

Part 1: Multiple Choice Questions (40 points)

Circle the right answer. Only one answer per question. No credit is given for multiple answers or additional explanations. Two points per question for correct answers.

- 1) Consider the regression model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$. Suppose you want to interpret β_1 and β_2 causally. Then you would have to assume that:
 - a. $E(u_i | X_{1i}, X_{2i}) = E(u_i | X_{2i})$.
 - b. $E(u_i | X_{1i}, X_{2i}) = E(u_i | X_{1i})$.
 - c. $E(u_i) = 0$.
 - d. $E(u_i | X_{1i}, X_{2i}) = 0$.

- 2) Consider the regression model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$. Suppose that X_{2i} is unobserved. Omitted variables bias is a concern if
 - a. X_{1i} and X_{2i} are correlated and $\beta_2 = 0$.
 - b. X_{1i} and X_{2i} are uncorrelated and $\beta_2 \neq 0$.
 - c. X_{1i} and X_{2i} are uncorrelated and $\beta_2 > 0$.
 - d. X_{1i} and X_{2i} are correlated and $\beta_2 \neq 0$.

- 3) If the estimates of the coefficient of interest changes substantially across specifications, then
 - a. this can be expected from sample variation.
 - b. you should change the scale of the variables to make the changes appear to be smaller.
 - c. this suggests that the original specification had omitted variable bias.
 - d. choose the specification for which your coefficient of interest is most significant.

- 4) The slope coefficient in the model $\ln Y_i = \beta_0 + \beta_1 X_i + u_i$ can approximately be interpreted as follows:
 - a. a change in X by one unit is associated with a $(100 \times \beta_1)$ percent change in Y .
 - b. a 1 percent change in X is associated with a β_1 percent change in Y .
 - c. a 1 percent change in X is associated with a change in Y of $(0.01 \times \beta_1)$.
 - d. a change in X by one unit is associated with a β_1 change in Y .

- 5) In the regression model $Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i) + u_i$, where X is a continuous variable and D is a binary gender dummy variable, β_3
 - a. indicates the slope of the regression when $D_i = 1$.
 - b. indicates the slope of the regression when $D_i = 0$.
 - c. indicates the difference between the genders in the slopes of the regression
 - d. has no meaning since $(X_i \times D_i)$ when $D_i = 0$.

- 6) Consider the regression model $Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i) + u_i$, where X_i is a continuous variable and D_i is a binary gender dummy variable. To test that the two regressions are identical across the genders, you must use the
- t -statistic separately for the hypotheses $\beta_2 = 0, \beta_3 = 0$.
 - F -statistic for the joint hypothesis that $\beta_2 = 0, \beta_3 = 0$.
 - t -statistic for the hypothesis $\beta_3 = 0$.
 - F -statistic for the joint hypothesis that $\beta_2 = 0, \beta_3 = 0$.
- 7) The following problems could be analyzed using probit and logit estimation with the exception of whether or not
- a college student decides to study abroad for one semester.
 - being a female has an effect on earnings.
 - a college student will attend a certain college after being accepted.
 - an applicant will default on a loan.
- 8) One of the following statements is not true. In the linear probability model,
- $\Delta \Pr(Y = 1) / \Delta X_1 \neq \beta_1$.
 - the errors are heteroskedastic.
 - predicted probabilities can be greater than unity.
 - predicted probabilities can be less than zero.
- 9) Errors-in-variables bias can be mitigated by
- having access to panel data.
 - having an alternative measure of the variable of interest.
 - specifying a non-linear regression.
 - increasing sample size sufficiently.
- 10) You are interested in the effects of participating in a training program (which may be of varying length). You have data on hourly wages after program completion for those who participated in the program and for a potential comparison group. A major concern for this study is:
- misspecification of the functional form.
 - sample selection bias.
 - bias caused by a so-called Hawthorne effect.
 - that you miss information on program length.
- 11) You want to estimate the price elasticity of cigarette demand. To do that you collect time series data on prices and quantities sold in the Stockholm area. The major concern for such a study is:
- simultaneous causality bias.
 - errors in variables bias.
 - wrong functional form.
 - sample selection bias.

