**P8100 – Applied Regression I**

**Homework #5 – due Thursday 3/23/2017 1pm the latest!**

**NO late homework will be accepted.**

If you upload the HW, please make sure that you upload PDF that is not a very large file.

* HW has to be TYPED, **indicating your name and UNI on top**, and STAPLED.
* Problems must be ordered, starting with Problem #1.
* Any question (a sentence ending with a question mark) must be answered with a full and meaningful sentence using the words of the problem!
* Don’t forget to **SHOW KEY STEPS OF YOUR WORK**! You get partial credit for showing work. A final answer is not enough for full credit.
* Any hypothesis test must include a null and alternative hypothesis, test statistics, degrees of freedom (if applicable), decision and interpretation.
* Write in proper and understandable English.
* You can work in groups or discuss the problems with your classmates. However, your **written answers and solutions must be strictly your own and cannot be an exact copy of your classmates answers!**

**Required readings for week 8 (3/23) discussion and quiz:**

1. Downs, J.S., J. Wisdom, B. Wansink, and G. Lowenstein, *Supplementing Menu Labeling with Calorie Recommendations to Test for Facilitation Effects.* Am J Public Health, 2013. **103**(9): p. 1604-1609.
2. Non, A.L., C.C. Gravlee, and C.J. Mulligan, *Education, genetic ancestry, and blood pressure in African Americans and Whites.* Am J Public Health, 2012. **102**(8): p. 1559-65.

**Problem 1 (similar to Problem 1 on HW5)**

You were asked to help with analysis of birth weights (BW) of 10,000 infants born in NYC during a certain period of time. The aim of the analysis is to see whether the birth weights of the infants are associated with mothers AGE at birth (continuous variable in years) and mothers smoking status (the maternal smoking status MSS contains 4 categories “Non-smoker”, “Past-smoker”, “Passive-smoker”, “Smoker”) and NYC boroughs (the BOROUGH variable contains 5 categories “Manhattan”, “Bronx”, “Brooklyn”, “Queens” and “State Island”),

1. Write down the population model that that estimates BW based on variables AGE, MSS and BOROUGH. Make sure that it is clear what each predictor means (same as HW5 Problem 1i) Please, use the same dummy variables as you did in HW5 (if you not sure how to do it, see solutions for HW5 online).
2. How many degrees of freedom will be in the numerator and denominator of the Partial F-test that tests the association between the predictor MSS and outcome BW when adjusted by AGE and BOROUGH?
3. How many degrees of freedom will be in the numerator and denominator of the Partial F-test that tests the association between the predictor BOROUGH and outcome BW when adjusted by AGE and MSS?
4. How many degrees of freedom will be in the numerator and denominator of the Partial F-test that tests the association between the predictor mothers AGE and outcome BW when adjusted by MSS and BOROUGH?
5. TRUE/FALSE: When testing the overall effect of MSS on the outcome BW when adjusted by AGE and BOROUGH, one does not have to use partial F-test.
6. TRUE/FALSE: When testing the overall effect of mother’s AGE on the outcome BW when adjusted by MSS and BOROUGH, one does not have to use partial F-test.
7. Write down the population model that assumes the effect of mother’s AGE at birth on infant weight BW is DIFFERENT for different levels of maternal smoking status. Specify the name of each predictor in the model and/or what it means.
8. How many degrees of freedom will be in the numerator and denominator of the Omnibus F-test of the model in part g).
9. How many degrees of freedom will be in the numerator and denominator of the Partial F-test for testing the significance of the interaction between AGE and MSS based on the model in part g).
10. How many regression lines are modeled in part g).

**Problems 2:** Run the following code to import data into the SAS (same as in HW5):

**data** HW6prob2;

input typ\_neighb tot\_pop prop\_child prop\_lunch prop\_change\_income crime\_rate;

datalines;

