Lecture 2-3: Regression models and matching estimators

P.J. Messe[1]

[1]Le Mans Université GAINS-TEPP, CEET, LEMNA

Master in Applied Econometrics

Some basics about regression

- The starting point: a Data Generating Process (DGP) that defines an infinite population

$$y_i = E(Y_i|X_i) + \epsilon_i$$

- $E(Y_i|X_i)$: the **Conditional Expectation Function** (CEF) of y_i given a set of characteristics (a $1xK$ vector X_i).

Some basics about regression

- The starting point: a Data Generating Process (DGP) that defines an infinite population

$$y_i = E(Y_i|X_i) + \epsilon_i$$

- $E(Y_i|X_i)$: the **Conditional Expectation Function** (CEF) of y_i given a set of characteristics (a $1xK$ vector X_i).

- In this course, X_i is a $1xK$ vector (different from Angrist and Pischke, 2008)
  - the number of rows (N) is the number of observations
  - the number of columns (K) corresponds to the number of regressors (**don't forget the constant!**)
  - Each X_i has one line: each column K is the value of each regressor K for the unit i

The researcher's challenge in regression models

- Issue: DGP (population CEF and residual) is not directly observable
  - The sample at hand is a **random draw** of y_i and X_i independtly and identically distributed (i.i.d.) from this population
  - Population moments (expectations of powers of random variable) are not directly observable

The researcher's challenge in regression models

- Issue: DGP (population CEF and residual) is not directly observable
  - The sample at hand is a **random draw** of y_i and X_i independtly and identically distributed (i.i.d.) from this population
  - Population moments (expectations of powers of random variable) are not directly observable

- The goal of the researcher
  - Finding unbiased (consistent estimator) of the population moments' distribution
  - With a variance as small as possible (to improve the precision of the estimators)

A quick reminder about OLS

- Let X be a NxK vector of covariates: Each line is a vector X_i
- Let Y be a $Nx1$ vector of values of the dependent variable: Each line is the value of Y_i
- Let ϵ be a $Nx1$ vector of residual values. Each line is the value of ϵ_i.

A quick reminder about OLS

- Let X be a NxK vector of covariates: Each line is a vector X_i
- Let Y be a $Nx1$ vector of values of the dependent variable: Each line is the value of Y_i
- Let ϵ be a $Nx1$ vector of residual values. Each line is the value of ϵ_i.

A quick reminder about OLS

- Let X be a NxK vector of covariates: Each line is a vector X_i
- Let Y be a $Nx1$ vector of values of the dependent variable: Each line is the value of Y_i
- Let ϵ be a $Nx1$ vector of residual values. Each line is the value of ϵ_i.

- The OLS estimator $\hat{\beta}$ writes as:

$$\underbrace{\hat{\beta}}_{Kx1} = \underbrace{(X'X)^{-1}}_{KxK} \underbrace{X'Y}_{Kx1}$$

A quick reminder about OLS: assumptions

Main assumptions (Gauss-Markov assumptions):

- Exogeneity of covariates: $cov(X_i, \epsilon_i) = 0$ or $E(X_i' \epsilon_i) = 0$
- $X'X$ is a full-rank (non-singular) matrix: inversible
  - No perfect multicolinearity between covariates
- $E(\epsilon_i) = 0$
- $E(\epsilon_i^2 | X_i) = \sigma_\epsilon^2$: homoskedasticity
- $E(\epsilon_i \epsilon_j | X_i, X_j) = 0$: no serial correlation between error-terms

A quick reminder about OLS: inference

Under Central Limit Theorem, we derive the limiting distribution of $\hat{\beta}$:

$$\hat{\beta} \xrightarrow{d} N(\beta, (X'X)^{-1}(\frac{N}{(N-K)}\sum \hat{\epsilon}_i^2 X_i X_i')(X'X)^{-1})$$

- where $\hat{\epsilon}_i = Y_i - X_i'\hat{\beta}$
- The variance of each $\hat{\beta}$ is given by the diagonal elements of this matrix

A quick reminder about OLS: inference

Under Central Limit Theorem, we derive the limiting distribution of $\hat{\beta}$:

$$\hat{\beta} \xrightarrow{d} N(\beta, (X'X)^{-1}(\frac{N}{(N-K)}\sum \hat{\epsilon}_i^2 X_i X_i')(X'X)^{-1})$$

- where $\hat{\epsilon}_i = Y_i - X_i'\hat{\beta}$
- The variance of each $\hat{\beta}$ is given by the diagonal elements of this matrix
- Under homoskedasticity assumption, the variance-covariance matrix of the estimator writes as:

$$(X'X)^{-1}\sigma_\epsilon^2$$

- where $\sigma_\epsilon^2 = \frac{1}{(N-K)}\sum \hat{\epsilon}_i^2$

A quick reminder about OLS: inference

What happens if errors are likely to be serially correlated

- For instance if the sample is drawn from different groups
- We could suspect correlation between units belonging to the SAME group

A quick reminder about OLS: inference

What happens if errors are likely to be serially correlated

- For instance if the sample is drawn from different groups
- We could suspect correlation between units belonging to the SAME group

- We have to use **cluster-robust** standard errors (clustering)
- In that case, given a number of G groups, the variance-covariance matrix writes as:

$$(X'X)^{-1}(\frac{G}{(G-1)}\frac{N-1}{(N-K)}\sum X_g\hat{\epsilon}_g\hat{\epsilon}'_gX'_g)(X'X)^{-1}$$

- where $\hat{\epsilon}_g$ is the vector of residuals for units of the group g and X_g is the vector of covariates for this units.
- It is **necessary to cluster standard-errors when using a macro variable** aggregated by group in a regression.

