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 22nd 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. (Variance of Mean Difference Estimator). Consider a sample of size N from which you randomly assign N_1 units to treatment. Let $D_i \in \{0,1\}$ indicate the treatment assignment of unit i. Denote with $\mathbf{D} = \{D_i\}_{i=1}^N$ an assignment, and with $\mathcal{D}$ the set of all possible assignments given the assignment mechanism.

a. What is the unit assignment probability of any observation i?

Suppose you now want to estimate the mean difference in the outcome between the treated and control, and you implement this with an OLS estimator of the outcome of interest on $\mathbf{X}' = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ D_1 & D_2 & \cdots & D_n \end{bmatrix}$ with associated coefficients $\beta' = \begin{bmatrix} \beta_1 & \beta_D \end{bmatrix}$.
Recall that the variance of the OLS estimator is $\mathbb{V}(\hat{\beta}|\mathbf{D}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega(\mathbf{D})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$, where $\Omega(\mathbf{D})$ has elements $\Omega_{ij}(\mathbf{D}) = \mathbb{E}[\epsilon_i\epsilon_j|\mathbf{D}]$, and we assume $\Omega_{ij}(\mathbf{D}) = 0$ if $i \neq j$ and $\Omega_{ij}(\mathbf{D}) = \sigma_i^2(D_i)$ if $i = j$.

b. Show that $\mathbb{V}(\hat{\beta}_D|\mathbf{D}) = \frac{\sum \sigma_i^2(1)D_i/N_1}{N_1} + \frac{\sum \sigma_i^2(0)(1-D_i)/(N-N_1)}{N-N_1}$.
(At some point useful: Notice that $\sum \sigma_i^2(D_i) = \sum \sigma_i^2(1)D_i + \sum \sigma_i^2(0)(1-D_i)$.)

c. Now assume that $\sigma_i^2(D_i) = \sigma^2(D_i)$ for all i, and show that $\mathbb{E}_{\mathbf{D}\in\mathcal{D}}[\mathbb{V}(\hat{\beta}_D|\mathbf{D})]$ coincides with the expression on slide 5 of lecture 4.

Question 2. (Stratification and Test Size).
Consider the following data generating process $Y_i(D_i, X_i) = \tau_i \times D_i + 5 \times X_i + \epsilon_i$, where $\epsilon_i \sim N(0,1)$, $\tau_i \sim N(1,1)$, $D_i \in \{0,1\}$ indicates treatment status of a treatment of interest, and $X_i \in \{0,1\}$ is some covariate. Both ϵ_i and τ_i are independent of D_i and X_i, and ϵ_i and τ_i are independent of each other.

a. What is the population average treatment effect?

b. Draw for sample of 100 observations τ_i and ϵ_i, and set $X_i = 1$ for 50 observations. Generate for each observation the potential outcome in case of treatment and non-treatment. What is the average treatment effect in your sample?

c. Consider the assignment mechanism where 25% of observations are randomly chosen to be assigned to treatment. Simulate 1000 such assignments, each time calculate the treatment effect from a regression of the observed Y_i on D_i, and record the coefficient estimate, the standard error and the p-value of the null that the coefficient estimate is equal to the sample ATE in (b).

d. Repeat the procedure of question (c), but assign 50% of observations to treatment.

e. Repeat the procedure of question (c), but stratify the treatment assignment on X_i.

f. Plot the distribution of the coefficient estimates in (c) together with the distribution of coefficient estimates in (d) and (e).

g. What is the standard deviation of estimates in (c), (d) and (e) what is the average standard error in both cases? What fraction of estimates has a p-value smaller than 0.05 in either case? Explain the patterns you find.

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

Question 3. (Bad Control).
Consider the same set-up as in Question 1, with one exception: X_i is no longer independent of D_i and ϵ_i. In particular,

$$X_i(D_i) = \begin{cases} \mathbf{1}(\epsilon_i > 1) & \text{if } D_i = 1 \\ \mathbf{1}(\epsilon_i > -1) & \text{if } D_i = 0, \end{cases}$$

where $\mathbf{1}(\cdot)$ is the indicator function.

a. Draw for a sample of 1000 observations τ_i and ϵ_i. Calculate the true average treatment effect of D_i in your sample – defined as the sample average of $Y_i(1, X_i(1)) - Y_i(0, X_i(0))$. Also report the sample average of τ_i.

Consider the assignment mechanism where 50% of observations are randomly chosen to be assigned to treatment. Simulate 1000 such assignments.

b. Each time run a regression of observed Y_i on D_i and a constant. Report the average coefficient estimate on D_i across all 1000 assignments.

c. Each time run a regression of observed Y_i on D_i and a constant, separately by sub-samples with $X_i = 0$ and $X_i = 1$, respectively. Report the average coefficient estimates on D_i across all 1000 assignments.

d. Explain the patterns you find.

Question 4. Prove that the distribution of covariates is balanced across treatment and control groups iff the propensity score is constant.

Question 5. (Matching by Regression).
Consider again the same set-up as in Question 1, with one exception: D_i is no longer independent of X_i. In particular, let

$$P(D_i = 1) = \begin{cases} 0.8 & \text{if } X_i = 1 \\ 0.5 & \text{if } X_i = 0. \end{cases}$$

a. In a sample of 10000 observations, set $X_i = 1$ for 4000 observations. Further draw τ_i and ϵ_i, and D_i. Generate for each observation the observed outcome Y_i.

b. Within the sub-samples with $X_i = 0$ and $X_i = 1$, respectively:

Run regression of Y_i on D_i and a constant and report the coefficient estimate on D_i.

c. Run a regression of Y_i on D_i, X_i and a constant and report the coefficient estimate on D_i.

d. Run a regression of Y_i on D_i and a constant and report the coefficient estimate on D_i.

e. Explain the patterns you find.