# Assignment 6

Samuel Fraley
Eric Gutierrez
Corneel Moons

November 17, 2025

## Contents

# Question 1: Randomized Control Trials

This question is based on Duflo, Hanna, and Ryan (2012), who evaluate whether teacher monitoring combined with financial incentives can reduce teacher absenteeism and improve learning in primary schools.

The NGO Seva Mandir operates non-formal primary schools in rural villages of Rajasthan (India). Before the program, teacher absenteeism was high (around 35%). In 2003, Seva Mandir introduced a teacher incentive program in 57 randomly selected schools. A camera system was installed to monitor teacher attendance, and teachers were paid according to a nonlinear function of valid teaching days (at least 5 hours of teaching with at least 8 students).

The program generated an immediate and persistent improvement in attendance in treated schools.

## Data

The dataset `ps1_q1.csv` is a simplified version of the original data collected for this RCT. Each observation corresponds to a visit to one of the study schools (identified by `schid`). The variable `time` equals 1 in the month before the program starts (baseline) and is greater than 1 in months after the program begins. Schools are randomly assigned to a treatment group (`treat`=1) or control group (`treat`=0). The main outcome variables are the number of students (`students`) and teacher attendance (`teacher_attendance`).

## 1.1 Baseline and Experiment Integrity

Under proper randomization, potential outcomes are independent of treatment status:

$$Y_{1i}, Y_{0i} \perp\!\!\!\perp D_i,$$

which implies that the average treatment effect on the treated (ATT) and the average treatment effect (ATE) coincide: $\alpha_{ATT} = \alpha_{ATE} = \beta$.

Before analyzing post-treatment outcomes, we should verify that treated and control schools are similar at baseline.

(a) Using only the observations from the month before the program starts (`time` = 1), compute the average teacher attendance and the average number of students per classroom separately for the treatment and control groups.

Replicate the information in Panels A and B of Table 1 in Duflo et al. (2012): produce baseline means by treatment status, report appropriate standard errors, and comment on whether the randomization appears to have produced comparable groups.

(b) Briefly discuss whether any observed baseline differences are economically and/or statistically significant. Explain why this check is important for interpreting the causal effect estimates later.

## 1.2 Results

When randomization is valid, the treatment effect can be estimated via the simple regression:

$$Y_i = \alpha + \beta D_i,$$

where $Y_i$ is the outcome of interest (e.g. teacher attendance) and $D_i$ is the treatment indicator.

(c) Using post-program data (`time` > 1), compute and compare average teacher attendance in treated and control schools. Replicate the first three columns of Panel A of Table 2 in Duflo et al. (2012), reporting the relevant means and differences.

(d) Based on your estimates, discuss whether the incentive program achieved its main goal of reducing teacher absenteeism. Comment on the magnitude and statistical significance of the estimated treatment effect.

# Question 2: Matching

Jacobson, LaLonde, and Sullivan (1993, JLS) study earnings losses following job displacement. Using administrative data from Pennsylvania, they document that workers involved in mass employment reductions suffer long-term earnings losses of roughly 25% per year. They distinguish between separations due to mass layoffs and other separations, and use stayers as a control group.

Their model can be written as:

$$w_{it}^A = \mu_i + \sum_{k \geq -4}^{6} \phi_k L_{it}^k + \sum_{l \geq -4}^{6} \psi_l M_{it}^l + \beta' X_{it} + \rho_t + \varepsilon_{it},$$

where $w_{it}^A$ is log annual earnings of worker $i$, $L_{it}^k$ and $M_{it}^l$ are sets of dummies indicating years relative to layoff and mass layoff, $X_{it}$ is a vector of covariates, and $\rho_t$ are time effects.

Couch and Placzek (2010, CP) revisit this question using matching estimators, arguing that displaced workers are systematically selected, so estimates based only on JLS-type comparisons may be biased upward.

