Linear Regression Application: Estimating Treatment Effects
This lesson: focus on estimating causal effect of a binary ‘treatment’
e.g., whether:

- Participation in a job training program increase earnings
- Receipt of transfer payment from a social program affects various social outcomes
- Implementation of new teaching method affects student outcomes
- Changes in financial sector rules/regulation affect behavior of financial entities

Objectives
- Introduction to “Treatment Evaluation” / “Program Evaluation” literature
- Introduction to “Potential Outcomes” framework
- Better understanding of regression analysis for causal inference
- Regression Discontinuity Design, Difference-in-Difference
Background

Suppose we have a sample of $N$ observations $\{Y_i\}_{i=1}^{N}$

- $N_1$ are ‘treated’
- $N_0$ are untreated $\quad N = N_1 + N_0$
- $Y_i$ represents outcome of interest
- $D_i$ indicator variable of whether subject $i$ was treated

\[
D_i = \begin{cases} 
1 & \text{if subject } i \text{ was treated} \\
0 & \text{if subject } i \text{ was not treated}
\end{cases}
\]

Can we use “difference-in-means” to estimate effect of treatment?

\[
\frac{1}{N_1} \sum_{\text{all treated } i} Y_i - \frac{1}{N_0} \sum_{\text{all untreated } i} Y_i
\]
Potential Outcomes Framework

- Careful way of thinking about this issue
- For each individual $i$ from population, define
  - $Y_i(1)$ ~ potential outcome if treatment is received
  - $Y_i(0)$ ~ potential outcome if treatment is not received

**Individual Treatment Effect:** $Y_i(1) - Y_i(0)$
- Not “before-after” effect, but “potential” effects

Outcome if treatment is received minus outcome if treatment is not received
i.e. “Counterfactuals” at same point in time

E.g., think of $Y_i(1)$ as potential annual earnings of an individual if she were to receive job training, $Y_i(0)$ as potential annual earning of the same individual if she were to not receive job training
Impossible to compute. For each $i$ we observe only $Y_i(1)$ or $Y_i(0)$, i.e.,

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

where $D_i = 1$ if subject received treatment, 0 otherwise

Two Measures of Treatment Effect:

**Average Treatment Effect**

$$\tau_{ate} = E[Y_i(1) - Y_i(0)]$$

*(average over the whole population)*

**Average Treatment Effect on the Treated**

$$\tau_{att} = E[Y_i(1) - Y_i(0) \mid D_i = 1]$$

*(average over segment of population who are treated)*

- Requires multiple observations
- Requires assumptions to ensure consistency of difference-in-mean estimator
“Stable Unit Treatment Value Assumption (SUTVA)”

(i) The potential outcomes for any “individual” do not vary with the treatments assigned to other units, and

(ii) Treatments given to “individuals” are the same (as opposed to different versions / efficacy of treatments across treated individuals leading to different potential outcomes).

(‘individual’ here refers to basic unit potential treated: a person, a worker, a firm, …)

If treatment of one individual bears externalities (positive or negative) on other individuals, then SUTVA is violated
**ATE, ATT, and Selection Biases**

Difference-in-mean estimator is sample version of

\[ E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] \]

We have

\[ E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] = E[Y_i(1) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0] \]

\[ = \underbrace{E[Y_i(1) \mid D_i = 1] - E[Y_i(0) \mid D_i = 1]}_{\tau_{att}} + \underbrace{E[Y_i(0) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0]}_{"Selection Bias"} \]

Highlights problem of comparing treated outcomes with untreated outcomes

- Average outcome of untreated \( E[Y_i(0) \mid D_i = 0] \) may not a good representation of potential earnings of the treated had they not been treated \( E[Y_i(0) \mid D_i = 1] \)
As a measure of ATE, \( E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] \) is affected by both

\[
E[Y_i(0) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0] \quad \text{and} \quad E[Y_i(1) \mid D_i = 1] - E[Y_i(1) \mid D_i = 0]
\]

selection biases. Let \( \pi = \Pr[D_i = 1] \). Then

\[
E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] \\
= \pi \{E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0]\} + (1 - \pi)\{E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0]\} \\
= \pi \left\{E[Y_i(1) \mid D_i = 1] - E[Y_i(0) \mid D_i = 1] + E[Y_i(0) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0]\right\} \\
+ (1 - \pi)\left\{E[Y_i(1) \mid D_i = 0] - E[Y_i(0) \mid D_i = 0] + E[Y_i(1) \mid D_i = 1] - E[Y_i(1) \mid D_i = 0]\right\} \\
= \underbrace{E[Y_i(1) - Y_i(0)]}_{\tau_{ate}} \\
+ \pi \left\{E[Y_i(0) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0]\right\} + (1 - \pi)\left\{E[Y_i(1) \mid D_i = 1] - E[Y_i(1) \mid D_i = 0]\right\}
\]
If you have

