**Revised QR Lesson (Correlation-study of relation between two quantities)**

**Paragraph describing the lesson and what the students and you will be doing.**

In this assignment students will conduct correlation analysis to study the relationship between two quantities. The concepts and methodology will be discussed in the class and will be followed up with a homework assignment. The work will involve studying the data given in the handout, drawing scatter plot to describe the relationship between the two variables, computing Karl Pearson correlation coefficient, and determine if the two variables are significantly linearly related by comparing their correlation value with the critical value obtained from the critical value table.

**Thinking and Other skills learning goal:**

Students should be able to discuss if the correlation implies causation in the given case. They should also be able to discuss the implications of having a significant correlation between two variables for further analysis such as, being able to predict the output value for a given value of the input. They should also be able to discuss the implication of case when correlation is not significant, such as, the prediction will not depend on the value of input variable and, regular average can be used for best prediction of the output value.

**Attitudes, values, dispositions and habits of mind learning goal:**

Students should be able to demonstrate the above mentioned skills in data analysis project. They should be able to use the above mentioned logic when conducting data analysis and writing reports for their project involving prediction using regression equation.

**The lesson starts here**

* **Homework reading assignment prior to the class session (one week time):**

Read about Climate Change in the following web link and the attached document (climate change-reading part 2).

<http://www.whrc.org/resources/primer_home.html>). Please take notes as you read. Consider the following questions as you prepare your notes.

* What are some notable points in the reading assignment?
* What is global warming?
* What are some of causes of it?
* Does global warming means warming up everywhere?
* **Class discussion /exercise with instructor: (time: 2 hours)**

The class session will involve preparing background and working through exercise problem to cover the topic of correlation analysis for bivariate data. The discussion consists of questions answer sessions based on reading assignment and general questions involving connection between two quantities. Some discussions questions are:

* What were some notable points in the reading assignment?
* What is global warming?
* What are some of causes of it?
* Does global warming means warming up everywhere?
* Can education make you happier?
* Do higher unemployment rates account for greater number of violent crime cases?
* What is the connection between educational attainment and unemployment?
* How do we study such relationships between quantities?
* If we graph on quantity on the x-axis and another quantity on the y-axis, what can we expect to see in the graph? (Note: such graph is called Scatter Plot)

Through the following exercise students will learn that a statistical method that allows one to investigate connections and determine the degree of association between two or more variables is called correlation analyses. It enables one to exploit these relations to make valid predictions using regression equation. To investigate the association between two variables we first obtain scatter plot to observe if any pattern is present. (Data in table 1 represent no association or linear relationships whereas data in table 2 represent quadratic relationships). Then we compute Pearson’s correlation coefficient *r* to quantify the linear association (if any). (The value of r will be significant when linear association exist and will not be significant for quadratic association exist). We compare this calculated value of r with the critical value of r obtain from the table to determine if the linear association is statistically significant. We will discuss the implications of association being present and association not present.

**Exercise**:

Table 1 shows the data for 10 students chosen at random from a college. The height (in centimeters), weight (in centimeters), waist (in centimeters), score in math exam (in percent), number of days spend on training for running, and time (in seconds) to run 100 meters after training are recorded. (Students will have access to electronic file of the data)

Table 2 shows data on median age and annual median income in 2006 in the U.S..

1. Use SPSS to draw scatter plots for the pairs of quantities listed below.
2. Study each graph and state if there seem to be a relationship between the two quantities.
3. If so, does the relationship appear to be linear?
4. If so, would you describe the linear relationship as positive or negative? For each graph state what pattern (increasing, decreasing, none) do you see.
   * + 1. *Weight versus height*
       2. *Math score versus height*
       3. *Waist versus height*
       4. *100 meter run time versus height*
       5. *Number of days in training versus 100 meter run time after the training*
       6. *Age versus income*

