Causal inference

Part I.b: randomized experiments, matching and regression

(this lecture starts with other slides on randomized experiments)

Frank Venmans

Example of a randomized experiment: Job Training Partnership Act (JTPA)

- Largest randomized training evaluation in US, started in 1983 at 649 site
- Sample: previously unemployed or low earnings
- D: assignment to on of 3 general service strategies
  - Classroom training in occupational skills
  - On the job traintin and/or job search assistance
  - Other services (eg. Probationary employment)
- Y: earnings 30 months following assignment
- X: characteristics measured befor assignment: age, gender, previous earnings, race, etc.

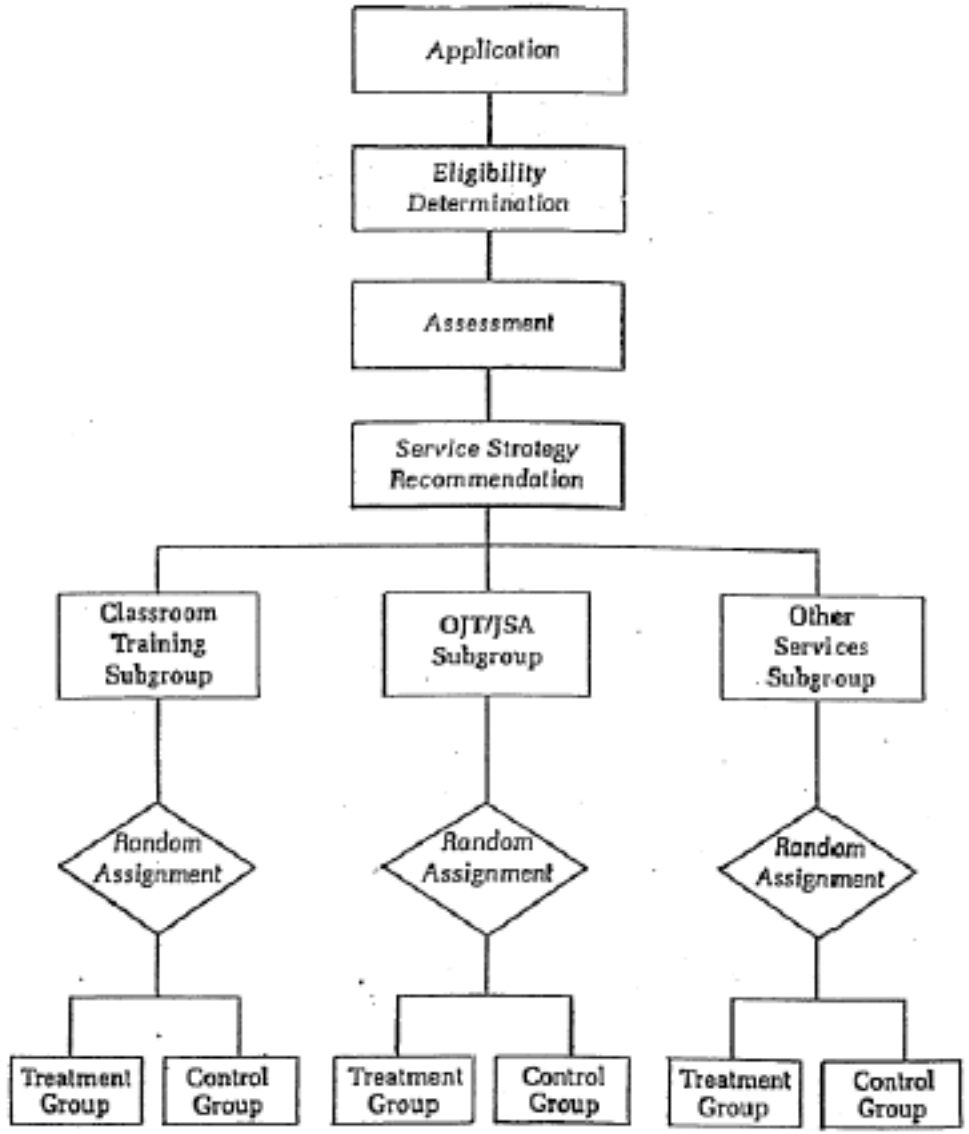

Application
Eligibility Determination
Assessment
Service Strategy Recommendation
Classroom Training Subgroup
OJT/JSA Subgroup
Other Services Subgroup
Random Assignment
Random Assignment
Random Assignment
Treatment Group
Control Group
Treatment Group
Control Group
Treatment Group
Control Group

