EPIDEMIOLOGY 640: SAS FOR EPIDEMIOLOGICAL RESEARCH
FINAL EXAM

This is an open-book test, take-home exam. You are welcome to use any documents including class notes, labs, cheat sheets and text books as well as electronic resources including SAS, SAS help, UCLA notes, etc. To test your knowledge, however, you must work alone. **DO NOT DISCUSS ANY ASPECT OF THIS EXAMINATION WITH ANYONE UNTIL AFTER THE DUE DATE.** All questions should be emailed copying your GSI (bdeepti@umich.edu or wenchaol@umich.edu) and both instructors (jpetrie@umich.edu and ratliffs@umich.edu).

- ☐ We will provide three datasets and one format file in the final exam assignment on Canvas. After downloading, please ensure that the data is ok (i.e., you can open it, attach formats etc.) and alert the instructors ASAP if there is a problem.

- ☐ All exams (code and questions) are due as an electronic copy on Canvas by **9:30 AM on Monday, December 18st. NO LATE EXAMS WILL BE ACCEPTED.**

- ☐ As always, your code will be graded on correct syntax, presence of comments, and clear formatting.

- ☐ Include only the requested output (**not all output produced**). Write-ups should be complete, concise, and written for an audience of epidemiologists with reasonable training in biostatistics. Be sure to explain what your results mean; do not simply report the numbers or state the statistical significance without interpretation. Assume an alpha of 0.05 and use two-sided tests.

The examination has **8** pages (including this cover sheet) and **8** questions for a total of **25 points**. Good Luck!

I affirm that I did not discuss the contents of this examination with anyone but the instructors or the GSIs.

Name (printed)__Stephanie Mecham_____

Student ID_____50217116_____

Signature_____Stephanie Mecham_____Date_____12/13/2017_____

In this exam, we will examine associations between perfluorooctanoic acid (PFOA) and high uric acid levels (hyperuricemia). PFOA is an environmentally persistent chemical used in a variety of manufacturing processes, and has been associated with a variety of adverse health outcomes. Hyperuricemia is metabolic condition that can lead to gout, diabetes, chronic kidney disease, cardiovascular disease and other chronic health conditions. We will examine the association between PFOA and hyperuricemia using NHANES data.

You will start by downloading three datasets from the Canvas site: DEMO_E.sas7bdat (be sure to download the formats file associated with this SAS dataset), ACTV_E.csv, and LBX_E.txt. You should also download the data dictionary that describes the variables in each of the datasets. Please confirm that your datasets are ok early in the week (i.e., you can open it, attach formats etc.). Send an email copying your GSI and both instructors ASAP if there are any problems.

1. Importing and preparing the ACTV_E.csv file for analysis.

   a) (1.5 pt) Import the ACTV_E.csv file into a temporary dataset **using PROC IMPORT.**

What is the delimiter of ACTV_E.csv?	comma
How many observations are in the dataset?	10108

   b) (1 pt) Create 3 new variables for analysis including BMI, BMI category, and an indicator variable for low physical activity. BMI is defined as weight in kilograms divided by height in meters squared (kg/m^2). Your BMI category variable should be numeric and indicate underweight (BMI less than 18.5), healthy weight (BMI greater than or equal to 18.5 but less than 25), overweight (BMI greater than or equal to 25 but less than 30), and obese (BMI greater than or equal to 30) BMI values. Your low physical activity variable should be equal to 1 if the subject responded "No" to both activity questions, equal to 0 if the subject responded "Yes" to either activity question, and missing otherwise. After checking your work, delete any records with missing values in any of your newly created BMI or activity variables from your final ACTV_E dataset.

Which variable had the most missing data after recoding, and how many missing observations did this variable have?	lowactivity had 3003 missing values, as opposed to the 1,247 missing values each in bmi and bmi_cat

How many observations does your dataset have after deleting records with missing data?	6744 observations

c) (1 pt) Drop all of the original, non-recoded variables from your dataset. Paste a copy of your log output after successfully dropping the original variables.

```
173  data recoded_bmi;
174  set bmi_final;
175  drop SEQN -- PAQ665;
176  run;

NOTE: There were 6744 observations read from the data set WORK.BMI_FINAL.
NOTE: The data set WORK.RECODED_BMI has 6744 observations and 3 variables.
NOTE: DATA statement used (Total process time):
      real time           0.03 seconds
      cpu time            0.03 seconds
```

d) (1 pt) Create formats corresponding to the coding you used to create your BMI category and low physical activity variables. Apply these formats to your variables. Run and paste a PROC CONTENTS into your exam to demonstrate that you successfully assigned these formats.

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format
1	bmi	Num	8	
2	bmi_cat	Num	8	BMICATEGORY.
3	lowactivity	Num	8	PHYSICAL_ACTIVITY.

2. Importing and preparing the LBX_E.txt file for analysis.

a) (1.5 pt) Import the LBX_E.txt file into a temporary dataset using a **DATA step**.

What is the delimiter of LBX_E.txt?	forwardslash (/)
How many observations are in the dataset?	8712

b) (0.5 pt) Add labels to all of your variables using the data dictionary for LBX_E.htm and paste your code below.

