**P8100 – Applied Regression I**

**Homework #7 – 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 10 (3/30) discussion and quiz:**

1. [Prabhakaran S](http://www.ncbi.nlm.nih.gov/pubmed?term=Prabhakaran%20S%5BAuthor%5D&cauthor=true&cauthor_uid=17687024), [Zarahn E](http://www.ncbi.nlm.nih.gov/pubmed?term=Zarahn%20E%5BAuthor%5D&cauthor=true&cauthor_uid=17687024), [Riley C](http://www.ncbi.nlm.nih.gov/pubmed?term=Riley%20C%5BAuthor%5D&cauthor=true&cauthor_uid=17687024), et al., *Inter-individual variability in the capacity for motor recovery after ischemic stroke.* [Neurorehabil Neural Repair.](http://www.ncbi.nlm.nih.gov/pubmed/17687024?dopt=Abstract) 2008. **22**(1): p. 64-71
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 and HW6)**

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 will properly test the effect of 3-way interaction between AGE, MSS, and BOROUGH on BW. 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 Omnibus F-test of the model in part a).
3. How many degrees of freedom will be in the numerator and denominator of the Partial F-test that tests whether the 3-way interaction is a significant part of the model in 1a)?
4. How many degrees of freedom will be in the numerator and denominator of the Partial F-test that tests whether the model including the 3-way interaction (based on part 1a) is significantly better than model that is including only the main effects of AGE, MSS, and BOROUGH?
5. **TRUE/FALSE:** If the 3-way interaction you specified in the model in part 1a) is significant and additionally the observed model has one of the included 2-way interactions that is not significant, it is OK to take the terms corresponding to the non-significant 2-way interaction out, and rerun the model without them while keeping the 3-way interaction term in the model.
6. How many regression lines are modeled in part a)?
7. How many regression lines are modeled in the model in HW6, Problem 1a)?
8. Which model estimates more lines? HW7 problem 1a) or HW6 Problem 1a)? Why?

**Problems 2:** Run the following code to import data into the SAS (same as in HW5 and HW6):

**data** HW7prob2;

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** hw7prob2;

set hw7prob2;

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 or HW 6, problem 2 a) 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. Add the following statement into your PROC REG code from part 2a).

plot r.\*prop\_change\_income;

Attach the plot that it created and describe what is on x-axis and what is on y-axis.

1. Add the following statement into your PROC REG code from part 2a).

plot r.\*p.;

Attach the plot that was created and describe what is on x-axis and what is on y-axis.

1. Add the following statement into your PROC REG code from part 2a).

plot h.\*prop\_change\_income;

Attach the plot that it created and describe what is on x-axis and what is on y-axis.

1. **TRUE/FALSE:** The closer is a datapoint to the center of all the datapoints, the smaller the leverage that particular datapoint will have.
2. Add the following statement into your PROC REG code from part 2a).

output out=HW7prob2out p=yhat r=resid h=lever student=intstud rstudent=extstud;

In SAS window, click on the Explorer tab in the left bottom corner, click on Libraries, then Work and then the new created file HW7prob2out. Describe what are the values outputted in the file in the last 5 columns.

1. Using the new file HW7prob2out, create histograms of the residuals, internally studentized residuals, and externally studentized residuals. You might need to adapt the following code:

**proc** **univariate** data=hw7prob2out;

histogram resid;

histogram XXX;

histogram XXX;

**run**;

Attach the plots that were created and describe what is on x-axis and what is on y-axis. Which of those 3 histograms can be compared to t-distribution?

**Problem 3:**

You are a researcher conducting research about chimpanzees’ cognitive processes. You have just spent a year of your life becoming familiar with the 5 new chimpanzees you purchased with your new grant money. Your first task for the first year was to make sure that all the chimpanzees behave the same way and none of them is either ‘too smart or too dumb’ of an outlier. The experiment you most recently worked on was finished and all you had to do was to analyze the data and see whether any new chimpanzee is an outlier. In your last experiment you were trying to see whether there is a relationship between the number of hours learning a task (continuous variable) and the task performance (scored using a continuous score between 0 and 15). Each chimpanzee was learning the task for a different number of hours and then each chimpanzee performed the task and its performance got scored. Yesterday, you ran a simple regression with intercept to see whether the number of learning hours can predict the chimpanzee’s task score. Until yesterday night, your life was going according to plan.

However, today everything has changed. There was a fire in your office and most of your important research data was lost. Luckily, the chimpanzees are ok. You are rummaging in tears through the ashes in your office and suddenly your face brightens. You have just found a part of a SAS output you ran last night. Granted, it is charred a bit on the edges, but all the important results are there. Now you know that you can finish your research investigating whether any chimpanzee is a significant outlier and then go on with your grant.

This is what you found:

|  |  |  |
| --- | --- | --- |
| **Observations** | **residuals** | **hii** |
| Chimpanzee Adam | 0 | 0.5439 |
| Chimpanzee Bob | 1 | 0.3421 |
| Chimpanzee Cecile | 1 | 0.2018 |
| Chimpanzee David | -4 | 0.2632 |
| Chimpanzee Eve | 2 | 0.6491 |

1. Write down the theoretical population model that you ran to investigate whether the number of learning hours can estimate the task score among your 5 chimpanzees.
2. Compute the SSE for your model.
3. Compute the MSE for your model.
4. Compute the standardized residuals corresponding to each chimpanzee.
5. Compute the internally-studentized residuals corresponding to each chimpanzee.
6. Which chimpanzee has the largest (positive or negative) internally-studentized residual and which chimpanzee has the internally-studentized residual closest to 0? Since the distribution of internally-studentized residuals is not ‘pretty’, is it easy to judge how large the outlier must be to be significant?
7. Compute the MSE(i) for each corresponding chimpanzee. There are 2 different ways to compute the MSE(i) . One way is to take the original data, omit the ith observation, rerun the regression on the rest of the data points and record the obtained MSE as MSE(i). However, since the data are lost in fire, we cannot do that. The second way is to use the direct formulas to compute MSE(i).

****

1. Compute the externally-studentized residuals corresponding to each chimpanzee.
2. Which chimpanzee has the largest (either positive or negative) externally-studentized residual and which chimpanzee has the externally-studentized residual closest to 0?
3. **Extra credit (5 points):** Test whether the largest (either positive or negative) externally-studentized residual is a significant outlier. Use level of significance 5% but make sure you use Bonferoni adjustment assuming that you are testing all residuals (one after another) and summarizing together. To determine the correct critical value based on the Bonferroni adjusted level of significance, you will have to use an online t distribution calculator. Please include the link you used in your HW solutions.