Exhibit 5 *Impacts on Total 30-Month Earnings: Assignees and Enrollees, by Target Group*

	Mean earnings		*Impact per assignee*		
	Treatment group (1)	*Control group (2)*	*In dollars (3)*	*As a percent of (2)*	*Impact per enrollee in dollars*
Adult women	$ 13,417	$ 12,241	$ 1,176***	9.6%	$ 1,837***
Adult men	19,474	18,496	978*	5.3	1,599*
Female youths	10,241	10,106	135	1.3	210
Male youth non-arrestees	15,786	16,375	-589	-3.6	-868
Male youth arrestees					
Using survey data	14,633	18,842	-4,209**	-22.3	-6,804**
Using scaled UI data	14,148	14,152	-4	0.0	-6

	Entire Sample	Assignment		Difference (t-stat.)
		Treatment	Control	
A. Men				
Number of observations	5,102	3,399	1,703	
Treatment				
Training	.42	.62	.01	.61
	[.49]	[.48]	[.11]	(70.34)
Outcome variable				
30 month earnings	19,147	19,520	18,404	1,116
	[19,540]	[19,912]	[18,760]	(1.96)
Baseline Characteristics				
Age	32.91	32.85	33.04	-.19
	[9.46]	[9.46]	[9.45]	(-.67)
High school or GED	.69	.69	.69	-.00
	[.45]	[.45]	[.45]	(-.12)
Married	.35	.36	.34	.02
	[.47]	[.47]	[.46]	(1.64)
Black	.25	.25	.25	.00
	[.44]	[.44]	[.44]	(.04)
Hispanic	.10	.10	.09	.01
	[.30]	[.30]	[.29]	(.70)
Worked less than 13 weeks in past year	.40	.40	.40	.00
	[.47]	[.47]	[.47]	(.56)

Policy outcome

- After the results of the JTPA study, funding for the youth were drastically cut.

Selection on observables

Observational studies

- Not always possible to randomize (eg. Effect of smoking)
- Main problem selection bias
- Goal is to design observational study to approximate an experiment

Smoking and Mortality (Cochran 1986)

TABLE 1
DEATH RATES PER 1,000 PERSON-YEARS

Smoking group	Canada	U.K.	U.S.
Non-smokers	20.2	11.3	13.5
Cigarettes	20.5	14.1	13.5
Cigars/pipes	35.5	20.7	17.4

TABLE 2
MEAN AGES, YEARS

Smoking group	Canada	U.K.	U.S.
Non-smokers	54.9	49.1	57.0
Cigarettes	50.5	49.8	53.2
Cigars/pipes	65.9	55.7	59.7

Subclassification

- Need to control for differences in age.
- Subclassification:
  - For each country, divide each group in different age subgroups
  - Calculate death rates within age subgroups
  - Average within age subgroup death rates using fixed weights (eg. Number of cigarette smokers)

Subclassification: example

	Death Rates Pipe Smokers	# Pipe-Smokers	# Non-Smokers
Age 20 - 50	15	11	29
Age 50 - 70	35	13	9
Age + 70	50	16	2
Total		40	40

- What is average death rate for pipe smokers?
- 15*(11/40)+ 35*(13/40)+50 (16/40)=35,5
- What is average death rate for pipe smokers if they had the same age distribution as non-smokers?
- 15*(29/40)+35*(9/40)+50*(2/40)=21,2

TABLE 3

ADJUSTED DEATH RATES USING 3 AGE GROUPS

Smoking group	Canada	U.K.	U.S.
Non-smokers	20.2	11.3	13.5
Cigarettes	28.3	12.8	17.7
Cigars/pipes	21.2	12.0	14.2