- \( E[Y_i(0) \mid D_i = 1] = E[Y_i(0) \mid D_i = 0] = E[Y_i(0)] \)

- \( E[Y_i(1) \mid D_i = 1] = E[Y_i(1) \mid D_i = 0] = E[Y_i(1)] \)

Then,

\[
\text{difference-in-mean } E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] = E[Y_i(1) \mid D_i = 1] - E[Y_i(0) \mid D_i = 1] = \tau_{att}
\]

\[
= E[Y_i(1) - Y_i(0)] = \tau_{ate}
\]

\[
\hat{\beta}_1 = \frac{1}{N_1} \sum_{\text{all treated } i} Y_i - \frac{1}{N_0} \sum_{\text{all untreated } i} Y_i \xrightarrow{p} \tau_{ate}
\]

This is the objective of Random Assignment in Randomized Controlled Trials
Randomized Controlled Trials

- Randomly assign eligible participants (who agree to participate) to either the
  - Treatment group
  - Control group
- Differences in outcome may depend on differences in characteristics of participants in the two groups
- Random assignment makes the ‘average characteristics’ of both groups the same

A few RCTs have been conducted in Economics

- Health, labor, development, behavioral economics
- Expensive, cannot double blind, requires agreement to participate, participants may adapt behavior when treated, substitution bias, ...
- Hard to get approval

But mostly we rely on observational data
Estimating $\tau_{ate}$ from Linear Regression Perspective:

Suppose

$$Y_i(0) = \alpha_0 + u_i(0), \ E[u_i(0)] = 0$$

$$Y_i(1) = \alpha_1 + u_i(1), \ E[u_i(1)] = 0$$

$$\tau_{ate} = \alpha_1 - \alpha_0$$

We have

$$Y_i = \alpha_0 + u_i(0) + (\alpha_1 + u_i(1) - \alpha_0 - u_i(0))D_i$$

$$= \alpha_0 + \tau_{ate}d_i + u_i(0) + (u_i(1) - u_i(0))D_i$$

$$Y_i = \alpha_0 + \tau_{ate}D_i + u_i$$
In the regression

\[ Y_i = \beta_0 + \beta_1 D_i + u_i \]

**OLS:**

\[ \hat{\beta}_1 = \frac{1}{N_1} \sum_{i \text{ treated}} Y_i - \frac{1}{N_0} \sum_{i \text{ untreated}} Y_i \]  

\[ (*) \]

**Exercise:** Show \( (*) \).

Is \( \hat{\beta}_1 \) consistent for ‘true causal effect of treatment’?

- If \( E[u_i \mid D_i] = 0 \)
- observations are not correlated

Then regression \( Y_i = \beta_0 + \beta_1 D_i + \epsilon_i \) consistently estimates \( E[Y_i \mid D_i] \), i.e.,

\[ \hat{\beta}_1 \xrightarrow{p} \tau_{ate} \]
If \( E[u_i(1) | D_i] = 0 \) and \( E[u_i(0) | D_i] = 0 \)

\[
E[u_i | D_i] = E[u_i(0) + (u_i(1) - u_i(0))D_i | D_i]
\]

\[
= E[u_i(0) | D_i] + \{E[u_i(1) | D_i] - E[u_i(0) | D_i]\} D_i = 0
\]

“Random Assignment Assumption”: \( d_i \) is independent of \( u_i(0) \) and \( u_i(1) \)

- Achieved if treatment is in fact randomly assigned
- Exactly what a RCT tries to achieve
- Equivalent to \( d_i \) is independent of \( y_i(0) \) and \( y_i(1) \)
  - selection not motivated by benefit of treatment
    - E.g., not the case that someone is selected because considered more likely to benefit from treatment
  - omitted factors that predict participation and outcome
Note: $u_i$ and $D_i$ are not independent even under Random Assignment Assumption

- $\text{var}[u_i \mid D_i = 0] = \text{var}[u_i(0)]$  
  \{ not necessarily \}

- $\text{var}[u_i \mid D_i = 1] = \text{var}[u_i(1)]$  
  \{ the same \}

Weighted Least Squares?

