**IC 252 Lab 6**

April 22, 2019

Consider the medical\_expenditure.csv dataset attached in this assignment. This dataset contains the average daily expenditure in USD ($) on a medicine named “A” across the United States of America between 2nd January 2011 and 15th April 2015. Table 1 provides a description of different columns in this dataset.

Table 1: Description of Attributes

|  |  |
| --- | --- |
| **Attributes Names** | **Attribute Descriptions** |
| AverageDailyExpenditure | Average daily expenditure on a medicine (USD/day) |
| Nmale | Number of male patients who consumed the medicine on a day |
| Nfemale | Number of female patients who consumed the medicine on a day |
| Nage-group1 | Number of patients from 0-17 age group who consumed the medicine on a day |
| Nage-group2 | Number of patients from 18-34 age group who consumed the medicine on a day |
| Nage-group3 | Number of patients from 34-44 age group who consumed the medicine on a day |
| Nage-group4 | Number of patients from 45-54 age group who consumed the medicine on a day |
| Nage-group5 | Number of patients from 55-65 age group who consumed the medicine on a day |
| Nregion1 | Number of patients from the southern region who consumed the medicine on a day |
| Nregion2 | Number of patients from the northeast region who consumed the medicine on a day |
| Nregion3 | Number of patients from the west region who consumed the medicine on a day |
| Nregion4 | Number of patients from the northcentral region who consumed the medicine on a day |
| Nregion5 | Number of patients from other regions who consumed the medicine on a day |
| Nhp0 | Number of patients using health plan 0 who consumed the medicine on a day |
| Nhp1 | Number of patients not using health plan 1 who consumed the medicine on a day |
| Nicd1 | Number of patients who went for total knee arthroplasty and consumed the medicine |
| Nicd2 | Number of patients who went for osteoarthrosis of secondary lower leg and consumed the medicine |
| Nicd3 | Number of patients who went for removal of foreign body from the eye and consumed the medicine |
| Nicd4 | Number of patients who went for total knee replacement and consumed the medicine |
| Nicd5 | Number of patients who went for osteoarthrosis of primary lower leg and consumed the medicine |
| Nicd6 | Number of patients who went for osteoarthrosis generalized lower leg and consumed the medicine |

Please answer the questions below.

1. **Independence**

Two events, A and B, are independent if P(A|B) = P(A) and P(B|A) = P(B).

Please check whether the following events are independent between 2nd January 2011 and 15th April 2015:

(A) “Number of males <= 93” and “Average daily expenditure on medicine A <= $504”.

(B) “Number of females <= 131” and “Average daily expenditure on medicine A <= $504”.

(C) “Number in age-group 2” <= 25” and “Average daily expenditure on medicine A > $504”.

(D) “Number in age-group 2” <= 25” and “Average daily expenditure on medicine A > $504”.

(E) “Number in region 5” <= 53” and “Average daily expenditure on medicine A > $504”.

1. **Covariance**

For a sample, the covariance is defined as the following:

Where:

– values of the X-variable

– values of the Y-variable

– mean (average) of the X-variable

– mean (average) of the Y-variable

– number of the data points

Please find the covariance between the following variables between 2nd January 2011 and 15th April 2015:

(A). Nmale and Average daily expenditure.

(B). Nfemales and Average daily expenditure.

(C). Nhp0 and Average daily expenditure.

(D). Nhp1 and Average daily expenditure.

(E). Nicd2 and Average daily expenditure.

(F). **[optional]** In Python, the numpy library has a cov() function, which could be used to compute the covariance between two variables. Please use the cov() function to verify the covariance computed in (A) to (E) above.

1. **Correlation**

The Pearson's correlation coefficient *r* (named after Karl Pearson) can be used to summarize the strength of a linear relationship between two variables and . The Pearson's correlation coefficient isdefined by the following equation:

Where, is the covariance between variables and and and are the standard deviations of variables and , respectively.

The table below shows the interpretation of values in the range [-1, +1]

|  |  |  |
| --- | --- | --- |
| range for a direct relationship between and | range for an indirect relationship between and | Relationship between and |
| 0.0 | 0.0 | None |
| (0.0, 0.1] | (-0.0, -0.1] | Weak |
| (0.1, 0.3] | (-0.1, -0.3] | Moderate |
| (0.3, 0.5] | (-0.3, -0.5] | Strong |
| (0.5, 1.0) | (-0.5, -1.0) | Very Strong |
| 1.0 | -1.0 | Perfect |

Please find the between the following variables between 2nd January 2011 and 15th April 2015 and categorize the relationship based upon the table above (categories could be None, Weak, Moderate, Strong, Very Strong, and Perfect):

(A). Nmale and Average daily expenditure.

(B). Nfemales and Average daily expenditure.

(C). Nhp0 and Average daily expenditure.

(D). Nhp1 and Average daily expenditure.

(E). Nicd2 and Average daily expenditure.

(F). **[optional]** In Python, the scipy.stats library has a pearsonr() function, which could be used to compute the correlation between two variables. Please use the pearsonr() function to verify the correlations computed in (A) to (E) above.