- Effect of cigarettes was underestimated because cigarette smokers were younger than average
- Effect of cigars was overestimated because cigar smokers are older than average

Covariates, outcomes and post-treatment bias

Predetermined Covariates:

- Variable X (ex age) is predetermined with respect to treatment D (smoking) if for each individual i, $X_{0i}=X_{1i}$
- This does not imply that X and D are independent
- Are often time invariant, but not necessarily

Outcomes

- Variables Y (ex death rate, lung cancer, color of teeth) that are (possibly) not predetermined are called outcomes if for some individual i, $Y_{0i} \neq Y_{1i}$
- In general, wrong to condition on outcomes, because this may induce post-treatment bias

Identification assumption

- ATE
- $(Y_1, Y_0) \perp D|$X (selection on observables)
  - For a given value of X, potential outcomes are the same for treated and control units
  - This means that all variables that affect the outcome and probability of being treated must be included in the model (X is a vector of covariates)!
- $0 < \Pr(D = 1|X) < 1 \; for \; (almost \; all)X$ (common support)
  - For a every value of X there is a non-zero probability to find treated and control units

ATET

- $Y_0 \perp D|$X (selection on observables)
  - For a given age, the death rate if they would have been non-smokers should be the same for smokers and non-smokers
- $\Pr(D = 1|X) < 1 \; (with \Pr(D = 1) > 0)$(common support)
  - For every value of X there is a non-zero probability to find control units. If for some values of X, there are no treated units, this is not a problem.

Subclassification estimator

- $\hat{\alpha}_{ATE} = \sum_{k=1}^{K}\left(\bar{Y}_1^k - \bar{Y}_0^k\right)\left(\frac{N^k}{N}\right); \hat{\alpha}_{ATET} = \sum_{k=1}^{k}\left(\bar{Y}_1^k - \bar{Y}_0^k\right)\left(\frac{N_1^k}{N_1}\right)$
- N^k is # of obs. and N_1^k is #of treated obs in cell k

X_k	Death rate smokers	Death rate non-smokers	Diff.	# smokers	# Obs.
Young	28	24	4	3	10
Old	22	16	6	7	10
Total	23,8	21,6	2,2	10	20

- ATE= 4 * (10/20) + 6 * (10/20)=5
- ATET= 4 * (3/10) + 6 * (7/10)= 5,4

Matching

- Calculate $\hat{\alpha}_{ATET}$ by « imputing » the missing potential outcome of eacht treated unit using the observed outcome from the « closest » control unit:
- $\hat{\alpha}_{ATET} = \frac{1}{N_1}\sum_{D_i=1}\left(Y_i - Y_{j(i)}\right)$
- With $Y_{j(i)}$ the outcome of an untreated observation such that $X_{i(j)}$ is the closest value to X_i among the untreated observations.
- Alternative: use the M closest matches
- $\hat{\alpha}_{ATET} = \frac{1}{N_1}\sum_{D_i=1}\left\{Y_i - \left(\frac{1}{M}\sum_{m=1}^{M} Y_{jm(i)}\right)\right\}$

Example

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	Y_{1i}	Y_{0i}	D_i	X_i
1	6	?	1	3
2	1	?	1	1
3	0	?	1	10
4	?	0	0	2
5	?	9	0	3
6	?	1	0	-2

- ATET= 1/3(6-9) + 1/3(1-0) + 1/3(0-9) = -3,7

Trade-off between bias and efficiency

Single matching vs multiple matching

- Single matching: only the best match is used => lower bias
- Multiple matching: a greater set of information is used=>more efficient (lower standard errors of estimate)

Matching with replacement vs without replacement

- Matching with replacement: the best match can be used several times => lower bias
- Matching without bias: the best match may not be picked because it served already as a match. Therefore the second best match is used. This increases the set of information that is used. More efficient.

Distance metric

- When there are multiple confounders, a distance metric needs to be specified.
- Euclidian distance: every variable (standardized to have the same variance) has the same weight. Ex if there are 3 variables, you would match points that are closest in a standardized 3D plot.
- Mahalanobis distance: takes into account correlations between variables. If two variables are highly correlated, they receive less weight. This is in many cases theoretically more appealing.
- You can impose an exact match on certain variables (for example country, or sector), combined with another distance metric for other variables.