Outline

Simple case of linear regression and potential outcomes framework

- A linear regression with one constant (the intercept) and one non-constant dummy variable (the slope)
  - Similar to the potential outcomes framework (cf lecture 1)

$$y_i = Y_{i0} + D_i(Y_{i1} - Y_{i0})$$

- Let Y_{i0} be denoted by α (the intercept), $Y_{1i} - Y_{0i} = ATE = \rho$ (the slope), assumed to be constant across individuals for simplicity, the corresponding regression writes as:

$$Y_i = \alpha + D_i\rho + \eta_i$$

- η_i because of DGP: here it is the random part of Y_{i0}: $\eta_i = Y_{i0} - E(Y_{i0})$.

Simple case of linear regression and potential outcomes framework

- The naive estimator of ATE writes as:

$$E(Y_i|D_i = 1) - E(Y_i|D_i = 0) = \underbrace{\rho}_{\text{ATE}}$$

$$+ \underbrace{E(\eta_i|D_i = 1) - E(\eta_i|D_i = 0)}_{\text{Selection bias}}$$

- The selection bias can be written:
  $E(Y_{i0}|D_i = 1) - E(Y_{i0}|D_i = 0)$
  - As in Lecture 1

Regression models and RCTs

- Randomized experiment (RCT) removes the selection bias (cf lecture 1)
  - So ρ is the **population Average Treatment Effect (ATE)**
  - Even if we relax we the assumption of constant effect, ρ is the **population Average Treatment Effect on the Treated (ATT)**

Regression models and RCTs

- Randomized experiment (RCT) removes the selection bias (cf lecture 1)
  - So ρ is the **population Average Treatment Effect (ATE)**
  - Even if we relax we the assumption of constant effect, ρ is the **population Average Treatment Effect on the Treated (ATT)**

- In that case, OLS estimator of ρ, $\hat{\rho}_{OLS}$ is a consistent estimator of population ATT (or ATE if the effect is constant): $\hat{\rho}_{OLS} \xrightarrow{p} \rho$ or $plim(\hat{\rho}_{OLS}) = \rho$ or $E(\hat{\rho}_{OLS}) = \rho$

Regression models and RCTs

- A simple bivariate regression case: only one non-constant regressor:

$$cov(Y_i, D_i) = cov(\alpha + D_i\rho + \eta_i, D_i)$$

$$= cov(\alpha, D_i) + cov(\rho D_i, D_i) + cov(D_i, u_i)$$

- Under exogeneity assumption, we can write the simple formula of OLS estimator $\hat{\rho}$ in the bivariate case:

$$\hat{\rho} = \frac{cov(Y_i, D_i)}{Var(D_i)}$$

Regression models and RCTs

- In this framework, it is even simpler because the unique non-constant regressor is binary
- Setting N_t the number of treated individuals and N_c the number of control ones, the OLS estimator writes as:

$$\hat{\rho}_{OLS} = \frac{1}{N_t} \sum_{i:D_i=1} Y_i - \frac{1}{N_c} \sum_{i:D_i=0} Y_i$$

Regression models and RCTs

- In this framework, it is even simpler because the unique non-constant regressor is binary
- Setting N_t the number of treated individuals and N_c the number of control ones, the OLS estimator writes as:

$$\hat{\rho}_{OLS} = \frac{1}{N_t} \sum_{i:D_i=1} Y_i - \frac{1}{N_c} \sum_{i:D_i=0} Y_i$$

- **So $\hat{\rho}_{OLS}$ is the simple difference in means of the outcome between the treated group and the control one**

Outline

Introducing other explanatory variables

- What happens to the coefficient of D_i in a multivariate regression model?
  - **What is a OLS coefficient in a multivariate regression model?**:
- In the case of a single non-constant covariate x_{ki}, the coefficient associated to this variable in OLS regression is:

$$\beta_k = \frac{Cov(Y_i, x_{ki})}{Var(x_{ki})}$$

Introducing other explanatory variables

- What happens to the coefficient of D_i in a multivariate regression model?
  - **What is a OLS coefficient in a multivariate regression model?**:
- In the case of a single non-constant covariate x_{ki}, the coefficient associated to this variable in OLS regression is:

$$\beta_k = \frac{Cov(Y_i, x_{ki})}{Var(x_{ki})}$$

- BUT what if we introduce more than one non-constant regressors?

$$Y_i = \alpha + \beta_1 x_{1i} + ... + \beta_k x_{ki} + ... + e_i$$

Introducing other explanatory variables

- Assume an auxiliary regression in which the variable x_{ki} is regressed on all the remaining regressors and let $\tilde{x_{ki}}$ be the residual from this auxiliary regression.

Introducing other explanatory variables

- Assume an auxiliary regression in which the variable x_{ki} is regressed on all the remaining regressors and let $\tilde{x_{ki}}$ be the residual from this auxiliary regression.