Let $D_i = 1$ if worker $i$ is displaced (due to a mass layoff or other separation) and $D_i = 0$ otherwise, and let $p(X_i)$ denote the propensity score. The average treatment effect on the treated (ATT) is:

$$\alpha_{TT} = \mathbb{E}\Big[\mathbb{E}[w_{1i}^A \mid D_i = 1, p(X_i)] - \mathbb{E}[w_{0i}^A \mid D_i = 0, p(X_i)] \;\Big|\; D_i = 1\Big].$$

To compare outcomes relative to a reference year $t_0$, CP consider a differenced ATT:

$$\alpha_{ATT}^D = \mathbb{E}\Big[\big(\mathbb{E}[w_{1it}^A \mid D_i = 1, p(X_i)] - \mathbb{E}[w_{1it_0}^A \mid D_i = 1, p(X_i)]\big) - \big(\mathbb{E}[w_{0it}^A \mid D_i = 0, p(X_i)] - \mathbb{E}[w_{0it_0}^A \mid D_i = 0, p(X_i)]\big) \;\Big|\; D_i = 1$$

Your task is to revisit CP's findings using a different dataset.

## Data

The dataset `ps1_q2.dta` is built from the Veneto Workers Histories (VWH), an administrative panel including all individuals working in the Italian region of Veneto from 1975–2001. The file `ps1_q2.dta` contains a subsample of workers who, in 1999, either:

- experienced a mass employment reduction,
- separated from the firm without being part of a mass layoff, or
- stayed with the same employer.

The panel covers the years 1995–2001. Mass layoffs are defined using the endogenous separation rate, following JLS and von Wachter, Song, and Manchester (2009). Displaced workers satisfy the standard requirements in this literature.

### 2.1 Propensity Score Index

(a) For a year of your choice, estimate the propensity score $p(X_i)$ using gender, decile of 1995 earnings, and decade of birth as covariates $X_i$. Use the `pscore` command (or an equivalent implementation) to compute the propensity scores.

(b) Check the balancing property of the propensity score: verify that, within propensity-score strata, the distribution of covariates is similar between displaced and non-displaced workers. Summarize and comment on the output.

## 2.2 Nearest Neighbor and Kernel Matching

(c) Using the estimated propensity scores, compute $\alpha_{ATT}$ with nearest neighbor (NN) matching for each year before and after displacement. Clearly indicate the reference year.

(d) Repeat the estimation using kernel matching. In this case, obtain standard errors via bootstrap (with at least 200 replications). Report the ATT estimates and their standard errors for each year.

(e) Compare the NN and kernel results. Comment on differences in the estimated effects and in the associated uncertainty.

## 2.3 Presenting Results

(f) Plot the time path of your estimated effects (before and after displacement) for both NN and kernel ATT estimates. Compare your figures to Figure 1 (included with the VWH data description) and discuss similarities and differences in the pattern of earnings losses.

# Question 3: Instrumental Variables

Angrist and Evans (1998) use an IV strategy to analyze how the number of children affects parents' labor supply. They find a sizable negative effect for mothers and essentially no effect for fathers. Here we focus on mothers and on employment (rather than hours worked).

## Data

The dataset `ps1_q3.dta` is a subset of the data used by Angrist and Evans (1998) and contains only mothers. The key variables are:

- `sexk`: sex of the first child,
- `kidcount`: total number of children,
- `agem`: age of the mother,
- `twin_latest`: indicator equal to 1 if the last birth was a twin birth,
- `blackm`, `hispm`, `othracem`: race dummies,
- `workedm`: indicator equal to 1 if the mother is employed.

## 3.1 Baseline Models

Consider the model:

$$y_i = \beta_0 + \beta_1 \text{kidcount}_i + X_i' \beta + \varepsilon_i,$$

where $y_i$ is the mother's employment status and $X_i$ is a vector of controls.