Bias correction

- If there are multiple continuous variables, matching estimators may behave badly.
- $\tilde{\alpha}_{ATET} = \frac{1}{N_1}\sum_{D_i=1}\left(Y_i - Y_{j(i)}\right) - \left(\hat{\mu}_0(X_i) - \hat{\mu}_0\left(X_{j(i)}\right)\right)$
- Where $\mu_0(X_i) = E[Y|X = X_i, D = 0]$ and $\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$ … estimated by OLS
- For example, if treated companies are much smaller than control companies, even if the matching algorithm searches among the smallest of control companies, the mean size of the control may still be greater than the mean size of the treated companies. Bias correction will calculate a size effect and deduce this from the estimated treatment effect (Abadie & Imbens, 2006).

Variance estimation (optional)

- Best with replacement to eliminate bias.
- But replacement will increase variance compared to a standard estimation of the form $\widehat{Var}(\hat{\alpha}_{ATET}) = \frac{1}{N_1^2}\sum_{D_i=1}\left(Y_i - Y_{i(j)} - \hat{\alpha}_{ATET}\right)^2$ (analytical solution not given)
- (Therefore bootstrap does not work)

Propensity score matching

- Propensity score is the probability of being treated conditional on the confounding variables: $\pi(X) = P(D = 1|X)$
- It can be shown that if $(Y_1, Y_0) \perp D|X \Rightarrow (Y_1, Y_0) \perp D|\pi(X)$
- If 2 individuals or companies are both as likely to be treated given the combination of their confounders X, then they are a good (unbiased) match.
- Ex: if both older and male individuals smoke more, a good match for a man would be a women that is a little bit older.
- Identification assumptions are the same: selection on observables and common support

Propensity score: estimation

- 1st step:
  - Estimate the propensity score $\pi(X) = P(D = 1|X)$ using logit/probit regression
- 2nd step:
  - Do matching (or sub-classification) on the propensity score
  - OR: multiply every observation by a weight based on the propensity score (no proof)
    - $\hat{\alpha}_{ATE} = \frac{1}{N}\sum_{i=1}^{N} Y_i \frac{D_i - \hat{\pi}(X_i)}{\hat{\pi}(X_i)\left(1-\hat{\pi}(X_i)\right)}$
    - $\hat{\alpha}_{ATET} = \frac{1}{N}\sum_{i=1}^{N} Y_i \frac{D_i - \hat{\pi}(X_i)}{1-\hat{\pi}(X_i)}$
    - Standard error estimation: need to adjust for first step estimation of propensity score.
      - Analytical solution: paramtetric first step: Neway & Mc Fadden (1994) or Newey (1994).
      - Alternative: Bootstrap