- Compute the covariance between the dependent variable Y_i and this residual $\tilde{x_{ki}}$

$$Cov(Y_i, \tilde{x_{ki}}) = Cov(\alpha + \beta_1 x_{1i} + ... + \beta_k x_{ki} + ... + e_i, \tilde{x_{ki}})$$

$$= Cov(\alpha, \tilde{x_{ki}}) + \beta_1 Cov(x_{1i}, \tilde{x_{ki}}) + ... + \beta_k Cov(x_{ki}, \tilde{x_{ki}}) + ... + Cov(e_i, \tilde{x_{ki}})$$

Introducing other explanatory variables

- $\tilde{x_{ki}}$ is the random part of x_{ki} so:
  - $E(\tilde{x_{ki}}) = 0$
  - $\tilde{x_{ki}}$ is uncorrelated with the set of other regressors
  - And $\tilde{x_{ki}}$ is uncorrelated with e_i.

Introducing other explanatory variables

- $\tilde{x_{ki}}$ is the random part of x_{ki} so:
  - $E(\tilde{x_{ki}}) = 0$
  - $\tilde{x_{ki}}$ is uncorrelated with the set of other regressors
  - And $\tilde{x_{ki}}$ is uncorrelated with e_i.

- So:

$$Cov(Y_i, \tilde{x_{ki}}) = \beta_k Cov(x_{ki}, \tilde{x_{ki}})$$

- x_{ki} can be written as: $E(x_{ki}|X_{-k}) + \tilde{x_{ki}}$, so:

$$\beta_k Cov(x_{ki}, \tilde{x_{ki}}) = \beta_k E[(E(x_{ki}|X_{-k}) + \tilde{x_{ki}})\tilde{x_{ki}}] + E(x_{ki})E(\tilde{x_{ki}})$$

Introducing other explanatory variables

- At last we have:

$$Cov(Y_i, \tilde{x_{ki}}) = \beta_k Var(\tilde{x_{ki}}) \Leftrightarrow \beta_k = \frac{Cov(Y_i, \tilde{x_{ki}})}{Var(\tilde{x_{ki}})}$$

- This is the Frisch-Waugh-Lovell theorem (1933, 1963) referred to as the **"regression anatomy"** by Angrist and Pischke (2008).

- Each coefficient in a multivariate regression is the bivariate slope coefficient for the corresponding regressor after partialling out all the other variables in the model.

Introducing other explanatory variables in RCTs

- In RCTs, D_i is random: $D_i = E(D_i|X_{-D}) + \tilde{D}_i = \tilde{D}_i$
- Let us denote ρ^X the coefficient associated to D_i if we include a set of other regressors:

$$\rho^X = \frac{Cov(Y_i, \tilde{D}_i)}{Var(\tilde{D}_i)} = \frac{Cov(Y_i, D_i)}{Var(D_i)} = \rho$$

- In the case of completely randomized experiment, **introducing other covariates should have no effect on the OLS coefficient** ρ
- The contrary should cast some doubts about the perfect randomness of treatment assignment.

Introducing other explanatory variables

- SO why should we introduce other covariates in RCTs?
  - Randomization does not ensure that the distribution of characteristics are balanced across the treated and the control group (RCTs may lead to undesirable assignments, cf lecture 1)
  - Looking at covariates allows to check this balancing property
  - And it also allows to **reduce the residual variance** (the "unexplained" part of Y_{i0}), improving the precision of estimates

The Conditional Independence Assumption

- If the treatment is not randomly assigned, there is a selection bias so $\hat{\rho}_{OLS}$ does not reflect necessarily the causal average treatment effect.

The Conditional Independence Assumption

- If the treatment is not randomly assigned, there is a selection bias so $\hat{\rho}_{OLS}$ does not reflect necessarily the causal average treatment effect.

- **A causal interpretation of the regression can be justified under the Conditional Independence Assumption (CIA)**
  - The naive estimator is a causal effect CONDITIONALLY on a set of covariates (a $1xK$ vector X_i of covariates), expected to be strong predictors of the outcome

$$(Y_{1i}, Y_{0i}) \perp D_i | X_i \Rightarrow E(Y_{0i}|X_i, D_i = 1) = E(Y_{0i}|X_i, D_i = 0)$$

The Conditional Independence Assumption

- Illustration: to estimate the average causal effect of education ($D_i = 1$ if the individual went to college) on earnings (Y_i).
  - Considering a $1xK$ vector X_i of strong predictors of earnings (ability or family backgrounds)

$$E(Y_i|X_i, D_i = 1) - E(Y_i|X_i, D_i = 0) = E(Y_{1i} - Y_{0i}) +$$

$$+ \underbrace{E(Y_{0i}|D_i = 1) - E(Y_{0i}|D_i = 0)}_{\text{Selection bias}}$$

- Even if we can expect that those who go to college would have earned more anyway (positive selection bias)
  - CIA allows to say that **this selection bias disappears given a set of observed characteristics** X_i.
  - This "selection-on-observables" assumption states that the difference in characteristics X_i is the only reason why η_i and D_i are correlated.

