DEVP252: Problem Set 2

Due March 13, 2024

This exercise requires data and regression analysis using sex worker and non-sex worker data from Ecuador. You may work in groups of up to 2 people, and then turn in just one copy for the group. Please provide very short answers to each of the questions and summarize the results answers for each question concisely using tables. Please submit your answers and code as a single pdf via Gradescope.

Part 1:

We will analyze the following two datasets:

1. Panel of sex workers
2. Cross section of sex workers and female non-sex workers

The sex worker panel is an extract of data from a survey conducted in 2003 in 8 cities in Ecuador. You have information on the last three transactions for each sex worker, making it a panel. The non-sex worker data in the cross section is an extract of data from the 2003 National Employment, Unemployment, and Underemployment Survey (ENEMDU) conducted by the national statistical office (INEC) of Ecuador from the same 8 cities.

Data notes:

Key variables in the panel

1. Sex worker characteristics: sex worker id number (`tp02`), age, education (`years_educ`), dummy if she has children (`children`), dummy if she is married (`married_civunion`), dummy if she is attractive (`beauty`);

2. Client characteristics: dummy for whether client is regular (`clnt_regular2`), dummy for client is rich (`clnt_rich2`), dummy for client is clean (`clnt_clean2`), dummy for handsome client (`clnt_guapo2`), dummy for foreign client (`cl_foreign`) , dummy for client very likely to have HIV (`very`);

3. Other transaction characteristics: no condom used (`noco`), price of transaction (`price`), log price of transaction (`ln_price_trans`), type of sex provided (`vaginal` , `anal` , `oral` , `non_sex`), city (`ciudad`)

Key variables in the cross section

1. Female characteristics: sex worker id number (`tp02`), age (`edad`), age squared (`edad2`), dummies for various age categories (`age_12_17` , `age_18_23` , `age_24_29` , `age_30_35` , `age_36_41` , `age_42_47` , `age_48_up`), dummy if woman is sex worker(`sex_worker`), education (`years_educ`), dummies for various education categories (`no_school` , `kinder` , `primary` , `secondary` , `sup_no_univ` , `university`), dummy if she has children (`children`), dummy if she is married (`married_civilunion`) , last week's log wages (`ln_wages`), never migrated (`always_lived_here`);
2. Geographic characteristics: city (`ciudad`), sexratio of the city (`sexratio`), level of urbanization of city (`urbanized`)

Note: `ciudad` takes on values 1-8 where 1 is Machala, 2 is Milagro, 3 is Daule, 4 is Esmeraldas, 5 is Santo Domingo, 6 is Quevedo, 7 is Quito, and 8 is Guayaquil.

Background: You have the first round of impact evaluation data from an HIV prevention intervention funded by the Bill and Melinda Gates Foundation. The intervention comprises of various educational activities targeting high risk groups like sex workers and teaches them about disease risk and the importance of condom use. The intervention or "treatment" is currently taking place and you will get the second round of data in one year. In the meantime, you are asked to do some analysis by the Gates Foundation as they are currently thinking of funding additional such projects in Africa. You obviously want to keep them happy so they will fund your future work!

Question 1

First they ask you to evaluate the effectiveness of the randomization. The randomization was conducted at the city level where Machala, Milagro & Daule were randomly selected as the control cities and Esmeraldas, Quevedo, and Santo Domingo were selected as the treatment cities. Guayaquil and Quito were included in the project due to high rates of HIV in the sex worker population, but they are not part of the actual impact evaluation (as they were not randomized in or out). Assess the validity of the randomization by generating a table of summary statistics by treatment and control status and test for differences in means, using the cross-sectional data. Please *only* include variables that you think are valid for testing pre-treatment balance across the two groups of cities.

Below is a function that may help you produce a nice table of summary statistics, although you should feel free to use other methods.

In [1]:
```r
ttable <- function(df, treatvar = 'treat', covnames = NULL) { # Arguments ar
    options(warn=-1)
    # Keep only covariate names if option is specified
    if (!is.null(covnames)) {  df <- df %>% select(treatvar, covnames)}
    # Otherwise covariate names are everything but the treatment variable
    else { covnames <- names(df %>% select(-treatvar)) }
    # Function for t-stats of individual variables
    niceT <- function(var, df,treatvar) {
        # Conduct a t-test of var by treatment, and extract relevant columns
        result <- (t.test(df[[var]]~df[[treatvar]]) %>% report() %>% as.data
        # Name the row fr
        rownames(result) <- var
        # Return renamed output (Note assumes Control < Treat)
        return(result %>% rename("Control Mean" = Mean_Group1, "Treat Mean"
        }
    # Get list of results
    result_list <- lapply(covnames, function(var) {
        niceT(var, covs, treatvar = treatvar)
    })
    # Turn it into a dataframe and round
    result_df <- do.call(rbind, result_list) %>% round(4)
    # Return results
    options(warn=0)
    return(result_df)
}
```

In [ ]:

Question 2

Non-condom use is a big policy concern. One of the proposed interventions Gates is considering funding is a conditional cash transfer program (CCT) aimed at reducing sex worker risk behavior. Sex workers will receive a CCT if they test disease free. The CCT should incentivize them to engage in condom use and say no to clients who offer additional money for not using a condom. For funding purposes, Gates needs to know on average, how much each woman would need to be compensated per transaction to forgo having non-condom sex. Please estimate that amount for them. Hint: Use the panel dataset and estimate the magnitude of the risk premium using sex worker fixed effects (FE) models where the dependent variable is log price and the coefficient of interest is on non-condom use. Control for client and transaction characteristics in both models. Interpret the results. You can use the Gertler et al. 2006 Journal of Political Economy paper if you need some background reference.

In [ ]:

Question 3

Another proposed intervention Gates is considering funding is to provide alternate employment opportunities for these women. In focus groups, sex workers have mentioned they would be willing to exit the sex market if these alternative job opportunities pay as much as they currently make. They also mention their most likely outside option is domestic work. Use the cross sectional data and estimate regressions where the dependent variable is log weekly wages and control variables include age, education, marital status, children, whether the person has always lived in the city, and city FEs. Include a sex worker dummy, and interpret the coefficient on the sex worker dummy. Is this the premium to sex work?

Question 4

Discuss why interpreting the coefficient on the sex worker dummy as the premium to sex work might be problematic. Suggest alternative empirical methods you might use to estimate the premium to sex work (either using data you currently have or other data that you might want to collect).