Matching in Stata

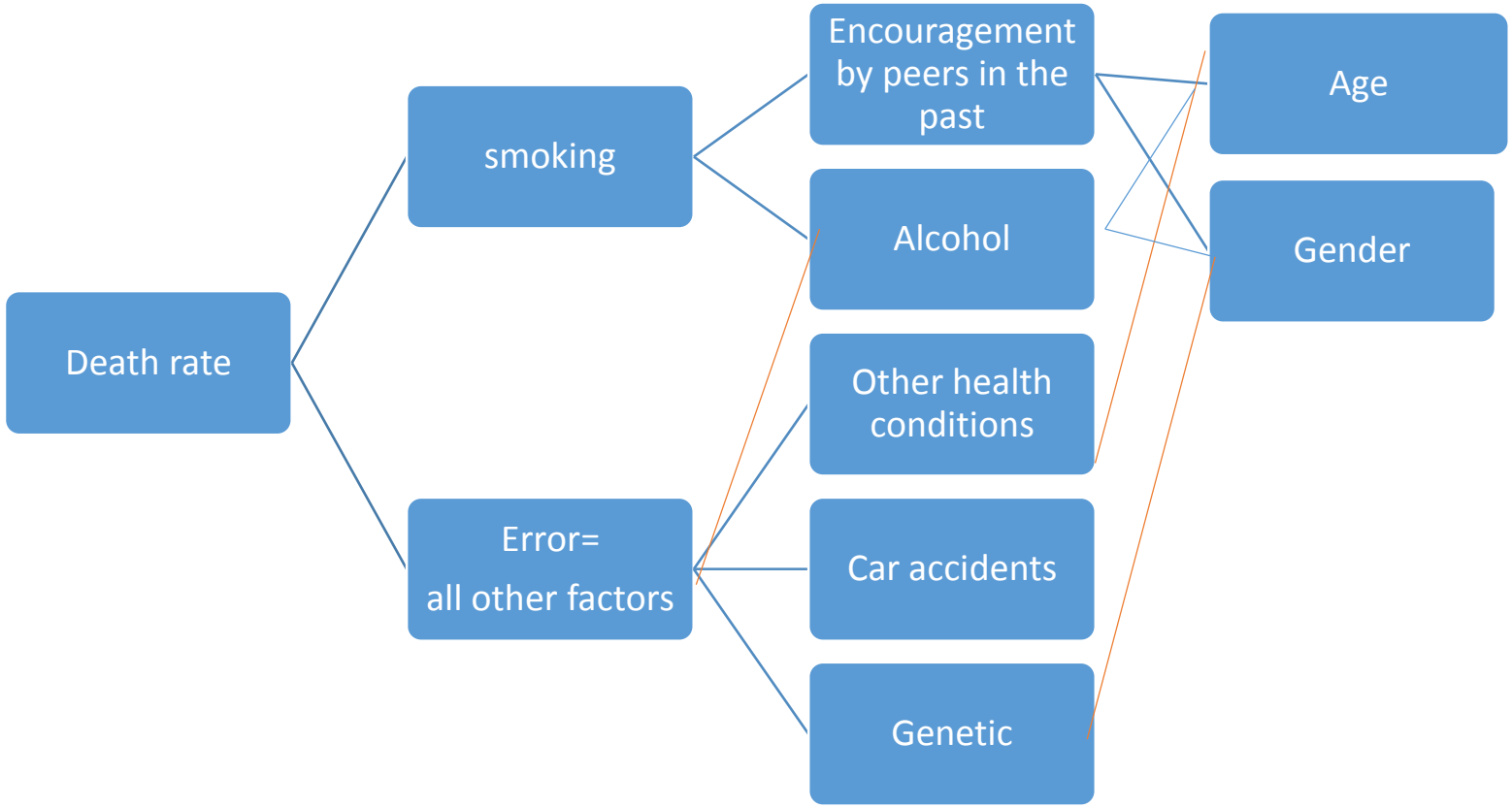
- Need to download package with command:
  - ssc install nnmatch, replace
  - ssc install psmatch2, replace
- Nnmatch, nearest neighbour matching doesn't do propensity score matching.
- PSmatch2 propensity score matching, but does also Mahalanobis matching. No exact matching.
- Matching in R has some more options compared to Stata

Ceteris paribus interpretation of regression

- 5 Gauss-Markov assumptions:
  - The true model is $Y = \alpha_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \epsilon$ with $\mathrm{E}[\epsilon] = 0$ (linearity)
  - No perfect collinearity (you cannot write X_1 as a linear combination of the other X_j's)
  - Homoscedastic errors $E(\epsilon_i^2) = \sigma^2$
  - Uncorrelated errors $E(\epsilon_i \epsilon_j) = 0$