The omitted variable bias (OVB) formula

- Is it sufficient to control for a large set of observable characteristics so that the CIA holds, ensuring the causality interpretation of $\hat{\rho}_{OLS}$?
- It is true that omitting strong predictors of outcomes may lead to the omitting variable bias.
- Let family background, intelligence and motivation be denoted by a vector A_i
  - The regression of earnings on the treatment indicator (going to college) controlling for this vector of covariates writes as:

$$Y_i = \alpha + \rho D_i + A_i\gamma + \epsilon_i$$

  - where γ is a $Kx1$ vector of population regression coefficients associated to each covariate.

The omitted variable bias formula

- If the CIA holds given the set of covariates A_i ρ is the causal population ATE and $\hat{\rho}_{OLS}$ is a consistent estimator of this ATE.
  - Let us call this regression the "long" regression

The omitted variable bias formula

- If the CIA holds given the set of covariates A_i ρ is the causal population ATE and $\hat{\rho}_{OLS}$ is a consistent estimator of this ATE.
  - Let us call this regression the "long" regression

- In practice A_i is hard to measure (how to measure motivation? intelligence? ability to understand courses? ...)
  - So the researcher can only estimate a "short" regression in which this strong predictors are omitted.
  - In that case the resulting "short regression" coefficient is related to what we would estimate in a "long" regression:

$$\rho^{SHORT} = \rho^{LONG} + \underbrace{\gamma' \delta_{Ad}}_{\textbf{The omitted variable bias}}$$

The omitted variable bias formula

- The anatomy of the omitted variable bias
  - γ: the resulting coefficients that we would have in the "long" regression
  - δ_{Ad} the vector of coefficients from the regression of A_i on D_i.

The omitted variable bias formula

- The anatomy of the omitted variable bias
  - γ: the resulting coefficients that we would have in the "long" regression
  - δ_{Ad} the vector of coefficients from the regression of A_i on D_i.

- If we omit intelligence/motivation ... when looking at causal effect of education on earnings
  - If we thing that these variables have positive effects on wages and that they are positively correlated with schooling, the omitting variable bias is positive
  - The coefficient from the "short" regression **overestimates** the average causal effect of schooling on earnings.

The omitted variable bias formula

- What if we introduce "bad" control variables in the regression?
  - In particular, covariates that may be the consequence of the treatment.
  - Illustration: introducing the occupational level ($W_i = 1$ for white-collar workers) in the earnings-education regression

The omitted variable bias formula

- What if we introduce "bad" control variables in the regression?
  - In particular, covariates that may be the consequence of the treatment.
  - Illustration: introducing the occupational level ($W_i = 1$ for white-collar workers) in the earnings-education regression

- Even in RCTs (ignorable treatment D_i), if you look at mean earnings between college graduates and others CONDITIONAL on working at a white-collar job
  - Treatment is ignorable but NOT W_i as it can result from schooling

$$E(Y_{1i}|W_{1i}=1, D_i=1) - E(Y_{0i}|W_{0i}=1, D_i=0) = E(Y_{1i}|W_{1i}=1) - E(Y_{0i}|W_{0i}=1)$$

$$= E(Y_{1i} - Y0i) + \underbrace{E(Y_{0i}|W_i=1) - E(Y_{0i}|W_i=0)}_{\text{Selection bias}}$$

The omitted variable bias formula

- A simple rule to choose a relevant set of characteristics: **timing matters**
  - Good controls: variables measured before the treatment
  - Bad controls: variables measured after the treatment (could be outcomes of the treatment)

- But in practice timing is uncertain or unknown
  - Requires additional assumptions about the timing or the absence of causal link between the treatment and the set of control variables.

The differents estimators adjusting for covariates

- Under CIA (selection-on-observables or unconfoundedness) assumption, treatment is ignorable
  - So we can compare outcomes between treated and non-treated individuals CONDITIONAL on a $1xK$ vector of characteristics X_i
  - In this respect, regression consists in controlling for this set of covariates, estimating:

$$Y_i = \alpha + \rho D_i + X_i\beta + \epsilon_i$$

  - where β is a $Kx1$ vector of population regression coefficients associated to each covariate.

The differents estimators adjusting for covariates

- Under CIA (selection-on-observables or unconfoundedness) assumption, treatment is ignorable
  - So we can compare outcomes between treated and non-treated individuals CONDITIONAL on a $1xK$ vector of characteristics X_i
  - In this respect, regression consists in controlling for this set of covariates, estimating:

$$Y_i = \alpha + \rho D_i + X_i\beta + \epsilon_i$$

  - where β is a $Kx1$ vector of population regression coefficients associated to each covariate.

- But there are many other estimators adjusting for covariates
  - Matching estimators
  - Subclassification estimators
  - Weighting estimators

The differents estimators adjusting for covariates

- All these estimators **(including regression)** rely on three key underlying assumptions
  - **CIA, unconfoundedness**: treatment is ignorable CONDITIONALLY on a set of observables X_i
  - **Common support**: For each covariate values $X_i = x^*$, we find both treated and control individuals.

$$0 < P(D_i = 1|X_i = x^*) < 1$$

  - **Stable Unit Treatment Value Assumption (SUTVA)**: the treatment D_i only affects the individual i: no spillover effects.

What is the main concern with regression

- A concern with regression estimators **in cases with limited overlap in covariate distributions**
  - Cases where characteristics are not well balanced between treatment and control groups.
  - We say: balancing property is not verified.