1 6.9 30.2 58.3 27.3 84.9

1 8.4 38.8 87.5 39.8 17.6

1 5.7 31.7 83.5 26 154.2

0 7.4 24.2 14.2 29.4 35.2

0 8.5 28.1 46.7 26.6 69.2

0 13.8 10.4 57.9 26.2 111

1 3.6 30 61.3 26.4 69.9

0 8.2 12.1 41 11.7 65.4

0 5 13.6 17.4 14.7 132.1

0 2.1 18.3 34.4 24.2 57.9

1 4.2 21.3 64.9 21.7 139.9

1 3.9 33.1 82 26.3 108.7

1 4.1 38.3 83.3 32.6 123.2

1 4.2 36.9 61.8 21.6 104.7

0 9.4 22.4 22.2 33.5 61.5

0 3.6 19.6 8.6 27 68.2

0 7.6 29.1 62.8 32.2 96.9

0 7.5 26.5 18.7 23.7 32

1 4.1 41.5 78.6 23.5 127

1 4.6 39 14.6 38.2 27.1

0 7.2 20.2 41.4 27.6 70.7

0 13.4 20.4 13.9 22.5 38.3

0 10.3 29.8 43.7 29.4 54

1 9.4 36 78.2 29.9 101.5

1 10.3 31.8 57.2 27.2 61.2

1 7.5 28.6 5.7 31.3 38.6

1 18.7 39.7 55.8 28.7 52.6

1 5.1 23.8 29 29.3 62.6

0 3.7 12.3 77.3 32 20.7

1 10.3 31.1 51.7 26.2 42.4

0 7.3 32.9 68.1 25.2 105.2

0 4.2 22.1 41.2 21.4 68.6

0 2.1 27.1 60 23.5 15.3

0 2.5 20.3 29.8 24.1 58.5

1 8.1 30 66.4 26 63.1

0 10.3 15.9 39.9 38.5 86.4

1 10.5 36.4 72.3 26 77.5

0 5.8 24.2 19.5 28.3 63.5

0 6.9 20.7 6.6 25.8 68.9

1 9.3 34.9 82.4 18.4 102.8

1 11.4 38.7 78.2 18.4 86.6

;

**run**;

The data were collected over random sample of Denver neighborhoods and contain variables with the following explanation:

typ\_neighb = ‘neighborhood type: 1=rural or 0=urban'

tot\_pop = ‘total population in thousands'

prop\_child = ‘% of children (under 18) in population'

prop\_lunch = '% of free school lunch participation'

prop\_change\_income = ‘% change in household income over past several years’

crime\_rate = ‘crime rate (per 1000 population)’

**Problem 2 (continues with the same data as in HW5)**

In the HW4, Problem 2, you were analyzing the effect of PROP\_CHANGE\_INCOME and PROP\_LUNCH ( continuous predictors) on crime rates for a sample of Denver neighborhoods. In this problem, we will analyze the effect of categorized PROP\_LUNCH on the variable CRIME\_RATE when accounting for PROP\_CHANGE\_INCOME. Use the data above. Use SAS to answer the following questions (if applicable). Attach the appropriate table from SAS to each question that requires SAS results.

The researchers decided to categorize the variable PROP\_LUNCH and used the following SAS code.

**data** hw6prob2;

set hw6prob2;

prop\_lunch\_cat="Lunch\_A";

if prop\_lunch>**33** then prop\_lunch\_cat="Lunch\_B";

if prop\_lunch>**66** then prop\_lunch\_cat="Lunch\_C";

**run**;

Run the researchers code.

1. (Same as HW 5, Problem 2 f) Using PROC REG, analyze the regression model for CRIME\_RATE (outcome) based on predictors PROP\_LUNCH\_CAT and PROP\_CHANGE\_INCOME (make sure that each predictor is clearly specified. Use PROP\_LUNCH\_CAT=”Lunch\_C” as a reference category), write down the estimated regression model and attach the appropriate SAS table. Don’t forget that you need to first create the dummy variables in SAS.
2. Test the significance of the effect of variable PROP\_LUNCH\_CAT on CRIME\_RATE when PROP\_CHANGE\_INCOME is also present in the model f). Don't forget to specify the name of the test, null and alternative hypothesis, test statistic, degrees of freedom, p-value, decision and conclusion in the words of the problem. Attach the relevant SAS table. Use level of significance 5%.
3. Test whether the crime rate in neighborhoods with 33% to 66% of schools participating in free school lunch program is significantly different than the crime rate in neighborhoods at least 66% of schools participating in free school lunch program. Don't forget to specify the name of the test, null and alternative hypothesis, test statistic, degrees of freedom, p-value, decision and conclusion in the words of the problem. Attach the relevant SAS table. Use level of significance 5%.
4. Test whether the crime rate in neighborhoods with less than 33% of schools participating in free school lunch program is significantly different than the crime rate in neighborhoods with at least 66% of schools participating in free school lunch program. Don't forget to specify the name of the test, null and alternative hypothesis, test statistic, degrees of freedom, p-value, decision and conclusion in the words of the problem. Attach the relevant SAS table. Use level of significance 5%.
5. Test whether the crime rate in neighborhoods with at less than 33% of schools participating in free school lunch program is significantly different than the crime rate in neighborhoods with 33% to 66% of schools participating in free school lunch program. Don't forget to specify the name of the test, null and alternative hypothesis, test statistic, degrees of freedom, p-value, decision and conclusion in the words of the problem. Attach the relevant SAS table.. Use level of significance 5%.
6. In PROC MIXED in SAS the model from part a) would be analyzed using the following code. Note that PROC MIXED is a modern procedure that utilizes CLASS statement which means that does not require user to create DUMMY variables and also computes the partial F-test directly. Run the code:

**proc** **mixed** data=hw6prob2;

class prop\_lunch\_cat;

model crime\_rate = prop\_lunch\_cat prop\_change\_income;

**run**;

Compare the results (F test statistics and DFs) for variable PROP\_LUNCH\_CAT displayed in “Type 3 Tests of Fixed Effects” with your partial F-test in part b). Are they similar? Which part of PROG REG output are you missing in output of PROC MIXED?

**Problem 3:**

Run the following code to import data into the SAS:

**data** HW6prob3;

input bldclot progindx enzyme liver age gender alcuse severealc survtime logsurv progcat $ ;

datalines;

6.7 62 81 2.59 50 0 1 0 695 6.544 med

5.1 59 66 1.7 39 0 0 0 403 5.999 med

7.4 57 83 2.16 55 0 0 0 710 6.565 low

6.5 73 41 2.01 48 0 0 0 349 5.854 med

5.8 38 72 1.42 65 1 1 0 348 5.852 low

5.7 46 63 1.91 49 1 0 1 518 6.25 low

3.7 68 81 2.57 69 1 1 0 749 6.619 med

6 67 93 2.5 58 0 1 0 1056 6.962 med

3.7 76 94 2.4 48 0 1 0 968 6.875 high

6.3 84 83 4.13 37 0 1 0 745 6.613 high

6.7 51 43 1.86 57 0 1 0 257 5.549 low

5.8 83 88 3.95 52 1 0 0 858 6.754 high

7.7 62 67 3.4 58 0 0 1 702 6.554 med

7.4 74 68 2.4 64 1 1 0 809 6.695 high

6 85 28 2.98 36 1 1 0 682 6.526 high

3.7 51 41 1.55 39 0 0 0 205 5.321 low

7.3 68 74 3.56 59 1 0 0 550 6.309 med

5.6 57 87 3.02 63 0 0 1 838 6.731 low

5.2 52 76 2.85 39 0 0 0 359 5.883 low

3.4 83 53 1.12 67 1 1 0 353 5.866 high

6.7 26 68 2.1 30 0 0 1 599 6.395 low

5.8 67 86 3.4 49 1 1 0 562 6.332 med

6.3 59 100 2.95 36 1 1 0 651 6.478 med

5.8 61 73 3.5 62 1 1 0 751 6.621 med

5.2 52 86 2.45 70 0 1 0 545 6.302 low

5.2 54 56 2.71 44 1 0 0 477 6.167 low

5.8 76 59 2.58 61 1 1 0 600 6.396 high

3.2 64 65 0.74 53 0 1 0 443 6.094 med

8.7 45 23 2.52 68 0 1 0 181 5.198 low

5 59 73 3.5 57 0 1 0 411 6.019 med

5.8 72 93 3.3 39 1 0 1 1037 6.944 med

5.4 58 70 2.64 31 1 1 0 482 6.179 med

5.3 51 99 2.6 48 0 1 0 634 6.453 low

2.6 74 86 2.05 45 0 0 0 678 6.519 high

4.3 8 119 2.85 65 1 0 0 362 5.893 low

4.8 61 76 2.45 51 1 1 0 637 6.457 med

5.4 52 88 1.81 40 1 0 0 705 6.558 low

5.2 49 72 1.84 46 0 0 0 536 6.283 low

3.6 28 99 1.3 55 0 0 1 582 6.366 low

6.5 56 77 2.85 41 0 1 0 538 6.288 low

3.4 77 93 1.48 69 0 1 0 482 6.178 high

6.5 40 84 3 54 1 1 0 611 6.416 low

4.5 73 106 3.05 47 1 1 0 960 6.867 med

5.1 67 77 2.86 66 1 0 0 581 6.365 med

3.9 82 103 4.55 50 0 1 0 1078 6.983 high

6.6 77 46 1.95 50 0 1 0 405 6.005 high

6.4 85 40 1.21 58 0 0 1 579 6.361 high

6.4 59 85 2.33 63 0 1 0 550 6.31 med

8.8 78 72 3.2 56 0 0 0 651 6.478 high

;

The variables have the following explanation:

progcat = 'prognostic category: 1=low, 2=med, 3=hi'

bldclot = 'blood clotting score'

progindx = 'prognostic index'

enzyme = 'enzyme fct test score'

liver = 'liver test score'

age = 'age (in years)'

gender = 'sex: 0=male, 1=female'

alcuse = 'moderate alcohol use dummy'

severealc = 'severe alcohol use dummy'

survtime = 'survival time (response)'

Use SAS to help answer the following. NEATLY present ***ONLY relevant SAS tables from the output***.

1. Write down the population model corresponding the SAS code:

**proc** **reg** data=HW6prob3;

model survtime = gender enzyme ;

**run**;

1. Write down the estimated model based on the results from SAS in part a). Attach appropriate SAS table.
