Edps/Psych/Stat 587
Spring 2022
C.J. Anderson

Final Exam Due Friday May 6, 2022

You may work with a partner on the final or one your own. If you work with a partner, you should **<u>NOT</u>** consult with other students on your work and you should turn in 1 copy of the exam with both partners names on the exam. If you work on your own, you should not consult other students on your work. If you have any questions or problems, you can ask me or Gamze. Exams should be e-mailed to edpsy587@gmail.com

I have created templates for grading the exams with sections labeled according to each analysis that you are asked to report. Using headings for these items/analyses is a very good idea. Feedback on your exam will be e-mailed to you.

For the final exam, you should **choose one of the following six options**:

1. data from the Progress in International Reading Literacy Study (PIRLS)

2. data from National Youth Longitudinal Study

3. NELS data

4. data on bullying.

5. data on fatal police shootings: set of research questions 1

6. data on fatal police shootings: set of research questions 2

Option 1: For this option, you will analyze data from the Progress in International Reading Literacy Study (PIRLS), which is an large scale international assessment of reading literacy of young students. You will use the PIRLS 2006 data from only the United States. To simplify the analysis, I have deleted cases with missing data, did not include sampling weights, and only provided one of the 5 plausible values for the reading scores. If you were doing the analysis for publication and not just as a course final exam, you would have to deal with the complex sampling design, missing data, and multiple estimates of reading scores.

The data are from 4th grade students where students are nested within schools. You can find the "pirls" data on the course web-site both in plain text form and with SAS code to create the dataset. The variables in the data set are:

reading is the response/outcome variables of interest.

schoolid is the school id number.

female equals 1 for female students and 0 for male.

bornUS equals 1 for those born in US and 0 for those born outside US.

watchTV is coded as
    1 no time
    2 up to 1 hour
    3 1–3 hours
    4 3–5 hours
    5 5 hours or more

econDisadv has levels that equal

$> 50\%$ economically disadvantaged

$26 - 50\%$ economically disadvantaged

$11 - 25\%$ economically disadvantaged

$< 10\%$ economically disadvantaged

enjoy is the response to a statement "I enjoy reading" where the responses options
were coded as
    1 disagree a lot
    2 disagree a little
    3 agree a little
    4 agree a lot

(100 points) I would like to see a complete and coherent analysis of the data. You should
report

1. Basic descriptive statistics, exploratory analyses, and what you learn from them.

2. Model fitting and refinement to arrive at a final model. Be sure to give tests and
   results for random and fixed effects.

3. Examination of assumptions for the model that you choose.

4. A complete interpretation of the model.

5. Answers to the following questions

   What variables explain the variability in reading scores and how to they
   related to reading scores? Based on your results, what suggestions
   might you make about ways to increase literacy among 4th grade
   students?

6. Write a paragraph of the results that might go in a newspaper article.

Option 2 For this option you will analyze data from the National Youth Longitudinal Study, which I obtained from the web. I have created the data set that you use for this option and made it is available on the course web-site. The description of the data follows.

1. (90 points) I would like to see a complete and coherent analysis of the data. You should report

   (a) Basic descriptive statistics, exploratory analyses, what you learn from them.

   (b) Model fitting and refinement to arrive at a final model. Give tests and results for random and fixed effects.

   (c) Examination of assumptions for the model that you choose.

   (d) A complete interpretation of the model.

   (e) Answers to the following questions

      How do kids' attitude toward deviant behavior change as they get older? Is change in attitude a function of the respondent's age, their parent's attitude toward deviant behaviors, and/or their gender?

   (f) Write a paragraph of the results that might go in a newspaper article.

2. (10 points) Due to the nature of the design, investigate whether cohort has an effect on the results. The evidence that you give to support your conclusion regarding whether there is or is not a cohort effect can be statistical and/or graphical in nature.

Hint: Consider non-linear relationships and serial correlation.

More details regarding the data, design, and variables are described on the next page.

The Design There were 4 collection times. The values in the following table correspond to the age of the respondent when they responded to the questionnaire.

	(Year data collected)			
	1976	1977	1978	1980
Cohort	Wave 1	Wave 2	Wave 3	Wave 4
1	11	12	13	15
2	12	13	14	16
3	13	14	15	17
4	14	15	16	18
5	15	16	17	19
6	16	17	18	20
7	17	18	19	21

The variables in the data file:

Variable	Value	Definition
ID	Range: 32 to 98952	Youth's identification number
Year	1976, 1977, 1978, 1980	Year of survey
Age	ranges from 11 to 21	Age of the youth
Gender	1= male 2= female	
Cohort	1–7	Data collected over 4 years.
Youth_dev	9–36	Sum of item scores regarding kid's attitude toward deviate behaviors
Parent_Att	10–25	Sum of parents' item scores regarding approval of deviate behaviors. This was measured at the beginning of the study (i.e., 1976)

For the youth deviate behavior scale, kid's were asked whether they thought each of the following behaviors was wrong:

- Cheat on tests at school
- Damage or destroy property
- Marijuana use
- Steal less than $5
- Hit someone
- Alcohol use
- Break into car
- Steal more than $50
- Sell hard drugs

The response options and scoring of the youth's response were
1   not wrong
2   a little bit wrong
3   wrong
4   very wrong

For the parent attitude scale, parents were asked whether they approved of the same behaviors listed above (excluding selling hard drugs). The response options were
1   strongly approve
2   approve
3   neither
4   disapprove
5   strongly disapprove

For even more information on this and other related surveys go to www.bls.gov/nls/home.htm or www.icpsr.umich.edu (and search for "National Youth Longitudinal" as the title).