*Part B: Labeling variables;

```
data PFOA;
set PFOA;
label
SEQN= 'ID Number'
LBXTC= 'Total Cholesterol (mg/dL)'
LBXCOT= 'Serum Cotinine (ng/mL)'
LBXSUA= 'Uric Acid (mg/dL)'
LBXPFOA= 'Perfluorooctanoic acid (ng/mL)';
run;
```

c) (1 pt) Delete all records with missing values (.) for PFOA.

How many missing values are there for each of your other variables after deleting records with missing PFOA?	There is 1 missing for total cholesterol, 3 missing for uric acid, and none for serum cotinine

d) (1 pt) Now create a binary dummy variable for hyperuricemia defined as uric acid levels greater than or equal to 7 mg/dL. Make sure that your code accounts for the possibility of missing values for uric acid measurements. Run the appropriate PROC to check your work. Paste the results into your exam.

The SAS System

The MEANS Procedure

Analysis Variable : LBXSUA Uric Acid (mg/dL)					
hyperuricemia	N Obs	N	N Miss	Minimum	Maximum
0	1794	1794	0	1.9000000	6.9000000
1	303	303	0	7.0000000	12.4000000

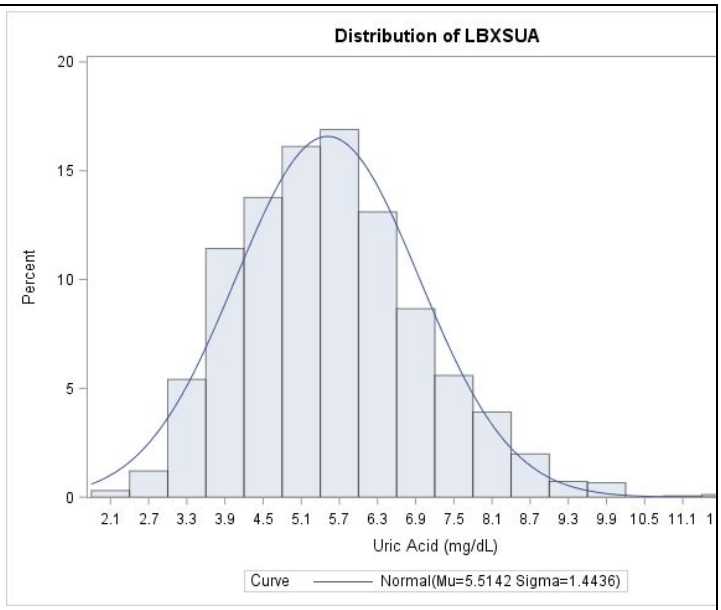
3. (2 pt) Merge your DEMO_E, ACTV_E, and LBX_E datasets together by study subject and save the resulting dataset in a permanent directory called final. <u>Retain only those records for persons that are in each of the three files.</u>

How many variables are in your merged dataset?	13
How many observations are in your merged dataset?	1665

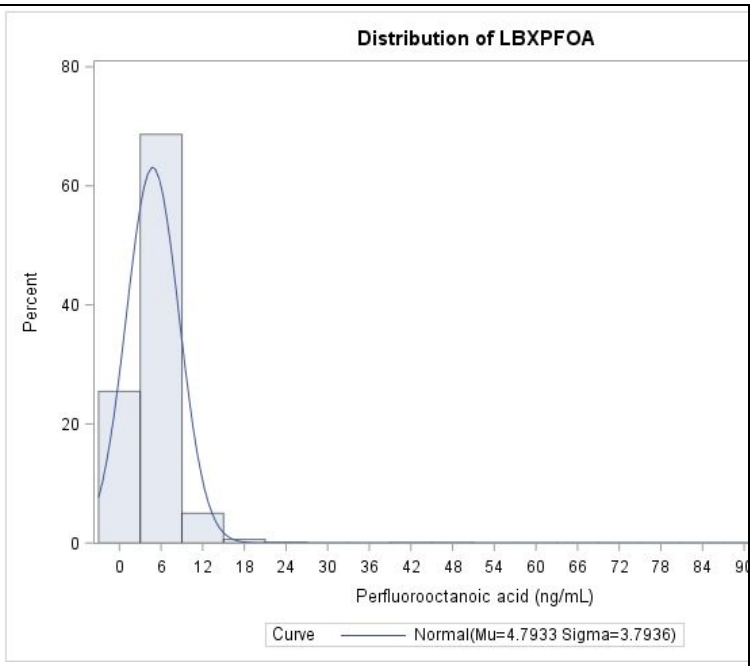
4. We will now summarize some of the key variables in your final merged dataset.

   a) (1 pt) Run a PROC to look at the distribution of uric acid and PFOA measurements and assess their normality.