Table 1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Subjects | height (CM) | weight (KG) | Math score(%) | waist (CM) | days in training for run | run time after training (Secs) |
| 1 | 180 | 87 | 56 | 101 | 100 | 10.2 |
| 2 | 176 | 55 | 29 | 71 | 90 | 11.7 |
| 3 | 144 | 52 | 45 | 62 | 20 | 13.9 |
| 4 | 195 | 94 | 93 | 113 | 50 | 14.3 |
| 5 | 159 | 87 | 67 | 88 | 10 | 17 |
| 6 | 185 | 79 | 38 | 87 | 20 | 16.5 |
| 7 | 166 | 59 | 85 | 71 | 10 | 18.4 |
| 8 | 173 | 64 | 77 | 83 | 60 | 11.3 |
| 9 | 149 | 45 | 56 | 58 | 60 | 12.7 |
| 10 | 168 | 77 | 71 | 85 | 60 | 13.1 |

Table 2

|  |  |
| --- | --- |
| Median Age | Median annual income ($) in 2006 |
| 20 | 10,964 |
| 30 | 32,131 |
| 40 | 42,637 |
| 50 | 45,693 |
| 60 | 41,477 |
| 70 | 23,500 |

1. Using SPSS command, compute Pearson’s correlation coefficient value in each case (I to VI) mentioned above. In each case, discuss if the correlation is significant. What are the implications of this.

We will discuss and learn that relationship between two variables does not always imply causal relationship. Following examples will be helpful in discussing this concept. We will also discuss the above cases to determine if they are causal relationships.

Example 1: The Ice cream sales and the number of shark attacks on swimmers are positively correlated. Can we say that there is a causal relationship between Ice cream sales and the number of shark attacks?

Example 2: The more firefighters fighting a fire, the more damage there is going to be. Can we say that there is a causal relationship between number of firefighters fighting a fire and amount of damage?

* Following exercise will help to synthesize the above discussed concepts and demonstrate the correlation analysis process step-by-step.

1. **Class exercise:** Following data on y = global average Temp (o F) and x = CO2 concentration (in parts per million (ppm)) will be used to study the above mentioned concepts and methodology.

|  |  |  |
| --- | --- | --- |
| Year | X = CO2 concentration in parts per million (ppm) | Y = Global average Temp (o F) |
| 1960 | 315 | 57.2 |
| 1965 | 320 | 57.1 |
| 1970 | 324 | 56.9 |
| 1975 | 334 | 57.0 |
| 1980 | 340 | 57.3 |
| 1985 | 348 | 57.2 |
| 1990 | 354 | 57.7 |
| 1995 | 361 | 57.7 |
| 2000 | 370 | 57.7 |
| 2005 | 375 | 58.0 |

* + 1. Study carefully the values of X and Y given in table. What connection do you see between the values of X and Y?
    2. Draw a scatter plot of temperature values versus CO2 concentration. Does the scatter plot support your conclusions from (1)?
    3. Does there appear to be a relationship between CO2 concentration and global average temperature? Why or why not?
    4. Discussion question: Can we say that CO2 concentration causes changes in the global average temperature? What are the rational for your answer? (reference/link for web resources and/or articles to support the arguments will be provided.)
    5. Does the relationship appear to be linear?
    6. Would you describe the relationship as positive or negative?
    7. Provide an estimate of the Pearson’s correlation coefficient value between temperature and CO2 concentration without performing the actual computation.
    8. Now, compute the Pearson’s correlation coefficient between temperature and CO2 concentration. Is this value close to what you guessed in (6)? Why or why not?
    9. Is the correlation statistically significant?
    10. Discussion question: How would we predict the global average temperature value when the CO2 concentration is 380 ppm? (will study the regression equation in next chapter)
    11. Discussion question: If the correlation was not significant, how would we predict the global average temperature value when the CO2 concentration is 380 ppm?
* The following homework assignment (project) will be assigned after regression chapter is discussed in the class. The concepts of unusual data values/outliers are also discussed in earlier chapters. Students are expected to retain the understanding from the correlation and earlier chapters and demonstrate that (attitude/habits of mind) in answering questions asked in the following project.

