Lecture 15: Causal inference

Economics 326 — Introduction to Econometrics II

Vadim Marmer, UBC

Motivation: causal questions

- Many important questions in economics are **causal**:
  - Does job training increase earnings?
  - Does a new drug improve health outcomes?
- The **fundamental challenge**: we can never observe the same individual both with and without treatment at the same time.
- This is modeled using the **potential outcomes framework**: each individual has two potential outcomes, one under treatment and one under control, but we only observe one of them.

Potential outcomes

- **Treatment status**: $D_i = 1$ if individual i receives treatment, $D_i = 0$ otherwise.

- For each individual i, define two **potential outcomes**:

  - $Y_i(1)$: outcome if individual i receives treatment $(D_i = 1)$,
  - $Y_i(0)$: outcome if individual i does not receive treatment $(D_i = 0)$.

- The **individual treatment effect** for person i is:

$$Y_i(1) - Y_i(0).$$

- Example: if Y_i is earnings and D_i indicates job training, then $Y_i(1) - Y_i(0)$ is the causal effect of training on earnings for person i.

The fundamental problem

- The **observed outcome** is:

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0).$$

- If $D_i = 1$, we observe $Y_i(1)$ but not $Y_i(0)$.

- If $D_i = 0$, we observe $Y_i(0)$ but not $Y_i(1)$.

- We can never observe both potential outcomes for the same individual. This is the **fundamental problem of causal inference**.

Causal parameters of interest

- **Average Treatment Effect (ATE):**

$$\text{ATE} = \text{E}\left[Y_i(1) - Y_i(0)\right].$$

- **Average Treatment Effect on the Treated (ATT):**

$$\text{ATT} = \text{E}\left[Y_i(1) - Y_i(0) \mid D_i = 1\right].$$

- ATE averages over the entire population; ATT averages only over those who actually receive treatment.

Selection bias

- Since we cannot observe the counterfactual for each individual, a natural idea is to use the outcomes of **other people** as stand-ins: compare average outcomes of the treated group to average outcomes of the untreated group.

- Can we estimate ATE by simply comparing group averages, $\text{E}\left[Y_i \mid D_i = 1\right] - \text{E}\left[Y_i \mid D_i = 0\right]$?

- Comparing group averages seems natural, but it conflates the treatment effect with pre-existing differences between groups:

$$\begin{aligned}
&\text{E}\left[Y_i \mid D_i = 1\right] - \text{E}\left[Y_i \mid D_i = 0\right] \\
&= \text{E}\left[Y_i(1) \mid D_i = 1\right] - \text{E}\left[Y_i(0) \mid D_i = 0\right].
\end{aligned}$$

- Add and subtract $\text{E}\left[Y_i(0) \mid D_i = 1\right]$:

$$\begin{aligned}
&= \underbrace{\text{E}\left[Y_i(1) - Y_i(0) \mid D_i = 1\right]}_{\text{ATT}} \\
&\quad + \underbrace{\text{E}\left[Y_i(0) \mid D_i = 1\right] - \text{E}\left[Y_i(0) \mid D_i = 0\right]}_{\text{Selection bias}}.
\end{aligned}$$

- **Selection bias** arises when the treatment and control groups differ in their baseline outcomes $Y_i(0)$.

- E.g., people who choose to enroll in job training may differ systematically from those who do not.

Random assignment

- **Randomization** solves the selection problem: by randomly assigning treatment, we ensure the treated and control groups are comparable in expectation.

- Under **random assignment**, treatment D_i is independent of potential outcomes:

$$\text{E}\left[Y_i(0) \mid D_i = 1\right] = \text{E}\left[Y_i(0) \mid D_i = 0\right] = \text{E}\left[Y_i(0)\right].$$

- The selection bias vanishes, and the simple difference in means equals the ATE:

$$\begin{aligned}
&\text{E}\left[Y_i \mid D_i = 1\right] - \text{E}\left[Y_i \mid D_i = 0\right] \\
&= \text{E}\left[Y_i(1) - Y_i(0)\right] = \text{ATE} = \text{ATT}.
\end{aligned}$$

- Randomized experiments (like the National Supported Work program) achieve this by randomly assigning individuals to treatment and control groups.

