P8130: Biostatistical Methods I
Final Project (Fall 2019)
Due, December 16th @ 5:00pm

Guidelines for Project Submission

This group project must be submitted through CourseWorks before the deadline. Email submissions WILL NOT be accepted and will receive a score of 'Zero' for all group members!!

All graphs, output and interpretations must be included in **ONE PDF** (not the R file), otherwise it will not be graded. In a separate attachment, you also have to submit the R/Rmd code used in your project.

General Writing Instructions

Your project should not exceed **5 double-spaced pages** using 11 or 12-point font, EXCLUDING figures and tables, references, appendix, that can be placed at the end of the five summary pages. Be selective in your output and visual displays!

Your report should be structured as a publishable research article containing the following sections:

- Abstract
- Introduction (context, background of the problem)
- Methods (data description and statistical methods)
- Results
- Conclusions/Discussion

Your findings should be written as for an informed (but non-statistical) audience (**no formulae!**). Each figure and table should be of publishable quality and well notated, i.e., labeled and/or captioned.

Grading Instructions

The rubric attached will be used to evaluate the project. This is a group project and collaborations within your group are essential and great practice for your career.

Academic dishonesty will be punished with a 'Zero' grade for this project.

A few years ago, the United States District Court of Houston had a case that arises under Title VII of the Civil Rights Act of 1964, 42 U.S.C. 200e et seq. The plaintiffs in this case were all female doctors at Houston College of Medicine who claimed that *the College has engaged in a pattern and practice of discrimination against women in giving promotions and setting salaries*. The lead plaintiff in this action, a pediatrician and an assistant professor, was denied for promotion at the College. The plaintiffs had presented a set of data to show that female faculty at the school were less likely to be full professors, more likely to be assistant professors, and earn less money than men, on average.

The main question that you are asked to address in this project is if the data support the claim of gender discrimination in setting salaries.

Dataset 'Lawsuit.csv' contains the following variables (N=261):

1	Dept	1= Biochemistry/Molecular Biology 2= Physiology 3= Genetics 4= Pediatrics 5= Medicine 6= Surgery
2	Gender	1= Male, 0= Female
3	Clin	1= Primarily clinical emphasis, 0= Primarily research emphasis
4	Cert	1= Board certified, 0= not certified
5	Prate	Publication rate (# publications on cv) / (# years between CV date and MD date)
6	Exper	# years since obtaining MD
7	Rank	1= Assistant, 2= Associate, 3= Full professor (a proxy for productivity)
8	Sal94	Salary in academic year 1994
9	Sal95	Salary after increment to Sal94

Note: Even though the response variable was recorded over a two-year period, you do not need to fit a model accounting for repeated measures. Thus, you may fit regression models using each year separately and/or a summary of the two years. In your multiple regression models, you should carefully choose relevant adjusting factors to support your final recommendation in this lawsuit.

Aspects to be addressed in your report:

- Data exploration: descriptive and visualization
    - You might want to include a *Table 1* summarizing all variables by *Gender* (the main covariate of interest)
    - Explore the distribution of the outcome(s) you chose and consider potential transformations
- Consider confounders and interaction of/with the main covariate of interest, and if the case, fit stratified models
- Model diagnostics
    - Heteroscedasticity, normality and multicollinearity
    - Functional form for continuous predictors
- Outliers/influential points