- 12) Consider the panel data model: $Y_{it} = \alpha_i + \beta_1 X_{it} + u_{it}$. Information on Y_{it} and X_{it} is available for 48 US states over 10 years. You can estimate β_1 in three ways: (i) Define a dummy variable for each US state and estimate the entire model using OLS; (ii) transform the model by “demeaning” the data and estimate the transformed model using OLS; (iii) transform the model by “first-differencing” the data and estimate the transformed model using OLS. Which of the following statements is correct?
- (i) and (ii) yield identical estimates of β_1 .
 - (i), (ii), and (iii) yield identical estimates of β_1 .
 - (i) and (iii) yield identical estimates of β_1 .
 - (ii) and (iii) yield identical estimates of β_1 .
- 13) Consider a standard panel data setting. Heteroscedasticity robust standard errors are invalid in large samples if
- the errors are homoskedastic
 - the error variance differs across units
 - the dependent variable is binary
 - the errors are serially correlated within unit over time
- 14) You consider adding individual gender (the effect is measured by β_1) and national unemployment (the effect is measured by β_2) to a panel data model with individual and time fixed effects. Then
- only β_1 is identified (i.e. can be estimated).
 - only β_2 is identified.
 - Both β_1 and β_2 are identified.
 - Neither β_1 nor β_2 are identified.
- 15) When there is a single instrument and a single (endogenous) regressor, the TSLS estimator for the slope can be calculated as follows ($\widehat{cov}(\cdot)$ ($\widehat{var}(\cdot)$) denotes estimated covariance (variance))
- $\hat{\beta}_1 = \widehat{cov}(X, Y) / \widehat{var}(X)$.
 - $\hat{\beta}_1 = \widehat{cov}(Z, X) / \widehat{cov}(Z, Y)$.
 - $\hat{\beta}_1 = \widehat{cov}(Z, Y) / \widehat{cov}(Z, X)$.
 - $\hat{\beta}_1 = \widehat{cov}(Z, Y) / \widehat{var}(Z)$.
- 16) Among other things a valid instrument should satisfy an “exclusion restriction”. If the exclusion restriction is not satisfied
- it is not possible to estimate the reduced form equation for Y .
 - the instrument is not exogenous.
 - the instrument is not relevant.
 - TSLS is consistent.

- 17) Experimental effects, such as the Hawthorne effect,
- typically do not arise in quasi-experiments.
 - typically require instrumental variable estimation in quasi-experiments.
 - can be dealt with using binary variables in quasi-experiments.
 - are the most important threat to internal validity in quasi-experiments.
- 18) You are interested in the effects of reducing the pay-roll tax. There are 2 comparable groups on the labor market. In one of the following examples you can apply a differences-in-differences approach:
- the pay-roll tax reduction applies to both group 1 and group 2.
 - the pay-roll tax reduction applies to group 1 but you cannot identify group 1 and 2 in the data.
 - the pay-roll tax reduction applies to group 1 and you can identify group 1 and 2 in the data.
 - the pay-roll tax reduction applies to group 1 and group 2 is affected by a lowering of the income tax at the same point in time.
- 19) In a sharp regression discontinuity design:
- the "common-trends" assumption is key.
 - you get the "treatment" if the assignment variable is above/below a known threshold.
 - being above/below a known threshold influences the probability of getting the treatment.
 - you must control for pre-determined characteristics.
- 20) With one exception, all of the following are reasons for adding control variables to a regression based on data from an experiment:
- efficiency.
 - providing a check for randomization.
 - providing an adjustment for "conditional" randomization.
 - adding control variables always reduces omitted variables bias.

Part 2: Discussion Questions (60 points)

Answer the following questions on separate sheets of paper. Answer clearly and concisely. Only legible answers will be considered. If you think that a question is vaguely formulated, specify the conditions used for answering it. Each question is worth 30 points.

Discussion Question 1

A recent paper analyzed the effects of payroll tax reductions on employment using Swedish data. The payroll tax reduction amounted to 10 percentage points and was implemented in 2002. Only firms located in a particular region of Sweden – Regional Support Area A (located in the northern inland of Sweden) – were *eligible* for the reduction; other firms were not. To get the tax reduction, eligible firms had to file an application to the tax authority (but all firms did not do that). The authors had access to annual firm-level data on employment and paid payroll taxes for the time period 2001–2004. The data set included all firms in Sweden during these years.

- a) Explain how you would estimate the effect of eligibility for the payroll tax reduction on employment in this setting. Indicate how you would specify the key regression(s) and be clear on how you define key variables of interest.
- b) Explain how you would estimate the effect of the payroll tax on employment in this setting. Indicate how you would specify the key regression(s) and be clear on how you define key variable(s) of interest.
- c) State the key identifying assumption(s). To what extent can you provide evidence on this (these) assumption(s) in this setting?

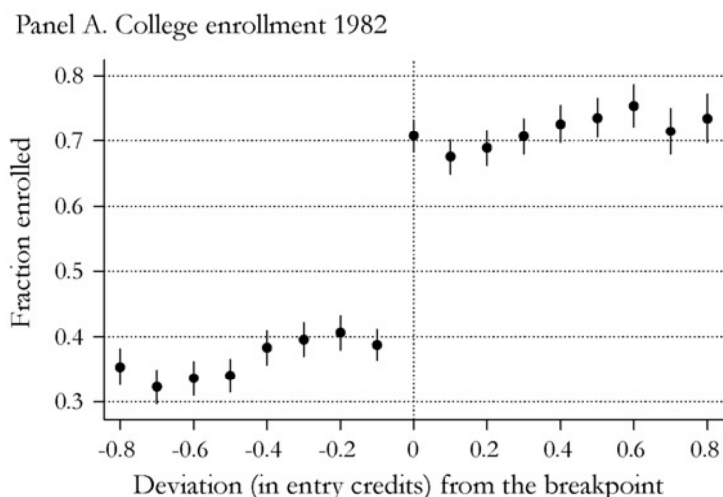
Discussion Question 2

Suppose you are interested in estimating the (causal) return to college education. Thus, you want to estimate the relationship:

$$\ln Y_i = \beta_0 + \beta_1 \text{College}_i + \beta_2 X_i + u_i$$

where i indexes individuals, $\ln Y$ denotes (the log of) annual earnings, College years of college education, and X a set of control variables.

