# Lecture 7: Identification of Causal Effects Under Treatment Ignorability

# Outline of Next 3 Lectures:

- Bias in observational regression
- Discussion of treatment *ignorability* assumption
- Examples with simulated data
- Identification of ATE under treatment ignorability
- Application: effect of ITQ adoption on the probability of fishery collapse
- Readings: A&P Chapter 3 (especially sections 3.2 and 3.3), I&W Lecture 1

# Recall the Rubin Causal Model

$Y_i(0)$ = potential outcome for unit i if untreated

$Y_i(1)$ = potential outcome for unit i if treated

$T_i$ = 1 if unit i treated, 0 if not

- $Y_i(1) - Y_i(0)$ = unit-level treatment effect. We seek to estimate its average (ATE)

- We observe $(Y_i, T_i)$, where $Y_i$ is the observed outcome:

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

$$= E[Y_i(0)] + [Y_i(1) - Y_i(0)]T_i + [Y_i(0) - E(Y_i(0))]$$

$$= \beta_0 + \beta_{1i}T_i + u_i$$

# OLS regression in an observational study

$$Y_i = \beta_0 + \beta_1 T_i + u_i$$

When $T_i$ not randomly assigned, it is potentially correlated with $Y_i(0)$ and $Y_i(1)$ so we cannot pull $T_i$ from expectations

$$\hat{\beta}_1 \xrightarrow{p} E[Y_i \mid T_i = 1] - E[Y_i \mid T_i = 0]$$

$$= E[Y_i(1) \mid T_i = 1] - E[Y_i(0) \mid T_i = 0] \qquad \pm E[Y_i(0) \mid T_i = 1]$$

$$= E[Y_i(1) - Y_i(0) \mid T_i = 1] + \{E[Y_i(0) \mid T_i = 1] - E[Y_i(0) \mid T_i = 0]\}$$

$$= ATT + \text{bias}$$

- Observational regression of Y on T converge to ATT + bias
- Bias term arises because treated individuals differ from non-treated individuals in the non-treated state of the world
  - Ex: the wage of college dropouts is not a good counterfactual for wage of college grads if they had not gone to college

# Identification of causal effects under "treatment ignorability" assumption

- Assumption of "treatment ignorability" conditional on pre-treatment characteristics $X_i$ (Rubin and Rosenbaum 1983)
  - Sometimes labeled "conditional independence assumption" or "unconfoundedness" or "selection on observables"):

$$T_i \perp (Y_i(0), Y_i(1)) | X_i$$

- Combined with a common support assumption, we can identify ATE and ATT under assumption of treatment ignorability

- Most of the empirical literature falls under this assumption
  - Treatment ignorability is a strong assumption in most cases, and not always a realistic assumption, but often there is no better alternative

# Interpretation of treatment ignorability

- Conditional on the vector of baseline (pre-treatment) covariates $X_i$, $T_i$ is statistically independent of potential outcomes
  - "$T_i \approx$ randomly assigned conditional on $X_i$"
  - Put another way, the distribution of the potential outcomes $(Y_{0i}, Y_{1i})$ is the same across levels of the treatment variable $T_i$ once we condition on $X_i$

- The key will be to determine if "treatment ignorability" is credible, on an application-by-application basis
  - 1: Do we know the selection/treatment assignment rule?
  - 2: Do we have all the relevant pre-treatment covariates?
  - ** Often there is no treatment assignment rule per se (ex: MPA designation), so how do we chose the covariates then?
  - 3: Is there overlap in X distribution over T=1 and T=0?
  - 4: How do we adjust for X? Linear regression, matching, blocking

# Simple model of treatment ignorability

- Consider the following treatment assignment rule as a function of some pre-treatment covariates $X_{1i}$ and $X_{2i}$

$$T_i^* = X_{1i}\pi_1 + X_{2i}\pi_2 + v_i$$

$$T_i = 1(T_i^* > 0)$$

- $\Rightarrow$ The treatment is ignorable if $v_i$ is independent of potential outcomes ($Y_i(0)$ , $Y_i(1)$)

- Importance of $v_i$: without it there is no "experiment", i.e. conditional on $X_{1i}$ and $X_{2i}$, random variable $v_i$ is what randomly assigns units to $T_i=0$ or $T_i=1$

- What is $v_i$ in your application of treatment ignorability?