URIC ACID

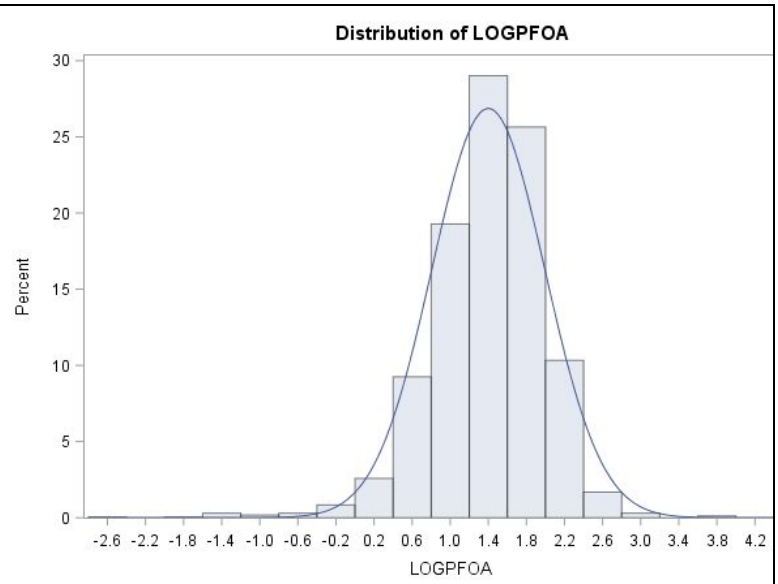
Paste your histogram.	
What was the test-statistic and p-value for one of the tests of normality? Be sure to include the name of the test you present.	A Kolmogorov-Smirnov test for uric acid resulted in a test statistic of 0.046 and a p-value of <0.010, which is less than our alpha value of 0.05. The Kolmogorov-Smirnov normality test was selected because of the relatively large sample size in this data set.
Were uric acid measurements normally distributed? Interpret your findings in a sentence.	Yes, because the shape of the histogram resembles a bell curve, the qq-plot was almost linear, and the p-value of the normality test was below an alpha level of 0.05. All of these are indicators of normal distribution.

PFOA

Paste your histogram.	
What was the test-statistic and p-value for one of the tests of normality? Be sure to include the name of the test you present.	A Kolmogorov-Smirnov test for uric acid resulted in a test statistic of 0.149 and a p-value of <0.010, which is less than our alpha value of 0.05. The Kolmogorov-Smirnov normality test was selected because of the relatively large sample size in this data set.
Were PFOA measurements normally distributed? Interpret your findings in a sentence.	No, because the shape of the histogram is pretty skewed and the qq-plot had a non-linear pattern. The p-value of the normality test was below an alpha level of 0.05, but our sample size is large enough that we cannot depend solely on that to determine normality.

b) (1 pt) Create a new variable called LOGPFOA in your final merged dataset. The values of LOGPFOA should be the natural log of your original PFOA measurement values. Assess the normality of this new variable.

LOG-Transformed PFOA

Paste your histogram.	
What was the test-statistic and p-value for one of the tests of normality? Be sure to include the name of the test you present.	A Kolmogorov-Smirnov test for uric acid resulted in a test statistic of 0.055 and a p-value of <0.010, which is less than our alpha value of 0.05. The Kolmogorov-Smirnov normality test was selected because of the relatively large sample size in this data set.
Were log-transformed PFOA measurements normally distributed? Interpret your findings in a sentence.	Yes, because the shape of the histogram resembles a bell curve, the qq-plot was relatively linear, and the p-value of the normality test was below an alpha level of 0.05. All of these are indicators of normal distribution.

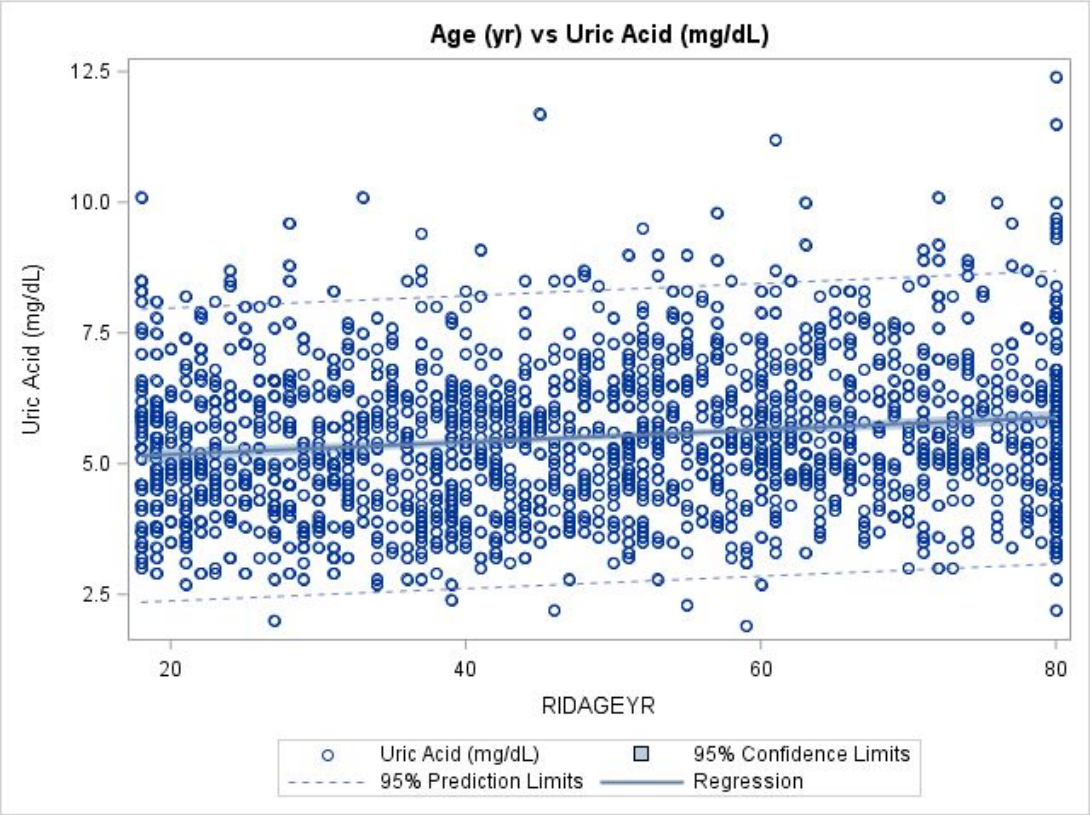
5. (2 pt) Next, you want to see if the proportion of persons with hyperuricemia varies by race, household income, low physical activity, and BMI categories. Run the appropriate statistical tests to answer these questions.