What is the main concern with regression

- A concern with regression estimators **in cases with limited overlap in covariate distributions**
  - Cases where characteristics are not well balanced between treatment and control groups.
  - We say: balancing property is not verified.
- We can assess the balancing property:
  - using a standard t-test
  - or looking at standardized mean differences

What is the main concern with regression

- Standard T-statistic

$$t = \frac{(\overline{X}_{t,k} - \overline{X}_{c,k})}{\sqrt{S^2_{X,c,k}/N_c + S^2_{X,t,k}/N_t}}$$

where: $S^2_{X,j,k}$ is the empirical variance of each characteristic k among the group j

What is the main concern with regression

- Standard T-statistic

$$t = \frac{(\overline{X}_{t,k} - \overline{X}_{c,k})}{\sqrt{S^2_{X,c,k}/N_c + S^2_{X,t,k}/N_t}}$$

  where: $S^2_{X,j,k}$ is the empirical variance of each characteristic k among the group j

- Problem: the $|t|$ may be large (small) simply because the sample is large (small), without reflecting substantial differences in X_i, k between the two groups.

What is the main concern with regression

- An alternative statistic to assessing overlap: standardized mean difference

$$\Delta_{X,k} = \frac{(\overline{X}_{t,k} - \overline{X}_{c,k})}{\sqrt{(S^2_{X,c,k} + S^2_{X,t,k}/2}}$$

What is the main concern with regression

- An alternative statistic to assessing overlap: standardized mean difference

$$\Delta_{X,k} = \frac{(\overline{X}_{t,k} - \overline{X}_{c,k})}{\sqrt{(S^2_{X,c,k} + S^2_{X,t,k}/2}$$

- Large values for standardized mean differences (>0.2) indicate substantial differences between the two sample average covariate values.

An illustration using the results of Lalonde study (1986)

- Lalonde (1986) analyzed data from a randomized experiment designed in the mid-1970's in US to evaluate the effect of a labor market training program: the National Supported Work (NSW)
  - Providing work experience for a period of 9 to 18 months to individuals with weal labor-force attachment or poor labor market histories (logic of "work-first" as in the "Garantie Jeunes").

An illustration using the results of Lalonde study (1986)

- Lalonde (1986) analyzed data from a randomized experiment designed in the mid-1970's in US to evaluate the effect of a labor market training program: the National Supported Work (NSW)
  - Providing work experience for a period of 9 to 18 months to individuals with weal labor-force attachment or poor labor market histories (logic of "work-first" as in the "Garantie Jeunes").

- **Lalonde question: could we have estimated this without randomized experiment?**

An illustration using the results of Lalonde study (1986)

- Lalonde (1986) constructed "non-experimental" control groups from data sets: the Current Population Survey (CPS) and the Panel Study of Income Dynamics (PSID)
  - Also considering subsets which resemble the treatment group in terms of pre-program characteristics

An illustration using the results of Lalonde study (1986)

- Lalonde (1986) constructed "non-experimental" control groups from data sets: the Current Population Survey (CPS) and the Panel Study of Income Dynamics (PSID)
  - Also considering subsets which resemble the treatment group in terms of pre-program characteristics
- He estimated the ATT using the treated group and each "non-experimental" comparison groups
- **Conclusion: Estimated ATT effects strongly diverge from experimental estimates: non-experimental approaches are not credible**
  - Influential conclusion in policy circles

Assessing covariate balance in the case of the Lalonde study (Imbens, 2015)

- Covariates are well balanced across both groups in experimental Lalonde data

	experimental controls (N_c=260)		trainees (N_t=185)			
Covariate	mean	(s.d.)	mean	(s.d.)	t-stat	nor-dif
Black	0.83	0.38	0.84	0.36	0.5	0.04
Hisp	0.11	0.31	0.06	0.24	-1.9	-0.17
Age	25.05	7.06	25.82	7.16	1.1	0.11
Married	0.15	0.36	0.19	0.39	1.0	0.09
Nodegree	0.83	0.37	0.71	0.46	-3.1	-0.30
Education	10.09	1.61	10.35	1.97	1.4	0.14
E'74	2.11	5.69	2.10	4.89	-0.0	-0.00
U'74	0.75	0.43	0.71	0.46	-1.0	-0.09
E'75	1.27	3.10	1.53	3.22	0.9	0.08
U'75	0.68	0.47	0.60	0.49	-1.8	-0.18

Assessing covariate balance in the case of the Lalonde study (Imbens, 2015)

- BUT substantially large differences between the two groups in non-experimental Lalonde data
- In that case, regression is not a recommended estimator

	CPS controls (N_c=15,992)		trainees (N_t=185)			
Covariate	mean	(s.d.)	mean	(s.d.)	t-stat	nor-dif
Black	0.07	0.26	0.84	0.36	28.6	2.43
Hisp	0.07	0.26	0.06	0.24	-0.7	-0.05
Age	33.23	11.05	25.82	7.16	-13.9	-0.80
Married	0.71	0.45	0.19	0.39	-18.0	-1.23
Nodegree	0.30	0.46	0.71	0.46	12.2	0.90
Education	12.03	2.87	10.35	2.01	-11.2	-0.68
E'74	14.02	9.57	2.10	4.89	-32.5	-1.57
U'74	0.12	0.32	0.71	0.46	17.5	1.49
E'75	13.65	9.27	1.53	3.22	-48.9	-1.75
U'75	0.11	0.31	0.60	0.49	13.6	1.19

