nhanes_multivariate_practice

November 11, 2020

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:

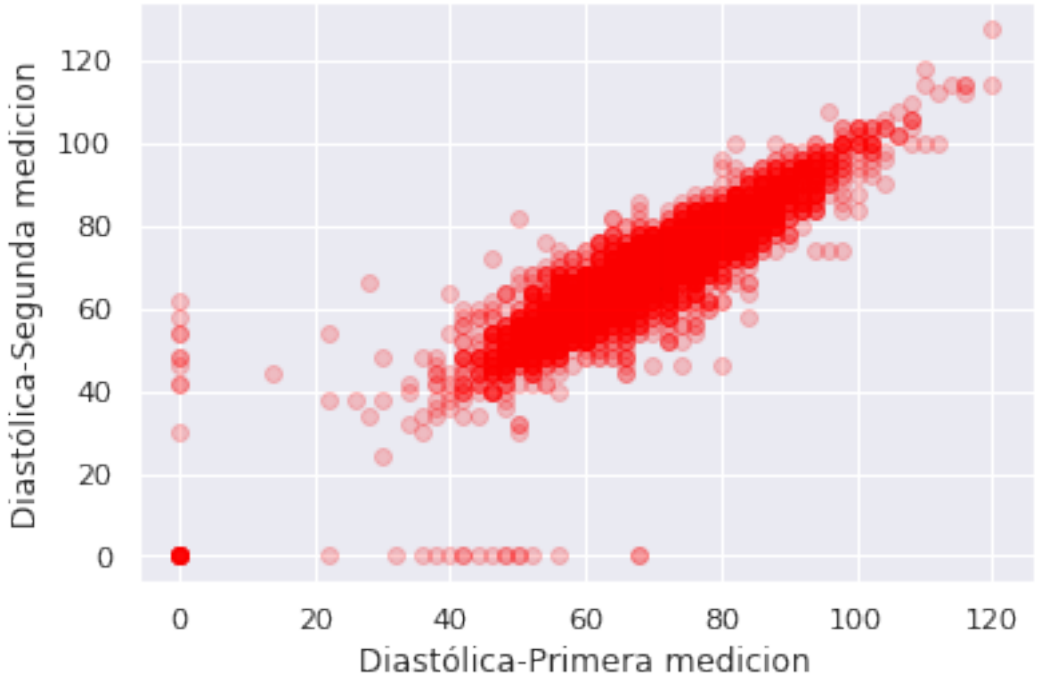
```
In [33]: %matplotlib inline
         import matplotlib.pyplot as plt
         import seaborn as sns
         import pandas as pd
         import statsmodels.api as sm
         import numpy as np

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

