Epidemiology 640 – Fall 2017

NAME: Stephanie Mecham

SECTION: 2

Homework 6: Statistical Procedures and Intermediate Programming

This homework assignment completes our replication of **Bansil et al., 2011** using the 2013-2014 NHANES data. You may discuss the assignment with others but you MUST run your own code and write up the answers in your own words. Please turn in a printed copy of your write-up that includes all relevant SAS code (attached as a printout to the back of your write-up) in lecture at **9:30AM on Friday, December 8th.** ONLY the relevant output that is **requested by the questions** should be included in the paper copy of your homework.

Be sure that: 1) **Your SAS code runs from start to finish.**

2) Your results make sense (check your sample size and look for unreasonable, unlikely, or impossible answers).

3) Your homework has your name on EACH page. Submit your write-up on Canvas as an attachment to Assignment 6. Your file should be turned in as either a .doc, .docx or .pdf in the following format:

LASTNAME_FIRSTNAME_HW6.docx

4) Your code is **well commented** (the top of your file should include the homework number and your name, each question should be identified in the code, and each new task should be described by comments) and formatted (use indentation and carriage returns to improve readability). Turn in your SAS code as an attachment to Assignment 6 on Canvas. Your file should be turned in as a .sas in the following format:

LASTNAME_FIRSTNAME_HW6.sas

5) 5% will be deducted if either of tasks 3 or 4 above is not completed.

Before beginning homework 6, create a new folder on your M: drive or flash drive specifically for Homework 6. Create a permanent library called **nhanes** that points to the folder you just created. Download the **sleepbp_final** dataset and **formats** file from Canvas and place it in the folder on your computer that corresponds to your nhanes library. The codebook below defines all variables and possible values in the **sleepbp_final** dataset.

NOTE: In this homework, you will be running models on continuous blood pressure values to test your knowledge on linear regression. This differs from the Bansil et al. paper, where they used logistic models to predict hypertension.

Variable	Variable Name	Possible Values (Defined)
Respondent sequence number	SEQN	
Gender	RIAGENDR	1: Male 2: Female
Age	RIDAGEYR	18-79: continuous age range 80: 80 years old and over
Ever told by doctor you have sleep disorder	SLQ060	1: Yes 2: No

Body Mass Index (BMI)	BMXBMI	12.1-82.9: continuous range of values
Covered by health insurance	HIQ011	1: Yes 2: No
Short sleep duration (<7 hrs)	shortsleep	0: No 1: Yes
Combination of sleep problems	sleepcombo	1: Sleep disorder and short sleep 2: Sleep disorder only 3: Short sleep only 4: None of the above
Average systolic BP	sysbp_ave	64.67-228.67: continuous range of values
Average diastolic BP	diabp_ave	0-128.00: continuous range of values
Current blood pressure medication use	bpmed	0: No 1: Yes
Hypertension	htn	0: No 1: Yes
Current smoking status	smoke	0: No 1: Yes
Diabetes	diab	0: No 1: Yes
Race	race	1: Mexican American 3: White 4: Black 5: Other
Education level	educ	1: Less than HS 2: HS/GED 3: At least some college
Poverty income ratio	pir	1: <1.00 2: 1.00 – 3.00 3: >3.00

Task 1: Testing Differences in Means and Proportions between Hypertension Groups

1. Complete table 1 from Homework 5 by running an appropriate statistical test to compare if there are differences in the distribution of subject characteristics by hypertension status. Fill in the p-values in Table 1 and list the statistical tests used for each variable in a footnote below the table. For those comparing means, comment on whether you used a parametric or nonparametric test and why.

Epidemiology 640 – Fall 2017 NAME: Stephanie Mecham

SECTION: 2

Table 1: Distribution of Demographic and Health Characteristics Among US Adults by Hypertension Status, National Health and Nutrition Examination Survey, 2013-2014