Regression with a treatment dummy

- With random assignment, the ATE can be estimated by regressing Y_i on a treatment dummy:

$$Y_i = \alpha + \tau D_i + U_i,$$

where $\mathrm{E}\left[U_i \mid D_i\right] = 0$ under randomization.

- The OLS estimate of τ equals the difference in sample means:

$$\hat{\tau} = \bar{Y}_1 - \bar{Y}_0,$$

where $\bar{Y}_1$ and $\bar{Y}_0$ are the sample averages for the treated and control groups.

Example: Lalonde data

- The **National Supported Work (NSW)** demonstration recruited disadvantaged workers (long-term unemployed, high-school dropouts, former drug users, ex-offenders).

- Among eligible applicants, some were **randomly assigned** to receive job training (treated group), and the rest formed the **experimental control group**. Both groups come from the same disadvantaged population.

- We use the `jtrain2` dataset from the `wooldridge` package, based on the Lalonde (1986) study:

```
library(wooldridge)
data(jtrain2)
set.seed(326)
jtrain2[sample(nrow(jtrain2), 10), c("train", "re78", "educ", "age", "black", "married")]
```

```
    train      re78 educ age black married
357     0   3.34322    9  21     1       0
320     0  10.79860    8  18     0       0
108     1   1.95327    9  17     1       0
364     0   0.00000   11  39     1       0
256     0   0.00000   12  24     1       0
415     0   0.00000    8  23     1       0
222     0   2.01550   10  20     1       0
100     1  26.81760    9  31     0       0
145     1   2.48455   12  31     0       0
244     0  12.89840    9  22     1       0
```

- `train`: 1 if randomly assigned to job training, 0 if assigned to control.

- `re78`: real earnings in 1978 (thousands of dollars).

Estimating the ATE

- Since treatment was randomly assigned, the coefficient on `train` estimates the ATE:

```
summary(lm(re78 ~ train, data = jtrain2))$coefficients
```

```
            Estimate Std. Error    t value      Pr(>|t|)
(Intercept) 4.554802  0.4080460 11.162474 1.154113e-25
train       1.794343  0.6328536  2.835321 4.787524e-03
```

- The estimated ATE is approximately $1,800 (`re78` is in thousands): on average, participation in the job training program increased 1978 earnings by about $1,800.

Observational studies

- In many settings, randomization is not feasible or ethical. Treatment is determined by individual choices or institutional rules.

- Self-selection generates **selection bias**: workers who voluntarily enroll in job training may be more motivated; workers at firms offering 401(k) plans tend to have higher incomes.

- **Key idea**: if we can identify covariates X_i that explain why some individuals are treated and others are not, then **after controlling for** X_i, treatment may be as good as random.

- This is the **selection on observables** assumption: all confounding factors are captured by observable covariates.

Potential outcomes with a covariate

- The selection on observables assumption can be formalized as follows: after conditioning on X_i, treatment assignment D_i is as good as random.

- Suppose the potential outcomes depend linearly on a covariate X_i:

$$Y_i(0) = \alpha_0 + \beta_0 X_i + U_i(0),$$
$$Y_i(1) = \alpha_1 + \beta_1 X_i + U_i(1),$$

where $\mathrm{E}\left[U_i(0) \mid X_i, D_i\right] = 0$ and $\mathrm{E}\left[U_i(1) \mid X_i, D_i\right] = 0$.

- These conditions are called **conditional mean independence**: while $Y_i(0)$ and $Y_i(1)$ depend on X_i (and thus on D_i), the residual terms after controlling for X_i, namely $U_i(0)$ and $U_i(1)$, are uncorrelated with both X_i and D_i.