```
Out[33]: 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

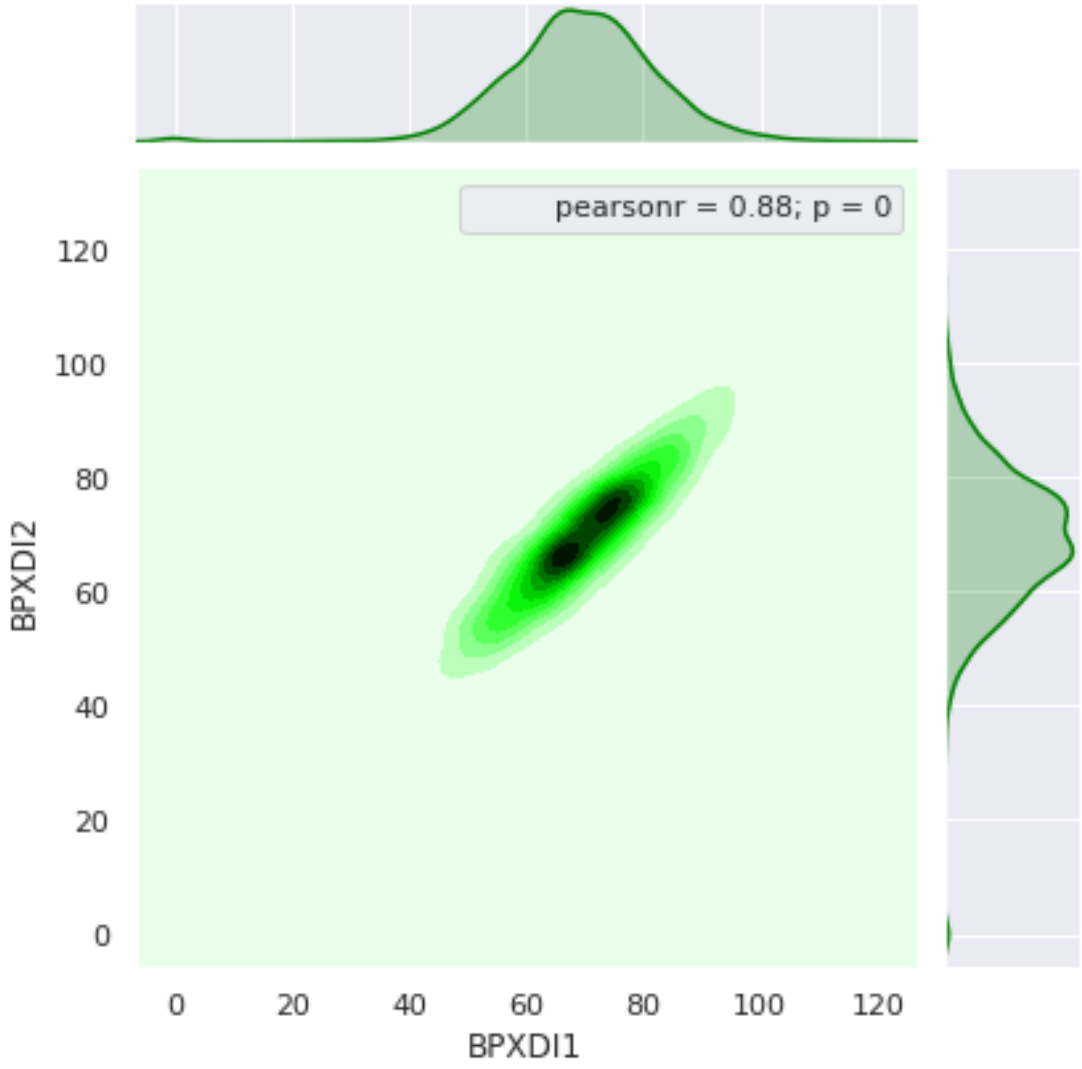
Haga un diagrama de dispersión que muestre la relación entre la primera y la segunda medición de la presión arterial diastólica ([BPXDI1] (https://wwwn.cdc.gov/Nchs/Nhanes/2015-2016/BPX_I.htm#BPXDI1) y [BPXDI2] (https://wwwn.cdc.gov/Nchs/Nhanes/2015-2016/BPX_I.htm#BPXDI2)). Obtenga también la matriz 4x4 de coeficientes de correlación entre las dos primeras medidas de presión arterial sistólica y las dos primeras diastólica.

In [34]:
```
sns.regplot(x="BPXDI1",y="BPXDI2",data=da, fit_reg=False, scatter_kws={"alpha": 0.2},
_ = plt.xlabel("Diastólica-Primera medicion")
_ = plt.ylabel("Diastólica-Segunda medicion")
```

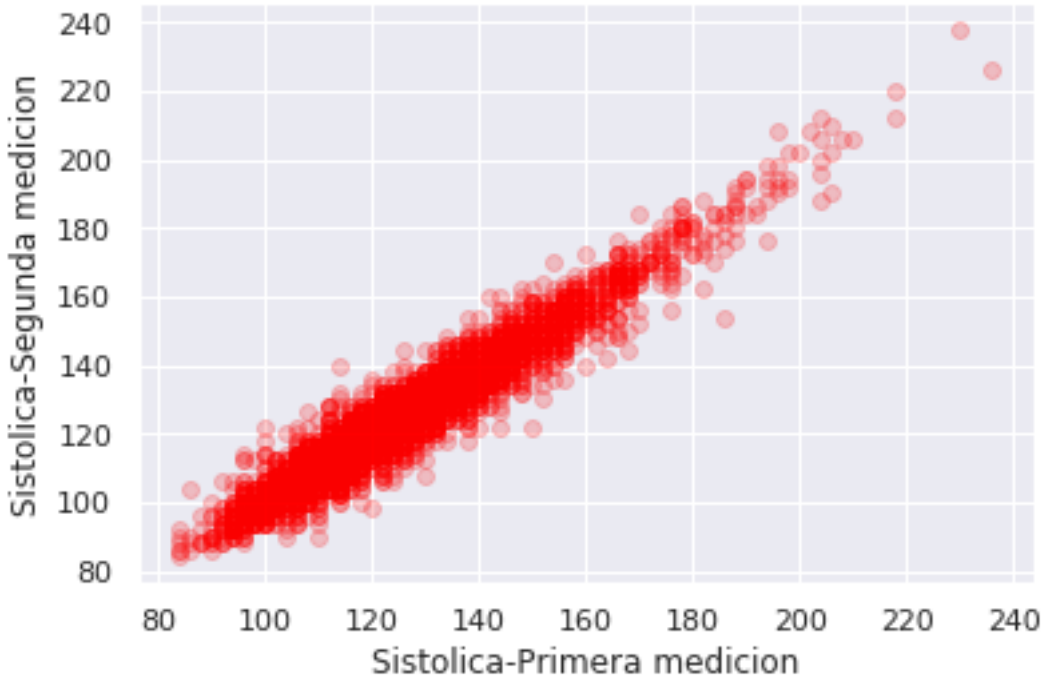


In [35]:
```
sns.set()
_ = sns.jointplot(x="BPXDI1", y="BPXDI2", kind='kde', data=da,color='green').annotate
plt.show()
```

```
/opt/conda/lib/python3.6/site-packages/seaborn/axisgrid.py:1847: UserWarning: JointGrid annotat
  warnings.warn(UserWarning(msg))
```

