nhanes_multivariate_practice

January 8, 2022

1 Practice notebook for multivariate analysis using NHANES data

This notebook will give you the opportunity to perform some multivariate analyses on your own using the NHANES study data. These analyses are similar to what was done in the week 3 NHANES case study notebook.

You can enter your code into the cells that say "enter your code here", and you can type responses to the questions into the cells that say "Type Markdown and Latex".

Note that most of the code that you will need to write below is very similar to code that appears in the case study notebook. You will need to edit code from that notebook in small ways to adapt it to the prompts below.

To get started, we will use the same module imports and read the data in the same way as we did in the case study:

```
In [2]: %matplotlib inline
        import matplotlib.pyplot as plt
        import seaborn as sns
        import pandas as pd
        import statsmodels.api as sm
        import numpy as np
        from scipy import stats

        da = pd.read_csv("nhanes_2015_2016.csv")
        da.columns
```

```
Out[2]: Index(['SEQN', 'ALQ101', 'ALQ110', 'ALQ130', 'SMQ020', 'RIAGENDR', 'RIDAGEYR',
               'RIDRETH1', 'DMDCITZN', 'DMDEDUC2', 'DMDMARTL', 'DMDHHSIZ', 'WTINT2YR',
               'SDMVPSU', 'SDMVSTRA', 'INDFMPIR', 'BPXSY1', 'BPXDI1', 'BPXSY2',
               'BPXDI2', 'BMXWT', 'BMXHT', 'BMXBMI', 'BMXLEG', 'BMXARML', 'BMXARMC',
               'BMXWAIST', 'HIQ210'],
              dtype='object')
```

1.1 Question 1

Make a scatterplot showing the relationship between the first and second measurements of diastolic blood pressure (BPXDI1 and BPXDI2). Also obtain the 4x4 matrix of correlation coefficients among the first two systolic and the first two diastolic blood pressure measures.

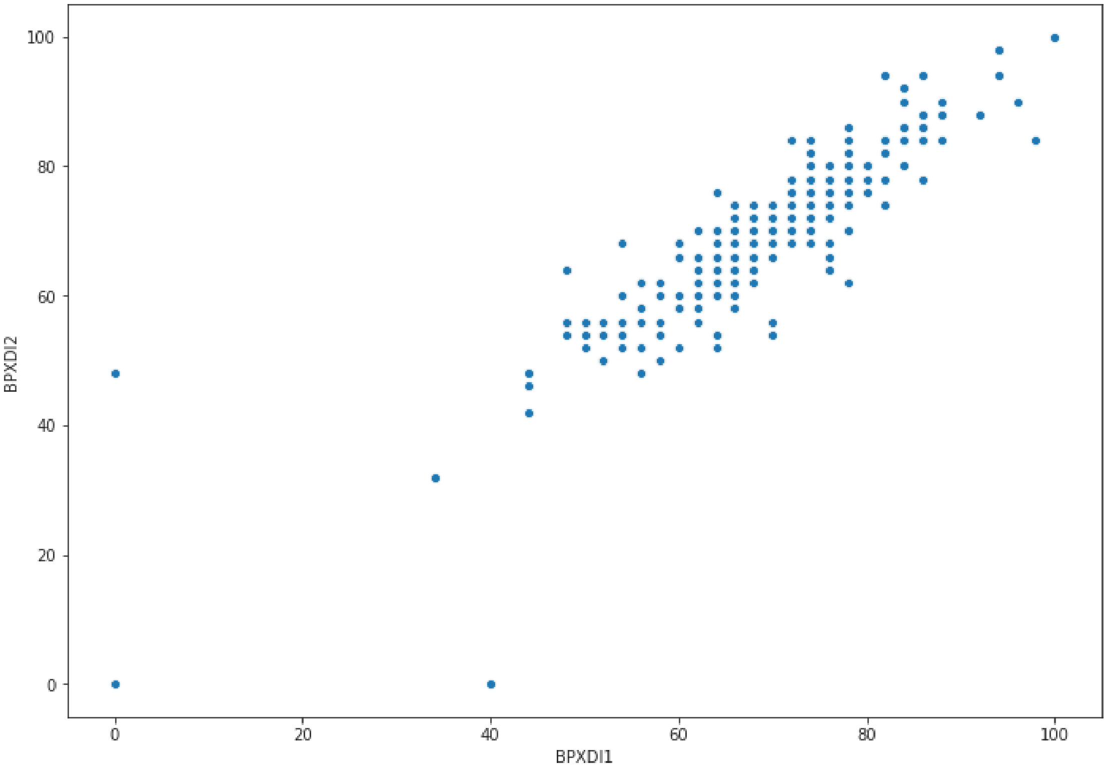
```
In [ ]: 'BPXSY1', 'BPXDI1', 'BPXSY2',
               'BPXDI2'
```

```
In [3]: da = da.dropna().astype(int)
```

```
In [4]: # plot
        sns.set_style('ticks')
        fig, ax = plt.subplots()

        # the size of A4 paper
        fig.set_size_inches(11.7, 8.27)
        sns.scatterplot(data= da, x="BPXDI1", y="BPXDI2")
```

```
Out[4]: <matplotlib.axes._subplots.AxesSubplot at 0x7f598dd019b0>
```



```
In [5]: x = pd.DataFrame(da[['BPXSY1', 'BPXDI1', 'BPXSY2','BPXDI2']])
```

```
In [42]: x.corr()
```

```
Out[42]:           BPXSY1    BPXDI1    BPXSY2    BPXDI2
         BPXSY1  1.000000  0.258749  0.954760  0.225749
         BPXDI1  0.258749  1.000000  0.297655  0.871593
         BPXSY2  0.954760  0.297655  1.000000  0.279470
         BPXDI2  0.225749  0.871593  0.279470  1.000000
```