(a) Estimate this equation using OLS.

```
-------------------------------------------------------------------------------------
      name:  <unnamed>
       log:  C:\Users\sffra\Downloads\BSE 2025-2026\econometrics\hw6\tables\ols.log
  log type:  text
 opened on:  13 Nov 2025, 13:32:26

. reg workedm kidcount agem blackm hispm othracem, robust

Linear regression                               Number of obs     =    400,169
                                                F(5, 400163)      =    3086.78
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0344
                                                Root MSE          =      .4871

------------------------------------------------------------------------------
             |               Robust
     workedm | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    kidcount |  -.0910744    .000951   -95.77   0.000    -.0929382   -.0892105
        agem |   .0146906   .0002209    66.50   0.000     .0142576    .0151236
      blackm |   .1506704   .0022648    66.53   0.000     .1462314    .1551093
       hispm |   -.008305   .0045639    -1.82   0.069    -.0172501    .0006402
    othracem |   .0275062   .0046564     5.91   0.000     .0183798    .0366326
       _cons |    .337658   .0069179    48.81   0.000     .3240992    .3512168
------------------------------------------------------------------------------

. log close
      name:  <unnamed>
       log:  C:\Users\sffra\Downloads\BSE 2025-2026\econometrics\hw6\tables\ols.log
  log type:  text
 closed on:  13 Nov 2025, 13:32:26
-------------------------------------------------------------------------------------
```

**Figure 1:** OLS regression output from Stata

(b) Estimate the same specification using a probit model.

```
------------------------------------------------------------------------------------------------
      name:  <unnamed>
       log:  C:\Users\sffra\Downloads\BSE 2025-2026\econometrics\hw6\tables\probit.log
  log type:  text
 opened on:  13 Nov 2025, 13:32:26

. probit workedm kidcount agem blackm hispm othracem

Iteration 0:  Log likelihood = -273933.15
Iteration 1:  Log likelihood = -266940.83
Iteration 2:  Log likelihood = -266932.47
Iteration 3:  Log likelihood = -266932.47

Probit regression                                       Number of obs =  400,169
                                                        LR chi2(5)    = 14001.37
                                                        Prob > chi2   =   0.0000
Log likelihood = -266932.47                             Pseudo R2     =   0.0256

------------------------------------------------------------------------------
     workedm | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
    kidcount |  -.2370727   .0025496   -92.99   0.000    -.2420698   -.2320757
        agem |   .0382601   .0005794    66.03   0.000     .0371244    .0393957
      blackm |   .4039446   .0064219    62.90   0.000     .3913579    .4165314
       hispm |  -.0203433   .0117697    -1.73   0.084    -.0434116    .0027249
    othracem |   .0717805   .0120918     5.94   0.000      .048081      .09548
       _cons |  -.4263508   .0178661   -23.86   0.000    -.4613677   -.3913339
------------------------------------------------------------------------------

. log close
      name:  <unnamed>
       log:  C:\Users\sffra\Downloads\BSE 2025-2026\econometrics\hw6\tables\probit.log
  log type:  text
 closed on:  13 Nov 2025, 13:32:27
------------------------------------------------------------------------------------------------
```

**Figure 2:** Probit regression output from Stata

(c) Discuss whether these approaches (OLS and probit) are appropriate for identifying the causal effect of the number of children on mothers' labor supply.

Just using OLS or probit alone is probably not the best approach to identifying the casual effect of number of children on labor supply due to endogeneity problems. For example, women with richer partners may not be required to work, and can decide to have more kids, thus showing that family income could influence both number of children and labor force participation. Various other endogeneity problems may arise as we have a relatively simple model (we already give household/family income, but what about access to childcare such as grandparents present, access to contraceptives, etc).

Be specific about possible sources of endogeneity and functional-form issues.

## 3.2 IV Probit

(d) Re-estimate the model using an IV probit specification, instrumenting `kidcount` with `twin_latest`. Clearly state the first-stage and structural equations.