**Exercise:** Describe how you would do (feasible) weighted least squares to estimate $\tau_{ate}$ in the context of the above example.

If omitted factors that predict participation and outcome are reflected in certain observable variables, we can control for these
Example

$Y_i \sim$ labor income

$D_i \sim$ participation in job training program

$x \sim$ education, age, current and past earnings, …

Suppose $E[x] = \eta$ and

\[
Y_i(0) = \alpha_0 + \gamma_{01}(x_1 - \eta_1) + \gamma_{02}(x_2 - \eta_2) + \cdots + \gamma_{0k}(x_k - \eta_k) + u_i(0)
= \alpha_0 + (x - \eta)' \gamma_0 + u_i(0)
\]

\[
Y_i(1) = \alpha_1 + \gamma_{11}(x_1 - \eta_1) + \gamma_{12}(x_2 - \eta_2) + \cdots + \gamma_{1k}(x_k - \eta_k) + u_i(1)
= \alpha_1 + (x - \eta)' \gamma_1 + u_i(1)
\]
\[ Y_i = Y_i(0) + (Y_i(1) - Y_i(0))D_i \]

\[ = \alpha_0 + (x - \eta)'\gamma_0 + u_i(0) + (\alpha_1 + (x - \eta)'\gamma_1 + u_i(1) - \alpha_0 - (x - \eta)'\gamma_0 - u_i(0))D_i \]

\[ = \alpha_0 + (\alpha_1 - \alpha_0)D_i + (x - \eta)'\gamma_0 + (x - \eta)'(\gamma_1 - \gamma_0)D_i + u_i(0) + (u_i(1) - u_i(0))D_i \]

\[ = \alpha_0 + \tau_{ate}D_i + (x - \eta)'\gamma_0 + (x - \eta)'\delta D_i + u_i \]

\[ = \alpha_0 + \tau_{ate}D_i + \gamma_{01}(x_1 - \eta_1) + \cdots + \gamma_{0k}(x_k - \eta_k) + \delta_1(x_1 - \eta_1)D_i + \cdots + \delta_k(x_k - \eta_k)D_i + u_i \]

\[ = \alpha^*_0 + \tau_{ate}D_i + \gamma_{01}x_1 + \cdots + \gamma_{0k}x_k + \delta_1(x_1 - \eta_1)D_i + \cdots + \delta_k(x_k - \eta_k)D_i + u_i \]

where

\[ \delta = \gamma_1 - \gamma_0, \ u_i = u_i(0) + (u_i(1) - u_i(0))D_i, \ \alpha^*_0 = \alpha_0 - \gamma_{01}\eta_1 - \cdots - \gamma_{0k}\eta_k \]

\[ \tau_{ate} = \alpha_1 - \alpha_0 \]
This suggests that to estimate $\tau_{ate}$, we can run the regression

$$y_i = \beta_0 + \beta_1 D_i + \beta_2 x_{1i} + \cdots + \beta_{k+2} x_{ki} + \beta_{k+3}(x_{1i} - \bar{x}_1)D_i + \cdots + \beta_{2k+2}(x_{ki} - \bar{x}_k)D_i + \varepsilon_i$$

and take $\hat{\beta}_1$ as the estimated ATE

Unbiased/consistent estimate of ATE can be obtained if

- $D_i$ independent of $u_i(0)$ and $u_i(1)$ conditional on $x_i$
- (equivalently) $D_i$ independent of $y_i(0)$ and $y_i(1)$ conditional on $x_i$

"Conditional Independence Assumption" which gives us:

$$E[u_i \mid D_i, x_i] = E[u_i(0) + (u_i(1) - u_i(0))D_i \mid D_i, x_i]$$

$$= E[u_i(0) \mid D_i, x_i] + \{E[u_i(1) \mid D_i, x_i] - E[u_i(0) \mid D_i, x_i]\} D_i$$

$$= E[u_i(0) \mid x_i] + \{E[u_i(1) \mid x_i] - E[u_i(0) \mid x_i]\} D_i$$

$$= 0$$

Intuition: For all members of population with same $x$, assignment is random
Remarks:

(a) The regression

\[ y_i = \beta_0 + \beta_1 D_i + \beta_2 x_{1i} + \cdots + \beta_{k+2} x_{ki} + \beta_{k+3} (x_{1i} - \bar{x}_1) D_i + \cdots + \beta_{2k+2} (x_{ki} - \bar{x}_k) D_i + \varepsilon_i \]

is sometimes called the “Regression Adjustment”

(b) Can consider more flexible forms of \((x - \eta)'\gamma_0\) and \((x - \eta)'\gamma_1\)

- Interaction effects among the \(x\) variables, polynomials
- More advanced applications consider forms that are nonlinear in parameters, or use ‘non-parametric estimates’

**Exercise:** Show that the interactions between \(x_{ji} - \bar{x}\) and \(D_i\) should be dropped from the regression adjustment if (i) \(\gamma_0 = \gamma_1\), or (ii) \(Y_i(1) - Y_i(0) = \tau_{ate}\).

*Remark:* (i) can be tested, of course. (ii) says that individual treatment effects are constant across individuals, which seems a rather strong assumption
(c) Regression adjustment can be used to explore how average treatment effect changes over different values of \( x \), i.e., define

\[
\tau_{ate}(x) = \tau_{ate} + (x - \eta)'\delta
\]

\[
\tilde{\tau}_{ate}(x) = \hat{\beta}_1 + \hat{\beta}_{k+3}(x_{1i} - \bar{x}_1) + \cdots + \hat{\beta}_{2k+2}(x_{ki} - \bar{x}_k)
\]

\( \hat{\tau}_{ate} = \hat{\beta}_1 \) is simply \( \tilde{\tau}_{ate}(x) \) averaged over all \( x \)

(d) To calculate \( \hat{\tau}_{att} \) (average treatment effect of the treated), take the average of \( \tilde{\tau}_{ate}(x) \) over all all observations with \( D_i = 1 \):

\[
\tilde{\tau}_{att}(x) = \hat{\beta}_1 + \frac{1}{\sum_{i=1}^{N} D_i} \sum_{i=1}^{N} D_i \left[ \hat{\beta}_{k+3}(x_{1i} - \bar{x}_1) + \cdots + \hat{\beta}_{2k+2}(x_{ki} - \bar{x}_k) \right]
\]
(e) “Regression Adjustment” equivalent to following:

i. Regress for $D_i = 0$ sample: 
$$y_i = \zeta_{00} + \zeta_{01}x_{1i} + \cdots + \zeta_{0k}x_{ki} + \varepsilon_{0i}$$
Predict over full sample: 
$$\hat{y}_i^{(0)} = \hat{\zeta}_{00} + \hat{\zeta}_{01}x_{1i} + \cdots + \hat{\zeta}_{0k}x_{ki}$$

ii. Regress for $D_i = 1$ sample: 
$$y_i = \zeta_{10} + \zeta_{11}x_{1i} + \cdots + \zeta_{1k}x_{ki} + \varepsilon_{1i}$$
Predict over full sample: 
$$\hat{y}_i^{(1)} = \hat{\zeta}_{10} + \hat{\zeta}_{11}x_{1i} + \cdots + \hat{\zeta}_{1k}x_{ki}$$

iii. Calculate $\hat{\tau}_{ate} = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i^{(1)} - \hat{y}_i^{(0)})$

Resulting $\hat{\tau}_{ate}$ will be equal to $\hat{\beta}_1$ in “adjusted regression”

Equivalent to predicting counterfactuals, what “would have been” had treated not been treated, and vice versa.
Example from Wooldridge Textbook (jtrain98.csv)

- 1130 observations on male workers
- to evaluate effect of a job training programme
- participation based on past labor income, partly voluntary

- **train**: =1 if in job training
- **age**: in years
- **educ**: years of schooling
- **black**: =1 if black
- **hisp**: =1 if Hispanic
- **married**: =1 if married
- **earn96**: earnings in 1996, $1000s
- **unem96**: =1 if unemployed all of 1995
- **earn98**: earnings in 1998, $1000s
- **unem98**: =1 if unemployed all of 1998
\[ \hat{\tau}_{ate} = -2.050 \]

