Practice notebook for regression analysis with NHANES

This notebook will give you the opportunity to perform some regression analyses with the NHANES data that are similar to the analyses done in the week 2 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 [ ]:
```
%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 = "https://raw.githubusercontent.com/kshedden/statswpy/master/NHANES/merged/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"]
da = da[vars].dropna()
```

Question 1:

Use linear regression to relate the expected body mass index (BMI) to a person's age.

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

Q1a. According to your fitted model, do older people tend to have higher or lower BMI than younger people?

Type *Markdown* and LaTeX: α^2

Q1b. Based your analysis, are you confident that there is a relationship between BMI and age in the population that NHANES represents?

Type *Markdown* and LaTeX: α^2

Q1c. By how much does the average BMI of a 40 year old differ from the average BMI of a 20 year old?

Type *Markdown* and LaTeX: α^2

Q1d. What fraction of the variation of BMI in this population is explained by age?

Type *Markdown* and LaTeX: α^2

Question 2:

Add gender and ethnicity as additional control variables to your linear model relating BMI to age. You will need to recode the ethnic groups based on the values in the codebook entry for RIDRETH1.

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

Q2a. How did the mean relationship between BMI and age change when you added additional covariates to the model?

Type *Markdown* and LaTeX: α^2

Q2b. How did the standard error for the regression parameter for age change when you added additional covariates to the model?

Type *Markdown* and LaTeX: α^2

Q2c. How much additional variation in BMI is explained by age, gender, and ethnicity that is not explained by age alone?

Type *Markdown* and LaTeX: α^2

Q2d. What reference level did the software select for the ethnicity variable?

Type *Markdown* and LaTeX: α^2

Q2e. What is the expected difference between the BMI of a 40 year-old non-Hispanic black man and a 30 year-old non-Hispanic black man?

Type *Markdown* and LaTeX: α^2

Q2f. What is the expected difference between the BMI of a 50 year-old Mexican American woman and a 50 year-old non-Hispanic black man?

Type *Markdown* and LaTeX: α^2

Question 3:

Randomly sample 25% of the NHANES data, then fit the same model you used in question 2 to this data set.

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

Q3a. How do the estimated regression coefficients and their standard errors compare between these two models? Do you see any systematic relationship between the two sets of results?

Type *Markdown* and LaTeX: α^2

Question 4:

Generate a scatterplot of the residuals against the fitted values for the model you fit in question 2.

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

Q4a. What mean/variance relationship do you see?

Type *Markdown* and LaTeX: α^2

Question 5:

Generate a plot showing the fitted mean BMI as a function of age for Mexican American men. Include a 95% simultaneous confidence band on your graph.

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

Q5a. According to your graph, what is the longest interval starting at year 30 following which the mean BMI could be constant? *Hint:* What is the longest horizontal line starting at age 30 that remains within the confidence band?

Type *Markdown* and LaTeX: α^2

Q5b. Add an additional line and confidence band to the same plot, showing the relationship between age and BMI for Mexican American women. At what ages do these intervals not overlap?

Type *Markdown* and LaTeX: α^2

Question 6:

Use an added variable plot to assess the linearity of the relationship between BMI and age (when controlling for gender and ethnicity).

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

Q6a. What is your interpretation of the added variable plot?

Type *Markdown* and LaTeX: α^2

Question 7:

Generate a binary variable reflecting whether a person has had at least 12 drinks in their lifetime, based on the ALQ110 variable in NHANES. Calculate the marginal probability, odds, and log odds of this variable for women and for men. Then calculate the odds ratio for females relative to males.

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

Q7a. Based on the log odds alone, do more than 50% of women drink alcohol?

Type *Markdown* and LaTeX: α^2

Q7b. Does there appear to be an important difference between the alcohol use rate of women and men?

Type *Markdown* and LaTeX: α^2

Question 8:

Use logistic regression to express the log odds that a person drinks (based on the binary drinking variable that you constructed above) in terms of gender.

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

Q8a. Is there statistical evidence that the drinking rate differs between women and men? If so, in what direction is there a difference?

Type *Markdown* and LaTeX: α^2

Q8b. Confirm that the log odds ratio between drinking and smoking calculated using the logistic regression model matches the log odds ratio calculated directly in question 6.

Type *Markdown* and LaTeX: α^2

Question 9:

Use logistic regression to relate drinking to age, gender, and education.

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

Q9a. Which of these predictor variables shows a statistically significant association with drinking?

Type *Markdown* and LaTeX: α^2

Q9b. What is the odds of a college educated, 50 year old woman drinking?

Type *Markdown* and LaTeX: α^2

Q9c. What is the odds ratio between the drinking status for college graduates and high school graduates (with no college), holding gender and age fixed?

Type *Markdown* and LaTeX: α^2

Q9d. Did the regression parameter for gender change to a meaningful degree when age and education were added to the model?

Type *Markdown* and LaTeX: α^2

Question 10:

Construct a CERES plot for the relationship between drinking and age (using the model that controls for gender and educational attainment).

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

Q10a. Does the plot indicate any major non-linearity in the relationship between age and the log odds for drinking?