- The effect of X_i on the potential outcomes can differ between the treated and control groups: β_1 may not equal β_0.

The ATE with a covariate

- Taking expectations of the potential outcomes:

$$\mathrm{E}\left[Y_i(1)\right] = \alpha_1 + \beta_1 \mathrm{E}\left[X_i\right],$$
$$\mathrm{E}\left[Y_i(0)\right] = \alpha_0 + \beta_0 \mathrm{E}\left[X_i\right].$$

- The ATE is:

$$\begin{aligned}
\mathrm{ATE} &= \mathrm{E}\left[Y_i(1) - Y_i(0)\right] \\
&= (\alpha_1 - \alpha_0) + (\beta_1 - \beta_0)\mathrm{E}\left[X_i\right].
\end{aligned}$$

- The ATE depends on the difference in intercepts **and** the difference in slopes, weighted by the population mean of X_i.

Two separate regressions

- One approach: estimate separate regressions for each group.

- **Control group** $(D_i = 0)$: $Y_i = \alpha_0 + \beta_0 X_i + U_i(0)$.

- **Treatment group** $(D_i = 1)$: $Y_i = \alpha_1 + \beta_1 X_i + U_i(1)$.

- The estimated ATE is:

$$\widehat{\text{ATE}} = (\hat{\alpha}_1 - \hat{\alpha}_0)$$
$$+ (\hat{\beta}_1 - \hat{\beta}_0)\bar{X},$$

  where $\bar{X}$ is the overall sample mean of X_i.

Combined regression with interactions

- The observed outcome $Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0)$ can be written as a single regression. Expanding:

$$Y_i = \alpha_0(1 - D_i) + \alpha_1 D_i$$
$$+ \beta_0 X_i(1 - D_i) + \beta_1 X_i D_i + U_i$$
$$= \alpha_0 + (\alpha_1 - \alpha_0)D_i$$
$$+ \beta_0 X_i + (\beta_1 - \beta_0)X_i D_i + U_i,$$

  where $U_i = (1 - D_i)U_i(0) + D_i U_i(1)$.
- This is a regression of Y_i on D_i, X_i, and the interaction $X_i D_i$.
- The coefficient on D_i is $\alpha_1 - \alpha_0$, which is **not** the ATE (unless $\beta_1 = \beta_0$).

The demeaning trick

- The coefficient on D_i is $\alpha_1 - \alpha_0$, not the ATE. To fix this, **add and subtract** $(\beta_1 - \beta_0)\text{E}\,[X_i]\,D_i$:

$$Y_i = \alpha_0 + (\alpha_1 - \alpha_0)D_i + \beta_0 X_i + (\beta_1 - \beta_0)X_i D_i + U_i$$
$$= \alpha_0 + [(\alpha_1 - \alpha_0) + (\beta_1 - \beta_0)\text{E}\,[X_i]]D_i$$
$$+ \beta_0 X_i + (\beta_1 - \beta_0)X_i D_i - (\beta_1 - \beta_0)\text{E}\,[X_i]\,D_i + U_i$$
$$= \alpha_0 + \underbrace{[(\alpha_1 - \alpha_0) + (\beta_1 - \beta_0)\text{E}\,[X_i]]}_{\text{ATE}}D_i$$
$$+ \beta_0 X_i + (\beta_1 - \beta_0)(X_i - \text{E}\,[X_i])D_i + U_i.$$

- Defining $\tau = \text{ATE}$ and $\delta = \beta_1 - \beta_0$:

$$Y_i = \alpha_0 + \tau D_i + \beta_0 X_i + \delta(X_i - \text{E}\,[X_i])D_i + U_i.$$

- The coefficient τ on D_i is exactly the **ATE**.

