Multivariate Statistics (**MARS6300**) - Homework 1 **Name:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Distributed:** 01/18/2018 **Due:** Tuesday 02/06/2018

**Instructions:** Copy and paste your answers below and turn in a word file and two spreadsheet files (e.g., Excel) before class via email to khyrenba@gmail.com using “MARS6300 hw#1” as the message subject. Label all files with your name (e.g., MARS6300\_hw1\_hyrenbach). You will be fined 0.5 points for failure to do so.

You are free to use any reference materials of your choice. While you are encouraged to work together, make sure you turn your own assignment. This homework is worth 5 points.

**Note:** You can perform all of these calculations with Excel or with SPSS or with R. If you use Excel, make sure you leave the formulas showing all of your calculations in the sheets, and explain your reasoning, to get partial credit.

**The objectives of this homework are:**

A) To review the following statistical principles:

* Calculation of full and partial correlation
* Calculation of partial regressions – similarity with partial correlations

 - Calculation of critical values and p values checking for significance

B) To use partial correlations / regressions to test a causal model of large scale drivers

of local upwelling dynamics in the California Current System.

C) To practice the documentation of the flow of data analysis.

D) To practice entering and evaluating data in PC-ORD.

To complete this homework, you will need:

* Instruction file: “MARS6300\_hw1.doc” (open with word file) – turn in
* Correlations data file: “stocks.xls” (open with excel) – turn in
* Causal model data file: “upwell.xls” (open with excel) – turn in

**Questions:**

* I) Use data in file “stocks.xls”, showing time series of three indices: NASDAQ, DOW and FTSE. The raw data are on sheet “raw”. Use the other sheets to make the required calculations and paste tables / figures of results into this word file. Rename the excel file with your results and turn in (e.g., stocks\_hyrenbach.xls)
1. Use the sheet “correlation\_calculations” to calculate the correlation coefficient for all three pair-wise combinations of two variables using the formula of the Pearson correlation coefficient discussed in class (Hint: -1 < r < +1). I want you to do it using the spreadsheet, and check it against the “correl” calculation from excel.

Copy and paste the results into the completed table 1, below:

|  |  |  |
| --- | --- | --- |
|  |  | Pairwise Correlation Coefficients ( r ) |
|  |  |  |  |  |
|  |  | ***NASDAQ*** | **DOW** | **FTSE** |
|  | **NASDAQ** | 1 | **0.924** | - |
| Significance (p value) | **DOW** |  - | 1 | - |
|  | **FTSE** |  - | -  | 1 |
|  |  |  |  |  |

1. Calculate the p values for these correlations using the table provided in the sheet “correlation\_results” and insert the results in the table above.

Answer these questions:

1. How many degrees of freedom are there?
2. What is the critical r value for alpha = 0.05?
3. What is the result: are these correlations statistically significant?

(Please explain Why / Why not ?)

(C) Calculate the partial pair-wise correlations using the sheet “partials”. Hint: I worked out the first example for you. Fill-in and paste Table 2, below. Answer these questions:

1. Which two variables are most strongly correlated, once you “remove” the

 effect of the third variable?

1. Interpret what influence the “control” (puppet-master) variable had on the other two variables, given the magnitude of the zero and the first order correlations.
2. Which two variables are most weakly correlated, once you “remove” the

effect of the third variable?

1. Interpret what influence the “control” (puppet-master) variable had on the other two variables, given the magnitude of the zero and the first order correlations.
* II) Use file “upwell.xls”, showing four time series of oceanographic indices spanning a 10-year period: SOI, NOI, Upwelling\_36, Upwelling\_39. The raw data are on sheet “raw”. Use the other sheets to make the required calculations and paste tables / figures of results into this word file. Rename the excel file with your results and turn in (e.g., upwell\_hyrenbach.xls)
1. Regress each variable against time (decimal years) to determine if there is a long-term significant trend. For each regression, report the following:

R-squared, Significance, Sum of the residuals. Discuss the result of each regression (is there a significant trend, how much of the variance is explained?)

For each regression, calculate the “predicted values” by plugging the calculated

(mean) coefficients into the equation of a straight line. Notice where the residuals are coming from.

NOTE: You can do this using the Excel “Data analysis” Add-in or a different

statistics software program. However, make sure you can paste and

report the residuals from each regression.

1. Use linear regressions to partial out the effect of one variable on the others, as shown in lecture, to determine the strength of the correlation between upwelling at 36 and upwelling at 39, given the influence of SOI and NOI on these two local upwelling indices.

