BIOS 6312 - Modern Regression Analysis
Spring 2021
Lab #3

Objective: To investigate the association between 6-month hemoglobin A1c and baseline hemoglobin A1c in a population of adults over the age of 45 years with uncontrolled diabetes. Please be sure to download the documentation for the REACH study.

Data pre-processing:

1. Load the REACH data set into Stata.
2. Just to get some practice with manipulating data in Stata, we're going to drop all subjects under 45 years of age for the purpose of this analysis.
3. Create a three-level treatment variable with the following groups: Control, REACH, and REACH + FAMS.

Exploratory data analysis:

1. Create a plot to display baseline and 6-month A1c values in each treatment group.

Primary data analysis:

1. Fit a linear regression model to quantify the association between baseline and 6-month A1c:

$$E[\texttt{a1c6}|\texttt{a1c0}] = \beta_0 + \beta_1 \texttt{a1c0}. \tag{1}$$

2. Develop a prediction interval for 6-month A1c among individuals with a baseline A1c of 7.2.
3. What assumptions would be required for the prediction interval you formed in 2 to be valid? Utilize a couple of diagnostic tools to evaluate how well they appear to hold.
4. Center the baseline A1c variable.
5. Fit a linear regression model to quantify the association between centered baseline and 6-month A1c (not centered):

$$E[\texttt{a1c6}|\texttt{a1c0}_c] = \beta_0^c + \beta_1^c \texttt{a1c0}_c. \tag{2}$$

   How do the interpretations of the intercepts in models (1) and (2) differ? What does not change? Which approach would be more advantageous and why?

6. Using separate regression models, quantify the association between centered baseline and 6-month A1c (not centered) in each treatment group.

$$Control: E[\texttt{a1c6}|\texttt{a1c0}_c] = \beta_0 + \beta_1 \texttt{a1c0}_c \tag{3}$$

$$REACH: E[\texttt{a1c6}|\texttt{a1c0}_c] = \alpha_0 + \alpha_1 \texttt{a1c0}_c \tag{4}$$

$$REACH + FAMS: E[\texttt{a1c6}|\texttt{a1c0}_c] = \gamma_0 + \gamma_1 \texttt{a1c0}_c. \tag{5}$$

7. What is the interpretation of the coefficients in models 3-5?

8. Can you use the results of the above three models to evaluate evidence that REACH + FAMS is differentially effective compared to REACH alone?

9. Suppose that we were hypothetically interested in comparing the geometric mean. What modification would you make to models (3)-(5) to accomplish this? Repeat Problems 6 and 7 with this modification in mind.

List of useful Stata commands:

- `import delimited`
- `drop`
- `replace`
- `destring`
- `group`
- `label define`
- `label values`
- `scatter`, `legend`
- `generate`/ `egenerate`
- `regress`, `robust`
- `predict`
- `lowess`

Model tables:

(1)

	Estimate	95% CI	p-value
Intercept			
`a1c0`			

(2)

	Estimate	95% CI	p-value
Intercept			
`a1c0`			

(3)

	Control		
	Estimate	95% CI	p-value
Intercept			
`a1c0`			

(4)

	REACH		
	Estimate	95% CI	p-value
Intercept			
`a1c0`			

(5)

	REACH + FAMS		
	Estimate	495% CI	p-value
Intercept			
`a1c0`			

(6)

	Control		
	Estimate	95% CI	p-value
Intercept			
`a1c0`			

(7)

	REACH		
	Estimate	95% CI	p-value
Intercept			
`a1c0`			

(8)

	REACH + FAMS		
	Estimate	495% CI	p-value
Intercept			
`a1c0`			