Estimating the ATE with covariates

- In practice, replace $\text{E}\,[X_i]$ with the sample mean $\bar{X}$ and run the regression:

$$Y_i = \hat{\alpha}_0 + \hat{\tau} D_i + \hat{\beta}_0 X_i$$
$$+ \hat{\delta}(X_i - \bar{X})D_i + \text{residual}.$$

- The coefficient $\hat{\tau}$ on D_i directly estimates the ATE.
- This works because demeaning the interaction "absorbs" the $(\beta_1 - \beta_0)\text{E}\,[X_i]$ part of the ATE into the coefficient on D_i.

Why not regress Y_i on D_i and X_i?

- A simpler regression omits the interaction:

$$Y_i = a + \tau D_i + b X_i + V_i.$$

- This is valid if $\beta_1 = \beta_0$ (the covariate has the same effect in both groups). Then $\delta = 0$, the interaction drops out, and the coefficient on D_i is the ATE.
- If $\beta_1 \neq \beta_0$ (the effect of X_i differs between groups), the omitted interaction creates bias: the interaction $X_i D_i$ affects Y_i and is correlated with D_i, so the coefficient on D_i does not equal the ATE.

Example: separate regressions

- Estimate separate regressions of `re78` on `educ` for the treatment and control groups:

```r
reg0 <- lm(re78 ~ educ, data = jtrain2, subset = (train == 0))
reg1 <- lm(re78 ~ educ, data = jtrain2, subset = (train == 1))
cbind(Control = coef(reg0), Treatment = coef(reg1))
```

```
              Control   Treatment
(Intercept) 3.80301658 -0.7821703
educ        0.07451936  0.6892860
```

- Compute the estimated ATE:

```r
xbar <- mean(jtrain2$educ)
a0 <- coef(reg0)[1]; b0 <- coef(reg0)[2]
a1 <- coef(reg1)[1]; b1 <- coef(reg1)[2]
ATE_manual <- (a1 - a0) + (b1 - b0) * xbar
cat("Sample mean of educ:", round(xbar, 2), "\n")
```
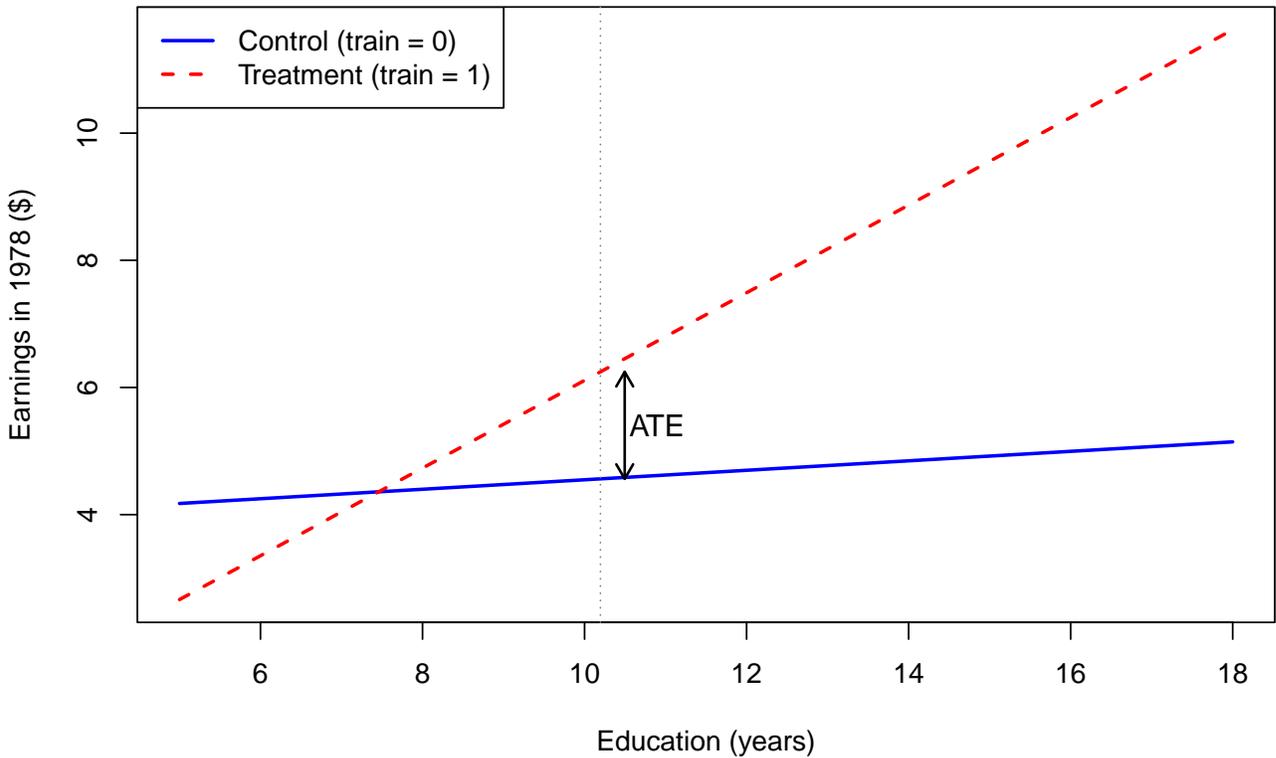
```
Sample mean of educ: 10.2
```

```r
cat("Estimated ATE:", round(ATE_manual, 2), "\n")
```

```
Estimated ATE: 1.68
```