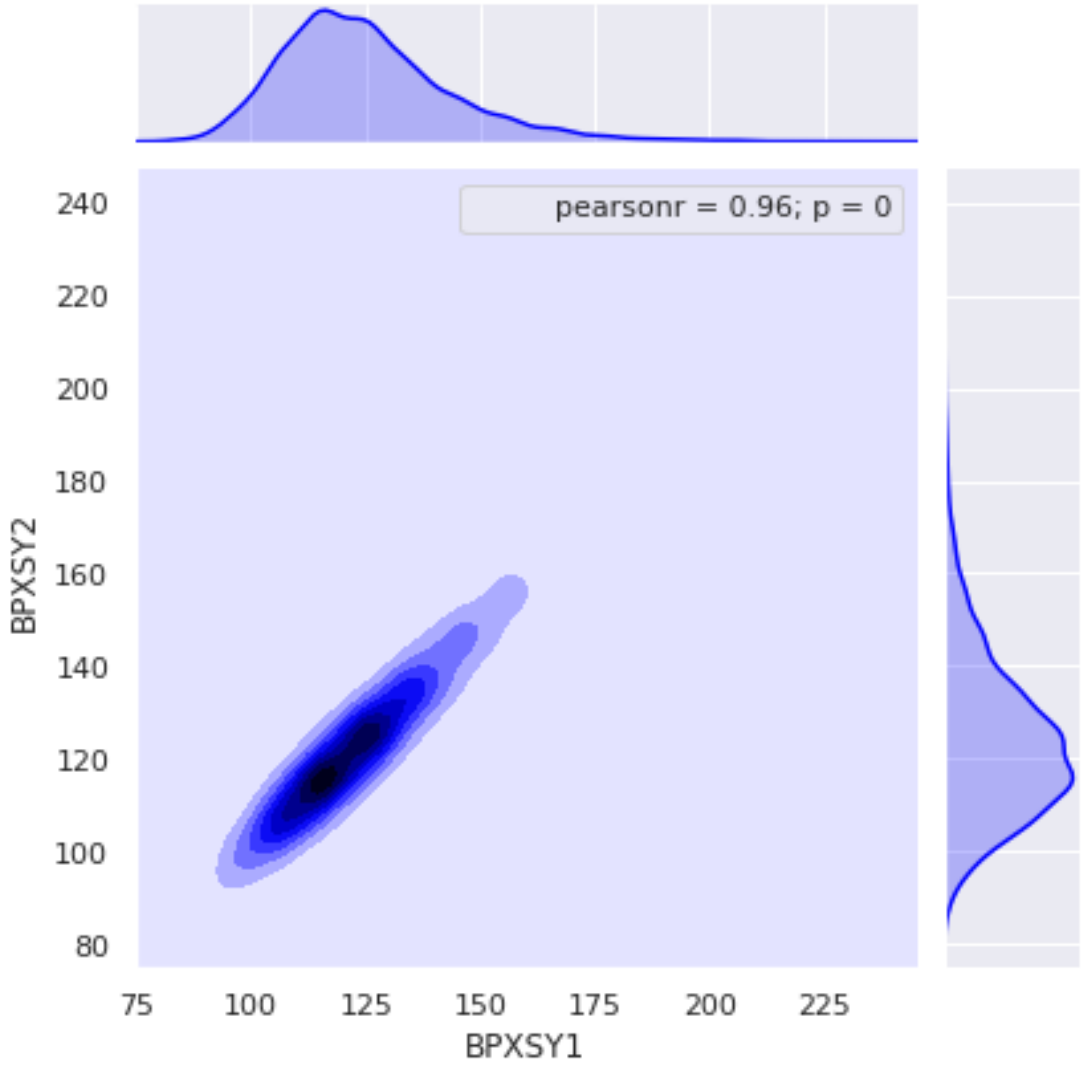



```
In [36]: sns.regplot(x="BPXSY1",y="BPXSY2",data=da, fit_reg=False, scatter_kws={"alpha": 0.2},
         _ = plt.xlabel("Sistolica-Primera medicion")
         _ = plt.ylabel("Sistolica-Segunda medicion")
```




```
In [37]: sns.set()
         _ = sns.jointplot(x="BPXSY1", y="BPXSY2", kind='kde', data=da,color='blue').annotate(s
         plt.show()
```

```
/opt/conda/lib/python3.6/site-packages/seaborn/axisgrid.py:1847: UserWarning: JointGrid annotat
  warnings.warn(UserWarning(msg))
```




```
In [38]: dfCorr = da.loc[:,["BPXSY1","BPXSY2","BPXDI1","BPXDI2"]]

         dfCorr.corr(method="pearson")
```