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>4.667042</td>
<td>1.145283</td>
<td>4.07501</td>
<td>0</td>
</tr>
<tr>
<td>TRAIN</td>
<td>2.410547</td>
<td>0.435262</td>
<td>5.538145</td>
<td>0</td>
</tr>
<tr>
<td>EARN96</td>
<td>0.372538</td>
<td>0.018626</td>
<td>20.00075</td>
<td>0</td>
</tr>
<tr>
<td>EDUC</td>
<td>0.362833</td>
<td>0.064047</td>
<td>5.665107</td>
<td>0</td>
</tr>
<tr>
<td>AGE</td>
<td>-0.181046</td>
<td>0.018875</td>
<td>-9.59183</td>
<td>0</td>
</tr>
<tr>
<td>MARRIED</td>
<td>2.481719</td>
<td>0.426263</td>
<td>5.822044</td>
<td>0</td>
</tr>
</tbody>
</table>

\[ \hat{\tau}_{ate} = 2.411 \]
\[ \hat{\tau}_{ate} = 3.106 \]
\[ (0.532) \]
$\hat{\tau}_{ate} = 3.106$

$(0.591)$
smpl @all if train=0
genr untrained_earn98 = earn98    ' = na for trained subjects
smpl @all if train=1
genr trained_earn98 = earn98      ' = na for untrained subjects
smpl @all

equation eq4a.ls trained_earn98 c earn96 educ age married

equation eq4b.ls untrained_earn98 c earn96 educ age married

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>5.779343</td>
<td>1.828301</td>
<td>3.161046</td>
<td>0.0017</td>
</tr>
<tr>
<td>EARN96</td>
<td>0.486116</td>
<td>0.050508</td>
<td>9.624475</td>
<td>0</td>
</tr>
<tr>
<td>EDUC</td>
<td>0.342964</td>
<td>0.112609</td>
<td>3.045622</td>
<td>0.0025</td>
</tr>
<tr>
<td>AGE</td>
<td>-0.137996</td>
<td>0.033972</td>
<td>-4.062058</td>
<td>0.0001</td>
</tr>
<tr>
<td>MARRIED</td>
<td>1.765867</td>
<td>0.697131</td>
<td>2.533049</td>
<td>0.0117</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>5.077574</td>
<td>1.386709</td>
<td>3.6616</td>
<td>0.0003</td>
</tr>
<tr>
<td>EARN96</td>
<td>0.353362</td>
<td>0.020004</td>
<td>17.66448</td>
<td>0</td>
</tr>
<tr>
<td>EDUC</td>
<td>0.377835</td>
<td>0.077778</td>
<td>4.857834</td>
<td>0</td>
</tr>
<tr>
<td>AGE</td>
<td>-0.196061</td>
<td>0.022744</td>
<td>-8.620472</td>
<td>0</td>
</tr>
<tr>
<td>MARRIED</td>
<td>2.759077</td>
<td>0.543997</td>
<td>5.071858</td>
<td>0</td>
</tr>
</tbody>
</table>
eq4a.fit trained_pred
eq4b.fit untrained_pred
genrate_i = trained_pred - untrained_pred
ate_i.hist
Matching

- Regression assumes
  - $E[Y_i(j) | x_i, D_i]$ does not depend on $D_i$, and
  - $E[Y_i(j) | x_i, D_i]$ is a linear function of $x_i$

- Matching is an alternative approach which avoids assumption that $E[Y_i(j) | x'_i, D_i]$ is a linear function of $x_i$

- Idea of matching is to estimate $E[Y_i(1) | x_i, D_i = 1] - E[Y_i(0) | x_i, D_i = 0]$ over observations with same (or similar $x_i$), the average these estimates using the distribution of $x_i$ for treated observations (to get $\tau_{att}$) or over both treated and untreated observations (to get $\tau_{ate}$)
Assume \( E[Y_i(j) | x_i, D_i] = E[Y_i(j) | x_i], j = 0,1 \)

Then

\[
\tau_{att} = E[Y_i(1) - Y_i(0) | D = 1] \\
= E_x \{ E[Y_i(1) | x_i, D_i = 1] - E[Y_i(0) | x_i, D_i = 1] | D = 1 \} \\
= E_x \{ E[Y_i(1) | x_i, D_i = 1] - E[Y_i(0) | x_i, D_i = 0] | D = 1 \}
\]

and

\[
\tau_{ate} = E[Y_i(1) - Y_i(0)] \\
= E_x \{ E[Y_i(1) | x_i, D_i = 1] - E[Y_i(0) | x_i, D_i = 1] \} \\
= E_x \{ E[Y_i(1) | x_i, D_i = 1] - E[Y_i(0) | x_i, D_i = 0] \}
\]
A simple example