Q1a. How does the correlation between repeated measurements of diastolic blood pressure relate to the correlation between repeated measurements of systolic blood pressure?

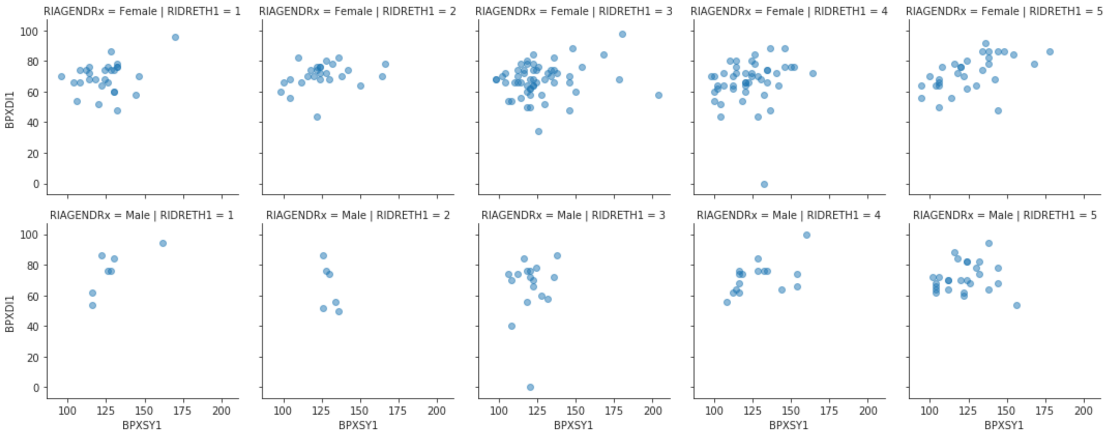
Q2a. Are the second systolic and second diastolic blood pressure measure more correlated or less correlated than the first systolic and first diastolic blood pressure measure?

1.2 Question 2

Construct a grid of scatterplots between the first systolic and the first diastolic blood pressure measurement. Stratify the plots by gender (rows) and by race/ethnicity groups (columns).

```
In [13]: da["RIAGENDRx"] = da.RIAGENDR.replace({1: "Male", 2: "Female"})
```

```
In [14]: strat = sns.FacetGrid(da, col = 'RIDRETH1', row = 'RIAGENDRx').map(plt.scatter, 'BPXS
```



Q3a. Comment on the extent to which these two blood pressure variables are correlated to different degrees in different demographic subgroups.

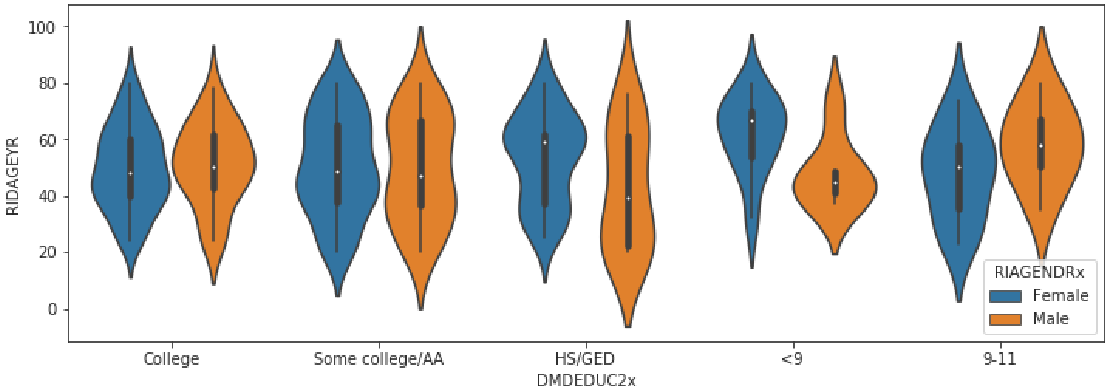
We almost see no correlation between gender, race, and systolic diastolic pressure