Out[38]:

	BPXSY1	BPXSY2	BPXDI1	BPXDI2
BPXSY1	1.000000	0.962287	0.316531	0.277681
BPXSY2	0.962287	1.000000	0.329843	0.303847
BPXDI1	0.316531	0.329843	1.000000	0.884722
BPXDI2	0.277681	0.303847	0.884722	1.000000

Q1a. £Cómo se relaciona la correlación entre las mediciones repetidas de la presión arterial diastólica con la correlación entre las mediciones repetidas de la presión arterial sistólica?

Podemos ver que la relacion entre la primera toma de presion diasolica esta fuermente relaciona con la segunda toma , sin embargo no tiene tanta relacion entre la primera toma de presion diasolica con la sistolica

Q2a. £Están la segunda medida de presión arterial sistólica y la segunda diastólica más correlacionada o menos correlacionada que la primera medida de presión arterial sistólica y la primera diastólica?

Gracias a nuestra tabla de correlacion , podemos nota que BPXSY1 esta mas realcionado con BPXD1 , esto significa que la presion sistolica tiene un coef de correlacion de 0.3165.. , mientras que la correlacion entre BPXD2 y BPXDi2 tiene menos correlacion con un coef de corr del 0.277 por lo tanto podemos concluir que la BPXSY2 tiene el mayor coef

1.2 Question 2

Construya una cuadrícula de diagramas de dispersión entre la **primera medición de presión arterial sistólica y la primera diastólica.** Estratifique las parcelas por género (filas) y por grupos de raza / etnia (columnas).

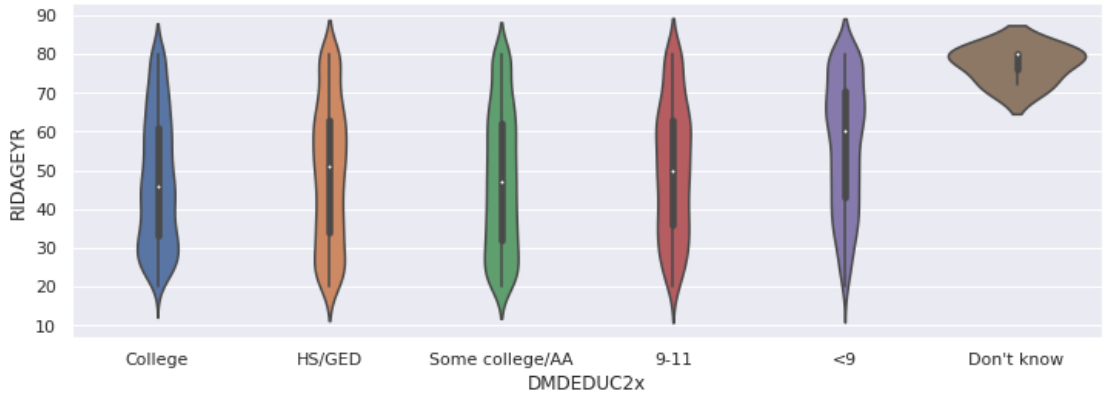
```
In [40]: da["RIAGENDRx"] = da.RIAGENDR.replace({1: "Male", 2: "Female"})
         sns.set()


         _ = sns.FacetGrid(da,col="RIDRETH1",row="RIAGENDRx").map(plt.scatter,"BPXSY1","BPXDI1
```

Q3a. Comente hasta qué punto estas dos variables de presión arterial se correlacionan en diferentes grados en diferentes subgrupos demográficos.

1.3 Question 3

Utilice "parcelas de violín" para comparar las distribuciones de edades dentro de los grupos definidos por género y nivel educativo.

```
In [42]: da["DMDEDUC2x"] = da.DMDEDUC2.replace({1: "<9", 2: "9-11", 3: "HS/GED", 4: "Some colle
                                                 7: "Refused", 9: "Don't know"})

         db = da.loc[(da.DMDEDUC2x != "Don't know") , :]
         plt.figure(figsize=(12, 4))
         a = sns.violinplot(da.DMDEDUC2x, da.RIDAGEYR)
```


90
80
70
60
50
40
30
20
10
RIDAGEYR
College
HS/GED
Some college/AA
9-11
<9
Don't know
DMDEDUC2x