	Total N (%)	Normotensive N (%)	Hypertensive N (%)	p-value
Mean systolic BP (SD)	122.92 (17.92)	115.43 (10.87)	137.05 (19.96)	<.0001
Mean diastolic BP (SD)	69.34 (12.47)	68.02 (10.14)	71.84 (15.66)	<.0001
Female gender	2941 (51.52)	1873 (51.19)	1068 (52.12)	0.4981
Mean age (SD)	47.98 (18.43)	40.38 (16.23)	61.54 (13.74)	<.0001
Race				<.0001
Non-Hispanic white	2417 (42.34)	1515 (41.40)	902 (44.02)	
Non-Hispanic black	1177 (20.62)	625 (17.08)	552 (26.94)	
Mexican American	787 (13.79)	566 (15.47)	221 (10.79)	
Other	1327 (23.25)	953 (26.05)	374 (18.25)	
Has health insurance	4515 (79.20)	2717 (74.36)	1798 (87.84)	<.0001
Education				<.0001
Less than high school	1147 (21.32)	631 (18.91)	516 (25.24)	
High school / GED	1215 (22.58)	702 (21.04)	513 (25.10)	
At least some college	3018 (56.10)	2003 (60.04)	1015 (49.66)	
Poverty level[a]				<.0001
<1.00	1221 (23.17)	821 (24.30)	400 (21.16)	
1.00 – 3.00	2085 (39.57)	1260 (37.29)	825 (43.65)	
>3.00	1963 (37.26)	1298 (38.41)	665 (35.19)	
Mean body mass index (SD)	28.89 (7.09)	27.79 (6.53)	30.91 (7.62)	<.0001
Has diabetes	704 (12.34)	199 (5.44)	505 (24.65)	<.0001
Current smoker	1139 (19.96)	767 (20.97)	372 (18.16)	0.0110
Has sleep disorder	544 (9.55)	256 (7.01)	288 (14.10)	<.0001
Short sleep duration (<7 hours)	2148 (37.69)	1361 (37.23)	787 (38.52)	0.3332
Combination of sleep problems[b]				<.0001
Sleep disorder and short sleep	281 (4.94)	136 (3.73)	145 (7.12)	
Sleep disorder only	261 (4.59)	120 (3.29)	141 (6.93)	
Short sleep only	1862 (32.74)	1222 (33.47)	640 (31.43)	
None of the above	3283 (57.73)	2173 (59.52)	1110 (54.52)	

Abbreviations: BP, blood pressure; SD, standard deviation. [a]The ratio of a family's income to the federally defined poverty threshold for a family of the same size in the same calendar year. [b]Mutually exclusive sleep problem categories. A chi-square test was performed for categorical variables: gender, sleep disorder, short sleep, combination of short sleep and sleep disorder, smoking status, diabetes status, race, and education. A parametric t-test was used for numeric variables that had a relatively normal distribution: A non-parametric t-test was used for numeric variables that did not have a normal distribution: age.

Task 2: Unadjusted Models and Checking Model Assumptions

2. Create a macro which performs a bivariate (simple) linear regression. Use it to predict **sysbp_ave** and **diabp_ave** from each of the main predictors (**SLQ060, shortsleep**, and

sleepcombo) and covariates (**RIDAGEYR, race, RIAGENDR, pir, BMXBMI, smoke**, and **diab**. This is a total of 20 models. For continuous predictors, have your macro use PROC REG, and for categorical predictors, use PROC GENMOD or GLM. I encourage you to do this in one macro, but you may use two, one for each predictor type. Be sure your macro produces graphics for checking model assumptions. Paste your macro code below.

```
*Task 2: Unadjusted Models and Checking Model Assumptions;

*Macro for categorical predictors;

%macro bp_model(bp,depvar,ref);
proc GLM data=nhanes.sleepbp_final plots(maxpoints=none);
	class &depvar (ref=&ref);
	model &bp = &depvar / solution clparm;
run;
quit;
%mend;

*Macro for continuous predictors;

%macro bp_model2(bp,depvar);
proc reg data=nhanes.sleepbp_final plots(maxpoints=none);
	model &bp = &depvar / clb;
run;
quit;
%mend;

*20 Models;

%macro model(bp,depvar);
proc GLM data=nhanes.sleepbp_final plots (maxpoints=none);
	class &depvar;
	model &bp = &depvar / solution clparm;
run;
quit;
%mend;

%model(sysbp_ave, slq060)
%model(sysbp_ave, shortsleep)
%model(sysbp_ave, sleepcombo)
%bp_model2(sysbp_ave, ridageyr)
%model(sysbp_ave, race)
%model(sysbp_ave, riagendr)
%model(sysbp_ave, pir)
%bp_model2(sysbp_ave, bmxbmi)
%model(sysbp_ave, smoke)
%model(sysbp_ave, diab)
%model(diabp_ave, slq060)
%model(diabp_ave, shortsleep)
%model(diabp_ave, sleepcombo)
%bp_model2(diabp_ave, ridageyr)
%model(diabp_ave, race)
%model(diabp_ave, riagendr)
```

```
%model(diabp_ave, pir)
%bp_model2(diabp_ave, bmxbmi)
%model(diabp_ave, smoke)
%model(diabp_ave, diab)
```

3. Fill out the unadjusted model columns in Table 2 below. Round to two decimal places.

4. Comment on any possible violations of model assumptions for the **RIDAGEYR** model predicting **diabp_ave**.

The assumption of homoscedasticity (equal variance across all values of x) is violated because RIDAGEYR had a clear non-linear pattern in the residual plot (looked like a curved upside down U shape).