Example: regression lines

- The two regression lines, with a vertical line at $\bar{X}$ and the ATE marked as the gap:

Separate regression lines by treatment group

Example: demeaned regression

- Create the demeaned interaction and run the combined regression:

```
jtrain2$educ_dm <- jtrain2$educ - mean(jtrain2$educ)
reg_dm <- lm(re78 ~ train + educ + I(train * educ_dm), data = jtrain2)
summary(reg_dm)$coefficients
```

```
                       Estimate Std. Error    t value    Pr(>|t|)
(Intercept)          3.80301658  2.5689136 1.4803988 0.13948090
train                1.68266981  0.6299604 2.6710725 0.00784062
educ                 0.07451936  0.2514521 0.2963561 0.76709766
I(train * educ_dm)   0.61476663  0.3472755 1.7702560 0.07737534
```

- The coefficient on `train` directly estimates the ATE, matching the result from the separate regressions approach.

Example: adding more covariates

- So far, we used only `educ` as the control. The dataset contains more pre-treatment characteristics: `age`, `black`, `married`.

- With random assignment, slopes are approximately equal across groups, so we can add controls directly without demeaned interactions:

```
reg_X <- lm(re78 ~ train + educ + age + black + married, data = jtrain2)
round(summary(reg_X)$coefficients, 1)
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.9        2.2     0.4      0.7
train              1.7        0.6     2.7      0.0
educ               0.4        0.2     2.4      0.0
age                0.1        0.0     1.2      0.2
```

```
black              -2.3        0.8    -2.7        0.0
married             0.2        0.8     0.2        0.9
```

Comparison: simple vs. controlled

- Compare the estimated ATE and its standard error across specifications, including one with demeaned inter-actions:

```r
m1 <- lm(re78 ~ train, data = jtrain2)
m2 <- lm(re78 ~ train + educ + age + black + married, data = jtrain2)
covs <- c("educ", "age", "black", "married")
for (v in covs) jtrain2[[paste0(v, "_dm")]] <- jtrain2[[v]] - mean(jtrain2[[v]])
m3 <- lm(re78 ~ train + educ + age + black + married +
            I(train * educ_dm) + I(train * age_dm) +
            I(train * black_dm) + I(train * married_dm), data = jtrain2)
make_row <- function(m) c(Estimate = coef(m)["train"],
                          SE = summary(m)$coef["train", "Std. Error"])
tab <- rbind(
  "No controls"            = make_row(m1),
  "Controls, equal slopes" = make_row(m2),
  "Demeaned interactions"  = make_row(m3)
)
round(tab, 2)
```

```
                        Estimate.train   SE
No controls                       1.79 0.63
Controls, equal slopes            1.68 0.63
Demeaned interactions             1.64 0.63
```

- All three estimates are nearly identical. The demeaned interactions barely matter because randomization ensures approximately equal slopes across groups.