# Example: Simulated data

- To help illustrate some of the methods, we use this simple model of potential outcomes and pre-treatment covariates:

  N=30,000

  $X_i$ = {1,2,3,4,5}, each with 20% probability

  $Y_i(0) = \mu_0(X_i) + \varepsilon_{0i}$

  $Y_i(1) = \mu_1(X_i) + \varepsilon_{1i}$

  Assume $(\varepsilon_{0i}, \varepsilon_{1i}) \sim$ Bivariate Normal(0,0,1,1,0.8)

$\mu_0(X_i) = 0.5*1(X=1) + 1*1(X=2) + 1.5*1(X=3) + 2*1(X=4) + 2.5*1(X=5)$

$\mu_1(X_i) = 1*1(X=1) + 2*1(X=2) + 3*1(X=3) + 4*1(X=4) + 5*1(X=5)$

- $\Rightarrow E[Y_i(0)]=1.5$ and $E[Y_i(1)]=3$ so ATE = 1.5

# Benchmark: Random assignment

- $T_i$ is randomly assigned, with $Pr(T_i=1)=0.5$
- Uncover 3 results:

1. $T_i$ uncorrelated with potential outcomes
   - (p-values = 0.40 and = 0.79)
2. Simple regression of $Y_i$ on $T_i$ identifies ATE
   - (estimated beta1 = 1.49 (std error = 0.02))
3. Distribution of $X_i$ balanced across treatment / control group:

```
. table X, c(sum T_rct sum C_rct);

----------------------------------
         X | sum(T_rct)   sum(C_rct)
-----------+----------------------
         1 |       2960         2945
         2 |       2998         2964
         3 |       3070         3077
         4 |       3025         2977
         5 |       3038         2946
```

**Covariate balance generally fails when treatment selected in part based on X**

**That is the source of bias we can eliminate with selection on observable methods**

# Now suppose treatment ignorable conditional on X

- Suppose everything same as before, but:

$$T_i = \mathbf{1}(-2*1(X=1) - 1*1(X=2) + 1*1(X=4) + 2*1(X=5) + v_i > 0)$$

- Where $v_i \sim N(0,1)$

- Importantly $v_i$ independent of $(\varepsilon_{0i}, \varepsilon_{1i})$ and $X_i$

# Result 1: Distribution of X no longer balanced

```
. table X, c(sum T_soo sum C_soo);

----------------------------------
         X | sum(T_soo)   sum(C_soo)
-----------+----------------------
         1 |        124         5781
         2 |        898         5064
         3 |       3119         3028
         4 |       5003          999
         5 |       5849          135
----------------------------------
```

**… Due to treatment assignment mechanism (previous page), observations with larger values of $X_i$ more likely to be in treatment group**

# Result 2:
# OLS estimator no longer consistent for ATE

```
. regress Y_soo T_soo, robust;

Linear regression                               Number of obs =   30000
                                                F(  1, 29998) =43832.88
                                                Prob > F      =  0.0000
                                                R-squared     =  0.5937
                                                Root MSE      =  1.2535

------------------------------------------------------------------------------
             |               Robust
       Y_soo |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       T_soo |   3.030628   .0144755   209.36   0.000     3.002256    3.059001
       _cons |   1.004894   .0090492   111.05   0.000     .9871572    1.022631
------------------------------------------------------------------------------
```

- Recall ATE = 1.5, so we are off by 100%

- How can we correct this?

# Normalized difference in pre-treatment covariates

- The normalized difference in a covariate X is defined as:

$$ND(X) = \frac{\overline{X}_1 - \overline{X}_0}{\sqrt{(V_{\overline{X}_1} + V_{\overline{X}_0})/2}}$$

- That is, the ratio of the difference in the sample means of covariate X by T=1 and T=0, divided by square root of sum of sample variances of X by T=1 and T=0

- This is an easy to compute statistic to assess overlap / balance in distribution of X by treatment status

- Similar to a t-statistic, but independent of sample size

# Normalized difference in pre-treatment covariates

- In the simulated data:

```
. normdiff X, over(T_soo) tstat;
Variables completed: ..1
          Mean:         Mean:  Difference:  Difference:
      T_soo==0     T_soo==1   Normalized       t-stat
X    1.4792481    3.5361878    .82398334      100.901
N        15007        14993            .            .
```

- A rule of thumb: if the normalized difference in one or more pre-treatment covariate exceeds 0.25, linear regression methods tend to be sensitive to the specification