5. Why are there only parameter estimates reported for three of the four levels of **race**?

   Because White race is used as a reference group, the parameter estimates are only reported for the three non-reference groups (Mexican American, Black, and Other).

6. Interpret the model parameters for the **pir** model predicting **sysbp_ave**.

The intercept estimate of 122.59 mmHg can be interpreted as the sysbp_ave value we would expect for a pir of 0. This is not a relevant measure because pir is limited to the values of 1, 2, and 3.

When the pir is from 1-3, the sysbp_ave will be 1.35 mmHg greater than the reference group of pir <1, on average.
When the pir is greater than 3, the sysbp_ave will 0.46 mmHg lower than the reference group of pir <1, on average.

7. Interpret the model parameters for the **BMXBMI** model predicting **diabp_ave**.

   The intercept estimate of 62.42 mmHg can be interpreted as the value we would expect for a BMI of 0. This is not relevant because humans cannot have a BMI of zero.

   For every unit increase in BMI, the diabp_ave will increase by 0.24 mmHg, on average.

8. Report and interpret the R^2 value for the **BMXBMI** model predicting **diabp_ave**.

   The R^2 value in this case is 0.0185, which we can interpret as 1.85% of the variation in the diabp_ave variable that can be explained by our linear model. This is pretty low, so we can conclude that the model does not fit our data very well in this case.

Epidemiology 640 – Fall 2017

NAME: Stephanie Mecham

SECTION: 2

Task 3: Multiple Linear Regression Predicting Systolic and Diastolic Blood Pressure

9. Run multiple linear regression models using PROC GENMOD or GLM. For each blood pressure outcome and each sleep problem predictor (6 models), adjust for the following covariates: **RIDAGEYR, RIDAGEYR^2, race, RIAGENDR, pir, BMXBMI, smoke**, and **diab.** Include options to output the parameter estimates and 95% confidence intervals for the parameter estimates, if necessary.

10. Complete Table 2 by filling out the adjusted model columns. Round to two decimal places.

Table2: Linear Regression Parameter Estimates Predicting Blood Pressure in US Adults with Individual and Combinations of Sleep Problems According to the National Health and Nutrition Examination Survey, 2013-2014

	Systolic BP		Diastolic BP	
	Unadjusted estimate (95% CI)	Adjusted[a] estimate (95% CI)	Unadjusted estimate (95% CI)	Adjusted[a] estimate (95% CI)
Sleep disorder				
Yes	2.65 (1.04, 4.25)	-1.74 (-3.23, -0.25)	1.37 (0.25, 2.49)	-0.71 (-1.83,0.40)
No	ref	ref	ref	ref
Sleep duration				
Short (<7 hours)	0.22 (-0.75, 1.19)	0.07 (-0.82, 0.96)	1.12 (0.45, 1.79)	0.04 (-0.62, 0.71)
Normal (≥7 hours)	ref	ref	ref	ref
Combination of sleep problems[b]				
Sleep disorder and short sleep	1.81 (-0.39, 4.02)	-1.49 (-3.51, 0.53)	2.09 (0.55, 3.63)	-0.97 (-2.48, 0.54)
Sleep disorder only	3.81 (1.52, 6.11)	-2.01 (-4.10,0.08)	1.38 (-0.23, 2.98)	-0.36 (-1.92, 1.21)
Short sleep only	0.35 (-0.68, 1.37)	0.16 (-0.78, 1.10)	1.13 (0.41, 1.84)	0.20 (-0.50, 0.91)
None of the above	ref	ref	ref	ref

Abbreviations: BP, blood pressure; CI, confidence interval; ref, reference category. [a]Models adjusted for age, age squared, race, gender, poverty to income ratio, bmi, smoking, and diabetes. [b]Mutually exclusive sleep problem categories.