Why do controls change little here?

- Under **random assignment**, D_i is independent of all covariates: the treatment and control groups are balanced in expectation.

- Because of balance, including X_i does not change the coefficient on D_i: there is no selection bias to remove.

- The NSW program recruited a narrow, disadvantaged population: long-term unemployed, high-school dropouts, former drug users, ex-offenders.

- Within this homogeneous group, covariates like `educ`, `age`, `black`, and `married` vary little and have limited predictive power for earnings.

- Both the point estimate and the standard error are nearly unchanged because the controls add almost no information beyond the treatment dummy.

Example: 401(k) eligibility and savings

- The `k401ksubs` dataset (from the `wooldridge` package) contains 9,275 households from the 1991 Survey of Income and Program Participation.

```r
data(k401ksubs)
cat("n =", nrow(k401ksubs), "\n")
```

```
n = 9275
```

```r
set.seed(326)
k401ksubs[sample(nrow(k401ksubs), 8), c("e401k", "nettfa", "inc", "age", "marr")]
```

```
    e401k nettfa     inc age marr
```

```
1893      1 24.200 47.328  55    1
876       0 -1.000 13.260  45    0
927       0  0.000 17.571  37    1
6878      1  7.999 57.843  45    1
4196      0  8.000 13.800  25    0
1887      1  6.450 40.980  39    0
4355      0 29.700 22.410  34    0
7596      1 18.500 60.300  35    1
```

- **Treatment**: e401k = 1 if the worker's employer offers a 401(k) plan, 0 otherwise.

- **Outcome**: nettfa = net total financial assets ($1000s).

- **Key covariate**: inc = family income ($1000s).

- This is **observational**: eligibility depends on the employer, and higher-income workers tend to work at firms offering 401(k) plans.

The selection problem

- Group means by eligibility status:

```
grp <- split(k401ksubs, k401ksubs$e401k)
tab <- rbind(
  "Ineligible (e401k=0)" = colMeans(grp[["0"]][, c("nettfa", "inc", "age")]),
  "Eligible (e401k=1)"   = colMeans(grp[["1"]][, c("nettfa", "inc", "age")])
)
round(tab, 1)
```

```
                      nettfa  inc  age
Ineligible (e401k=0)    11.7 34.1 40.8
Eligible (e401k=1)      30.5 47.3 41.5
```

- Eligible workers have much higher income and much higher savings. The raw gap in savings is **not** a treatment effect — it partly reflects income differences.

Separate regressions

- Estimate separate regressions of nettfa on inc for each group:

```
reg0 <- lm(nettfa ~ inc, data = k401ksubs, subset = (e401k == 0))
reg1 <- lm(nettfa ~ inc, data = k401ksubs, subset = (e401k == 1))
cbind(Ineligible = coef(reg0), Eligible = coef(reg1))
```

```
              Ineligible    Eligible
(Intercept) -14.7353612 -25.105194
inc           0.7753202   1.176382
```

- Compute the estimated ATE from the two-regression formula:

```
xbar <- mean(k401ksubs$inc)
a0 <- coef(reg0)[1]; b0 <- coef(reg0)[2]
a1 <- coef(reg1)[1]; b1 <- coef(reg1)[2]
ATE_manual <- (a1 - a0) + (b1 - b0) * xbar
cat("Sample mean of inc:", round(xbar, 2), "\n")
```

```
Sample mean of inc: 39.25
```

```
cat("Estimated ATE:", round(ATE_manual, 2), "\n")
```

```
Estimated ATE: 5.37
```

