# DS-UA 201: Final Exam

Due December 20, 2023 at 5pm

## Instructions

*You should submit your write-up (as a knitted .pdf along with the accompanying .rmd file) to the course website before 5pm EST on Wednesday, Dec 20th Please upload your solutions as a .pdf file saved as* `Yourlastname_Yourfirstname_final.pdf`*.In addition, an electronic copy of your .Rmd file (saved as* `Yourlastname_Yourfirstname_final.Rmd`*) should accompany this submission.*

*Late finals will not be accepted,* ***so start early and plan to finish early****.*

*Remember that exams often take longer to finish than you might expect.*

*This exam has* ***3*** *parts and is worth a total of* ***100 points****. Show your work in order to receive partial credit.*

*Also, we will penalize uncompiled .rmd files and missing pdf or rmd files by 5 points.*

*In general, you will receive points (partial credit is possible) when you demonstrate knowledge about the questions we have asked, you will not receive points when you demonstrate knowledge about questions we have not asked, and you will lose points when you make inaccurate statements (whether or not they relate to the question asked). Be careful, however, that you provide an answer to all parts of each question.*

*You may use your notes, books, and internet resources to answer the questions below. However, you are to work on the exam by yourself. You are prohibited from corresponding with any human being regarding the exam (unless following the procedures below).*

*The TAs and I will answer clarifying questions during the exam. We will not answer statistical or computational questions until after the exam is over. If you have a question, send email to all of us. If your question is a clarifying one, we will reply. Do not attempt to ask questions related to the exam on the discussion board.*

# Problem 1 (100 points)

In this problem, you will examine whether family income affects an individual's likelihood to enroll in college by analyzing a survey of approximately 4739 high school seniors that was conducted in 1980 with a follow-up survey taken in 1986.

This dataset is based on a dataset from

> Rouse, Cecilia Elena. "Democratization or diversion? The effect of community colleges on educational attainment." Journal of Business & Economic Statistics 13, no. 2 (1995): 217-224.

The dataset is `college.csv` and it contains the following variables:

- `college` Indicator for whether an individual attended college. (Outcome)
- `income` Is the family income above USD 25,000 per year (Treatment)
- `distance` distance from 4-year college (in 10s of miles).
- `score` These are achievement tests given to high school seniors in the sample in 1980.
- `fcollege` Is the father a college graduate?
- `tuition` Average state 4-year college tuition (in 1000 USD).
- `wage` State hourly wage in manufacturing in 1980.
- `urban` Does the family live in an urban area?

## Question A (35 points)

Draw a DAG of the variables included in the dataset, and explain why you think arrows between variables are present or absent. You can use any tool you want to create an image of your DAG, but make sure you embed it on your compiled .pdf file. Assuming that there are no unobserved confounders, what variables should you condition on in order to estimate the effect of the treatment on the outcome, according to the DAG you drew? Explain your decision in detail. In your explanation, provide a definition of confounding.

## Question B (35 points)

Choose one of the methodologies we learned in class to calculate a causal effect under conditional ignorability. What estimand are you targeting and why? Explain why you made your choice, and discuss the assumptions that are needed to apply your method of choice to this dataset. State if and why you think these assumptions hold in this dataset. In addition, choose a method to compute variance estimates (i.e., robust standard errors or bootstrapping), and discuss the reasons behind your choice in the context of this dataset.

## Question C (30 points)

Using the methodology you chose in Question B to control for the confounders you have selected in Question A, as well as the relevant R packages, provide your estimate of the causal effect of the treatment on the outcome. Using your variance estimator of choice, report standard errors and 95% confidence intervals around your estimates. Interpret your results and discuss both their statistical significance and their substantive implications. Be as specific and detailed as possible.