

Week 3 Assessment

This Jupyter Notebook is auxiliary to the following assessment in this week. To complete this assessment, you will complete the 5 questions outlined in this document and use the output from the python cells to answer them.

Run the following cell to initialize your environment and begin the assessment.

```
In [1]: ##### RUN THIS

import warnings
warnings.filterwarnings('ignore')

import numpy as np
import statsmodels.api as sm
import pandas as pd

url = "nhanes_2015_2016.csv"
da = pd.read_csv(url)

# Drop unused columns, drop rows with any missing values.
vars = ["BPXSY1", "RIDAGEYR", "RIAGENDR", "RIDRETH1", "DMDEDUC2", "BMXBMI",
        "SMQ020", "SDMVSTRA", "SDMVPSU"]
da = da[vars].dropna()

da["group"] = 10*da.SDMVSTRA + da.SDMVPSU

da["smq"] = da.SMQ020.replace({2: 0, 7: np.nan, 9: np.nan})

np.random.seed(123)
```

Question 1: What is clustered data? (You'll answer this question within the quiz that follows this notebook)

Question 2: (You'll answer this question within the quiz that follows this notebook)

Utilize the following output for this question:

```
In [2]: for v in ["BPXSY1", "SDMVSTRA", "RIDAGEYR", "BMXBMI", "smq"]:
        model = sm.GEE.from_formula(v + " ~ 1", groups="group",
        cov_struct=sm.cov_struct.Exchangeable(), data=da)
        result = model.fit()
        print(v, result.cov_struct.summary())

BPXSY1 The correlation between two observations in the same cluster is 0.030
SDMVSTRA The correlation between two observations in the same cluster is 0.959
RIDAGEYR The correlation between two observations in the same cluster is 0.035
BMXBMI The correlation between two observations in the same cluster is 0.039
smq The correlation between two observations in the same cluster is 0.026
```

Which of the listed features has the highest correlation between two observations in the same cluster?

Question 3: (You'll answer this question within the quiz that follows this notebook)

What is true about multiple linear regression and marginal linear models when dependence is present in data?

Question 4: (You'll answer this question within the quiz that follows this notebook)

Multilevel models are expressed in terms of _.

Question 5: (You'll answer this question within the quiz that follows this notebook)

Which of the following is NOT true regarding reasons why we fit marginal models?