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 $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$, where X_{1i} is a variable of interest and X_{2i} a control variable. To causally interpret β_1 and β_2 you must assume that:
   a. $E(u_i|X_{1i}) = E(u_i|X_{2i})$.
   b. $E(u_i|X_{1i}, X_{2i}) = E(u_i|X_{1i})$.
   c. $E(u_i|X_{1i}, X_{2i}) = 0$.
   d. $E(u_i|X_{2i}) = E(u_i)$.

2) Consider the regression model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$, where $\beta_1 > 0$. Suppose that X_{2i} is unobserved and X_{1i} and X_{2i} are positively correlated. The OLS estimate of β_1 is biased upwards if
   a. $\beta_2 > 0$.
   b. $\beta_0 < 0$.
   c. $\beta_2 < 0$.
   d. $\beta_0 > 0$.

3) Consider the regression model $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$. The effect of changing X (marginally) on Y is given by
   a. $\beta_1 + \beta_2$
   b. β_1
   c. β_2
   d. $\beta_1 + 2\beta_2 X$

4) To test whether or not the population regression function is linear or whether it is better described by the regression in 3), you estimate the regression in 3) and
   a. check whether the regression R^2 is higher than that of the linear regression.
   b. test whether $\beta_1 = 0$ and $\beta_2 = 0$ using an F-test.
   c. test whether $\beta_2 = 0$ using a t-test.
   d. check whether the Total Sum of Squares (TSS) is higher than that of the linear regression.

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

6) Consider two regression models: $Y_i = \gamma_0^m + \gamma_1^m X_i + e_i$ and $Y_i = \gamma_0^f + \gamma_1^f X_i + e_i$. The first model applies to males, the second to females, and X_i is a continuous variable. Define M_i as a male dummy variable and pool the two regressions $Y_i = \beta_0 + \beta_1 X_i + \beta_2 M_i + \beta_3 (M_i \times X_i) + u_i$. In the pooled regression, β_3
   a. indicates the slope of the regression for a male.
   b. indicates the male-female difference in the slopes of the two regressions.
   c. indicates the slope of the regression for a female.
   d. shows the male-female difference in the mean of the dependent variable.

7) The following problems could be analyzed using probit or logit estimation <u>with the exception</u> of whether or not
   a. a college student decides to study abroad for one semester.
   b. a college student will attend a certain college after being accepted.
   c. being a male has an effect on earnings.
   d. an applicant will default on a loan.

8) In the Logit model,
   a. predicted probabilities are linear in the parameters.
   b. $\Delta \Pr(Y = 1) / \Delta X_1 \neq \beta_1$.
   c. predicted probabilities can be greater than unity.
   d. predicted probabilities can be less than zero.

9) Errors-in-variables bias
   a. vanishes when sample size is very large.
   b. arises from error in the measurement of the independent variable.
   c. can be mitigated with panel data.
   d. arises from error in the measurement of the dependent variable.

10) Sample selection bias occurs when:
   a. data are missing at random.
   b. data are missing based on the values of the control variables.
   c. the choice between two samples is made by the researcher.
   d. data are missing based on the values of the dependent variable.

11) External validity
   a. is guaranteed in an ideal randomized experiment.
   b. is guaranteed if you have access to quasi-experimental variation.
   c. is threatened if there is omitted variables bias.
   d. is threatened if there is measurement error in the dependent variable.

12) 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:

a. simultaneous causality bias.
b. errors in variables bias.
c. wrong functional form.
d. sample selection.

13) Consider the panel data model: $Y_{it} = \alpha_i + \beta_1 X_{it} + u_{it}$. Information on *Y* and *X* is available for 48 US states over 2 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?

a. (i) and (ii) yield identical estimates of β_1.
b. (i), (ii), and (iii) yield identical estimates of β_1.
c. (i) and (iii) yield identical estimates of β_1.
d. (ii) and (iii) yield identical estimates of β_1.

14) Consider a standard panel data setting. Heteroscedasticity robust standard errors are invalid in large samples if

a. the errors are homoskedastic
b. the error variance differs across units
c. the dependent variable is binary
d. the errors are serially correlated within unit over time

15) The panel data model with entity and time fixed effects

a. handles any kind of omitted variables bias.
b. always produces internally valid estimates.
c. deals with simultaneous causality bias.
d. requires that the variable of interest varies over entities and time.

16) In one of the following cases, TSLS estimation is not possible:

a. the number of instruments equals the number of endogenous regressors.
b. the model is over-identified.
c. the model is under-identified.
d. the model is exactly identified.

17) Consider the simple regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$. Suppose *X* and *u* are correlated, and that you have one instrument that you can use to estimate β_1. In this setting

a. the TSLS is unbiased (provided the instrument is valid).

b. the TSLS estimator is consistent (provided the instrument is valid).

c. You can use the *J*-statistic to test the validity of your model.

d. OLS and TSLS produce the same estimate in large samples.