2. **TRUE/FALSE:** The model in Problem 3a) will result in parallel regression lines (one for males and one for females).
3. Write down the population model corresponding the SAS code:

**data** hw6prob3;

set hw6prob3;

gender\_enzyme = gender\*enzyme;

**run**;

**proc** **reg** data=HW6prob3;

model survtime=gender enzyme gender\_enzyme;

**run**;

1. Write down the estimated regression model based on the results from SAS in part d). Attach appropriate SAS table.
2. **TRUE/FALSE:** The model in Problem 2d) will result in parallel regression lines (one for males and one for females).
3. Write down the estimated regression model for males based on the results from SAS in part d). Interpret the computed slope for variable ENZYME for males.
4. Write down the estimated regression model for females based on the results from SAS in part d). Interpret the computed slope for variable ENZYME for females.
5. Compare the slope for variable ENZYME in part g) and part h). Are they different? By how much they are different?
6. Test whether the difference between slopes in part g) and part h) is significant. Use level of significance 5%.
7. **TRUE/FALSE:** In part j) we tested whether the effect of interaction between ENZYME and GENDER is significant.

1. What is the estimated survival time for male with enzyme level of 50 and male with enzyme level of 60? How does it compare to your answer in part g)?
2. What is the estimated survival time for female with enzyme level of 50 and female with enzyme level of 60? How does it compare to your answer in part h)?

1. Using what we have already learned, first create dummy variable(s) for variable PROGCAT (use category ‘low’ as your reference group). Then, create an interaction terms with ENZYME for each dummy variable (refer to Lecture 7). Using the variables that you have created, run the regression model for analyzing whether the effect of ENZYME on survival time is different for different prognostic groups. Write down the estimated regression model based on the results from SAS and attach the corresponding SAS table.
2. Test whether the meaningfulness of the model analyzed in part n) using level of significance 5%. Attach appropriate SAS tables.
3. Test whether the effect of interaction between ENZYME and PROGCAT is significant using level of significance 5%. You might have to run another regression model to answer the questions. Attach appropriate SAS tables.

1. PROC MIXED is a modern procedure that utilizes CLASS statement which means that does not require user to create DUMMY variables and specific INTERACTION TERMS, and it also computes the partial F-test directly. Run the code:

**proc** **mixed** data=HW6prob3;

class progcat;

model survtime= enzyme progcat progcat\*Enzyme /s ;

**run**;

Compare the results (F test statistics and DFs) for variable ENZYME\*PROGCAT displayed in “Type 3 Tests of Fixed Effects” with your partial F-test in part p). Are they similar?

1. Based on your answer in part p) would you further investigate whether there are significant differences in effect of enzyme on survival time between high and low (reference group) category? Why?
2. Based on your answer in part p) would you keep the interaction terms in the model or should they be omitted from the final model?