Calculate the correlation between Upwelling at 36 and Upwelling at 39. Report a Pearson correlation coefficient and a p value (Note: you can use the “correl” command in the Excel “Data Analysis” Add-In, or another program of your choice). Hint: use the same table of critical Pearson values from question 1.

1.Using linear regressions, determine the correlation between Upwelling at 36 and Upwelling at 39, once you remove the effect of SOI on both variables. Paste the residuals into the sheet “residuals” and calculate their sum. Report the R squared and the p value. Finally, paste two plots showing the best-fit regression line and the residuals. To check that the regression and the correlation give you the same result, calculate the Pearson correlation between the residuals of Upwelling at 36 and the residuals of Upwelling at 39. Is the partialed (1-order) correlation stronger or weaker than the simple (0-order) correlation you calculated before?

2.Using linear regressions, determine the correlation between Upwelling at 36 and Upwelling at 39, once you remove the effect of NOI on both variables. Paste the residuals into the sheet “residuals” and calculate their sum. Report the R squared and the p value. Finally, paste two plots showing the best-fit regression line and the residuals. To check that the regression and the correlation give you the same result, calculate the Pearson correlation between the residuals of Upwelling at 36 and the residuals of Upwelling at 39. Is the partialed (1-order) correlation stronger or weaker than the simple (0-order) correlation you calculated before?

III) Data Interpretation: Draw (using paint or powerpoint, or pen and paper and take a digital photo) a conceptual diagram (with bubbles and arrows) of the cause and effect relationship between the two upwelling indices (the explanatory variables) and the two driver variables (NOI, SOI).

 1.Show the following: partial correlation coefficient between both upwelling indices, the 0-order correlation between the two driver variables (NOI, SOI), and the correlation coefficients between the explanatory and the driver variables, calculated using the partial regressions. (Hint: your diagram should have 4 bubbles, and 6 arrows. Make sure you correctly label one-headed (cause-effect) and two-headed (covariation) arrows).

2.Finally, use the diagram to interpret your results. Do you agree with the Schwing et al. (2002) assertion that the NOI is a better descriptor of upwelling dynamics in the California Current than the SOI? Explain Why / Why Not?

IV) Data Documentation: Create a flow chart (using paint or powerpoint, or using pen and paper and taking a digital photo or scanning) where you illustrate all of the steps you took to calculate the correlation values from each of the six arrows in the plot from figure 3. Make sure you include the 0-order / 1-order correlations.

V) Data Manipulation: For this part of the homework, you will use PC-ORD.

Load the data from the file “upwell.xls”, located in the PC-ORD sheet. Note: you can use the “import” command –in the “File” menu to open a single sheet from excel. Note: This is the “main matrix”; there is no “second matrix”.

Explore the data using the following tools. Copy and paste screen shots (use print screen), submit requested “result” files, and answer specific questions.

1. **Command: ADVISOR > Show Current Profile**

Run this command (using Euclidean distance) to check the profile of your data.

Focus on “skewness”. Define what it means?

Copy and paste the skewness results below:

 Columns

 5 vars

 Skewness

|  |  |
| --- | --- |
|  Average  |  |
|  Maximum  |  |
|  Minimum  |  |

Copy and paste the outliers below:

 **Potential Outliers:**

 **SD Item**

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

 What does SD mean? How can we can use SD to measure the outliers?

Use the Help button to read about what is the “Relativized Euclidean Distance”. Copy and paste explanation below:

1. **Command: SUMMARY > Row and Column Summary**

Run this command to summarize the data by columns (vars) only.

Copy and paste output below:

**C) Command: SUMMARY> Outlier Analysis**

Run this command to identify outliers. Copy and paste output below:

What is the largest outlier in this sample? What sample (month-year) is it?

What does it mean to be an outlier, according to PC-ORD?

 (Hint – use the program’s Help)

**D) Command: GRAPH > Scatterplot**

Run this command to create scatterplots of pairs of variables.

Create a plot of Upwell\_36 versus Upwell\_39, reformat the figure (add tick marks and labels, change the axes labels, and add a figure title). Label the largest outlier with the sample number – What sample (month-year) is it?. Copy and paste figure below:

**E) Command: GRAPH > Scatterplot Matrix**

Run this command to create scatterplots of all pairs of variables. Copy and paste figure below:

**Extra-credit:** What does “jittering” do? Briefly explain?