18) In a difference-in-differences approach you estimate the causal effect of a policy change by:

a. comparing the change in the outcome across the treatment and the comparison group.

b. calculating the change in the outcome for the treatment group.

c. comparing the mean of the outcome after the policy change across the treatment and the comparison group.

d. calculating the change in the outcome for the comparison group.

19) In the fuzzy Regression Discontinuity design:

a. you get the "treatment" if the assignment variable is above/below a known threshold.

b. the threshold is unknown.

c. being above/below a known threshold influences the probability of getting the treatment.

d. you must control for pre-determined characteristics.

20) In the ideal randomized experiment

a. you can estimate the individual causal effects for all individuals participating in the experiment.

b. you must control for variables that are correlated with the dependent variable.

c. self-selection bias is a serious issue.

d. you can estimate the average causal effect for individuals participating in the experiment.

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

The current Swedish government has lowered payroll taxes for youths. A recent paper has analyzed the employment effects of such payroll tax reductions using Swedish data.[1] The authors analyzed a payroll tax reduction implemented in 2007. This reform reduced payroll taxes by 11 percentage points for all individuals who were aged 19-25. The authors had access to annual individual-level data for the time period 2001-2010. The data contain information on employment and standard individual characteristics, such as education, gender, birth year, and immigrant status. The data set included all individuals living in Sweden during these years.

a) Explain how you would estimate the effect of 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) State the key identifying assumption(s). What are the threats to identification? And how would you provide evidence on this (these) assumption(s) in this setting?

[1] Egebark, J. and N. Kaunitz (2013). Do payroll tax cuts raise youth employment? IFAU Working Paper 2013:27.

Discussion Question 2

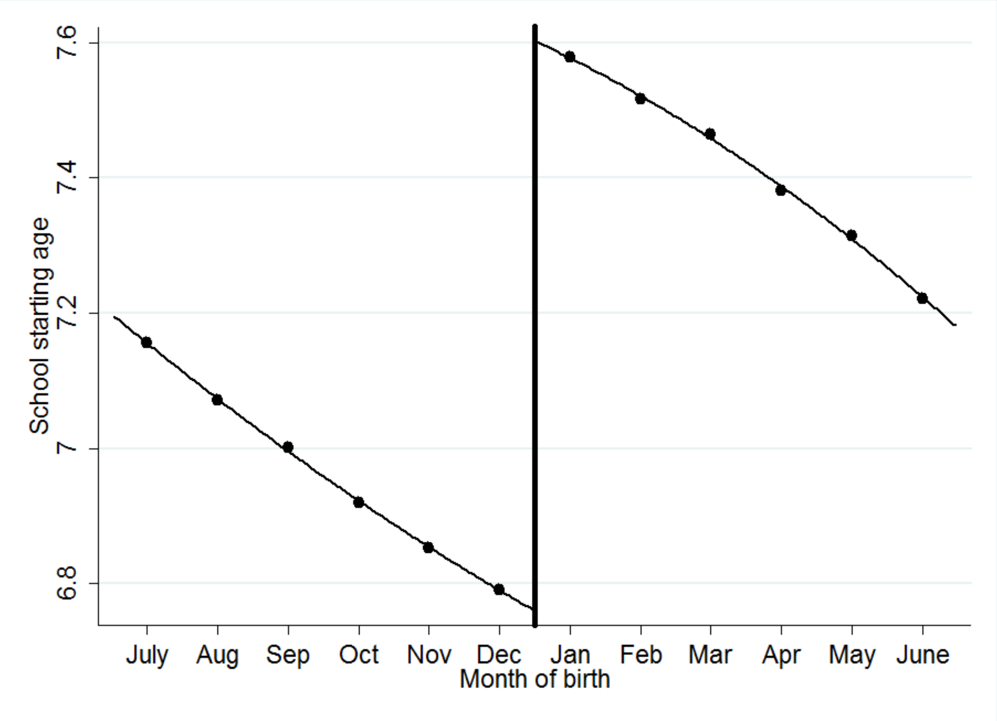
Suppose you are interested in estimating the causal effect of school starting age on educational attainment. You want to estimate the relationship:

$$ED_i = \beta_0 + \beta_1 SSA_i + \beta_2 X_i + u_i$$

where *i* indexes individuals, *ED* denotes years of schooling, *SSA* school starting age, and *X* a set of control variables.

a) 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?

b) Some researchers have noted that school entry age rules can be useful for identifying the causal effect of school starting age. In Sweden, you typically start formal schooling the year you turn 7. Those born in January are generally older when they start school than those born in December. Note that the rule is not binding; children can start earlier or later than normal. The figure below plots the relationship between school starting age and month of birth using data for individual born in the late 1940s.



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

c) A regression of school starting age on an indicator for being born January-June (as opposed to July-December) and relevant control variables yields an estimate on the indicator of 0.858 (with a standard error of 0.018). Interpret the estimate. Why does it deviate from 1? What does the estimate tell you about the validity of the research design?

d) A regression of individual years of schooling on mother's years of schooling yields an estimate of 0.230 (with a standard error of 0.004), suggesting that the education of the child is positively correlated with the education of the parent. Does this imply that you should control for the education of the mother in the research design you outlined in b)? Why or why not?