Variable	Test Used	P-value
Race	Wilcoxon Rank Sum	0.2103
Household Income	Wilcoxon Rank Sum	0.7680
Low Physical Activity	Wilcoxon Rank Sum	0.8703
BMI	Wilcoxon Rank Sum	<0.0001

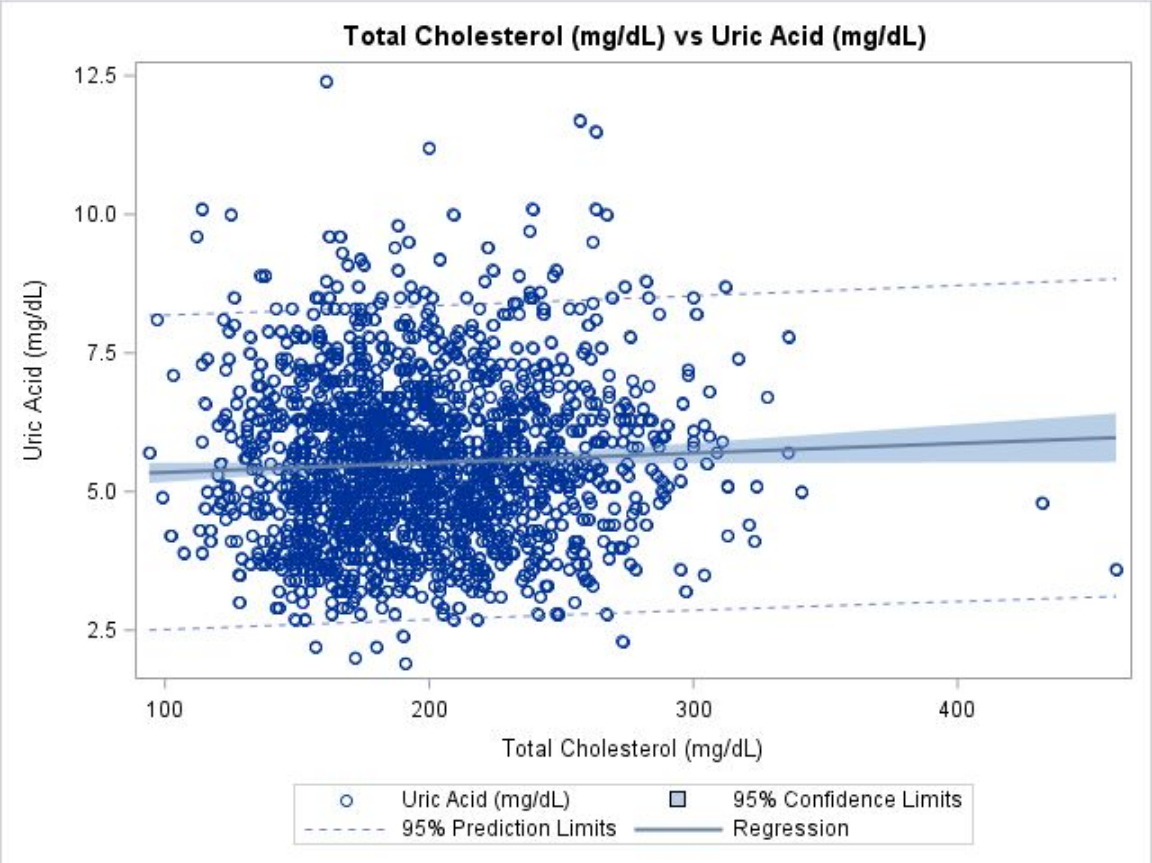
Briefly interpret your findings in 4 or fewer sentences:
I chose to run Wilcoxon Rank Sum tests for each of these variables because they are all categorical (some ordinal) and therefore are not normally distributed. As we can see from the test output, the only variable with a p-value lower than our alpha threshold of 0.05 with respect to hyperuricemia is BMI category. Therefore, we can conclude that the levels of hyperuricemia are significantly different between BMI categories, and none of the other tested variable are strong predictors of hyperuricemia.

