**UNIVERSITY OF TORONTO**
**DEPARTMENT OF SOCIOLOGY**
**PH.D. COMPREHENSIVE EXAMINATION IN QUANTITATIVE METHODS**
**August 22-26, 2022**

---

You are required to answer THREE (3) QUESTIONS (ONE QUESTION FROM EACH OF PARTS A, B & C). Each answer should be 10-12 pages (12-point Times New Roman font, standard margins, and double-spaced) in length. The complete exam should not total more than 36 pages in length (12-point Times New Roman font, standard margins, and double-spaced), not including references. THE COMMITTEE WILL NOT READ PAST THE 36TH PAGE.

## Part A

1. Imagine you are a researcher who has been tasked with describing the disparities in PhD program experiences of students from different racial backgrounds. You have access to data from a major university on a cohort of students who entered a PhD program in the same year. The data consists of information collected at the start of the PhD program, including the self-identified race of each student and other demographic information, as well as many relevant individual-level measures of previous educational and family backgrounds, personal characteristics, research interests, and so on. The dataset also contains three outcome variables collected eight years after the students began their program. These include: a binary measure of whether or not the student reported being satisfied with their PhD experience, a count of the number of publications the student had during the eight years, and the date of their successful dissertation defense (for the 80% of students who completed the program by year eight; assume there were no drop-outs).

For each outcome variable explain why linear regression may be inappropriate and suggest alternative modeling approaches. For each alternative modeling approach, discuss the key assumptions and common, practical challenges of estimating and interpreting the model results. Finally, critically assess what you can expect to learn about your research topic (disparities in PhD program experiences) from the analyses of each variable.

2. Individual-level observational data in the social sciences are often organized in groups at higher levels: students attend schools, patients receive treatment from doctors who work in specific hospitals, and citizens in a particular nation live in some city and neighborhood. Longitudinal observations are also inherently nested within individual units of observation.

Sometimes social scientists treat the grouped structure of data as a nuisance that needs to be adjusted or corrected. Other times, however, social scientists view the hierarchical structure as a feature of the data that is of substantive interest. Discuss these two scenarios. What are the motivating goals in each? What is gained and lost in each approach/interpretation? Finally, identify and discuss the commonly used modeling approaches in each scenario, including key assumptions, data requirements, and common practical challenges.

**Part B**

3. In recent years counterfactual reasoning has been applied to the mediation of causal effects. Suppose a sociologist is interested in examining how the effect of a composite measure of parental occupational prestige (the average occupational prestige of both parents) on a set of respondents' occupational prestige is mediated by educational attainment measured in years. Assume that all three variables are continuous. Using this simple example, explain the conceptual and mathematical differences between a total causal effect, natural direct and indirect effects, and the controlled direct effect. State clearly the assumptions required to estimate these quantities and the pros/cons of each in understanding causal effects. In your response, make sure to explain how these quantities can be estimated using conventional regression models and how counterfactual-based mediation analysis is related to path analysis as it has developed in sociology.

4. The editor of a prominent sociology journal decides to issue a series of guidelines for quantitative researchers submitting articles. The guidelines are as follows:

> i. A cut-off of $\alpha = 0.10$ and two-tailed statistical tests will be used to determine whether or not findings are statistically significant.
>
> ii. Submitted articles with statistically insignificant findings do not warrant publication in the journal.
>
> iii. Results should be presented primarily using tables, not graphical displays.
>
> iv. Structural equation models and hierarchical/multilevel models should only be used to present descriptive results.
>
> v. Except in rare circumstances, complex models are preferred over simple models.
>
> vi. If conducting an observational study, causal effects should be estimated using fixed effects regression, regression discontinuity design, and/or instrumental variables.
>
> vii. Estimated causal effects from randomized experiments are privileged over those from all other techniques.

Critically evaluate the above guidelines, discussing the pros and/or cons of each.

**Part C**

5. You are interested in studying the characteristics of people's friendship groups and how those characteristics relate to individual-level outcomes, particularly economic measures. You have at your disposal individual-level data sourced from a social media website, which contains information about social interactions (comments on posts, tags, etc.) on the website, as well as a wide variety of individual-level characteristics.

a) While the social media website is very popular, not everyone in the population you are interested in has an account, and not everyone that has an account is active on the website. Given you are interested in economic measures, what are some possible issues with using these data to make inferences about the broader population?

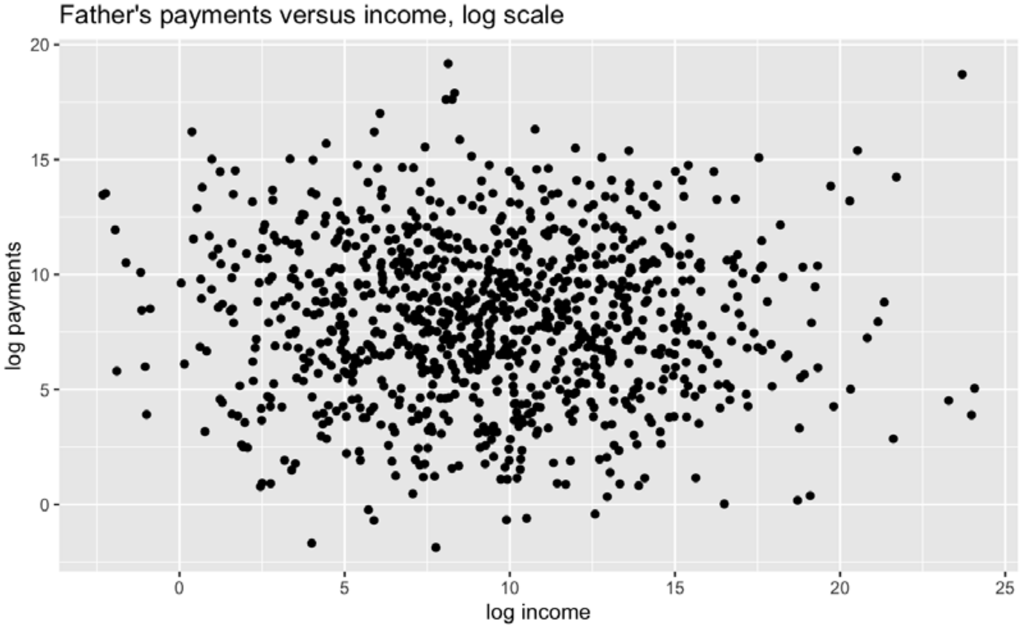
b) The data do not contain information on individual-level income; however, for around 20% of the sample you have information on location of residence (down to 'census block', a very small area) and you know the median income of each census block. As such, you decide to estimate individual level income as follows:

