



Practice notebook for regression analysis with dependent data in NHANES

This notebook will give you the opportunity to perform some analyses using the regression methods for dependent data that we are focusing on in this week of the course.

Enter the code in response to each question in the boxes labeled "enter your code here". Then enter your responses to the questions in the boxes labeled "Type Markdown and LaTeX".

This notebook is based on the NHANES case study notebook for regression with dependent data. 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 [1]: %matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
import statsmodels.api as sm
import numpy as np

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

# Drop unused columns, drop rows with any missing values.
vars = ["BPXSY1", "RIDAGEYR", "RIAGENDR", "RIDRETH1", "DMDDEDUC2", "BMXBMI",
        "SMQ020", "SDMVSTRA", "SDMVPSU"]
da = da[vars].dropna()

# This is the grouping variable
da["group"] = 10*da.SDMVSTRA + da.SDMVPSU
```

Question 1:

Build a marginal linear model using GEE for the first measurement of diastolic blood pressure (BPXD1), accounting for the grouping variable defined above. This initial model should have no covariates, and will be used to assess the ICC of this blood pressure measure.

```
In [2]: # enter your code here
```

Q1a. What is the ICC for diastolic blood pressure? What can you conclude by comparing it to the ICC for systolic blood pressure?

Type *Markdown* and LaTeX: α^2

Question 2:

Take your model from question 1, and add gender, age, and BMI to it as covariates.

```
In [3]: # enter your code here
```

Q2a. What is the ICC for this model? What can you conclude by comparing it to the ICC for the model that you fit in question 1?

Type *Markdown* and LaTeX: α^2

Question 3:

Split the data into separate datasets for females and for males and fit two separate marginal linear models for diastolic blood pressure, one only for females, and one only for males.

```
In [4]: # enter your code here
```

Q3a. What do you learn by comparing these two fitted models?

Type *Markdown* and LaTeX: α^2

Question 4:

Using the entire data set, fit a multilevel model for diastolic blood pressure, predicted by age, gender, BMI, and educational attainment. Include a random intercept for groups.

```
In [5]: # enter your code here
```

Q4a. How would you describe the strength of the clustering in this analysis?

Type *Markdown* and LaTeX: α^2

Q4b: Include a random intercept for age, and describe your findings.

Type *Markdown* and LaTeX: α^2