- Suppose relevant control variables are $X_{1i}$ and $X_{2i}$, both binary. Suppose we have $N$ observations, spread over four categories

<table>
<thead>
<tr>
<th></th>
<th>$X_1 = 0$</th>
<th>$X_1 = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_2 = 0$</td>
<td>$N_{00}$</td>
<td>$N_{10}$</td>
</tr>
<tr>
<td>$X_2 = 1$</td>
<td>$N_{01}$</td>
<td>$N_{11}$</td>
</tr>
</tbody>
</table>

- In each category $(X_1, X_2) = (j, k)$, there are $N_{jk1}$ treated and $N_{jk0}$ untreated observations (call the categories $C_{jk}$, $j,k = 0,1$)
For each category, calculate

\[ \Delta_{jk} = \frac{1}{N_{jk1}} \sum_{i \in C(j,k)} D_i Y_i - \frac{1}{N_{jk0}} \sum_{i \in C(j,k)} (1 - D_i) Y_i \]

where \( Y_i = Y_i(0) + (Y_i(1) - Y_i(0)) D_i \)

Then calculate

\[ \tau_{ate} = \frac{N_{00}}{N} \Delta_{00} + \frac{N_{10}}{N} \Delta_{10} + \frac{N_{01}}{N} \Delta_{01} + \frac{N_{11}}{N} \Delta_{11} \]

and

\[ \tau_{att} = \frac{N_{001}}{N_T} \Delta_{00} + \frac{N_{101}}{N_T} \Delta_{10} + \frac{N_{011}}{N_T} \Delta_{01} + \frac{N_{111}}{N_T} \Delta_{11} \]

where \( N_T = N_{001} + N_{101} + N_{011} + N_{111} \), the number of treated individuals
- Can be hard to find matched treated and untreated individuals, especially when $x$
  variables are continuous
- Have to define ‘closeness’ so as to match individuals with ‘similar’ $x$
- Lots of details omitted, left for further reading / study in interested
  - Nearest neighbor matching
  - Kernel matching
  - Propensity Score matching
  - Stratification / Interval Matching
  - Radius Matching
  - …
Regression Discontinuity Design

- Strategy for estimating treatment effects in non-experimental setting where assignment to treatment based on whether an assignment variable exceeds a known threshold, treatment assigned to $i$ if assignment variable $A_i > c$

Example

- $A_i$ may be a test score
- Treatment: financial award, given if $A_i > c$ for some fixed threshold
- $Y_i$ may be some measure of future outcomes of interest (earnings? CGPA? enrollment in post-graduate programmes?)
- Does financial award improve future academic outcome?
Example

- Do incumbents have advantage in election
- $A \sim$ difference in vote share in previous election ($A>0$ means incumbent)
- $Y \sim$ vote share in current election

Method

- If outcome $Y_i$ is linear function of $A_i$, run regression

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 A_i + \varepsilon_i$$

$$\hat{\beta}_1 = \hat{\tau}_{ate}$$
Validity of RD

- subjects do not have precise control over whether they receive treatment or not
e.g., good students cannot guarantee that they meet the criterion of award
- there is no reason to expect outcome to be a discontinuous function of any variable except the assignment variable
- other factors determining $Y$ evolve smoothly with respect to $A$
- Any discontinuity in $Y_i$ at $A_i = c$ is due only to treatment
- RD as “Local Randomization”
  - subjects around cut off will be similar in characteristics
  - RDD does not require unconfoundedness with covariates (but can add if available to reduce standard errors)
If precise control is possible, then expect discontinuity in distribution of forcing variable at cut-off

- Rough check: plot histogram of $X$ with focus around assignment threshold

**Example**  

- Do incumbents have advantage in elections?
- Election data from 6558 Districts (US)

$X$ “Assignment” or “Forcing” Variable: difference in vote share between Democratic and Republican in one election. Treatment: Incumbent if $X_i > 0$.

