (i.e., values 15 through 21 for EDUC1)?

1. What is the percentage of married people?
2. What is the percentage of people covered by medical insurance?
3. (6.5 points)

Based on the dataset ‘person\_injury\_11’, create a new dataset ‘injury2011’.

In the ‘injury2011’ 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. Here, you can create intermediate variable(s) of your choosing before creating the variable injury\_mul. In addition, create a format for the variable ‘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. Using the format above, 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 2011 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 ‘injury2011’ (in part-d), 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. Here, make sure to ***exclude*** respondents who had missing cause of injury episode (i.e., ICAUS = . ) **and** who did not sustain injuries in 2011 (i.e., had zero injuries based on the diagnosis codes ICD9\_1 – ICD9\_8 ) – In other words, include those **without** “ICAUS = . ” and had one or more injuries. However, ensure **not to exclude** the numbers 97, 98, and 99 for the variable ICAUS in your analysis.

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

1. Based on the dataset ‘injury2011’ (in part-d), 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. Here, make sure to ***exclude*** respondents who had missing cause of injury episode (i.e., ICAUS = . ) **and** who did not sustain injuries in 2011 (i.e., had zero injuries based on the diagnosis codes ICD9\_1 – ICD9\_8) – In other words, include those **without** “ICAUS = . ” **and** had one or more injuries. However, ensure **not to exclude** the numbers 97, 98, and 99 for the variable ICAUS.

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

1. (8.5 points)

Based on the dataset ‘injury2011’ (from part-d), perform a chi-square test (and using *alpha* = 0.05) to test null hypothesis: Insurance coverage status is NOT associated with whether a person has at least 3 injuries or not.

**Note**: Here, **exclude** respondents who had unknown coverage status (i.e., NOTCOV = 9) in the analysis.

Apply the formats you created for the variables NOTCOV (in part-c) and ‘injury\_mul’ (part-(d)) in your analysis so that your output will be more readable.

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 ‘injury2011’ (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)

OPTIONS NOFMTERR;

Libname project "C:\Users\iasan\Desktop\Final project";

**PROC** **CONTENTS** DATA=project.nhanes\_6yr;

**RUN**;

**DATA** project.person\_12\_17;

SET project.personsx12 project.personsx13 project.personsx14 project.personsx15 project.personsx16 project.personsx17;

**RUN**;

**DATA** project.family\_12\_17;

SET project.familyxx12 project.familyxx13 project.familyxx14 project.familyxx15 project.familyxx16 project.familyxx17;

**RUN**;

**PROC** **SORT** DATA=project.person\_12\_17;

BY HHX FMX;

**RUN**;

**PROC** **SORT** DATA=project.family\_12\_17;

BY HHX FMX;

**RUN**;

**DATA** project.person\_family12\_17;

MERGE project.person\_12\_17 (IN=a) project.family\_12\_17;

IF a=**1**;

BY HHX FMX;

**RUN**;