Option 3: For this option I downloaded some NELS data from
https://nces.ed.gov/EDAT, but some of the same variables we used in class. I did list
wise deletion so that there is no missing data and I have also not included weights.
Neither of these are "good" practice, but this is a final. This include data for all
schools, which means results could be very different than what we found in class.

The variables that you have are

variable	coding/description
sch_id	school id number
reading	student's estimated reading ability
math	student's estimated math ability
race	1 = Asian or Pacific Islander
	2 = Hispanic
	3 = Black
	4 = White
	5 = American Indian or Alaska Native
gender	1 = male
	2 = female
public	1 = public school
	0 = Catholic or other private school
pminority	percent of minority students in school

I would like to see a complete and coherent analysis of the data. You should report

1. Basic descriptive statistics, exploratory analyses, what you learn from them.

2. Model fitting and refinement to arrive at a final model. Be sure to explain
   details your process of coming up with a final model, including giving tests and
   results for random and fixed effects.

3. Examination of assumptions for the model that you choose.

4. A complete interpretation of the model.

5. Answer the following question

   What variables explain the variability in reading ability. Based on your
   results, are there any measures that could be taken to increase literacy
   among high school students? What other variables might you include to
   improve this analysis?

6. Write a paragraph of the results that might go in a newspaper article.

Option 4 The data used for this option come from Dorothy Espelage's research on the effects of bullying. In this data set, children are nested within peer groups. The variables in the data set are

peer The peer group to which a child belongs

bully A child's score on a bully scale. Higher values indicate more of a bully.

gender A child's gender

fight A child's score on a scale that measures the tendency to get into fights

empathy A child's score on the Davis empathy measured

use the data provided to answer the following questions:

Questions:

- Do children differ in terms of being a bully base on the child's gender, tendency to fight, and their empathy for others?
- Are there differences between children's levels of bullying due to the peer group to which they belong? If there are differences, describe them.

You should report a complete and coherent analysis of the data that includes

1. Exploratory analyses and what you learned from them (don't forget descriptive statistics).
2. Model fitting and refinement to arrive at a final model. Give tests and results for fixed and random effects.
3. Examination of the assumption for the model that your choose.
4. A complete interpretation that addresses the questions given above.
5. Write a paragraph of the results that might go in a newspaper article.

Option 5 The data for this option is from https://github.com/washingtonpost/data-police-shootings, which was retrieved 11/12/2021. The data contain information about incidents of deaths due to police shootings. The data online were complied by the Washington Post and cover 2015 through today. For this option, you only need to use data from 2019.

The data from the internet was modified for this final as follows:

- A subset of variables were retained
- Missing values were deleted (relative small number)

- Only victims who were Black, White, or Hispanic were retained.
- Victims age range from 1477 (very young and very old victims were deleted)

The variables that you will be using are

- age
- gender (M=male, Female)
- race (B =black, W=white, H=Hispanic)
- signs .01inof .01inmental .01inillness (TRUE, FALSE)
- threat .01inlevel (attack, other, undetermined)
- flee (whether or how fleeing)
- body camera (whether a body camera was used)
- state

The goal of this analysis is to describe who gets shot and under what circumstances with particular attention paid to the race of the victim. Using these data answer the **ONE** of the following questions.

Research Questions 1:

- Are victims more likely to be non-white and under what circumstance (i.e., binary logistic regression)
- Are there differences between states? If so, describe them.

OR s Research Questions 2:

- Are Black, Hispanic, or White more likely be the victim and under what circumstance (i.e., multicategory logistic regression).
- Are there differences between states? If so, describe them.

You should report a complete and coherent analysis of the data that includes the following:

- Which set of research questions that you chose to answer.
- Exploratory analysis (e.g., descriptive statistics, plots, etc) and what you learned from these.

- For each question report your model fitting and refinements to arrive at a final model. Give tests and results. Include enough information so that I can see the process, logic, and your decision making. I want to know how you arrived final models.

- For your final model(s), examine the assumptions of the model. If needed, make change to you analyses.

- Report the final model(s).

- A complete interpretation that answers each of the questions above.

- Write a paragraph that might go into an article in the newspaper.

If you want to include R code or scripts, you may do so in an Appendix; however, I may not look at these. What I dont want to see in the body of your exam, is page after page of output from R, especially without some sort of annotation.