INTL 601 Research Methods I
Exercise #1

Dataset: gg_fake.dta
This dataset contains **5,000 simulated observations of individual voters** and is designed to mimic a **field experiment on voter mobilization** in the spirit of Gerber & Green.

- **Y** is the outcome variable: **turnout** (1 = voted, 0 = did not vote).
- **Z** is the **randomized assignment** to be canvassed (1 = assigned, 0 = not assigned).
- **T** is the **actual treatment received** (1 = actually contacted, 0 = not contacted). Because of imperfect compliance, assignment (**Z**) affects contact (**T**) but does not determine it perfectly.

The dataset also includes several **pre-treatment covariates**: age, education (educ), past turnout (pastvote), partisan strength (party_id), and district competitiveness (competitive). The data are generated so that assignment (**Z**) influences contact (**T**), contact (**T**) increases the probability of turnout (**Y**), and the covariates affect both contact and turnout. This structure allows you to practice **OLS,** study **confounding and control variables**, and estimate **treatment effects**, intention to treat **(ITT)** in a controlled setting.
Note: This is a simulated dataset for teaching purposes only.

Load the data and show the main descriptive properties of this dataset.
What is: The turnout rate? The assignment rate? The contact rate? The contact rate among those assigned vs not assigned? Analyze the experiment using OLS and the basic difference-of-means test.

Use STATA's margins at(T=(0 1)). What are the predicted probabilities for T=0 and T=1? What is the difference? How does this relate to the regression coefficient?

Estimate the same OLS regression for the experiment now, including the control variables in the dataset (age educ pastvote party_id competitive). Report the new coefficient on T. Compare it to the first univariate result earlier. Did it change a lot or a little? By how much?

What is the marginal effect of T? How does it compare to the raw difference in means from the univariate result earlier? Explain **numerically** which control variable has the biggest association with Y? How can you see that from the regression output?

Estimate the following causal structure :

Z (random assignment) → T (actual contact) → Y (turnout)

with other covariates also affecting **T** and **Y**. What does this model help us infer? What kind of questions does it allow us to answer? Think of direct and indirect effects.

Now **pretend** this is no longer a real experiment.
Generate a “targeted treatment” variable:
gen T_target = (runiform() < invlogit(-1 + 1.2*pastvote + 0.5*party_id))

A **targeted treatment variable** is a treatment indicator that is **not randomly assigned**, but instead is **given based on people’s characteristics**—that is, treatment is **targeted** to certain types of units. It’s a treatment given selectively to people who look more (or less) likely to benefit, respond, or achieve the outcome anyway.

What does the above-generated targeted treatment do?

Check:
tab T_target pastvote, row

Now estimate:

reg Y T_target
reg Y T_target age educ pastvote party_id competitive

Compare the coefficient on T_target **with** and **without** controls. Which one is bigger? By how much? Now compare the coefficient on T from the real experimental treatment, and the coefficient on T_target with no controls. Which one is more biased? How can you see this **in the numbers**? Explain using the results s to which variable is doing most of the confounding here? (Hint: look at what predicts T_target and what predicts Y.)