6. Now you want to examine the relationship of age, total cholesterol, and serum cotinine on uric acid concentration as continuous variables.

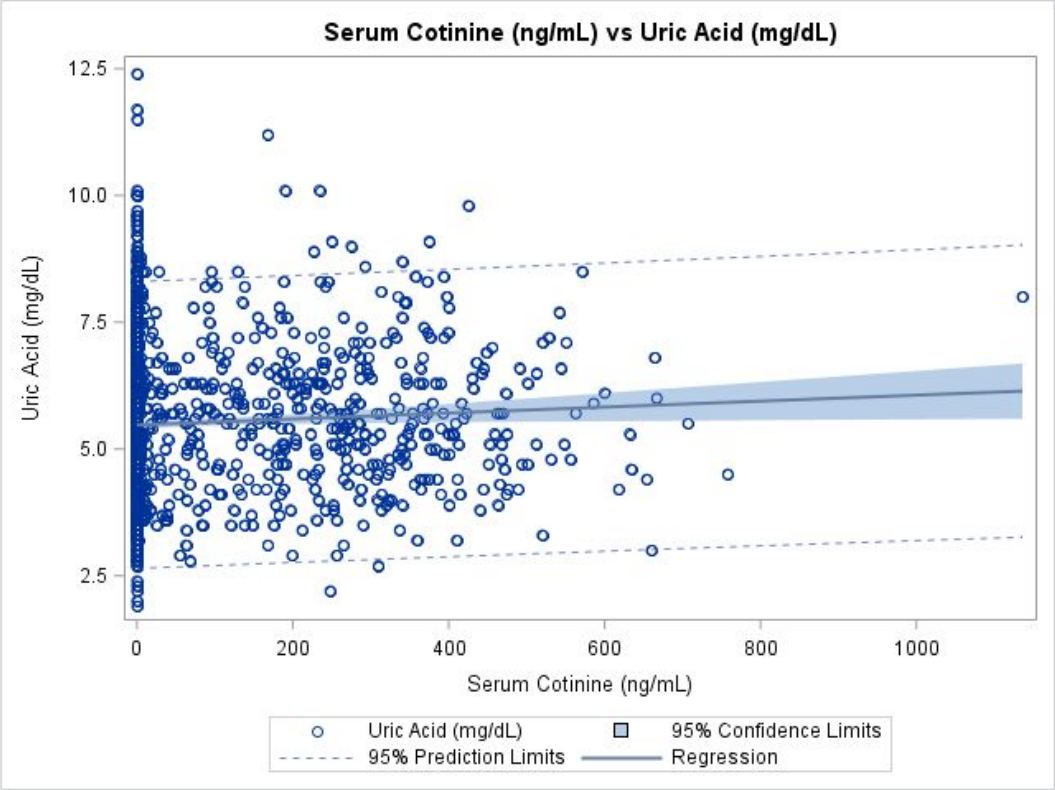
   a) (1 pt) Create a scatter plots to illustrate the crude relationship of uric acid concentration versus age, total cholesterol, and serum cotinine. Include a regression line and the appropriate titles in each plot and briefly interpret the findings below.



This scatterplot and regression line shows a very slight positive relationship between age and uric acid concentration. However, this trend is minimal and we can interpret this to mean that although is a pattern in the graph, age is not a very strong predictor of uric acid concentration.



This scatterplot and regression line shows no relationship between total cholesterol and uric acid. However, we can tell that most of our data points are clumped between 100 and 300 mg/dL total cholesterol and 2.5 and 7.5 mg/dL uric acid, with some outliers.