Outline

Matching estimators: the intuition

- Matching each individual i with treatment D_i with an individual j with treatment $(1 - D_i)$ **closest (in terms of characteristics)** to the individual i

Matching estimators: the intuition

- Matching each individual i with treatment D_i with an individual j with treatment $(1 - D_i)$ **closest (in terms of characteristics)** to the individual i

$$\hat{Y}_{i1} \begin{cases} Y_i & \text{if } D_i = 1 \\ Y_{j(i)} & \text{if } D_i = 0 \end{cases}$$

- and:

$$\hat{Y}_{i0} \begin{cases} Y_{j(i)} & \text{if } D_i = 1 \\ Y_i & \text{if } D_i = 0 \end{cases}$$

Matching estimators: the intuition

- And look at their average differences in outcomes to estimate the ATT:

$$\hat{\rho}^{sm}_{ATT} = \frac{1}{N_t} \sum_{i:D_i=1} (\hat{Y}_{i1} - \hat{Y}_{i0})$$

- Or the ATE:

$$\hat{\rho}^{sm}_{ATE} = \frac{1}{N} \sum_{i=1}^{N} (\hat{Y}_{i1} - \hat{Y}_{i0})$$

- This is a simple one-to-one matching (sm) estimator

Matching estimators: the intuition

- It is recommended to use matching WITH replacement: individual j could be the best match for SEVERAL individuals i
  - Matching without replacement does not allow to estimate the ATE

Matching estimators: the intuition

- It is recommended to use matching WITH replacement: individual j could be the best match for SEVERAL individuals i
  - Matching without replacement does not allow to estimate the ATE
- But this comes at the cost of higher variance
  - Estimator is potentially based on less information

Matching estimators: the intuition

- Considering this sample:

Id	D	Y	Y(0)	Y(1)	x
1	0	10	10	??	12
2	0	15	15	??	16
3	0	5	5	??	4
4	1	14	??	14	12
5	1	18	??	18	16

- The naive estimator of ATT would be:
$(14 + 18)/2 - (10 + 15 + 5)/3 = 6$

Matching estimators: the intuition

- Using one-to-one simple matching estimator consists in pairing each individual with treatment $D_i = 1$ with the closest individual (in terms of x) with treatment $1 - D_i$.

Id	D	Y	Y(0)	Y(1)	x
1	0	10	10	**14**	12
2	0	15	15	**18**	16
3	0	5	5	**??**	4
4	1	14	**10**	14	12
5	1	18	**15**	18	16

- The one-to-one simple matching estimator is : $[(14 - 10) + (18 - 15)]/2 = 3.5 < 6$

Matching estimators: the intuition

- One can improve the precision of the estimator using more than one match
  - In the case where only treated units are matched and the pool of control units is large relative to the number of treated ones

Matching estimators: the intuition

- One can improve the precision of the estimator using more than one match
  - In the case where only treated units are matched and the pool of control units is large relative to the number of treated ones

- Let $\delta_M(i) = \{j_{1i}, j_{2i}..., j_{Mi}\}$ be the set of indices for the first M matches for unit i:
- Define

$$\hat{Y}_{i1}^M \begin{cases} Y_i & \text{if } D_i = 1 \\ \sum_{j \in \delta_M(i)} Y_j / M & \text{if } D_i = 0 \end{cases}$$

- and:

$$\hat{Y}_{i0}^M \begin{cases} \sum_{j \in \delta_M(i)} Y_j / M & \text{if } D_i = 1 \\ Y_i & \text{if } D_i = 0 \end{cases}$$

Matching estimators: the intuition

- And look at their average differences in outcomes to estimate the ATT:

$$\hat{\rho}_{ATT}^{sm-M} = \frac{1}{N_t} \sum_{i:D_i=1} (\hat{Y}_{i1}^{M} - \hat{Y}_{i0}^{M})$$

- Or the ATE:

$$\hat{\rho}_{ATE}^{sm-M} = \frac{1}{N} \sum_{i=1}^{N} (\hat{Y}_{i1}^{M} - \hat{Y}_{i0}^{M})$$

- This is a simple one-to-M matching ($sm - M$) estimator

Matching estimators: the intuition

- Why not using too many matched units?
  - OK it improves the precision
  - BUT the higher M, the lower the quality of matched units
  - On average the difference between characteristics for a unit and and its average match increases : additional bias

- It is recommended to use a small number of multiple matches: between 1 and 4 (Imbens and Rubin, 2012)

Matching estimators: the intuition

- To make CIA credible, need to match on many observable variables
  - difficult to find perfectly similar i and j on all X (exact matching)

Matching estimators: the intuition

- To make CIA credible, need to match on many observable variables
  - difficult to find perfectly similar i and j on all X (exact matching)

- Other methods:
  - Propensity-score based matching: similarity between i and j is measured in terms of differences in a single index: the propensity score
  - Distance-based matching: A need to define a metric for the distance between two $1 * K$ vectors of covariates X_i and X_j