# Identification of causal effects under treatment ignorability

- 1. Multivariate matching approach ("exact" matching)
- 2. Traditional (linear) regression methods
- 3. Propensity score methods (including matching)
- $\Rightarrow$ **If treatment ignorability and overlap assumptions are satisfied, choice of estimator is not so important**
- $\Rightarrow$ Evidence of variability in estimates across estimation methods 1-3 may be indicative of failure of treatment ignorability assumption (or lack of overlap)

# Identification of causal effects under treatment ignorability (ctd)

- There is a common structure to all methods to solve the identification problem: find suitable counterfactuals $Y_i(0)$ (for the treated) and $Y_i(1)$ (for the non-treated)

- Some notation:

$$\hat{Y}_i(0) = \text{counterfactual for } Y_i(0) \text{ for } T_i = 1$$

$$\hat{Y}_i(1) = \text{counterfactual for } Y_i(1) \text{ for } T_i = 0$$

- The various methods (matching, linear regression, propensity score) use different approaches to "fill in" the counterfactuals (or their means)

# Counterfactuals construction:

- **Matching**: Construct $\hat{Y}_i(0)$ and $\hat{Y}_i(1)$ from observations with the "closest" value of the vector of pre-treatment characteristics $X_i$
  - Impute $Y_i(0)$ for the treated using $Y_i=Y_i(0)$ of non-treated observations with "closest" $X_i$
  - Impute $Y_i(1)$ for the non-treated using $Y_i=Y_i(1)$ of treated observations with "closest" $X_i$
- **Propensity score methods**: Construct $\hat{Y}_i(0)$ and $\hat{Y}_i(1)$ from observations with the "closest" value of the scalar variable $p(X_i)$, the propensity score, which is the probability of receiving the treatment, conditional on pre-treatment characteristics $X_i$ $(Pr(T_i=1)|X_i)$
  - Similar approach to matching. Next lecture

# Counterfactuals construction (ctd)

- **Linear regression:** use a linear function of $X_i$ to impute the counterfactual means:

$$E[\hat{Y}_i(0) \mid T_i = 1] = \overline{Y}_0 + \hat{\gamma}_0(\overline{X}_1 - \overline{X}_0)$$

$$E[\hat{Y}_i(1) \mid T_i = 0] = \overline{Y}_1 + \hat{\gamma}_1(\overline{X}_0 - \overline{X}_1)$$

- Linearity assumption is key to result here. Also note how the linear regression solves the overlap problem

# 1. Multivariate / Exact matching

- Consider the following contrast for the vector $X_i = x$

$$\Delta(x) \equiv E[Y_i \mid T_i = 1, X_i = x] - E[Y_i \mid T_i = 0, X_i = x]$$
$$= E[Y_i(1) \mid T_i = 1, X_i = x] - E[Y_i(0) \mid T_i = 0, X_i = x]$$
$$= E[Y_i(1) \mid X_i = x] - E[Y_i(0) \mid X_i = x] = ATE(X_i = x)$$

- Simply a difference in mean outcome for those with the same vector of covariates, by treatment/control status

- For a given vector $X_i = x$, ATE and ATT are the same (since conditional on $X_i = x$, treatment and potential outcomes are independent)
  - $\Rightarrow$ Unconditional ATE and ATT differ since the distribution of $X_i$ in the overall population and the $T_i = 1$ population may differ

# Estimation of unconditional ATE and ATT

$$\hat{\Delta}(x) = \overline{Y}_1(x) - \overline{Y}_0(x)$$

$$\hat{ATE} = \sum_x w_{ATE}(x)\hat{\Delta}(x), \quad \hat{ATT} = \sum_x w_{ATT}(x)\hat{\Delta}(x)$$

- $\Rightarrow$ Weighted sum of ATE(x), ATT(x)
- Applications in Angrist (Econometrica, 1998), Card and Sullivan (Econometrica, 1988).
- Issue: small cells, no common support
  - Solution: nearest neighbor matching (Abadie & Imbens)
  - Solution: propensity score matching – to come

# Multivariate matching with simulated data

| $X$ | $N(T=1)$ | $N(T=0)$ | $\hat{\Delta}(x)$ | $w_{ATE}(x)$ | $w_{ATT}(x)$ |
|---|---|---|---|---|---|
| 1 | 124 | 5,781 | 0.65 | 0.2 | 0.01 |
| 2 | 898 | 5,064 | 0.97 | 0.2 | 0.06 |
| 3 | 3,119 | 3,028 | 1.48 | 0.2 | 0.21 |
| 4 | 5,003 | 999 | 1.98 | 0.2 | 0.33 |
| 5 | 5,849 | 135 | 2.39 | 0.2 | 0.39 |