(e) Explain the economic intuition behind using twin births as an instrument. Discuss the relevance and validity (exclusion restriction) of this instrument, and provide a critical assessment of potential threats to its validity.

```
------------------------------------------------------------------------------------------------
      name:  <unnamed>
```

```
       log:  C:\Users\sffra\Downloads\BSE 2025-2026\econometrics\hw6\tables\firststage.log
  log type:  text
 opened on:  13 Nov 2025, 13:32:27

. reg kidcount twin_latest agem blackm hispm othracem, robust

Linear regression                               Number of obs     =    400,169
                                                F(5, 400163)      =    2754.01
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0404
                                                Root MSE          =     .79889

------------------------------------------------------------------------------
             |               Robust
    kidcount | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
 twin_latest |   .3850951   .0108008    35.65   0.000     .3639259    .4062643
        agem |   .0309692   .0003467    89.33   0.000     .0302897    .0316487
      blackm |   .3236812   .0046599    69.46   0.000     .3145479    .3328146
       hispm |   .4370501   .0097859    44.66   0.000       .41787    .4562303
    othracem |   .1209274   .0084557    14.30   0.000     .1043545    .1375003
       _cons |   1.564569   .0103037   151.85   0.000     1.544374    1.584764
------------------------------------------------------------------------------

. log close
      name:  <unnamed>
       log:  C:\Users\sffra\Downloads\BSE 2025-2026\econometrics\hw6\tables\firststage.log
  log type:  text
 closed on:  13 Nov 2025, 13:32:27
------------------------------------------------------------------------------------------------


------------------------------------------------------------------------------------------------
      name:  <unnamed>
       log:  C:\Users\sffra\Downloads\BSE 2025-2026\econometrics\hw6\tables\ivprobit.log
  log type:  text
 opened on:  13 Nov 2025, 13:32:27

. ivprobit workedm agem blackm hispm othracem ///
>     (kidcount = twin_latest), vce(robust)

Fitting exogenous probit model:
Iteration 0:  Log likelihood = -273933.15
Iteration 1:  Log likelihood = -266932.81
Iteration 2:  Log likelihood =  -266924.4
Iteration 3:  Log likelihood =  -266924.4

Fitting full model:
Iteration 0:  Log pseudolikelihood = -744887.07
Iteration 1:  Log pseudolikelihood = -744887.07

Probit model with endogenous regressors                Number of obs = 400,169
                                                       Wald chi2(5)  = 5172.18
Log pseudolikelihood = -744887.07                      Prob > chi2   =  0.0000

--------------------------------------------------------------------------------------------
                           |               Robust
                           | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
---------------------------+----------------------------------------------------------------
                  kidcount |   -.071591   .0413186    -1.73   0.083    -.1525739     .009392
                      agem |   .0328902   .0015182    21.66   0.000     .0299145    .0358659
                    blackm |    .347367   .0161035    21.57   0.000     .3158046    .3789293
                     hispm |  -.0915387   .0210812    -4.34   0.000    -.1328572   -.0502203
                  othracem |   .0515407   .0131275     3.93   0.000     .0258113    .0772702
                     _cons |  -.6800457   .0640531   -10.62   0.000    -.8055874   -.5545039
---------------------------+----------------------------------------------------------------
 corr(e.kidcount,e.workedm)|  -.1310747   .0322676                     -.1937034   -.0673821
             sd(e.kidcount)|   .7988864   .0016005                      .7957557    .8020295
--------------------------------------------------------------------------------------------
Wald test of exogeneity (corr = 0): chi2(1) = 16.12       Prob > chi2 = 0.0001
Endogenous: kidcount
Exogenous:  agem blackm hispm othracem twin_latest

. log close
      name:  <unnamed>
       log:  C:\Users\sffra\Downloads\BSE 2025-2026\econometrics\hw6\tables\ivprobit.log
  log type:  text
 closed on:  13 Nov 2025, 13:32:33
-------------------------------------------------------------------------------------------------
```

### 3.3 Marginal Effects

(f) Using your preferred IV probit specification, estimate the marginal effect of an additional child on the probability that a mother is employed.

(g) Plot how this marginal effect varies with relevant covariates (e.g. age or baseline number of children), or report marginal effects evaluated at meaningful covariate profiles. Comment on the pattern of these effects and what they imply about the impact of family size on mothers' employment.

# References

- Angrist, J.D. and Evans, W.N. (1998). “Children and Their Parents’ Labor Supply: Evidence from Exogenous Variation in Family Size.” *American Economic Review*, 88(3): 450–477.
- Couch, K.A. and Placzek, D.W. (2010). “Earnings Losses of Displaced Workers Revisited.” *American Economic Review*, 100(1): 572–589.
- Duflo, E., Hanna, R., and Ryan, S.P. (2012). “Incentives Work: Getting Teachers to Come to School.” *American Economic Review*, 102(4): 1241–1278.
- Jacobson, L.S., LaLonde, R.J., and Sullivan, D.G. (1993). “Earnings Losses of Displaced Workers.” *American Economic Review*, 83(4): 685–709.