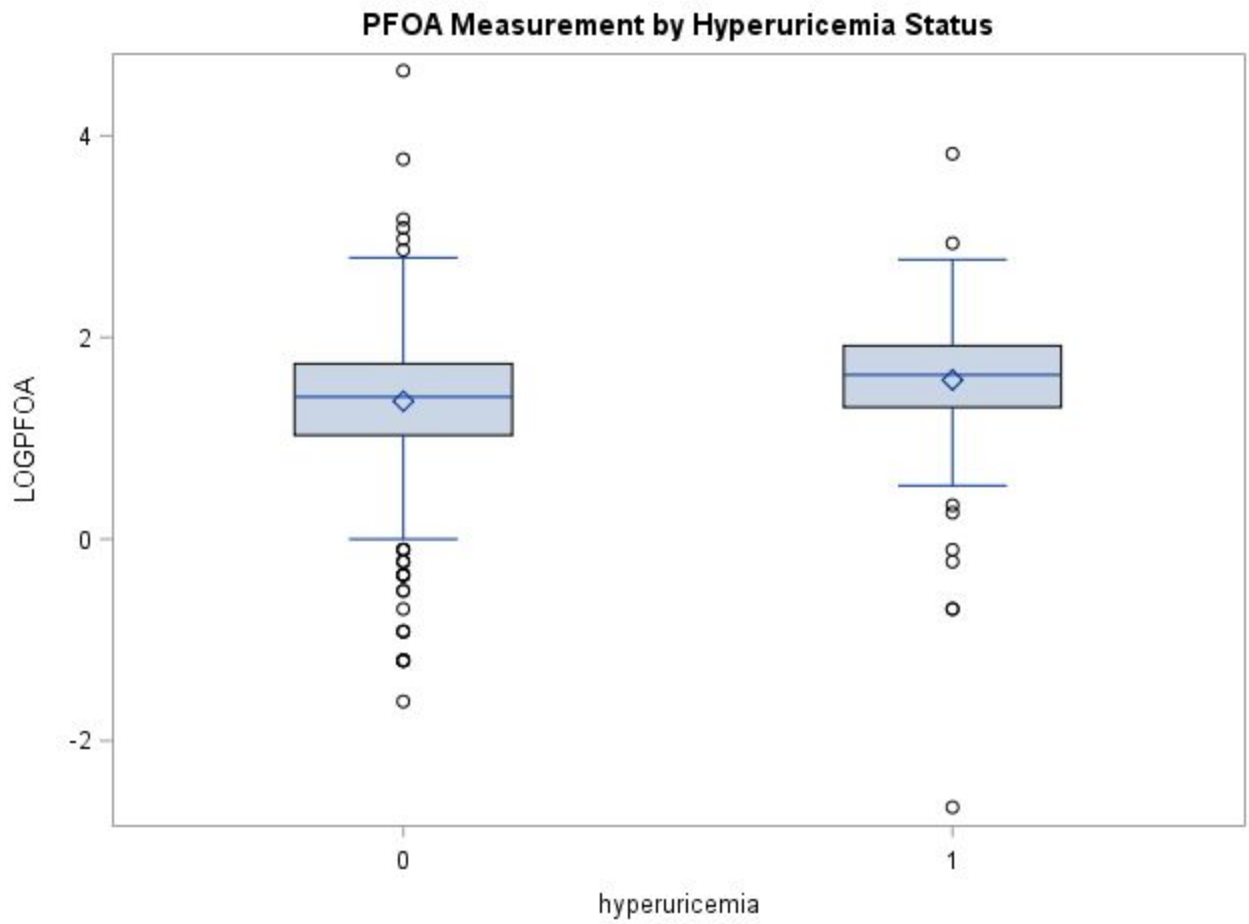
This scatterplot and regression line shows no relation between serum cotinine and uric acid levels. We have quite a few data points with zero ng/mL serum cotinine, with one very extreme outlier. Uric acid concentration is relatively normally distributed, with the majority of data points falling between 2.5 and 8.5 mg/dL.

b) (1 pt) Run the appropriate procedure to calculate the correlation coefficient between age, total cholesterol, and serum cotinine and uric acid concentration; report each of the 3 correlation coefficients and their respective *p*-values.

Variable	Correlation Coefficient (rounded to the nearest hundredth decimal point)	*P*-value
Age	0.153	<.0001
Total Cholesterol	0.051	0.0390
Serum Cotinine	0.056	0.0234
Briefly interpret your findings in 3 or fewer sentences: The correlation coefficients for all three above variables are positive and relatively close to zero, indicating weak positive linear relationships to uric acid concentration. All of the p-values are less than alpha=0.05, meaning that all of them have statistically significant relationships with uric acid; the relationship between age and uric acid is the strongest of the ones tested because it has the lowest p-value (p<0.0001).		

7. Now you want to see if people with hyperuricemia have higher PFOA exposure than those without hyperuricemia.

   a) (1.5 pt) Create a box plot displaying the PFOA measurements among persons with and without hyperuricemia. Use the PFOA measurement variable form (untransformed vs. log-transformed) that you found to be the most normally distributed in question 4. Present this graphic with the title "PFOA Measurement by Hyperuricemia Status" and paste your box plots below.



b) (2 pt) Run a t-test using the most appropriate PFOA measurement variable form to test if people with hyperuricemia have a different mean PFOA measurement than those without hyperuricemia.

Which PFOA measurement variable form did you use, and why?	I used the log-transformed PFOA variable, because the proc univariate and goodness of fit tests from question 4 showed that the log-transformed variable was more normally distributed than the original PFOA variable. T-tests are more accurate for more normally distributed data, so using the log-transformed PFOA will result in a more accurate result.
What consideration, if any, must you make about the variances?	The p-value for the equality of variances test here is 0.565, which is above our alpha=0.05. Therefore we fail to reject the null that the variances significantly differ, and can assume equal variances among the hyperuricemia categories (pooled ouput).
Report the difference between the means and the 95% CI of the differences.	The difference between means is -0.2115 with a 95% CI of (-0.2898, -0.1332).
Interpret your results in a sentence.	People who do not have hyperuricemia have, on average, 0.2115 ln ng/mL less PFOA concentration than people who do have hyperuricemia; since our p-value is less than our alpha level of 0.05 (<0.0001), we can conclude that this relationship is statistically significant.

