1. **Multiple choice (20 points, 4 points each)** Please tick (*Kryssa för*) the correct answer. Only one answer is correct.

   (a) (4 points) In the naïve regression $Y_i = \beta_0 + \beta_1 X_i + u_i$, where the causal relationship of interest is the effect of X_i on Y_i, what can β_1 capture?

   A) The causal effect of X_i on Y_i
   B) Reverse causality (the effect of Y_i on X_i)
   C) Omitted variables bias (the correlation between Y_i and X_i that results from an omitted variable W_i affecting both X_i and Y_i)
   D) All of the above

   (b) (4 points) Say we estimate the naïve regression $Y_i = \beta_0 + \beta_1 X_i + u_i$, where the causal relationship of interest is the effect of X_i on Y_i. If there is an omitted variable W_i which is *negatively* correlated with X_i and has a *negative* effect on Y_i, what is the sign of the omitted variables bias in β_1?

   a) Positive
   b) Negative
   c) It does not create bias in β_1
   d) There is not enough information to determine the sign of the bias.

   (c) (4 points) Say I randomly assign a treatment variable X_i to individuals, and measure the outcome Y_i. If I run the regression $\ln Y_i = \beta_0 + \beta_1 X_i + u_i$, how do I interpret the coefficient β_1?

   a) A 1 unit change in X_i yields a change in Y_i of β units.
   b) A 1 unit change in X_i yields a change in Y_i of $\beta \times 100\%$.
   c) A 1% change in X_i yields a change in Y_i of $\beta \times 0.01$ units.
   d) A 1% change in X_i yields a change in Y_i of $\beta\%$ units.

   (d) (4 points) Say I randomly assign a treatment variable X_i to individuals, and measure the outcome Y_i. If I run the regression $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + u_i$, what is the expected change in Y_i from changing from $X_i = 1$ to $X_i = 2$?

   a) $\beta_1 + \beta_2 + \beta_3$
   b) $2 \times \beta_1 + \beta_2 + 5 \times \beta_3$
   c) $\beta_1 + 3 \times \beta_2 + 8 \times \beta_3$
   d) $\beta_1 + 4 \times \beta_2 + 9 \times \beta_3$

   (e) (4 points) Say I randomly assign a treatment program to individuals. If they receive treatment, $T_i = 1$. Otherwise, $T_i = 0$. I measure the outcome Y_i. Y_i is a binary outcome variable i.e. it can only take the values 0 and 1. If I run the regression $Y_i = \beta_0 + \beta_1 T_i + u_i$, how do I interpret the coefficient β_1?

   a) $P(Y_i = 1)$ in the control group (for whom $T_i = 0$)
   b) $P(Y_i = 1)$ in the treated group (for whom $T_i = 1$)
   c) $P(Y_i = 1)$ in the treated group (for whom $T_i = 1$) minus $P(Y_i = 1)$ in the control group (for whom $T_i = 0$)
   d) $P(Y_i = 1)$ in the treated group (for whom $T_i = 1$) plus $P(Y_i = 1)$ in the control group (for whom $T_i = 0$)

2. **Multiple choice (20 points, 4 points each)** Please tick (*Kryssa för*) the correct answer. Only one answer is correct.

   (a) (4 points) I ran an experiment in 45 *Gymnasieskolor* in Stockholm, randomly selecting half the students in each school to receive after school tuition. I collected data on exam results for all the students in those schools. In total, there were 18,000 students in my study. I want to evaluate the effect of the program. How should I treat my standard errors?

   a) I can assume standard errors are homoskedastic, and use the STATA default standard errors.
   b) I should assume standard errors are heteroskedastic, and use STATA's robust option.
   c) I should assume students who study at the same school are more similar to each other than students from different schools, and I should *not* assume independence among observations and I should cluster standard errors by school.
   d) I should assume that students who receive the tuition program will differ from those who don't receive the tuition program, so I should *not* assume independence among the treated and control group and I should cluster standard errors by whether or not the student received the treatment program.

   (b) (4 points) In analyzing my after-school tuition program experiment, I first regress *baseline* characteristics on a treatment dummy. I find that students who received the program were richer, more likely to be born in Sweden, and more likely to be female. All these differences were statistically significant at the 1% level. How should I interpret these regressions?

   a) The tuition program had an effect on income, ethnic origin and gender.
   b) Something could have gone wrong with my randomization design, as people with these characteristics were more likely to get the tuition program.
   c) Income, ethnic origin and gender might be bad controls.
   d) Income, ethnic origin and gender predict test results.

