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 $\beta_1$ 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 positively correlated with $X_i$ and has a negative effect on $Y_i$, what is the sign of the omitted variables bias in $\beta_1$?

   a) Positive
   b) Negative
   c) It does not create bias in $\beta_1$
   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 $\ln Y_i = \beta_0 + \beta_1 \ln X_i + u_i$, how do I interpret the coefficient $\beta_1$?

   a) A 1 unit change in $X_i$ yields a change in $Y_i$ of $\beta$ 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 $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$, what is the expected change in $Y_i$ from changing from $X_i = 2$ to $X_i = 3$?

   a) $\beta_1$
   b) $\beta_2$
   c) $\beta_1 + 5 \times \beta_2$
   d) $2 \times \beta_1 + \beta_2$

   (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 $\beta_0$?

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

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 90 *Gymnasieskolor* in Stockholm, providing voluntary after school tuition programs in 45 of those schools, which I randomly selected. I collected data on exam results for all the students in those schools. In total, there were 36,000 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 since all the students at a school have the same treatment status, 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 school received the treatment program.

   (b) (4 points) In analyzing my after-school tuition program experiment, I include a control for hours studied each week. Before I include this control, I see a strong positive effect of the tuition program on test scores. When I include this control, the result disappears. How should I interpret these regressions?

   a) The tuition program didn't really have any effect on test scores.

   b) Hours studied each week is a bad control, because the tuition program could have affected the number of hours studied. These results shouldn't change our beliefs about whether the program increased test scores or not.

   c) Leaving out hours studied each week creates omitted variable bias.

   d) There was a problem with my randomization design.

   (c) (4 points) I want to estimate the effects of a randomly assigned program to increase access to safe drinking water. At baseline, 21% of households in the control group have access to safe drinking water, and 23% of households in the treated group have access to safe drinking water. At follow-up, 22% of households in the control group have access to safe drinking water, while 45% of households in the treated group have access to safe drinking water. Calculate the difference-in-difference estimate of the effect of the program on access to safe drinking water.

   a) 21%

   b) 22%

   c) 23%

   d) 45%

   (d) (4 points) Oh no, termites ate my data! Some of my paper questionnaires got eaten by termites before I was able to enter the data into the computer. Before the termites got in, I had a representative sample of the population. Assuming the termites ate questionnaires at random, what is the consequence?

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

   b) The results of the study will be biased by measurement error, but the direction of the bias is unclear.

   c) The results of the study will be unbiased, but less precisely estimated, because of having a smaller sample size.

   d) The results of the study will be unbiased and more precisely estimated.

(e) (4 points) I am interested in the causal effect of $X_i$ on $Y_i$. I have a valid and relevant instrument $Z_i$. I use $Z_i$ to predict $X_i$ by estimating the first stage equation: $X_i = \gamma_0 + \gamma_1 Z_i + \nu_i$. The estimate of $\gamma_1$ is 4. Then I estimate the reduced form equation $Y_i = \pi_0 + \pi_1 Z_i + \epsilon_i$. The estimate of $\gamma_1$ is 8. What is the IV estimate of the effect of $X_i$ on $Y_i$?

a) 0.5

b) 2

c) 12

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:

$$GRADE_{it} = \beta_0 + \beta_1 FINAL_{it} + \beta_2 TREATED_{it} + \beta_3 FINAL_{it} \times TREATED_{it} + u_{it}$$

   where $FINAL_{it}$ is equal to 0 for the preliminary exam, and 1 for the final exam, and $TREATED_{it}$ 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) $\beta_0$?

      ii. (2 points) $\beta_1$?

      iii. (2 points) $\beta_2$?

      iv. (2 points) $\beta_3$?

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

| Coefficient | Estimate |
|---|---|
| $\beta_0$ | 0.47 |
| $\beta_1$ | 0.16 |
| $\beta_2$ | -0.02 |
| $\beta_3$ | 0.14 |

   Calculate the following:

      i. (2 points) The mean score on the preliminary exam in the control group.

      ii. (2 points) The mean score on the preliminary exam in the treated group.

      iii. (2 points) The mean score on the final exam in the control group.

      iv. (2 points) The mean score on the final exam 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. (2 points) $\alpha_0$

      ii. (2 points) $\alpha_1$

4. **Panel data (20 points)** Say I would like to understand more about the effect of temperature on economic production. In particular, I would like to know about the effect of temperature on agricultural production. I have data on agricultural production (measured in dollars) for most countries in the world, for more than 40 years. I also have yearly mean temperature in each country.

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

   $$AGPROD_{it} = \beta_0 + \beta_1 TEMP_{it} + u_{it}$$

   where $AGPROD_{it}$ is agricultural production in country $i$ in year $t$, and $TEMP_{it}$ is the mean temperature in country $i$ in year $t$. Does the coefficient $\beta_1$ have a causal interpretation? Why, or why not?

   (b) (4 points) Write down an alternative regression that you think would give a causal estimate.

   (c) (4 points) What must be true (what assumption needs to hold) in order for us to interpret the coefficient from your regression as a causal relationship? You may give your answer in words or using mathematical notation.

   (d) (4 points) Do you think it's reasonable to believe that this assumption holds in this case? Why or why not?

   (e) (4 points) Referring to a paper we discussed in class, why might it be important to consider nonlinearity in this analysis? (You do not need to give the names of the authors or title of the paper if you do not remember them).

5. **Analysis using policies with thresholds (20 points)** Say I enroll the 600 students in Econ 101 in a second study. I use the results of the preliminary exam I asked them to take in the first week. All students who score below 50% on the exam will receive 2 hours of extra tuition every week. Students who score higher than 50% on the exam will not receive any extra tuition. I also have data on student characteristics at the start of class, including but not limited to grades in high school, parental income, gender, and date of birth.

   (a) (3 points) How might I measure the causal effect of the tuition program on student grades in the final?

   (b) (3 points) Explain using words why the method you propose should recover a causal estimate.

   (c) (6 points) What graphical (visual) evidence could I provide to demonstrate the effect of the program? You may provide a sketch or give a description in words of what you would do.

   (d) (4 points) What evidence could I provide to support the case that the effect I measured was a causal relationship?

   (e) (4 points) What specific concerns would you have about whether this was a causal relationship or not, and how could you provide evidence to evaluate the concern? *Hint: Assume I announced the program before I administered the test, and think about bunching.*