$Y$ “Outcome” Variable: vote share in next election
Scatterplot and Distribution of Assignment Variable

Discontinuity visible from scatterplot

Distribution of $X$ does not appear to demonstrate discontinuity at $X = 0$
<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.442736</td>
<td>0.003168</td>
<td>139.7447</td>
<td>0</td>
</tr>
<tr>
<td>X&gt;0</td>
<td>0.113728</td>
<td>0.005528</td>
<td>20.57227</td>
<td>0</td>
</tr>
<tr>
<td>X</td>
<td>0.330533</td>
<td>0.005989</td>
<td>55.1863</td>
<td>0</td>
</tr>
</tbody>
</table>

**Extensions**

- Can include other covariates to reduce standard errors?
- Non-linear specifications?
- Narrow the “bandwidth” around threshold (observations with $X$ very close to -1 or 1 may not be so relevant)
- “Local” estimates
- “Fuzzy” Regression Discontinuity Designs
Difference-in-Difference Estimators

In some cases, you have data for both treated and control groups before and after treatment is given.

In such cases, can use a Difference-in-Difference Estimator to estimate causal effect of treatment.

**Example 13.3 Wooldridge, Effect of Garbage Incinerator’s Location on House Prices**

1981 ~ data on house prices and distance to incinerator  
*First rumors that there would be an incinerator built at that location began after 1978, construction began 1981*

1978 ~ data on house prices and distance to same location of future incinerator
Using 1981 data only

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>11.47852</td>
<td>0.03428</td>
<td>334.8438</td>
<td>0</td>
</tr>
<tr>
<td>NEARINC</td>
<td>-0.402571</td>
<td>0.064589</td>
<td>-6.232829</td>
<td>0</td>
</tr>
</tbody>
</table>

LPRICE ~ log of house prices in 1978 dollars

NEARINC ~ indicator variable 1 if house is within 3 miles of incinerator location

Houses NEARINC are about 40% lower in value than houses outside the 3 mile radius

*Did building the incinerator result in house values falling?*

- Maybe houses NEARINC are systematically different from houses not NEARINC
- Maybe incinerator location was chosen because house prices in that area are low
Add controls?

NBH is a ‘neighborhood number’ 1 to 6

After controls added, houses NEARINC are about 11% lower in value than houses not NEARINC. Other factors missed out?

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>7.484311</td>
<td>0.441523</td>
<td>16.95113</td>
<td>0</td>
</tr>
<tr>
<td>NEARINC</td>
<td>-0.110785</td>
<td>0.040598</td>
<td>-2.728852</td>
<td>0.0072</td>
</tr>
<tr>
<td>AGE</td>
<td>-0.01256</td>
<td>0.002149</td>
<td>-5.845956</td>
<td>0</td>
</tr>
<tr>
<td>AGESQ</td>
<td>7.24E-05</td>
<td>2.08E-05</td>
<td>3.477651</td>
<td>0.0007</td>
</tr>
<tr>
<td>LAREA</td>
<td>0.534182</td>
<td>0.05667</td>
<td>9.426128</td>
<td>0</td>
</tr>
<tr>
<td>NBH</td>
<td>-0.029138</td>
<td>0.007696</td>
<td>-3.786279</td>
<td>0.0002</td>
</tr>
</tbody>
</table>
What does 1978 data say?

- Pre-incinerator, houses ‘NEARINC’ were already 34% lower than houses not NEARINC
- Building of incinerator, house values NEARINC became even lower (by about 6%) relative to houses not NEARINC
- “Difference-in-difference” seems a more credible estimate of the causal effect of the building of the incinerator on house values
- Controls for other factors that may affect house prices

\[ Y_i = \beta_0 + \beta_1 \text{nearinc}_i + \beta_2 y81 + \beta_3 \text{nearinc}_i \cdot y81 + \varepsilon_i \]
Sample ave. of prices of houses in 1978 that are not nearinc = \( \hat{\beta}_0 \)

Sample ave. of prices of houses in 1978 that are nearinc = \( \hat{\beta}_0 + \hat{\beta}_1 \)

Sample ave. of prices of houses in 1981 that are not nearinc = \( \hat{\beta}_0 + \hat{\beta}_2 \)

Sample ave. of prices of houses in 1981 that are nearinc = \( \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 \)

\[
\hat{\beta}_3 = (\bar{Y}_{81,\text{nearinc}} - \bar{Y}_{81,\sim\text{nearinc}}) - (\bar{Y}_{78,\text{nearinc}} - \bar{Y}_{78,\sim\text{nearinc}})
\]

