**P8100 – Applied Regression I**

**Homework #2 – due Thursday 2/2/2017 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 3 (2/2) discussion and quiz:**

1. Bures, R.M., *Childhood residential stability and health at midlife.* Am J Public Health, 2003. **93**(7): p. 1144-8.
2. Ewing, R., R.A. Schieber, and C.V. Zegeer, *Urban sprawl as a risk factor in motor vehicle occupant and pedestrian fatalities.* Am J Public Health, 2003. **93**(9): p. 1541-5.

**Problem 1:**

Use the fictional data below to answer the following questions. Assume Type I error to be bounded by 5%.

Consider the data in the following table:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| i | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Xi | 2 | 2 | 2 | 4 | 4 | 4 | 6 | 6 | 6 | 14 |
| Yi | 2 | 4 | 6 | 2 | 4 | 6 | 2 | 4 | 6 | 15 |

1. Calculate the sample average of X and Yby hand using a calculator or using Excel. Show key steps.
2. Calculate the sample correlation coefficient *r* by hand using a calculator or using Excel. Show key steps. Interpret your result in terms of the strength of the relationship.
3. By hand using a calculator or using Excel, test whether the correlation between variables X and Y is significant.

1. Based on your results in b) are you willing to conclude that the variables are linearly related? If yes, answer why, if not, list some of the reasons why not.
2. Using SAS code provided below, compute the correlation and plot the scatterplot of x and y data and present it in the homework. Based on the scatterplot, are you willing to conclude that the variables are linearly related? Notice that the SAS code below had only first 3 observations entered. You will have to make sure that all 10 observations are included in the cards statement of the SAS code.

**data** HW2Prob1;

input x y;

cards;

2 2

2 4

2 6

;

ods graphics on;

**proc** **corr** data=hw2prob1 plots=scatter(ellipse=none);

**run**;

ods graphics off;

**Problem 2:**

Run the following code to import data into the SAS:

**data** HW2Prob2;

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 answer the following. NEATLY present relevant evidence from the output.

1. Plot the relationships between the variable ENZYME and each of the following variables: BLDCLOT, SURVTIME, LIVER, and AGE. Use the following SAS code. Notice that you will have to change the code for the rest of the scatterplots. Cut and paste the scatterplot into your HW.

**proc** **gplot** data=HW2Prob2;

plot enzyme\*bldclot;

title 'Scatterplot of enzyme by bldclot';

**run**;

1. Are there any variable among BLDCLOT, SURVTIME, LIVER, and AGE that appears to have non-linear relationships (quadratic, for instance) with ENZYME?
2. Compute the correlation coefficients between ENZYME and each appropriate (does not appear to have non-linear relationship) variable among BLDCLOT, SURVTIME, LIVER, and AGE. Use the SAS code below

**proc** **corr** data=HW2Prob2;

var enzyme bldclot survtime liver age;

**run**;

1. Compare the computed correlations with the plots you created in 2a. Using the plots and the correlations, rank the variables from those with the strongest (least spread) to the weakest (most spread) linear relationship with ~~SURVTIME~~ ENZYME.
2. Test whether the computed correlation between ENZYME with each of the variables BLDCLOT, SURVTIME, LIVER, and AGE is significantly different than 0. Don't forget to state the null and alternative hypothesis, test statistics, df (if applicable), critical value and conclusion in the sentence using words of the problem.
3. What percent of the variance in ENZYME can be explained by each of the variables BLDCLOT, SURVTIME, LIVER, and AGE separately?
4. Create a plot between the variable SURVTIME and AGE. How would you describe the nature of this relationship?
5. Compute the correlation coefficient between SURVTIME and AGE. The standard error of the sample correlation can be computed as:



Compute the 95% confidence interval for correlation between ~~ENZYME and PROGINDX~~ SURVTIME and AGE. Don’t forget to use proper t-value with appropriate degrees of freedom.

**Problem 3:**

For this problem, think of 2 continuous variables/questions that you think might be related (Such as: “How many cups of coffee did your have yesterday?” And “How many hours/minutes did you sleep last night?”. Your questions should be different) and ask them at least 30 of your friends and collegues. Obviously, your sample will be a convenient sample, not a random sample.

Use SAS to analyze whether there is a significant linear association between the responses to the first continuous question (outcome variable) and the second continuous question (predictor variable). Also, test whether the association is significant. For example, whether there is a significant linear relationship between the number of coffee cups drank yesterday and sleep last night .

In order to import your data into the SAS, you might want to use the following syntax (this is only example of 2 subjects):

**data** HW2Prob3;

input coffee\_cups sleep\_h;

datalines;

3 5.5

0 8

;

Or you can use the Import Data option in the menu of SAS to import your excel file.

Based on what you learned in the class write three paragraphs about your small study. Your write up should resemble a research study write up. The first paragraph should explain what are you talking about and how were your two question operationalized and description of your sample. The second paragraph should describe the data that are provided and their statistical (means and SD for each question) and graphical summaries (scatterplot). The third paragraph should include the results of your analysis (correlation and significance) and interpretation of the results.

**Use the following format:**

Paragraph 1 [Study Description]:

Sentences 1 and 2: Describe what questions are being investigated and how exactly are operationalized.

Sentence 3: Describe your sample (proportions male and female, anything else you know about your sample?).

Sentence 4: Are there any potential issues you are worried about because of your sampling strategy?

Paragraph 2 [Data description]:

Sentence 1: What are the overall and group-specific means and standard deviations for each of your question? (You can use PROC MEANS in SAS or other sofware)

**proc** **means** data= HW2Prob3;

var coffee\_cups sleep\_h;

**run**;

Sentence 2-3: Using SAS, plot scatterplot of your data. Attach the graph at the end of your write up as Figure 1 and describe it in this paragraph.

Paragraph 3 [Hypothesis test and interpretation]:

Sentence 1: What is the estimated correlation between your two questions?

Sentence 2: Is the estimated correlation significant? Write out the conclusion of your hypothesis test but also include the test statistic, df, and p-value.

Sentence 3: Discuss possible validity of your hypothesis test results.

Sentence 4: Discuss whether your results can be extrapolated to some larger populations.

Sentence 5: Discuss whether correlation is a valid measure of association between your two questions (Linearity? Outliers?).

Each paragraph should have no more than 5-6 sentences (we will count them). Be brief but informative. Do not deviate from the structure outlined above--we will subtract points for including too many sentences with unnecessary or irrelevant information.

**You are expected to have different paragraphs (write ups) than your classmates.**