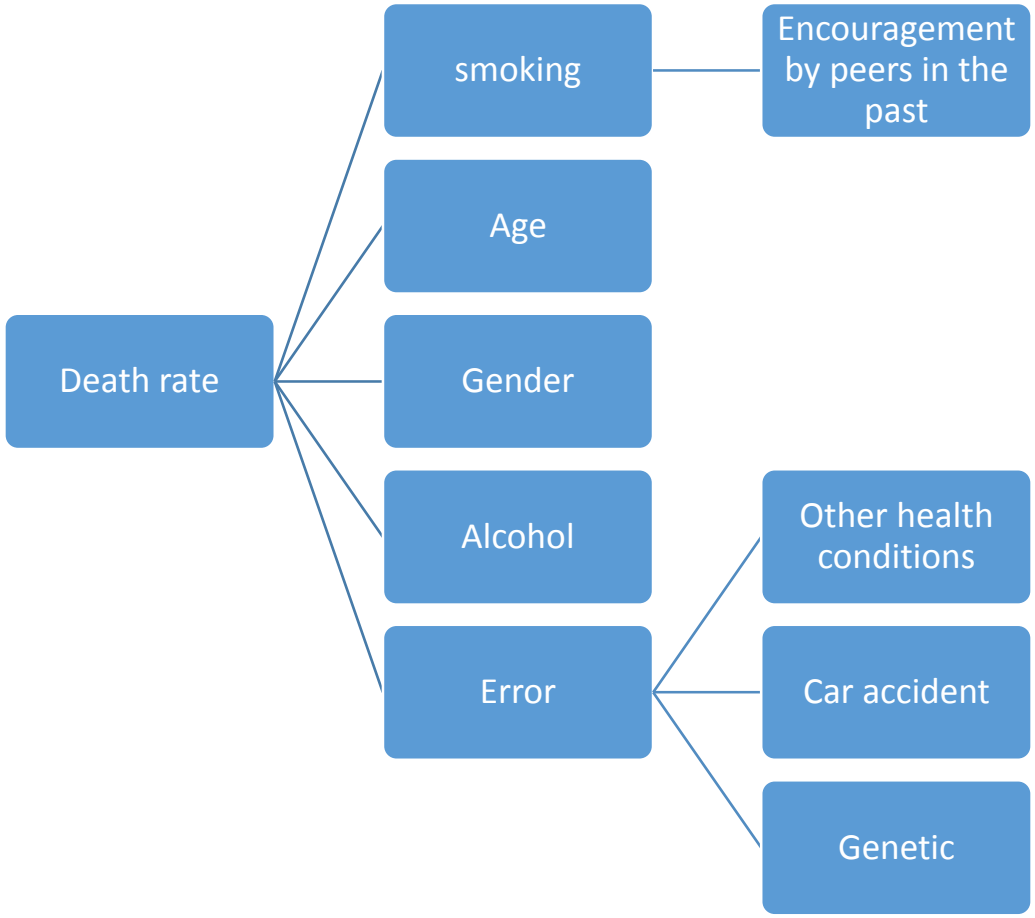
  $matrix\ notation\ E[\epsilon\epsilon'] = \sigma I$
  - $E[\epsilon | X_1, X_2] = 0$ (exogenous explainatory variables, no endogeneïty)
- Conditions 1 and 5 imply $E(Y|X_1, X_2) = \alpha_0 + \beta_1 X_1 + \beta_2 X_2$.
- This allows a ceteris paribus interpretation of beta's: all other relevant factors being equal, an increase of X1 by one unit will increase Y by β_1.
- For a causal interpretation of regression, some extra caution is needed
  - The relationship must be specified in the correct way, i.e. X causes (precedes) Y and not the inverse
    - Remark: If there is a causation in 2 senses, (think of a feedback loop), condition 5 will be violated (simultaneïty).
  - There may not be a causal relationship between treatment and covariates.
    - Ex: the effect of parents education on school results of their children keeping income constant underestimates the causal effect of parents' educations on their children's results. Because the indirect effect is not included.

The effect of smoking « all else being equal »



$E[\epsilon|X\ ] \neq 0 \Rightarrow cov(\epsilon, X) \neq 0 \Rightarrow \epsilon$ and Xare driven by common factors.

