**DAILY ASSESSMENT FORMAT**

|  |  |  |  |
| --- | --- | --- | --- |
| **Date:** | **21/07/2020** | **Name:** | **PADMINI M** |
| **Course:** | **Basic statistics** | **USN:** | **4AL17EC066** |
| **Topic:** | **Week 2** | **Semester & Section:** | **6th Bsec** |
| **GitHub Repository:** | **Padmini** |  |  |

|  |
| --- |
| **FORENOON SESSION DETAILS (9.00am to 1.00pm)** |
| **Image of session** |
| **Report – Report can be typed or hand written for up to two pages.**  **Correlation**  **The first two videos in this module discuss the concept of correlation. In the first video, we'll talk about how we can display the correlation between two variables using tables and graphs. First we'll look at categorical variables and discuss contingency tables. In a next step we look at how we can best display the relationship between two quantitative variables. Here we'll introduce the scatterplot.**  **In the second video we'll discuss the Pearson's r - one of the most frequently used measures of correlation. It is an appropriate measure if the variables under analysis are measured on a quantitative level and if they are linearly related to each other. The Pearson's r expresses the direction and strength of the correlation**  **Regression**  **Regression analysis is one of the most frequently employed statistical methods. In the next three videos we'll discuss the basics of regression analysis. In the first video we'll explain how we can find the regression line (the line that best represents the linear correlation between two quantitative variables in a scatterplot). You'll learn that the best fitting line is the line for which the sum of the squared residuals (vertical distances of the cases in your scatterplot to the line) is the smallest. We therefore talk about ordinary least squares (OLS) regression.**  **In the next video foucses how we can describe what the regression line looks like. This is very useful because it can help us make predictions about our dependent variable. We can make these predictions by means of the regression equation of which important ingredients are the regression coefficient and the regression slope. In the final part of this video we'll show you how you can also find the regression line by means of two rather simple formulas.**  **The third video in this section focuses on the question how we can assess how well a regression line fits the data under analysis. Here we'll introduce the so-called r-squared. It tells you how much better a regression line predicts your dependent variable than the mean of that variable, and it shows you how much of the variance in your dependent variable is explained by your independent variable.**  **Although many people like eating chocolate, most people are slightly cautious with their chocolate consumption, because they know that there is a strong correlation between the amount of chocolate you eat and your body weight. However, a recent study shows that it might actually be a good idea to eat a lot of chocolate. This scatterplot shows that a country’s annual chocolate consumption per person (so, how much chocolate someone eats in a year) is positively related to the number of Nobel Prize winners per 10 million people in a country. Notice that in this scatterplot chocolate consumption is displayed as the independent variable and the number of Nobel prize winners as the dependent variable. The units of analysis in this scatterplot are countries. You can see that the correlation is pretty high. In fact, the Pearson’s r correlation coefficient here is 0.93.**  **Let me give you an example based on a random line. Say… this one. You measure the vertical distance between Japan and the line, the distance between Spain and the line, and so on, until you know the distance to the line of every case in your study. Every distance is called a residual. You end up with positive residuals (the distances from cases above the line to the line, displayed in blue) and negative residuals (distances from cases below the line to the line, displayed in red). You measure these residuals for every possible line through the scatterplot. So, not only for this line, but also for this line, this line and this line. And for every other possible line through the scatterplot. Eventually, you choose the line for which the sum of the squared residuals is the smallest. That’s this one. Why the squared residuals? Because positive and negative residuals cancel each other out: the sum of the length of the positive residuals (the blue lines) is exactly as big as the sum of the length of the negative residuals (the red lines). The best fitting line is called the regression line, and the name of the method of analysis is called ordinary least squares regression, which refers to the way we have found the line.