   (c) (4 points) I want to estimate the effects of a randomly assigned program to increase access to safe drinking water. Working with a Bangladeshi NGO, I randomly assigned 50 villages to receive treatment, and 50 villages to control status. However, the NGO ran out of money for the project towards the end of the program, and they only managed to treat 45 out of the 50 villages we planned to treat. So if I run a regression of *actually receiving treatment* T_i on being *assigned to treatment* D_i, $T_i = \gamma_0 + \gamma_1 D_i + \epsilon_i$, the estimate of $\hat{\gamma_1}$ is 0.9.

   I also run the reduced form regression of the change in access to safe drinking water on being assigned to treatment (which I know is randomly assigned and therefore the estimate is unbiased) i.e. $\Delta ACCESS_i = \pi_0 + \pi_1 D_i + u_i$. The estimated coefficient $\hat{\pi_1}$ is 0.18. Calculate the IV estimate of the local average treatment effect of the program.

   a) 0.2
   b) 0.72
   c) 1.08
   d) 5

   (d) (4 points) To evaluate the experimental program to increase access to safe drinking water, I ran a household survey, in a random sample of households in the village. I sometimes found that the households I had randomly sampled were not available to be interviewed, because the adult members of the household were not at home when my enumerators arrived. My enumerators reported to me that, according to neighbours, these households were often the households of single mothers, or particularly young families with no children. Luckily, I had given the enumerators a list of randomly selected replacement households to interview if they failed to locate the households on the main list, so they replaced these households with other households from that list. If I found a household at baseline, I had no trouble locating them at followup, so attrition was very low. What is the consequence for my study?

a) The results of the study will be biased towards zero by measurement error.

b) The results of the study will be biased because the households who weren't interviewed differ from the households who were interviewed.

c) The results of the study will be unbiased but less precisely estimated.

d) The results of the study will be unbiased, but the population for whom I estimate the treatment effect is not the full village, but those households who were available for interview on the occasions my enumerators were in the village.

(e) (4 points) I am interested in the causal effect of a treatment T_i on Y_i. I have a measure of a baseline variable W_i for all individuals. Individuals with $W_i > w_0$ were eligible to receive treatment T_i, but not everyone took up treatment. I construct the dummy variable D_i which measures whether or not an individual was eligible for treatment or not i.e. is equal to 1 if $W_i > w_0$, and 0 otherwise. I run a first stage equation to predict treatment: $T_i = \gamma_0 + \gamma_1 D_i + g(W_i) + \epsilon_i$. The estimate of $\hat{\gamma_1}$ is 0.3. Then I estimate the reduced form equation $Y_i = \pi_0 + \pi_1 D_i + f(W_i) + u_i$. The estimate of $\hat{\pi_1}$ is 3. What is the fuzzy R.D. estimate of the effect of T_i on Y_i (the local average treatment effect)?

a) 0.1

b) 0.9

c) 10

d) There is not enough information to calculate the answer.

3. **Interpreting the results of a regression (20 points)** Say I carried out an experiment among the 600 students of Econ 101. I assigned half of them to specific study groups, and asked them to meet each week to study together. I assigned the rest of the students to the control group.

   I gave all the students a preliminary exam in the first week of the semester ($t = 0$), before I assigned treated students to study groups, and then collected all the students grades on the finals ($t = 1$). For each student, I therefore measure their pre-treatment grade, $GRADE_{i0}$ and their follow-up grade $GRADE_{i1}$. So I have two observations per student. Then I run the following regression:

$$GRADE_{it} = \alpha_i + \beta_1 FINAL_t + \beta_2 FINAL_t \times TREATED_i + u_{it}$$

   where $FINAL_t$ is equal to 0 for the preliminary exam, and 1 for the final exam, and $TREATED_i$ is equal to 0 for students in the control group, and 1 for students in the treated group.

   (a) How do I interpret the following coefficients?
      i. (2 points) α_i?
      ii. (3 points) β_1?
      iii. (3 points) β_2?