The effect of smoking « all else being equal »



Sources of endogeneity

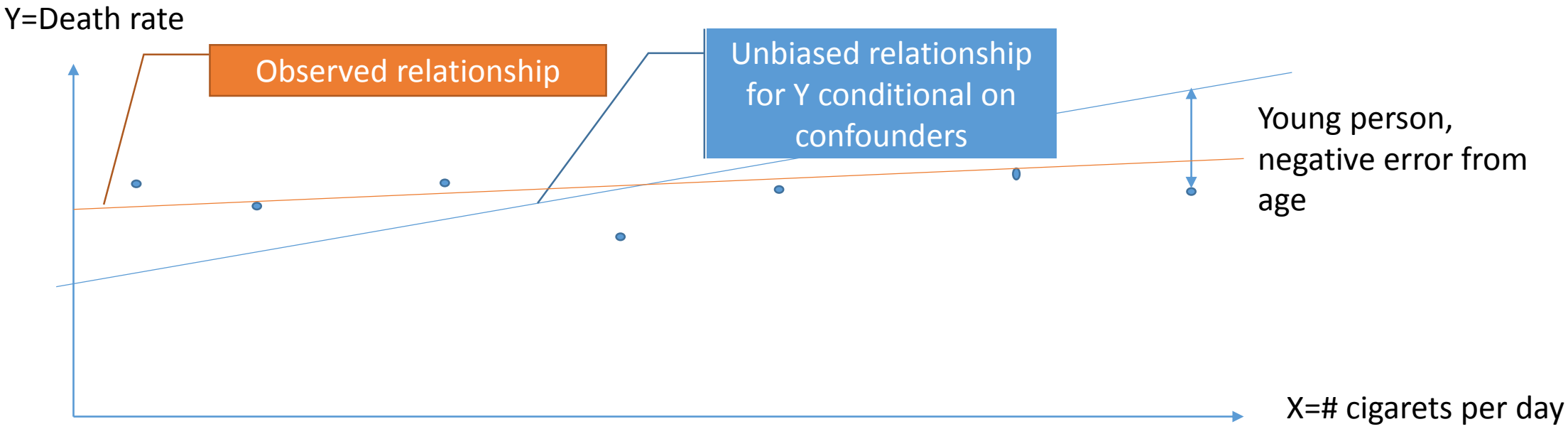
- Assume that the real model is: $Y = \alpha_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
- Assume that we omit X_2 and estimate $Y = \alpha_0 + \beta_1 X_1 + u$ with $u = \beta_2 X_2 + \epsilon$
- $E[X_1 u] = E[X_1(\beta_2 X_2 + \epsilon)] = \underbrace{E[X_1 X_2]\beta_2}_{\neq 0 \; if \; X1 \; and \; X2 \; correlated \; and \; Y \; and \; X2 \; correlated} + \underbrace{E[X_1 \epsilon]}_{0} \neq 0$

$$\Rightarrow \mathrm{E}[\mathrm{u}|\mathrm{X}_1] \neq 0$$

- Leaving out a confounder creates an endogeneity bias (to all betas).
- Other causes of endogeneity (which can be framed as an omitted variable problem)
  - Measurement errors correlated with X and Y.
  - Simultaneity: X causes Y but also Y causes X (ex. prices as a function of concentration in aviation sector; supply and demand function)
  - Y_{t-1} as a regressor when errors are serially correlated (ϵ_{t-1} affects ϵ_t and y_{t-1} as well).

Smoking example again

- Assume we want to know the relationship between death rates and number of cigarettes per day, but we omit age =>error correlated with X



Matching vs regression

- Consider a regression of the form: $Y = \alpha_0 + \alpha_1 D + \beta_1 X_1 + \beta_2 X_2 \ldots + \epsilon$
- The standard condition of exogenous X's $E[\epsilon|D, X_1, X_2] = 0$ boils down to the condition $Y_0, Y_1 \perp D|X$
  - all variables affecting the outcome and treatment probability at the same time are included in the model.
- The same holds if D is not a dummy, but a continuous variable representing a continuum of treatments and a continuum of counterfactual scenarios.
- The estimated $\hat{\alpha}_1$ is a variance-weighted average treatment effect (~ a variance- weighted multiple matching without replacement).
- Matching allows for a balance check (check that treated and controls have the same mean X_1, X_2…) which is an advantage over OLS.
- Matching only needs the linearity of the model for its bias correction. When exact matching, linearity of the model is not necessary.
- In general, OLS will be more efficient and at a higher risk of bias
- Both are based on the following assumptions
  - Selection on observables
  - Control variables may not have a causal relation with the treatment variable (else => use structural equation modeling)
  - Common support