Epidemiology 640 – Fall 2017

NAME: Stephanie Mecham

SECTION: 2

Code:

```
*Homework 6: Statistical Procedures and Intermediate Programming | Stephanie Mecham |
Section 2;


LIBNAME nhanes "C:\Users\smecham\Desktop\nhanes";
options fmtsearch= (nhanes);

*Task One: Testing Differences in Means and Proportions between Hypertension Groups;

*Chi-Square Test for Categorical Variables;

proc freq data=nhanes.sleepbp_final;
tables riagendr*htn SLQ060*htn shortsleep*htn sleepcombo*htn smoke*htn diab*htn
race*htn educ*htn HIQ011*htn pir*htn / chisq;
run;

*Testing for Normality for Numeric Variables;

proc univariate data=nhanes.sleepbp_final;
var ridageyr bmxbmi sysbp_ave diabp_ave;
histogram ridageyr bmxbmi sysbp_ave diabp_ave / normal;
qqplot;
run;

*Parametric T-Test for Relatively Normal Numeric Variables;

proc ttest data=nhanes.sleepbp_final
alpha= .05;
class htn;
var bmxbmi;
run;

proc ttest data=nhanes.sleepbp_final
alpha= .05;
class htn;
var sysbp_ave;
run;

proc ttest data=nhanes.sleepbp_final
alpha= .05;
class htn;
var diabp_ave;
run;
```

```
*Non-parametric T-Test for Non-Normal Numeric Variables;

proc npar1way data= nhanes.sleepbp_final
wilcoxon;
class htn;
var ridageyr;
run;

*Task 2: Unadjusted Models and Checking Model Assumptions;

*Macro for categorical predictors;

%macro bp_model(bp,depvar,ref);
proc GLM data=nhanes.sleepbp_final plots(maxpoints=none);
        class &depvar (ref=&ref);
        model &bp = &depvar / solution clparm;
run;
quit;
%mend;

*Macro for continuous predictors;

%macro bp_model2(bp,depvar);
proc reg data=nhanes.sleepbp_final plots(maxpoints=none);
        model &bp = &depvar / clb;
run;
quit;
%mend;

*20 Models;

%macro model(bp,depvar);
proc GLM data=nhanes.sleepbp_final plots (maxpoints=none);
        class &depvar;
        model &bp = &depvar / solution clparm;
run;
quit;
%mend;

%model(sysbp_ave, slq060)
%model(sysbp_ave, shortsleep)
%model(sysbp_ave, sleepcombo)
%bp_model2(sysbp_ave, ridageyr)
%model(sysbp_ave, race)
```

```
%model(sysbp_ave, riagendr)
%model(sysbp_ave, pir)
%bp_model2(sysbp_ave, bmxbmi)
%model(sysbp_ave, smoke)
%model(sysbp_ave, diab)
%model(diabp_ave, slq060)
%model(diabp_ave, shortsleep)
%model(diabp_ave, sleepcombo)
%bp_model2(diabp_ave, ridageyr)
%model(diabp_ave, race)
%model(diabp_ave, riagendr)
%model(diabp_ave, pir)
%bp_model2(diabp_ave, bmxbmi)
%model(diabp_ave, smoke)
%model(diabp_ave, diab)

*Task 3: Multiple Linear Regression Predicting Systolic and Diastolic Blood Pressure;

proc glm data=nhanes.sleepbp_final
plots (maxpoints=none)=(diagnostics residuals(smooth));
        class sleepcombo race RIAGENDR pir smoke diab;
        model sysbp_ave = sleepcombo RIDAGEYR RIDAGEYR*RIDAGEYR race RIAGENDR pir
BMXBMI smoke diab  / solution clparm;
run;

proc glm data=nhanes.sleepbp_final
plots (maxpoints=none)=(diagnostics residuals(smooth));
        class shortsleep race RIAGENDR pir smoke diab;
        model sysbp_ave = shortsleep RIDAGEYR RIDAGEYR*RIDAGEYR race RIAGENDR pir
BMXBMI smoke diab  / solution clparm;
run;

proc glm data=nhanes.sleepbp_final
plots (maxpoints=none)=(diagnostics residuals(smooth));
        class SLQ060 race RIAGENDR pir smoke diab;
        model sysbp_ave = SLQ060 RIDAGEYR RIDAGEYR*RIDAGEYR race RIAGENDR pir
BMXBMI smoke diab  / solution clparm;
run;

proc glm data=nhanes.sleepbp_final
plots (maxpoints=none)=(diagnostics residuals(smooth));
        class sleepcombo race RIAGENDR pir smoke diab;
        model diabp_ave = sleepcombo RIDAGEYR RIDAGEYR*RIDAGEYR race RIAGENDR pir
BMXBMI smoke diab  / solution clparm;
run;
```

```
proc glm data=nhanes.sleepbp_final
plots (maxpoints=none)=(diagnostics residuals(smooth));
          class shortsleep race RIAGENDR pir smoke diab;
          model diabp_ave = shortsleep RIDAGEYR RIDAGEYR*RIDAGEYR race RIAGENDR pir
BMXBMI smoke diab  / solution clparm;
run;

proc glm data=nhanes.sleepbp_final
plots (maxpoints=none)=(diagnostics residuals(smooth));
          class SLQ060 race RIAGENDR pir smoke diab;
          model diabp_ave = SLQ060 RIDAGEYR RIDAGEYR*RIDAGEYR race RIAGENDR pir
BMXBMI smoke diab  / solution clparm;
run;

*end of Homework 6;
```