Propensity-score based matching estimators

- To solve this curse of dimensionality, we aggregate all differences in X_i in only one index: **the propensity score**
- **The propensity score** (Rosenbaum and Rubin, 1983): the conditional probability of assignment to a treatment given a vector of observed covariates
  - Let $p(X_i)$ be the propensity score: $p(X_i) = E(D_i|X_i)$

Propensity-score based matching estimators

- To solve this curse of dimensionality, we aggregate all differences in X_i in only one index: **the propensity score**
- **The propensity score** (Rosenbaum and Rubin, 1983): the conditional probability of assignment to a treatment given a vector of observed covariates
  - Let $p(X_i)$ be the propensity score: $p(X_i) = E(D_i|X_i)$

- **The propensity score theorem**: under CIA, potential outcomes are independent of treatment CONDITIONAL on the propensity score

$$\{Y_{0i}, Y_{1i}\} \perp D_i|p(X_i)$$

Propensity-score based matching estimators

Before matching we have to

1. Estimate the propensity score
2. Check the common support assumption: $0 < p(X_i) < 1$ for all i

How to estimate the propensity score?

How to estimate the propensity score?

- Introducing a set of covariates and some higher-order terms (including interaction terms) a priori viewed as important for explaining the assignment and plausibly related to the outcome
- This is the specification $h(X_i)$ of your model for predicting the pscore

How to estimate the propensity score?

How to estimate the propensity score?

- Introducing a set of covariates and some higher-order terms (including interaction terms) a priori viewed as important for explaining the assignment and plausibly related to the outcome
- This is the specification $h(X_i)$ of your model for predicting the pscore

- Fit the Probit (or Logit) model: $P(D_i = 1|X_i) = \Phi\{h(X_i)\}$ where $\Phi(.)$ denotes the normal (or logistic) c.d.f.

How to estimate the propensity score?

How to estimate the propensity score?

- Introducing a set of covariates and some higher-order terms (including interaction terms) a priori viewed as important for explaining the assignment and plausibly related to the outcome
- This is the specification $h(X_i)$ of your model for predicting the pscore
- Fit the Probit (or Logit) model: $P(D_i = 1|X_i) = \Phi\{h(X_i)\}$ where $\Phi(.)$ denotes the normal (or logistic) c.d.f.
- BUT in practive how to choose the good specification for the pscore?
  - Not a ideal way of doing the choice
  - Some algorithms exist but are not so commonly used

How to estimate the propensity score? : the Imbens and Rubin (2012) procedure

1. Introduce linearly a first set of covariates a priori viewed as important for explaining the assignment and plausibly related to the outcome (as starting specification $h(X_i)$)
2. Among the remaining covariates, add one covariate at a time and calculating for each of these specifications a Likelihood Ratio (LR) statistic $-2(l^r - l^{ur})$ for assessing the null hypothesis that the newly included variable has a zero coefficient
3. If at least one of the LR stat is greater than an arbitrary threshold (e.g. 1), add the covariate with the largest LR-stat
4. Among the new set of remaining variables, repeat step 2 and 3 until none of the remaining LR-stat exceed the threshold.

How to estimate the propensity score? : the Imbens and Rubin (2012) procedure

5 From our set of K_L covariates, decide which of the $K_L(K_L+1)/2$ quadratic and interaction terms involving these covariates to include repeeating similar steps as 2 and 3 but now adding one quadratic/interaction term at a time and considering another threshold (e.g. 2.71)

How to estimate the propensity score? : the Imbens and Rubin (2012) procedure

5 From our set of K_L covariates, decide which of the $K_L(K_L+1)/2$ quadratic and interaction terms involving these covariates to include repeeating similar steps as 2 and 3 but now adding one quadratic/interaction term at a time and considering another threshold (e.g. 2.71)

- At last: K_L covariates, K_Q second-order terms and an intercept (a vector of $1 + K_L + K_Q$ functions of covariates)

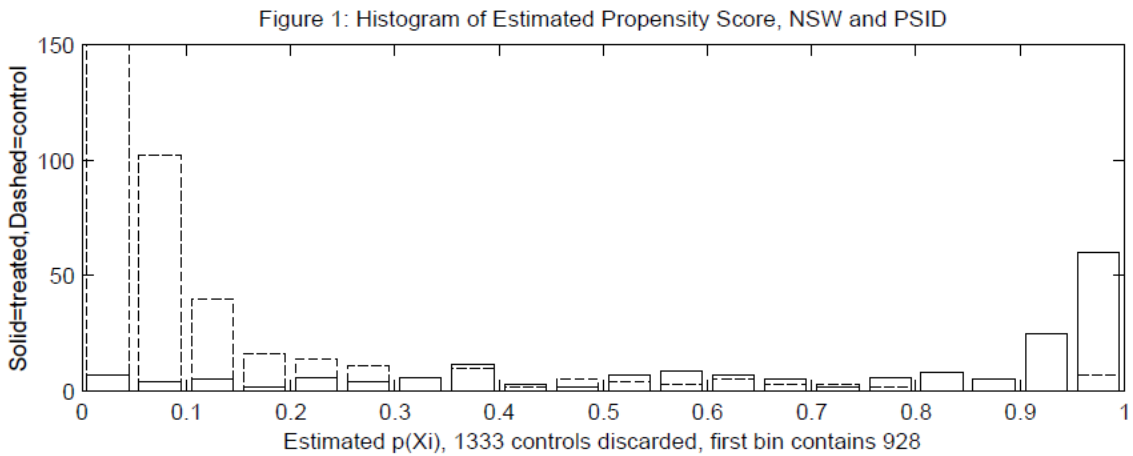
Checking the common support and addressing lack of overlap