**Homework Assignment (Project)**

***Direction: Please type your responses using MSWord, font size 12, double spacing. Copy and paste any relevant graph, SPSS output tables, etc on your WORD document. Print a copy and submit for grading.***

* + - * 1. Since the 1950s, both the atmospheric CO2 level and crime levels have increased sharply. Hence, atmospheric CO2 causes crime. Is this causal relationships conclusion flawed? Explain your reasoning.
        2. The number of cavities an elementary school child has and the child's vocabulary size has a strong positive correlation. Is this causal relationships conclusion flawed? Explain your reasoning.
        3. Is there a relation between educational attainment and crime rate? In the following table, the first variable (labeled X) is the Educational Attainment in 2009, determined as a percentage of persons aged 25 years and older living in a state and holding at least a Bachelors degree. The second variable (Y) is the Deadly Violence Rate in 2010, obtained as a combined suicide and homicide rate and measured for each state as a number of fatal incidents per 100000 persons. The values of X and Y are given for the 50 states of U.S.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **State** | **X** | **Y** | **State** | **X** | **Y** |
| Alabama | 22 | 19.9 | Montana | 27.4 | 28.1 |
| Alaska | 26.6 | 27.4 | Nebraska | 27.4 | 13.6 |
| Arizona | 25.6 | 23.5 | Nevada | 21.8 | 26.1 |
| Arkansas | 18.9 | 19.9 | New Hampshire | 32 | 15.9 |
| California | 29.9 | 15.3 | New Jersey | 34.5 | 12.4 |
| Colorado | 35.9 | 19.8 | New Mexico | 25.3 | 26.9 |
| Connecticut | 35.6 | 13.6 | New York. | 32.4 | 12.5 |
| Delaware | 28.7 | 17.5 | North Carolina | 26.5 | 17.3 |
| Florida | 25.3 | 20 | North Dakota | 25.8 | 17.3 |
| Georgia | 27.5 | 17.4 | Ohio | 24.1 | 16.7 |
| Hawaii | 29.6 | 17 | Oklahoma | 22.7 | 21.7 |
| Idaho | 23.9 | 19.9 | Oregon | 29.2 | 20.4 |
| Illinois | 30.6 | 14.7 | Pennsylvania | 26.4 | 17.5 |
| Indiana | 22.5 | 17.4 | Rhode Island | 30.5 | 15.1 |
| Iowa | 25.1 | 13.4 | South Carolina | 24.3 | 19.5 |
| Kansas | 29.5 | 17.5 | South Dakota | 25.1 | 20 |
| Kentucky | 21 | 18.8 | Tennessee | 23 | 20.5 |
| Louisiana | 21.4 | 23.3 | Texas | 25.5 | 16.4 |
| Maine | 26.9 | 15.8 | Utah | 28.5 | 19 |
| Maryland | 35.7 | 16.1 | Vermont | 33.1 | 18 |
| Massachusetts | 38.2 | 12.4 | Virginia | 34 | 16.7 |
| Michigan | 24.6 | 18.7 | Washington | 31 | 16.5 |
| Minnesota | 31.5 | 13.2 | West Virginia | 17.3 | 18.2 |
| Mississippi | 19.6 | 20 | Wisconsin | 25.7 | 16.6 |
| Missouri | 25.2 | 21.3 | Wyoming | 23.8 | 24.6 |

* + - 1. Study carefully the values of X and Y given in table. What connection do you see between the values of X and Y?
      2. Draw a scatter plot of Y versus X. Does the scatter plot support your conclusions from (1)?
      3. Does there appear to be a relationship between educational attainment and deadly violence rate? Why or why not?
      4. Can we say that there is a causal relationship between educational attainment and deadly violence rate? Provide rational for your answer. (You may use web resources or articles to support your arguments. Please provide references for the resource you use.)
      5. If we imagine a linear relationship between X and Y, would you describe the relationship as positive or negative? Why?
      6. Provide an estimate of the Pearson’s correlation coefficient between X and Y without performing the actual computation.
      7. Compute the Pearson’s correlation coefficient between X and Y.
      8. Is the correlation statistically significant? What can we conclude about the relationship between educational attainment and deadly violence rate?
    1. Does the value of correlation suggest a weak or strong linear relationship between educational attainment and deadly violence rate? How does the data or graph support your answer?
    2. Does there seem to be any outlier in the data? (Use 2-standard deviation limit of the y-values to identify outliers.)
    3. Find the correlation after removing the pair of data for which y-values were outliers. How does your conclusions about relationship between X and Y and strength of linear relationship change?
    4. Discuss the implications of having a significant value of Pearson’s correlation on the prediction of crime rate for a given value of educational attainment.
    5. If the correlation value was not significant, how would we predict the crime rate for a given value of educational attainment? Explain why?