- Estimated ATE = 1.50 (multiply columns 4 and 5 and sum)
- Estimated ATT = 1.96 (multiply columns 4 and 6 and sum)

# Multivariate matching as regression estimator

- Angrist (1998) shows that the multivariate matching estimator can be obtained by the following regression:

$$Y_i = \beta_0 + \beta_1 T_i + \sum_{x \in X} d_{xi} \gamma_x + u_i$$

$$d_{xi} = 1(X_i = x)$$

- $d_{xi}$ defined over set of all possible "permutations" of vector $X_i$ (i.e. all values of $X_i$ for which we compute $\Delta(X_i=x)$ in prior case)

- This can lead to a high-dimensional regression. In Angrist's (1998) application: $X_i$=age (17-22), schooling (1-20), AFQT score (1-100)
  - Then the full set is 6*20*100 = 12,000 variables $d_{xi}$ !

# Multivariate matching as regression estimator with simulated data

```
. regress Y_soo T_soo X2 X3 X4 X5, robust;

Linear regression                               Number of obs =   30000
                                                F(  5, 29994) =16890.62
                                                Prob > F      =  0.0000
                                                R-squared     =  0.7358
                                                Root MSE      =   1.011

------------------------------------------------------------------------------
             |               Robust
       Y_soo |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       T_soo |   1.493656   .0182623    81.79   0.000     1.457861    1.529451
          X2 |   .4350547   .0187044    23.26   0.000     .3983932    .4717162
          X3 |   1.005929   .0204649    49.15   0.000     .9658172    1.046041
          X4 |   1.935344   .0240678    80.41   0.000      1.88817    1.982518
          X5 |   2.976642   .0258754   115.04   0.000     2.925926    3.027359
       _cons |   .4995081   .0131929    37.86   0.000     .4736495    .5253668
------------------------------------------------------------------------------
```

- Note: simple set of control are dummy variables for all value of $X_i = \{1,2,3,4,5\}$. Real applications will necessarily be more computational demanding

# Matching estimator in practice: Measuring the effect of college selectivity on earnings

- Influential study by S. Dale and A. Krueger (Quarterly Journal of Economics, 2002)
- Q: What is the economic return to attending a highly selective college?
  - Sample: top private universities (Princeton, Yale), top liberal arts colleges (Swarthmore, Williams), strong public universities (Michigan, Penn State), and others (all pretty selective)
  - The average (1978) SAT scores at these schools ranged from a low of 1,020 at Tulane to a high of 1,370 at Bryn Mawr. In 1976, tuition rates were as low as $540 at the University of North Carolina and as high as $3,850 at Tufts
- A: Combine data on college attended (mid 1970s) with earnings data from the mid 1990s

# Identification problem:

- Selectivity of college attended is not randomly assigned
  - On average, more talented students are more likely to attend more selective college

- Observational regression measure of the college selectivity wage premium may reflect unmeasured individual ability (in addition to the causal effect of college selectivity)

- Recall that the observational regression bias arises because treated individuals differ from non-treated individuals in the non-treated state of the world
  - Ex: the wage of a Canyon State U grad is not a good counterfactual for the wage of a UCSB Gaucho if he/she did not attend UCSB

**Research strategy: match individuals who applied to the same colleges, were admitted by the same colleges, but attended different colleges (exact matching)**

TABLE 2.1
The college matching matrix

| Applicant group | Student | Private | | | Public | | | 1996 earnings |
|---|---|---|---|---|---|---|---|---|
| | | Ivy | Leafy | Smart | All State | Tall State | Altered State | |
| A | 1 | | Reject | Admit | | Admit | | 110,000 |
| | 2 | | Reject | Admit | | Admit | | 100,000 |
| | 3 | | Reject | Admit | | Admit | | 110,000 |
| B | 4 | Admit | | | Admit | | Admit | 60,000 |
| | 5 | Admit | | | Admit | | Admit | 30,000 |
| C | 6 | | Admit | | | | | 115,000 |
| | 7 | | Admit | | | | | 75,000 |
| D | 8 | Reject | | | Admit | Admit | | 90,000 |
| | 9 | Reject | | | Admit | Admit | | 60,000 |

