

BIOS 525 - 2020 Final Exam

Instructions:

1. Submit report by noon (12 pm) on Friday, December 4th.
2. Include *all* analytic code in an appendix.
3. Do not include any output from statistical software directly in the report.
4. All work must be COMPLETELY independent.
5. Email all questions to the instructor. Aggregated responses will be sent to all students.

Question 1: Diarrheal Prevalence

[Suggested time: 1-2 hours]

The dataset *diarrhea.csv* contains a cross-sectional survey of children at different schools. The objective is to estimate the **difference in diarrhea prevalence** between schools with and without programs for safe water, adequate sanitation, and improved hygiene. We are particularly interested in whether associations between hygiene programs and diarrhea prevalence were modified by age and sex.

Variable Codebook:

- | | |
|------------------|--|
| 1. <i>School</i> | School ID |
| 2. <i>Sex</i> | Student's sex: 0 = boy, 1 = girl |
| 3. <i>Age</i> | Student's age at the time of survey |
| 4. <i>Z</i> | Indicator variable for whether the child had diarrhea over the last three days |
| 5. <i>treat</i> | Indicator variable for the presence of a school-level program to improve sanitation and hygiene. |

Address the scientific questions using a **generalized estimation equation** approach. Only describe a single statistical model you decide to use.

- a) [10 points] Report relevant descriptive statistics (1 paragraph, tables/figure optional).
- b) [15 points] Write down your statistical model. Justify why and how the model is useful for answering the scientific questions.
- c) [15 points] Summarize your findings and relevant model parameters (1-2 paragraphs, tables/figure optional).

Question 2: Chlamydia Incidence

[Suggested time: 1-2 hours]

The dataset (*CDC.RData*) contains annual number of chlamydia cases reported at the county-level during years 2003 to 2010. Here we will only analyze data from counties with population greater than 500,000.

Several variables on county-level population characteristics are also obtained from Census 2000. We are interested in identifying **population characteristics that are associated** with chlamydia incidence rates.

Variable Codebook:

1. FIPS	Unique county identifier (Federal Information Processing Standards)
2. Area	County name
3. State	Two-letter state abbreviation
4. Year	Reporting year (see note*)
5. Population	At-risk population size
6. Cases	Reported cases of chlamydia
7. HHIncome2000	Median household income (in \$1,000) from Census 2000
8. PBlack2000	Percent black population from Census 2000
9. PHisp2000	Percent Hispanic population from Census 2000

*Using the year variable directly may result in convergence issues due to scaling (i.e. coefficients being too small). Consider reparametrizing the time variable (e.g. centering or subtracting a reference year) when fitting the regression model.



Address the scientific questions using a **random-intercept modeling** approach. Only describe a single statistical model you decide to use.

- [10 points] Report relevant descriptive statistics (1 paragraph, tables/figure optional).
- [15 points] Write down your statistical model. Justify why and how the model is useful for answering the scientific questions.
- [15 points] Summarize your findings and relevant model parameters (1-2 paragraphs, tables/figure optional).

Question 3: Power Calculations

[Suggested time: 1-2 hours]

A researcher is interested in evaluating the effect of a household intervention to reduce blood pressure among adults. The intervention involves replacing household biomass cookstoves to those that use cleaner energy sources. To ensure that all participants eventually receive the intervention, the cookstove replacement will be “phased in” *randomly* to 750 households over a period of 3 years as described below.

	Year 1	Year 2	Year 3
Group 1 (250 households)	Intervention		
Group 2 (250 households)	Control	Intervention	
Group 3 (250 households)	Control	Control	Intervention

Cookstove replacement will begin at the start of each year and blood pressure for a single adult within each household will be measured at the end of the year, longitudinally across 3 years. Hence at the end of the study, participants in Group 1, 2, and 3 will have experienced the intervention for 1, 2, and 3 years, respectively.

a) [5 points] Assume that the intervention does not change over time, write down a **random-intercept model** for testing whether the intervention has an effect on blood pressure.

b) [10 points] Assume blood pressure has a standard deviation of 12 and a within-individual intraclass correlation of 0.30, estimate the minimal detectable effect size with 80% power and a type I error rate of 0.05 based on the model in part (a).

c) [5 points] The investigator is further interested in estimating how the duration of the intervention may impact its effect. For example, perhaps longer intervention duration will result in larger reduction in blood pressure. Repeat part (a) and part (b) to address this scientific question. [Note: there are multiple ways to accomplish this based on the study design and different solutions will be accepted.]