8. Finally, you want to examine the crude and adjusted associations between PFOA exposure and uric acid concentration as a continuous variables.

a.) (2 pt) Run a linear regression to look at the crude association between uric acid concentration and PFOA exposure. Use the most appropriate PFOA measurement variable form.

Report and interpret the intercept in a sentence (round to the nearest tenth place decimal point).	The intercept of 0.85 represents the average value of PFOA we would expect in an individual with zero mg/dL uric acid.

Report and interpret the effect estimate and confidence interval (round to the nearest hundredth place decimal point).	For every one unit increase in uric acid, there will be a .099 increase in PFOA, on average.

b) (1 pt) What are the assumptions of linear regression? Based on results from previous questions and diagnostics produced for this model, describe whether or not your model meets each of these assumptions.

The assumptions are linearity of the relationship between the independent and dependent variable, independence of observations, normality of error distribution, and homoscedasticity. Also, extreme outliers would skew our results.

Linearity: The uric acid vs log-transformed PFOA output scatterplot confirms a linear relationship.

Independence: Our Pr > ChiSq statistic of 0.0879 > alpha 0.05 and our Durbin-Watson statistic (2.019) around 2 signals that our data are independent.

Normality of error distribution: The histogram and qq-plot of the residuals from the diagnostic output both indicate a normal distribution. The histogram is shaped like a bell curve and the qq-plot is close to linear.

Homoscedasticity: The residual vs predictor plot shows a random distribution around zero and don't have a specific shape that would lead us to believe there is not equally variance.

There **does** appear to be an extreme outlier, upon looking at the residual vs leverage plot and the Cook's D plot, so this model may be skewed because of that anomalous data point.

c) (2 pt) Run a linear regression to look at the association between uric acid concentration and PFOA exposure adjusted for age, gender, race, BMI category, household income, low physical activity, total cholesterol, and serum cotinine. Use the most appropriate PFOA measurement variable form and make sure that categorical variables are modeled appropriately.

Report the adjusted effect estimate for PFOA and confidence interval (round to the nearest hundredth place decimal point).	After adjusting for all the demographic variables listed above, for every one unit increase in uric acid, there will be a 0.059 increase in PFOA, on average, with a

	95% confidence interval of (0.038, 0.081).
How is this effect estimate different than the one you obtained from the unadjusted model?	It is a little bit weaker than the effect estimate from the unadjusted model (0.059 compared to 0.099), signaling that part of the effect estimate in the unadjusted model could actually be attributed to the demographic variables we just adjusted for, instead of solely due to uric acid.
What is your overall conclusion about the association between PFOA exposure and uric acid?	Overall, uric acid has a positive, relatively linear relationship with PFOA exposure. However, given the relatively low R-square value of 0.123, the relationship between these two variables is relatively weak.

THE END.
HAVE A GREAT WINTER BREAK!