- Consider estimating the above equation by OLS. Why is the OLS-estimate of β_1 likely biased? What is the likely sign of the bias?
- Some researchers have noted that admission rules can be useful for identifying the causal effect of years of college. Admission to a given program is typically based on entry credits (for instance the grade point average from upper secondary school). Those who are above the entry credit breakpoint are admitted to the program; those who are below the breakpoint are not admitted to the program (but may be admitted to another program). The figure below comes from a recent paper using admission data from 1982

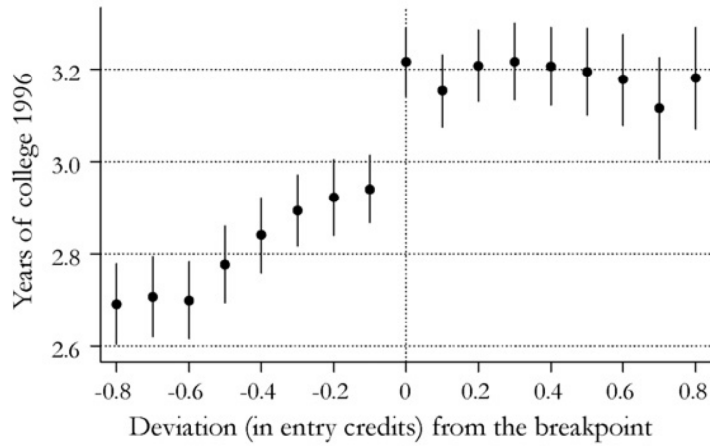


Note: Circles are local averages. Solid lines around each circle show the 95% confidence interval.

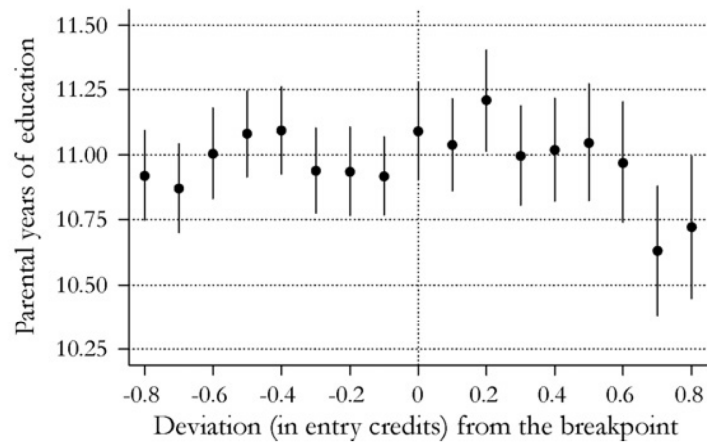
Explain how the admission rule may help you in estimating the causal effect of years of college. How would you specify the regression(s) that you would use to estimate the causal effect of interest? Please specify the regressions explicitly.

- The three figures below come from the same paper. Explain (briefly) why each of the three figures is important and what you would conclude from each of the three figures. (Please make sure to specify which figure you are talking about. For instance: "Panel B shows...It is important because...From Panel B I conclude...")

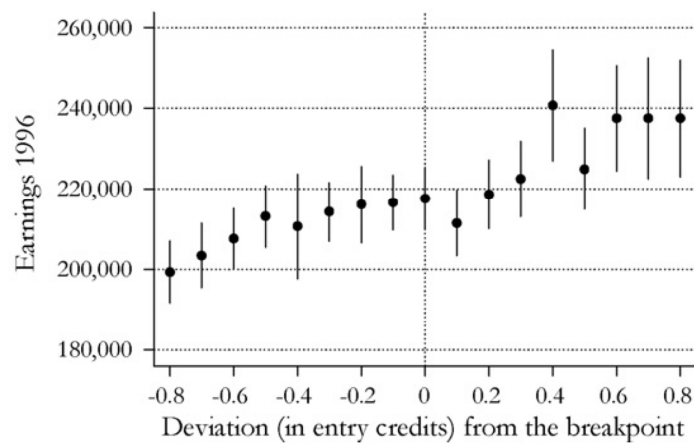
Panel B. Years of college 1996



Panel C. Parental years of education



Panel D. Earnings 1996



Note: Circles are local averages. Solid lines around each circle show the 95% confidence interval.