   1) Regress median income of each census block on a series of individual level characteristics (such as age, education, marital status, gender…)
   2) Use these estimates to predict the income of individuals that do not have location information

   Briefly discuss the advantages and disadvantages of this approach, particularly in how it could affect the study of income characteristics of friendship groups.

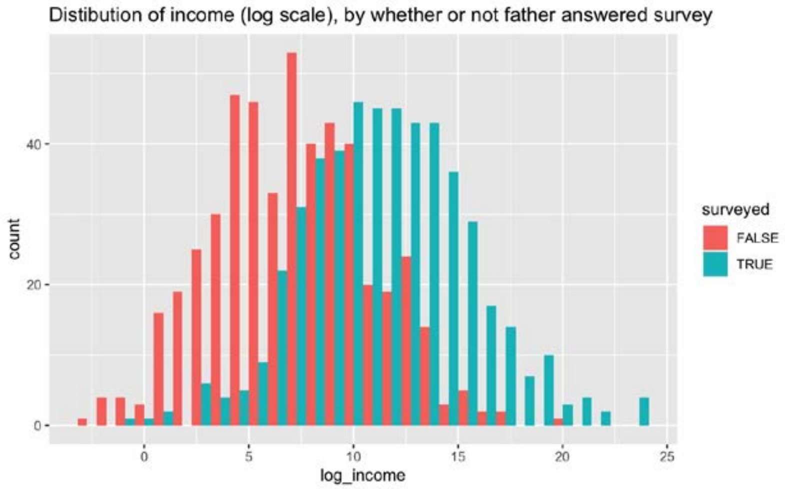
6. This question relates to a dataset of 1000 fathers who are divorced and required to pay child support payments. The `income` variable refers to the father's income, the `payment` variable refers to the amount of child support payments paid monthly. The fathers were asked to respond to a survey, and the `surveyed` variable is TRUE if they responded to the survey and FALSE otherwise.

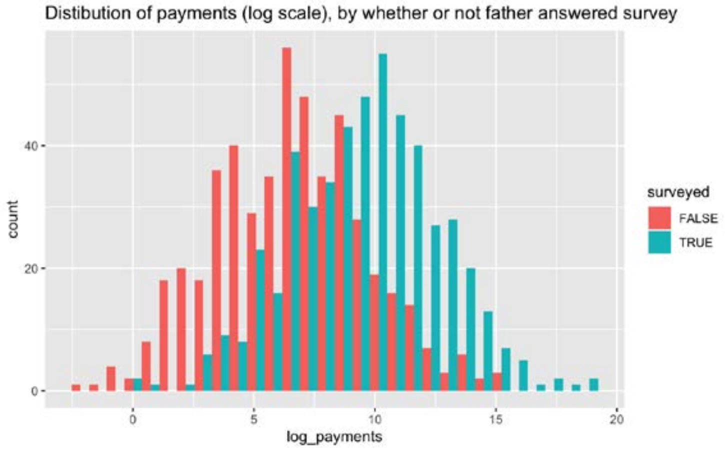
a) Below is a scatterplot of log income (x axis) and log payment (y axis), and the R output results of a simple linear regression model of log(payments) versus log(income). Interpret what you observe.



```
Call:
lm(formula = log_payments ~ log_income, data = d)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.9986 -2.2685  0.0538  2.3860 11.0476 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.32649    0.25572  32.561   <2e-16 ***
log_income  -0.02476    0.02459  -1.007    0.314    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.451 on 998 degrees of freedom
Multiple R-squared:  0.001015,	Adjusted R-squared:  1.389e-05 
F-statistic: 1.014 on 1 and 998 DF,  p-value: 0.3142
```

b) Below are histograms of log income and log payments by whether or not the fathers answered the survey, as well as the R output results of a simple linear regression model of log(payments) versus log(income), restricted just to fathers who answered the survey. Interpret what you observe.





```
Call:
lm(formula = log_payments ~ log_income, data = d %>% filter(surveyed))

Residuals:
    Min      1Q  Median      3Q     Max 
-8.3669 -1.9692 -0.0274  1.9008 12.0502 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  12.3878     0.4038  30.679  < 2e-16 ***
log_income   -0.2420     0.0330  -7.333 9.01e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.933 on 504 degrees of freedom
Multiple R-squared:  0.09642,	Adjusted R-squared:  0.09462 
F-statistic: 53.78 on 1 and 504 DF,  p-value: 9.014e-13
```

c) Comment briefly on what conclusions you make from this analysis, keeping in mind that usually we would only have data (and therefore be making inferences) based on surveyed fathers. Are all fathers equally likely to respond to the survey? Why or why not? How does survey response affect the conclusions we make from our analysis?