<table>
<thead>
<tr>
<th>Dependent Variable: LRPRICE</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>11.28542</td>
<td>0.030514</td>
<td>369.8383</td>
<td>0</td>
</tr>
<tr>
<td>NEARINC</td>
<td>-0.339922</td>
<td>0.054555</td>
<td>-6.230749</td>
<td>0</td>
</tr>
<tr>
<td>Y81</td>
<td>0.193094</td>
<td>0.045321</td>
<td>4.260607</td>
<td>0</td>
</tr>
<tr>
<td>NEARINC*Y81</td>
<td>-0.06265</td>
<td>0.083441</td>
<td>-0.750829</td>
<td>0.4533</td>
</tr>
</tbody>
</table>

Advantage of ‘pooling cross-sections’ and using regression framework
- Can compute standard errors (and can make s.e. robust against heteroskedasticity)
- Can include controls (to get more precise estimators)

| Dependent Variable: LRPRICE | | | |
|----------------------------|-------------------------------|
| Sample: 1 321              | | | |
| Variable                  | Coefficient          | Std. Error | t-Statistic | Prob.  |
| C                         | 7.131928             | 0.308577   | 23.11233    | 0      |
| NEARINC                   | -0.001326            | 0.040595   | -0.032673   | 0.974  |
| Y81                       | 0.133124             | 0.029797   | 4.467718    | 0      |
| NEARINC*Y81               | -0.115264            | 0.053929   | -2.137316   | 0.0333 |
| AGE                       | -0.01172             | 0.001202   | -9.753429   | 0      |
| AGESQ                     | 5.76E-05             | 7.81E-06   | 7.375807    | 0      |
| LAREA                     | 0.558947             | 0.040324   | 13.86126    | 0      |
| NBH                       | -0.013403            | 0.005661   | -2.367393   | 0.0185 |

- Note: $\hat{\beta}_3$ now is the ‘difference-in-difference’ of average prices after controlling for $AGE$, $AGE^2$, $\log(AREA)$ and $NBH$. 
- We are assuming that there is nothing else that changes relative prices between 1978 and 1981 (or that any other factor that may change relative prices have been controlled for)

- ‘Parallel trends’ assumption
  - Trends for ‘near’ and ‘far’ houses would have been the same in absence of the treatment

In some examples, presence of another control group may permit a ‘difference-in-difference-in-difference’ estimate
E.g. \( Y \sim \text{measure of health outcomes potentially impacted by a certain Policy} \)

**State A**

\[
Y_A,\text{low,t} \text{ vs } Y_A,\text{mid,t}
\]

**Policy instituted**

\[
Y_A,\text{low,t+1} \text{ vs } Y_A,\text{mid,t+1}
\]

\[
\text{DID} = (\bar{Y}_{A,\text{mid,t+1}} - \bar{Y}_{A,\text{low,t+1}}) - (\bar{Y}_{A,\text{mid,t}} - \bar{Y}_{A,\text{low,t}})
\]

**State B**

\[
Y_B,\text{low,t} \text{ vs } Y_B,\text{mid,t}
\]

**Policy not instituted**

\[
Y_B,\text{low,t+1} \text{ vs } Y_B,\text{mid,t+1}
\]

\[
\text{DID} = (\bar{Y}_{B,\text{mid,t+1}} - \bar{Y}_{B,\text{low,t+1}}) - (\bar{Y}_{B,\text{mid,t}} - \bar{Y}_{B,\text{low,t}})
\]

\[
\text{DIDID} = [(\bar{Y}_{A,\text{mid,t+1}} - \bar{Y}_{A,\text{low,t+1}}) - (\bar{Y}_{A,\text{mid,t}} - \bar{Y}_{A,\text{low,t}})] - [(\bar{Y}_{B,\text{mid,t+1}} - \bar{Y}_{B,\text{low,t+1}}) - (\bar{Y}_{B,\text{mid,t}} - \bar{Y}_{B,\text{low,t}})]
\]

**DIDID** will measure the effect of policy if health trends for low vs middle income groups if do not differ across States A and B
Exercise

What regression would you estimate to compute DIDID in the health outcomes example, and which parameter captures the DIDID?

Further Readings


Wooldridge Textbook – Sections 1-4, 2-7, 3-7e, 4-7, 7-6, 13-1, 13-2