# Log earnings regression:

"Private" is intended to mean "more selective"

| | No selection controls | | | Selection controls | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Private school | .135 | .095 | .086 | .007 | .003 | .013 |
| | (.055) | (.052) | (.034) | (.038) | (.039) | (.025) |
| Own SAT score ÷ 100 | | .048 | .016 | | .033 | .001 |
| | | (.009) | (.007) | | (.007) | (.007) |
| Log parental income | | | .219 | | | .190 |
| | | | (.022) | | | (.023) |
| Female | | | −.403 | | | −.395 |
| | | | (.018) | | | (.021) |
| Black | | | .005 | | | −.040 |
| | | | (.041) | | | (.042) |
| Hispanic | | | .062 | | | .032 |
| | | | (.072) | | | (.070) |
| Asian | | | .170 | | | .145 |
| | | | (.074) | | | (.068) |
| Other/missing race | | | −.074 | | | −.079 |
| | | | (.157) | | | (.156) |
| High school top 10% | | | .095 | | | .082 |
| | | | (.027) | | | (.028) |
| High school rank missing | | | .019 | | | .015 |
| | | | (.033) | | | (.037) |
| Athlete | | | .123 | | | .115 |
| | | | (.025) | | | (.027) |
| Selectivity-group dummies | No | No | No | Yes | Yes | Yes |

These indicator variables form an equivalent to the high-dimensional exact matching regression estimator

# Other multivariate matching estimators

- There exist other matching estimators, using different criteria for defining "closeness" between 2 matches

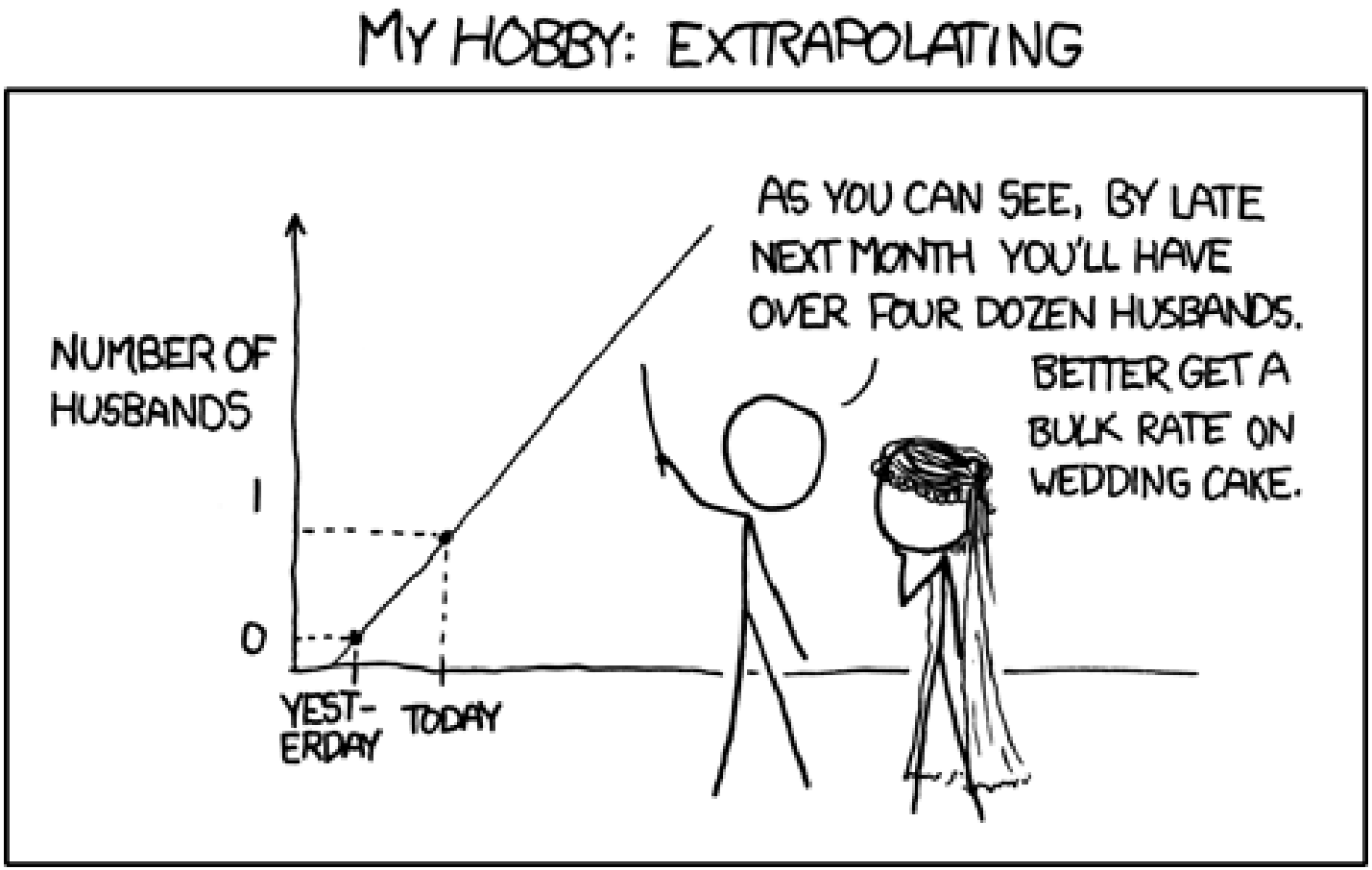
- Requiring exact matches typically creates empty cells (no treatment and control observations with the same vector X)

