Please collect the answers to the questions below (including graphs) in a `.pdf` and upload them on Athena; also upload the `.do` files that generate your answer. Both need to be done by the 29th of April, 4pm.
Please remember to lay out your results as clearly as possible, and to comment your code in a way that makes it easily accessible to others.

Question 1. Consider the following data generating process $y = \beta_x x + \epsilon$, $z = \nu$, and $x = \beta_{FS} z + \epsilon + \mu$, where $\epsilon_i \sim N(0, 3)$, $\nu_i \sim N(0, 1)$, and $\mu_i \sim N(0, 1)$, all three are i.i.d. and mutually independent, and set $\beta_x = \beta_{FS} = 1$. The estimand of interest is the causal effect of x on y, which is 1. Draw 10.000 samples of 20 observations each.

a. In each sample estimate the causal effect of x on y through a two-stage-least-squares (2SLS). We refer to that estimate as $\hat{\beta}_{2SLS}$. Further calculate $\tilde{x} = \beta_{FS} z$, i.e. imposing your knowledge of the true first stage coefficient, and run a regression of y on $\tilde{x}$ calculated that way. Plot the distribution of the latter estimate and $\hat{\beta}_{2SLS}$ in the same graph, and report their means.

b. In each first stage regression, calculate the F-statistic of a test of the null of irrelevance of the excluded instruments. Calculate the absolute second stage bias, $|\hat{\beta}_{2SLS} - \beta_x|$. How big is the absolute second stage bias on average in case that $\hat{\beta}_{FS} < \beta_{FS}$ and in case $\hat{\beta}_{FS} > \beta_{FS}$. Focussing on the cases where the first stage F-statistic is greater than 10, do you find that the second stage bias decreases as the first stage F-statistic increases?

c. Calculate the ratio of the residual sum of squares in the first stage over the true error sum of squares $(\mu + \epsilon)'(\mu + \epsilon)$. Plot, in three separate bin-scatter plots always using 50 bins, the residual sum of squares, the error sum of squares and the their ratio over $\hat{\beta}_{FS}$. What is the highest ratio of residual sum of squares over error sum of squares across all 10.000 simulations?

d. Create a bin-scatter of the absolute second stage bias over the first stage coefficient estimate, $\hat{\beta}_{FS}$, using 50 bins. For which values of $\hat{\beta}_{FS}$ do you find the largest absolute second stage bias? Also create a bin-scatter of the first stage F-statistic over the first stage coefficient estimate.

d. Now focus on cases where $\hat{\beta}_{FS} > 0.5$. Repeat the exercise of question (d.) For which values of $\hat{\beta}_{FS}$ is the absolute bias smallest, and does the first stage F-statistic appear a useful diagnostic to detect bias?

e. Explain the patterns you uncovered.

(Make sure to use the `seed` command in STATA so your results replicate. You might also find the `binscatter` command helpful.)

Question 2. (Exam 2020) Consider the (purely hypothetical) case where you know the fraction of always-takers, never-takers and compliers in your sample to be 30%, 30%, and 40%, respectively. Further, 50% of the full sample is assigned to treatment. You observe that 20% of the full sample are individuals who were assigned to treatment, yet they did not take up treatment; a further 20% of the sample was not assigned to treatment, yet did take up treatment; and 30% of the sample was assigned to treatment and did take up treatment.

1. Why is this scenario necessarily 'purely hypothetical'?

2. Was the randomization successful in the sense that the fraction of never-takers, compliers and always-takers, respectively, is the same amongst those who were assigned to treatment, and amongst those who were not assigned to treatment?

3. What is the coefficient estimate of the first stage regression of a dummy indicating treatment status on a dummy indicating treatment assignment and a constant? Is it identical to the true fraction of compliers in the sample? Please explain the relation between the two numbers.

Now imagine that the outcome of interest of never-takers is 1, the outcome of always-takers is 5, the outcome of compliers in case of treatment is 4 and the treatment effect for compliers is homogenous and 2. (So there is no heterogeneity in potential outcomes amongst never-takers, and similarly for always-takers and compliers.)

4. You run a regression of the outcome of interest on a dummy indicating treatment status and a constant instrumented by a dummy indicating treatment assignment and a constant. What is the second stage coefficient estimate on the dummy indicating treatment status? Explain how this compares to the true 'local average treatment effect' in your sample.

Question 3. Consider the following setting. A subset of a population has been treated, and actual treatment status is denoted by $D \in \{0, 1\}$. Additionally a binary treatment assignment $Z \in \{0, 1\}$ exists such that $\mathbb{E}[D|Z = 1] > \mathbb{E}[D|Z = 0]$. The treatment assignment has been fully randomized. Denote with $D(Z)$ a function that maps treatment assignment Z into treatment status D. So individuals with $D(0) = 0$ and $D(1) = 1$ are 'compliers', individuals with $D(1) = 0$ and $D(0) = 1$ are 'defiers' and so on. Assume that no defiers exist. Additionally a covariates X exists, which is heterogenous across individuals.
Imagine you are asked to characterize the complier subpopulation by calculating their average covariate value. Proof that this can be achieved by regressing the interaction of treatment status and the covariate, DX, on D, instrumented by Z.
Hint: The corresponding (population) Wald estimator is given by

$$\frac{\mathbb{E}[XD|Z = 1] - \mathbb{E}[XD|Z = 0]}{\mathbb{E}[D|Z = 1] - \mathbb{E}[D|Z = 0]}. \tag{1}$$

You need to show that (1) can indeed be rewritten as $\mathbb{E}[X|D(1) = 1, D(0) = 0]$.
Note: This is alternative method to characterise compliers to what we discussed in the lecture.