Exercise 1 Suppose you have data for an outcome $y$ and covariates $X$ and you estimate the following linear model using OLS:

$$y_i = x_i' \beta + \epsilon_i$$

with $i = 1, \ldots, n$, where $x_i$ and $\beta$ are $k \times 1$ and $y_i$ and $\epsilon_i$ are $1 \times 1$. Importantly, the first element of $x_i$ is 1 for all $i$, which means that you are including an intercept in your regression.

The predicted value for observation $i$, $\hat{y}_i$, is defined by

$$\hat{y}_i = x'_i \hat{\beta}$$

Prove that if you include an intercept in this OLS estimation, the following is true:

$$\frac{1}{N} \sum_{i=1}^{N} y_i = \frac{1}{N} \sum_{i=1}^{N} \hat{y}_i$$

In your proof, you should ONLY use algebraic properties of the OLS estimator. This is, your proof should not rely on any of the classical assumptions or its implications (HINT: if in your proof you use expectations, $E(.)$, you are on the wrong path).

Exercise 2 Perform simulations to assess whether OLS confidence intervals have the correct coverage. This is, simulate data, estimate the parameters with OLS, and test the null hypothesis that
your estimator is equal to the true population parameter. Repeat this a large number times (at least 1000) and calculate the proportion of times that you reject the true null hypothesis using an $\alpha$-level test. If you have correct coverage, you should observe that your confidence intervals include the true value of the parameter $1-\alpha\%$ of the times. Do your confidence intervals have correct coverage? Why should you expect/not expect to observe correct coverage in this case?

**Exercise 3** Do the same simulations you did in the previous exercise, but this time estimate the ATT using a matching algorithm. This is, change only the estimation method with respect to Exercise 2, and keep constant the model that generates the data. Do confidence intervals based on A-I standard errors have correct coverage? Why should you expect/not expect to observe correct coverage in this case?

**Exercise 4** Now repeat these simulations using the same model to generate the data that you used in both Exercise 2 and Exercise 3, but combine both procedures. This is, in each iteration, after you simulate the data, match on at least one covariate, and then estimate the same linear model that you estimated in Exercise 2 on the matched data. Do confidence intervals based on the OLS standard errors that you obtained after matching have correct coverage in this case? Why should you expect/not expect to observe correct coverage in this case?