- Instead some methods use mathematical definitions of distance between 2 vectors to form the matches. The most prominent is the Mahalanobis distance:

$$\| X_i, X_j \| = (X_i - X_j)' \Omega_X^{-1} (X_i - X_j)$$

- See Abadie & Imbens 2006 for the method

# Causal effects with linear regression




$$E[\hat{Y}_i(0) \mid T_i = 1] = \overline{Y}_0 + \hat{\gamma}_0(\overline{X}_1 - \overline{X}_0)$$

$$E[\hat{Y}_i(1) \mid T_i = 0] = \overline{Y}_1 + \hat{\gamma}_1(\overline{X}_0 - \overline{X}_1)$$

# Linear regression

- Very common approach to estimate ATE with assumption of treatment ignorability
- Important to note that linear regression makes functional form assumption in addition to treatment ignorability conditional on $X_i$
- **Model:**

  $Y_i(0) = m_0 + \varepsilon_{0i}$

  $Y_i(1) = m_1 + \varepsilon_{1i}$
- **Observed outcomes:**

  $Y_i = m_0 + \beta_{1i} T_i + \varepsilon_{0i}$

  $Y_i = m_0 + \beta_1 T_i + T_i[(\varepsilon_{1i}-\varepsilon_{0i})] + \varepsilon_{0i}$ where $\beta_1 = m_1 - m_0 = ATE$
- What is the implied regression function $E[Y_i | T_i , X_i]$?

# Assumption: CEFs for potential outcomes linear in X

- **Suppose:**

$$E[\varepsilon_{0i}|X_i] = \alpha_0 + X_i'\gamma_0$$
$$E[\varepsilon_{1i}|X_i] = \alpha_1 + X_i'\gamma_1$$

**Scale the $X_i$ so that they have mean 0 $\Rightarrow \alpha_0 = \alpha_1 = 0$**

- Implied linear regression function:

$$Y_i = m_0 + \beta_1 T_i + (X_i - \overline{X})'\gamma_0 + T_i(X_i - \overline{X})'(\gamma_1 - \gamma_0) + u_i$$

- $\Rightarrow$ ATE identified by linear regression of outcome on treatment status indicator, controls for X, and interactions between treatment indicator status and X

- Other functional forms could be used for control functions
  - Quadratics, interactions
  - Exact matching "high-dimensional" indicator variable modeling

# Linear regression estimator with simulated data

```
. regress Y_soo T_soo X TX, robust;

Linear regression                               Number of obs =   30000
                                                F(  3, 29996) =26542.43
                                                Prob > F      =  0.0000
                                                R-squared     =  0.7275
                                                Root MSE      =  1.0266

------------------------------------------------------------------------------
             |               Robust
       Y_soo |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       T_soo |   1.619511   .0169724    95.42   0.000     1.586244    1.652778
           X |   .4633827   .0081726    56.70   0.000     .4473641    .4794013
          TX |   .4450819   .0118523    37.55   0.000     .4218509    .4683129
       _cons |   1.481247   .0117056   126.54   0.000     1.458303     1.50419
------------------------------------------------------------------------------
```

- Linear regression does an “OK” job, but it is clearly off from the ATE of 1.5 (equality rejected at <1% level)