Tutorial 2: Work on Data and Inferences

PHQ-9, PSS and Loneliness Scale dataset analysis

Work on Data

1. Replaced the string data with globally used numerical values.
   - PHQ-9
     - Consists of 9 questions each having 4 options:
       - Not at all
       - Several days
       - More than half the days
       - Nearly every day
     - Score corresponding to every option:
       - 0 = Not at all
       - 1 = Several days
       - 2 = More than half the days
       - 3 = Nearly every day
   - Loneliness Scale
     - Consists of 20 questions each having 4 options:
       - Often
       - Sometimes
       - Rarely
       - Never
     - Score corresponding to every option:
       - 4 = Often
       - 3 = Sometimes
       - 2 = Rarely
       - 1 = Never
   - Perceived Stress Scale
     - Consists of 20 questions each having 5 options:
       - Never
       - Almost Never
       - Sometimes
       - Fairly Often
       - Very Often
     - Score corresponding to every option:
       - 0 = Never
       - 1 = Almost Never

        - ♦ 2 = Sometimes
        - ♦ 3 = Fairly Often
        - ♦ 4 = Very Often

        Reverse the scores for questions 4,5,7,8.

        - ♦ 4 = Never
        - ♦ 3 = Almost Never
        - ♦ 2 = Sometimes
        - ♦ 1 = Fairly Often
        - ♦ 0 = Very Often

2. Get the total score of each participant.
    - PHQ-9
        - ▪ Get the total of all the answer's scores.
        - ▪ Range of total score is 0-27.
        - ▪ Scores of 5, 10, 15, and 20 represent cutpoints for mild, moderate, moderately severe and severe depression respectively.
    - Loneliness Scale
        - ▪ Get the total of all the answer's scores.
        - ▪ Range of total score is 20-80.
        - ▪ Higher scores indicate higher degree of loneliness.
    - Perceived Stress Scale
        - ▪ Get the total of all the answer's scores.
        - ▪ Range of total score is 0-40.
        - ▪ Scores of 13 and 26 represent cutpoints for low, moderate, and high perceived stress respectively.
3. Compared pre and post semester scores using mean and standard deviation.
    - Got the mean and standard deviation of pre and post semester scores of participants in PHQ=9, Loneliness scale and PSS datasets.

Inferences

1. Depression and stress were seen a little higher and loneliness was seen a little lower in the post survey as compared to the pre semester survey.

Linear Regression

Performed linear regression between:

- Stress and sleep hours
- Stress and no. of calls
- Stress and no. of deadlines

- Stress and GPA
- Stress and no. of active days on piazza
- Depression and sleep hours
- Depression and no. of calls
- Depression and no. of deadlines
- Depression and GPA
- Depression and no. of active days on piazza

Steps to perform linear regression

1. Created datasets of stress and sleep hours, stress and no. of calls, stress and no. of deadlines, stress and GPA, and no. of active days on piazza.
2. In Google Colab Python, import the libraries pandas, numpy, linear_model from sklearn, matplotlin and io.

   ```
   import io
   import pandas as pd
   import numpy as np
   from matplotlib import pyplot as plt
   from sklearn import linear_model
   ```

3. Write import statement for uploading files in google colab.

   ```
   from google.colab import files
   up=files.upload()
   ```

4. Write upload statement for the required .csv file.

   ```
   df = pd.read_csv(io.BytesIO(up['Sleep Regression.csv']))
   df
   ```

5. Plot the scatter plot of independent variable (e.g. sleep hours) versus stress(Y-axis).

   ```
   %matplotlib inline
   plt.xlabel('sleep')
   plt.ylabel('stress')
   plt.scatter(df.sleep,df.stress,color='orangered',marker='+')
   ```

6. Create a new dataframe consisting of only the independent variable.

   ```
   dfnew = df.drop('stress',axis='columns')
   dfnew
   ```

7. Create a new dataframe consisting of only stress values.

   ```
   stress = df.stress
   stress
   ```

8. Create a linear regression object.

   ```
   linreg = linear_model.LinearRegression()
   ```

```
linreg.fit(dfnew,stress)
print(linreg.fit(dfnew,stress))
```

9. Get regression score.

```
linreg.score(dfnew, stress) #R squared value
```

10. Plot the linear regression graph.

```
%matplotlib inline
plt.xlabel('sleep')
plt.ylabel('stress')
plt.scatter(df.sleep,df.stress,color='indigo',marker='*')
plt.plot(df.sleep,linreg.predict(dfnew),color='aqua')
```

11. Get regression coefficient and intercept.

```
print(linreg.coef_)
linreg.intercept_
```

12. Repeat the same process for all the other variables with dependent variable as stress.
13. Repeat the same process for all the other variables with dependent variable as depression.

Inferences

- Stress increases as sleeping hours increase.
- Depression decreases as sleeping hours increase.
- Stress is seen higher with more number of calls.
- There’s not much correlation between call logs and depression.
- Stress level and depression is seen higher with more deadlines.
- Stress level is seen higher with higher GPA.
- Depression is seen lower when the student gets higher GPA.
- Stress level seems to decrease with an increase in number of active days on piazza.
- Depression seems to increase with an increase in number of active days on piazza.