   (b) I estimate this regression, and recover the following coefficients:

Coefficient	Estimate
β_1	0.24
β_2	-0.08

   Calculate the following:
      i. (3 points) The average change in scores between preliminary and final exams in the control group.
      ii. (3 points) The average change in scores between preliminary and final exams in the treated group.

   (c) Say that I calculate, instead, the change in scores between the preliminary exam and final exam, for each student i.e. $\Delta GRADE_i = GRADE_{i1} - GRADE_{i0}$. Then I run the following regression:

$$\Delta GRADE_i = \alpha_0 + \alpha_1 TREATED_i + \epsilon_{it}$$

   What will be my estimate of:
      i. (3 points) α_0
      ii. (3 points) α_1

4. **Difference-in-difference analysis (20 points)** I want to evaluate the effects of a change in legislation on air pollution in the United States. The legislation only affects coal-fired power stations. I have air pollution data for all counties in the United States, for a period of thirty years centred on the year of the reform i.e. fifteen years before, and fifteen years after the reform. I consider the counties with coal-fired power stations to be the treated counties ($TREATED_i = 1$) and the counties without coal-fired power stations to be the control counties ($TREATED_i = 0$).

   (a) (4 points) I write down the following regression:

   $$POLL_{it} = \beta_0 + \beta_1 POST_t + \beta_2 TREATED_i + \beta_3 POST_t \times TREATED_i + u_{it}$$

   where $POLL_{it}$ is a measure of air pollution in country i in year t, and $TEMP_{it}$ is the mean temperature in country i in year t. Under what circumstances does the coefficient β_3 have a causal interpretation?

   (b) (4 points) Intuitively, why might you expect or not expect the circumstances you describe in part a) to apply in this case? Give your answer in words.

   (c) (4 points) How could I test whether or not the circumstances you describe in part a) apply in this case?

   (d) (4 points) Assume that I make the test you describe in part c), and I am satisfied that the circumstances you describe in part a) apply. How should I construct standard errors when I estimate the above regression, and why?

   (e) (4 points) Assume that I make the test you describe in part c), and I am satisfied that the circumstances you describe in part a) apply. However, when I estimate β_3, constructing the standard errors as you describe in part d), the standard errors are quite large. Write down an alternative regression that might yield more precise estimates, and explain why.

5. **Instrumental variables analysis (20 points)** A long standing question in labour economics and development is how the number of children a family has affects female labour supply. The structural equation of interest is the following:

$$LABOUR_i = \beta_0 + \beta_1 FERTILITY_i + u_i$$

where $LABOUR_i$ is a measure of a women's labour supply (e.g. hours worked per week) and $FERTILITY_i$ is a measure of the number of children a women has.

(a) (4 points) If I estimated the above equation by OLS using US census data, would the coefficient β_1 have a causal interpretation? Why or why not?

(b) (4 points) Say I would like to find an instrumental variable to estimate the above equation. What are the two conditions that the instrument needs to fulfil? Name the conditions, and explain either in a sentence or mathematically what they mean.

(c) (4 points) Angrist and Evans (1998) proposed using the gender of the first two births in a family as an instrument for family size. Families with two children of the same gender (two boys or two girls) are 6% (standard error: 0.2%) more likely to go on to have a third child than families with two children of opposite genders. (Angrist and Evans think this is because people have a preference for having a child of each gender). They construct a dummy variable $SAMESEX_i$, which is 1 if a family's first two children are of the same sex, and 0 otherwise. They propose using this as an instrument for $FERTILITY_i$ in the following first equation, limiting the sample to families who have at least two children:

$$FERTILITY_i = \alpha + \beta_1 SAMESEX_i + u_i$$

Why would $SAMESEX_i$ be a plausible instrument for $FERTILITY_i$? Discuss both conditions you named in part b).

(d) (4 points) What type of evidence could you provide in support of whether or not the instrument is valid or not? *Hint: you also have census data on a range of other household characteristics, such as age, age at first birth, ethnicity, income, marital status etc.*

(e) (4 points) Another instrument that has been proposed in the past is twin births. At least some families who have twins end up having larger families than they intended. Angrist and Evans also estimated the effect of fertility on labour supply using an indicator for having twins at second birth as an instrument for fertility. The estimated effect of fertility on labour supply was smaller when they used the "twins" instrument than when they used the "same sex" instrument, in the same sample of families with at least two children. Why might this be so?