**  **When we do a regression analysis, we assume that the independent variable X explains the dependent variable Y. Building on that assumption, we can make a scatterplot and let the computer draw the line that best describes the linear relationship between the two variables. With this line and the corresponding regression equation we can predict the values of the dependent variable based on the values of the independent variable. Moreover, with r-squared we can also assess how well the line fits our data. However, for at least two reasons, we need to be very careful when we interpret the results. The first reason is that on the basis of a regression analysis, we can never prove that there is a causal relationship between two variables. We can, in other words, never be certain that one variable is the cause of another variable. This translates to one single and not very complicated, but extremely important message: correlation is no causation. For instance, research suggests that eating a lot of chocolate makes you fat. This scatterplot shows that the more chocolate people eat, the larger their body weight tends to be. However, we need to be careful here. It might also be the case that causality runs in the opposite direction. The correlation between the two variables could also have another reason. It might for instance be the case that people with more body weight are more hungry and therefore eat more chocolate. This means that your X variable becomes your Y variable, and your Y variable becomes your X variable. This changes your scatterplot and your regression equation. This is the old scatterplot and this is the new one. This is the old equation and this is the new one. The most likely explanation of the relation between chocolate consumption and body weight, however, is that causality runs in both ways. The more chocolate you eat, the heavier you get, and the heavier you get, the more you crave chocolate.**  **abline() to add any line we like, as long as the first argument is the intercept and the second is the slope**  **These are the two lines you plotted in the last assignment. One line shows the mean, and one shows the regression line. Clearly, there is less error when we use the regression line compared to the mean line. This reduction in error from using the regression line compared to the mean line tells us how well the independent variable (money) predicts the dependent variable (prosocial behaviour).**  **4 days Online Workshop on ‘How to develop a Pythonic coding rather than Python coding – Logic Perspective’**  **Introduction to Python**  • Python - a general-purpose,Interpreted,  interactive, object-oriented and high-level  programming language.  • Fastest growing open source Programming  language  • Dynamically typed  • Versatile and can be adapted in DA,  ML, GUI,Software &Web development  • It was created by Guido van Rossum during  1985-1990.  **Python IDEs**  **• IDLE**  **• Pycharm**  **• Spyder**  **• Thonny**  **• Atom**  **• Anaconda -Jupyter Notebook, I python**  **for larger project in different domains.**  **• Google colab**  **Indentation and Blocks**  **• Python doesn't use braces ({}) to indicate blocks of code for class and function definitions or flow control.**  **• Blocks of code are denoted by line indentation, which is rigidly enforced.**  **• All statements within the block must be indented the same level**  **Conditional Execution**  **• if and else**  **if v == c:**  **#do something based on the**  **condition**  **else:**  **#do something based on v != c**  **• elif allows for additional branching**  **if condition:**  **…...**  **elif another condition:**  **…**  **else: #none of the above**  **13**  **14**  **# python program for finding greater of two numbers**  **a=int(input(‘Enter the first number’))**  **b=int(input(‘Enter the second number’))**  **if a>b:**  **print("The greater number is",a)**  **else:**  **print("The greater number is",b)**  **# for satisfying equality condition**  **if a>b:**  **print("The greater number is",a)**  **elif a==b:**  **print(“both numbers are equal",a)**  **else:**  **print(“The greater number is",b)**  **Variables, expressions, and statements**  **python**  **>>> print(4)**  **4**  **If you are not sure what type a value has, the**  **interpreter can tell you.**  **>>> type('Hello, World!')**  **<class 'str'>**  **>>> type(17)**  **<class 'int'>**  **>>> type(3.2)**  **<class 'float'>**  **>>> type('17')**  **<class 'str'>**  **>>> type('3.2')**  **<class 'str'>**  **for loop**  **for var in <collection>:**  **<statements>**  **where collection is iterable obj like list, tuple, dictionary, string and range**  **while loop**  **while condition:**  **<Statements>**  **Converting anything to a String**  **The built-in str() function can convert an**  **instance of any data type into a string**  **Ex:**  **print(“Hello ” + str(2))**  **Slicing strings**  **A segment of a string is called a slice.**  **Selecting a slice is similar to selecting acharacter:** |