Demeaned regression

- Create the demeaned interaction and run the combined regression:

```
k401ksubs$inc_dm <- k401ksubs$inc - mean(k401ksubs$inc)
reg_dm <- lm(nettfa ~ e401k + inc + I(e401k * inc_dm), data = k401ksubs)
summary(reg_dm)$coefficients
```

```
                     Estimate Std. Error     t value     Pr(>|t|)
(Intercept)       -14.7353612 1.47212820 -10.009564 1.819941e-23
e401k               5.3737011 1.30596972   4.114721 3.910019e-05
inc                 0.7753202 0.03654010  21.218339 1.320126e-97
I(e401k * inc_dm)   0.4010617 0.05286179   7.586988 3.589988e-14
```

- The coefficient on `e401k` directly estimates the ATE, matching the result from the separate regressions.

Comparison across specifications

- Four specifications for the effect of 401(k) eligibility on net financial assets:

```
m_none <- lm(nettfa ~ e401k, data = k401ksubs)
m_inc  <- lm(nettfa ~ e401k + inc, data = k401ksubs)
m_int  <- lm(nettfa ~ e401k + inc + I(e401k * inc), data = k401ksubs)

make_row <- function(m) c(Estimate = coef(m)["e401k"],
                          SE = summary(m)$coef["e401k", "Std. Error"])
tab <- rbind(
  "No controls"            = make_row(m_none),
  "Income, no interaction" = make_row(m_inc),
  "Interaction (demeaned)"    = make_row(reg_dm),
  "Interaction (not demeaned)" = make_row(m_int)
)
round(tab, 2)
```

```
                           Estimate.e401k   SE
No controls                         18.86 1.35
Income, no interaction               6.06 1.31
Interaction (demeaned)               5.37 1.31
Interaction (not demeaned)         -10.37 2.53
```

- **No controls**: selection bias inflates the estimate — eligible workers earn more and save more regardless of 401(k) access.

- **Income, no interaction**: removes much of the bias but constrains the model to equal slopes.

- **Interaction (demeaned)**: the coefficient on `e401k` is the ATE evaluated at mean income, showing 401(k) eligibility increases savings by about $5,400.

- **Interaction (not demeaned)**: the coefficient on `e401k` is the **CATE** (conditional average treatment effect) evaluated at `inc = 0`:

$$\text{CATE}(x) = \text{E}\left[Y_i(1) - Y_i(0) \mid X_i = x\right] = (\alpha_1 - \alpha_0) + (\beta_1 - \beta_0)x.$$

At $x = 0$ (zero income), this is an extrapolation far outside the data range — the **sign flips** to negative.

- Contrast with `jtrain2`: there, controls barely mattered because randomization already eliminated selection bias. Here, controls and the demeaning trick are essential.

Summary

- The **fundamental problem of causal inference** is that we never observe both potential outcomes $Y_i(1)$ and $Y_i(0)$ for the same individual.

- Comparing group averages $\mathrm{E}\,[Y_i \mid D_i = 1] - \mathrm{E}\,[Y_i \mid D_i = 0]$ does not generally yield the ATE because of **selection bias**: treated and control groups may differ in their baseline characteristics.

- **Randomization** eliminates selection bias by making treatment independent of potential outcomes. The simple difference in means then equals the ATE, which can be estimated by regressing Y_i on a treatment dummy D_i.

- In **observational studies**, treatment is not randomly assigned. Under **selection on observables**, controlling for covariates X_i restores the "as good as random" property of treatment assignment.

- With covariates, $\text{ATE} = (\alpha_1 - \alpha_0) + (\beta_1 - \beta_0)\mathrm{E}\,[X_i]$ depends on both intercept and slope differences.

- The **demeaning trick** makes the ATE appear directly as the coefficient on D_i: interact D_i with $(X_i - \bar{X})$ instead of X_i. This absorbs the $(\beta_1 - \beta_0)\mathrm{E}\,[X_i]$ component into the D_i coefficient.