1.3 Question 3

Use "violin plots" to compare the distributions of ages within groups defined by gender and educational attainment.

```
In [16]: da["DMDEDUC2x"] = da.DMDEDUC2.replace({1: "<9",
                                                 2: "9-11",
                                                 3: "HS/GED",
                                                 4: "Some college/AA",
                                                 5: "College",
                                                 7: "Refused",
                                                 9: "Don't know"})
```

```
In [21]: plt.figure(figsize = (12,4))
         v = sns.violinplot(da.DMDEDUC2x, da.RIDAGEYR, hue = da.RIAGENDRx)
```



Q4a. Comment on any evident differences among the age distributions in the different demographic groups.

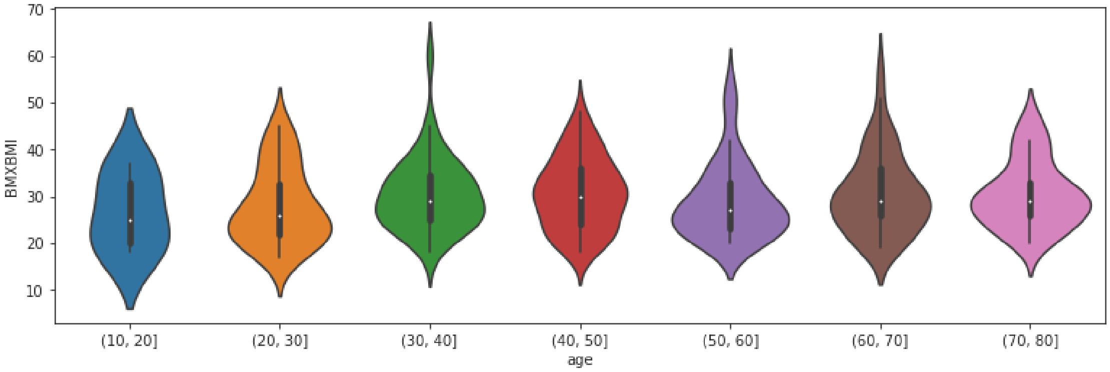
The age ranges in various education background are mostly similar except for <9 education.

1.4 Question 4

Use violin plots to compare the distributions of BMI within a series of 10-year age bands. Also stratify these plots by gender.

```
In [25]: da['age'] = pd.cut(da.RIDAGEYR,[10,20,30,40,50,60,70,80])
```

```
In [26]: plt.figure(figsize=(13,4))
         z = sns.violinplot(da.age, da.BMXBMI)
```



Q5a. Comment on the trends in BMI across the demographic groups.

For all age group, the median of BMI is around 20-30.

1.5 Question 5

Construct a frequency table for the joint distribution of ethnicity groups (RIDRETH1) and health-insurance status (HIQ210). Normalize the results so that the values within each ethnic group are proportions that sum to 1.

da[“RIDRETH1x”] = da.DMDEDUC2.replace({1: “<9”, 2: “9-11”, 3: “HS/GED”, 4: “Some college/AA”, 5: “College”, 7: “Refused”, 9: “Don’t know”}) da[“HIQ210”] = da.DMDMARTL.replace({1: “Married”, 2: “Widowed”, 3: “Divorced”, 4: “Separated”, 5: “Never married”, 6: “Living w/partner”, 77: “Refused”}) db = da.loc[(da.DMDEDUC2x != “Don’t know”) & (da.DMDMARTLx != “Refused”), :]

```
In [27]: c = pd.crosstab(da.RIDRETH1, da.HIQ210)
         c
```

```
Out[27]: HIQ210    1   2
         RIDRETH1
         1         6  29
         2         4  28
         3         4  69
         4         8  54
         5         4  53
```

```
In [28]: c.apply(lambda c: c/c.sum(), axis=1)
```

```
Out[28]: HIQ210           1         2
         RIDRETH1
         1         0.171429  0.828571
         2         0.125000  0.875000
         3         0.054795  0.945205
         4         0.129032  0.870968
         5         0.070175  0.929825
```

Q6a. Which ethnic group has the highest rate of being uninsured in the past year?