CODE:

```
*EPID 640 Take Home Final Exam | Stephanie Mecham | Section 2;

*Question 1: Importing and preparing the ACTV_E.csv file for analysis;

*Part A: importing via PROC import;
proc import
datafile = 'C:\Users\smecham\Desktop\final\ACTV_E.csv'
out= actv_e
dbms=csv
replace;
getnames=yes;
datarow=2;
run;

*Part B: Creating new variables;

data bmi;
	set actv_e;
	if BMXWT =. then bmi=.;
	if BMXHT=. then bmi=.;
	else bmi= BMXWT/(BMXHT/100)**2;
	run;

data bmi;
	set bmi;
	if bmi =. then bmi_cat=.;
	else if bmi < 18.5 then bmi_cat=1;
	else if bmi >= 18.5 and bmi <25 then bmi_cat=2;
	else if bmi >= 25 and bmi <30 then bmi_cat=3;
	else if bmi >= 30 then bmi_cat=4;
	run;

data bmi;
	set bmi;
	if PAQ650 =. or PAQ650=7 or PAQ650=9 and PAQ665=. or PAQ665=7 or
PAQ665=9 then lowactivity=.;
	else if PAQ650=2 and PAQ665=2 then lowactivity=1;
	else if PAQ650=1 then lowactivity=0;
	else if PAQ665=1 then lowactivity=0;
	run;
```

```
*Checking work;

proc means data= bmi n nmiss min max;
var bmi bmi_cat lowactivity;
run;

*Delete missing variables;

data bmi_final;
        set bmi;
        if bmi=. then delete;
        if bmi_cat=. then delete;
        if lowactivity=. then delete;
run;

*Part C: Drop non-recoded variables;

data recoded_bmi;
set bmi_final;
drop BMXWT -- PAQ665;
run;


*Part D: Create and apply formats;

proc format;
        value bmicategory
        1= 'underweight'
        2= 'healthy weight'
        3= 'overweight'
        4= 'obese';

        value physical_activity
        0='normal'
        1='low';
        run;

data recoded_bmi;
set recoded_bmi;
format bmi_cat bmicategory. lowactivity physical_activity.;
run;
```

```
*Check work via PROC CONTENTS;

proc contents data=recoded_bmi;
run;

*Question 2;

*Part A: Importing file via DATA step;

data PFOA;
infile 'C:\Users\smecham\Desktop\final\LBX_E.txt'
DLM= '/'
firstobs=2
DSD
MISSOVER;
input SEQN LBXTC LBXCOT LBXSUA LBXPFOA;
run;

*Part B: Labeling variables;

data PFOA;
set PFOA;
label
SEQN= 'ID Number'
LBXTC= 'Total Cholesterol (mg/dL)'
LBXCOT= 'Serum Cotinine (ng/mL)'
LBXSUA= 'Uric Acid (mg/dL)'
LBXPFOA= 'Perfluorooctanoic acid (ng/mL)';
run;


*Part C: Delete records with missing PFOA values;

data PFOA_final;
        set PFOA;
        if LBXPFOA=. then delete;
run;

*finding remaining missing values;

proc means data= PFOA_final n nmiss min max;
var LBXTC LBXCOT LBXSUA;
```

```
run;


*Part D: Creating a Hyperuricemia Dummy Variable;

data PFOA_final;
set PFOA_final;
if LBXSUA = . then hyperuricemia=.;
else if LBXSUA < 7 then hyperuricemia=0;
else hyperuricemia=1;
run;

*Checking work;

proc means data= PFOA_final n nmiss min max;
var LBXSUA;
class hyperuricemia;
run;


*Question 3: Merging datasets into a permanent library;

libname final 'C:\Users\smecham\Desktop\final';
options fmtsearch = (final);

proc sort
data=final.DEMO_E;
by SEQN;
run;

proc sort
data=recoded_bmi;
by SEQN;
run;

proc sort
data=PFOA_final;
by SEQN;
run;

data final.combined;
merge final.DEMO_E (in=demo) recoded_bmi (in=actv) PFOA_final (in= pfoa);
```

```
by SEQN;
if demo and actv and pfoa;
run;


*Question 4: Summaries of Key Variables in Combined Dataset;

*Part A: Assessing normality;

proc univariate data=final.combined;
var LBXSUA;
histogram LBXSUA / normal;
qqplot;
run;


proc univariate data=final.combined;
var LBXPFOA;
histogram LBXPFOA / normal;
qqplot;
run;

*Part B: Log-transform PFOA variable;

data final.combined;
        set final.combined;
        LOGPFOA= log(LBXPFOA);
        run;

*Assessing normality;
proc univariate data=final.combined;
var LOGPFOA;
histogram LOGPFOA / normal;
qqplot;
run;

*Question 5: Evaluating hyperuricemia by various characteristics;

*Assessing normality of variables via Wilcoxon Rank-Sum Test;

proc npar1way data=final.combined wilcoxon;
class hyperuricemia;
```

```
var RIDRETH INDHHIN lowactivity bmi_cat;
exact wilcoxon;
run;

*Question 6: Evaluating uric acid concentration by various characteristics;


*Part A: Creating scatterplots;

proc sgplot data=final.combined;
scatter x=RIDAGEYR y=LBXSUA;
reg x=RIDAGEYR y=LBXSUA / cli clm;
title 'Age (yr) vs Uric Acid (mg/dL)';
run;

proc sgplot data=final.combined;
scatter x=LBXTC y=LBXSUA;
reg x=LBXTC y=LBXSUA / cli clm;
title 'Total Cholesterol (mg/dL) vs Uric Acid (mg/dL)';
run;

proc sgplot data=final.combined;
scatter x=LBXCOT y=LBXSUA;
reg x=LBXCOT y=LBXSUA / cli clm;
title 'Serum Cotinine (ng/mL) vs Uric Acid (mg/dL)';
run;

*Part B: Finding Pearson coefficient and p-values;

proc corr data=final.combined;
var RIDAGEYR LBXTC LBXCOT LBXSUA;
run;


*Question 7: Comparing PFOA exposures by hyperuricemia status;

*Part A: Creating boxplots;

proc sgplot data=final.combined;
vbox LOGPFOA / category=hyperuricemia;
title 'PFOA Measurement by Hyperuricemia Status';
run;
```

```
*Part B: T-test;

proc ttest data=final.combined
CI=equal
alpha=0.05;
class hyperuricemia;
var LOGPFOA;
run;

*Question 8: Examine association between PFOA and uric acid;

*Part A: Linear Regression;

proc reg data=final.combined plots(maxpoints=none);
model LOGPFOA=LBXSUA /
dw
spec;
run;
quit;

*Part C: Adjust association between PFOA and uric acid for other characteristics;

proc glm data=final.combined
plots (maxpoints=none)=(diagnostics residuals(smooth));
class RIAGENDR RIDRETH bmi_cat INDHHIN lowactivity;
model LOGPFOA=LBXSUA RIDAGEYR RIAGENDR RIDRETH bmi_cat INDHHIN
lowactivity LBXTC LBXCOT / solution clparm;
run;

*END OF CODE;
```