Plotting the density distribution of the estimated propensity score for both groups and checking whether they sufficiently overlap

- Alternative ways to ensure sufficient overlap:
  - Discard control units with an estimated pscore $\hat{p}$ less than the minimum (or greater than the maximum) estimated pscore for treated units: the Min-Max method (Dehejia and Wahba, 1999)
  - Do some trimming: in addition to the Min-Max method, remove all the treated observations with $\hat{p}$ larger than the k-smallest value among control units
  - Picking only observations for which $0.1 < \hat{p} < 0.9$ (Crump et al., 2009)

Checking the common support and addressing lack of overlap

A visual analysis of the density distribution of the estimated pscore for both groups using the Lalonde data (Dehejia and Wahba, 1999)



- After estimating the pscore and checking overlap, some propensity-based matching methods can be used
  - Nearest neighbor matching
  - k-th nearest-neighbors matching estimator
  - Radius (caliper) matching estimators
  - Kernel-matching estimators

- After estimating the pscore and checking overlap, some propensity-based matching methods can be used
  - Nearest neighbor matching
  - k-th nearest-neighbors matching estimator
  - Radius (caliper) matching estimators
  - Kernel-matching estimators

- Nearest neighbor matching:
  - Each unit i with treatment D_i is matched with replacement to the unit j with treatment $(1 - D_i)$ and displaying the closest propensity score.

  - Let $j(i)$ be the index for the closest match defined as: $j(i) = \underset{j:D_j \neq D_i}{\text{argmin}} \hat{p}_i - \hat{p}_j$

Propensity-score based matching estimators

- Define:

$$\hat{Y}_{i1} \begin{cases} Y_i & \text{if } D_i = 1 \\ Y_{j(i)} & \text{if } D_i = 0 \end{cases}$$

- and:

$$\hat{Y}_{i0} \begin{cases} Y_{j(i)} & \text{if } D_i = 1 \\ Y_i & \text{if } D_i = 0 \end{cases}$$

Propensity-score based matching estimators

- Define:

$$\hat{Y}_{i1} \begin{cases} Y_i & \text{if } D_i = 1 \\ Y_{j(i)} & \text{if } D_i = 0 \end{cases}$$

- and:

$$\hat{Y}_{i0} \begin{cases} Y_{j(i)} & \text{if } D_i = 1 \\ Y_i & \text{if } D_i = 0 \end{cases}$$

- We find the same *sm* estimators as in the previous section:

$$\hat{\rho}^{sm}_{ATT} = \frac{1}{N_t} \sum_{i:D_i=1} (\hat{Y}_{i1} - \hat{Y}_{i0})$$

- Or the ATE:

$$\hat{\rho}^{sm}_{ATE} = \frac{1}{N} \sum_{i=1}^{N} (\hat{Y}_{i1} - \hat{Y}_{i0})$$

Propensity-score based matching estimators

- k-th nearest-neighbors matching estimator:
  - Each unit i with treatment D_i is matched with replacement to the k unit j with treatment $(1 - D_i)$ displaying the closest propensity score.

Propensity-score based matching estimators

- k-th nearest-neighbors matching estimator:
  - Each unit i with treatment D_i is matched with replacement to the k unit j with treatment $(1 - D_i)$ displaying the closest propensity score.

- Let $\delta_M(i) = \{j_{1i}, j_{2i}..., j_{Ki}\}$ be the set of indices for the first M matches for unit i:

- The parallel can be done with the simple one-to-M matching $(sm - M)$ estimator

Different propensity-score based matching estimators

- Radius (caliper) matching estimators
  - Sometimes, even the closest match has a value of pscore too different from the unit i.
  - So we impose a maximal distance (caliper or radius) between $\hat{p}_i$ and $\hat{p}_j$

Different propensity-score based matching estimators

- Radius (caliper) matching estimators
  - Sometimes, even the closest match has a value of pscore too different from the unit i.
  - So we impose a maximal distance (caliper or radius) between $\hat{p}_i$ and $\hat{p}_j$
- Kernel-matching estimators
  - For each unit j a weight is placed and is defined by $K(\frac{\hat{p}_i - \hat{p}_j}{h})$
  - where K is the Kernel estimator and h is the bandwith.
  - The higher the bandwidth, the lower the variance but the higher the bias.

Matching and overlap

- Nearest or k-nearest neighbor matching discards unmatched units j
  - We can check whether matching improves overlap
  - Plotting distributions of estimated pscore for both groups after matching
  - Or assessing the balance of covariates across both groups after matching

Matching and overlap

- Nearest or k-nearest neighbor matching discards unmatched units j
  - We can check whether matching improves overlap
  - Plotting distributions of estimated pscore for both groups after matching
  - Or assessing the balance of covariates across both groups after matching

- Kernel matching estimators assign a weight to each unit j according to their distance in terms of pscore between unit i
  - We can check whether matching improves overlap
  - Plotting distributions of estimated pscore for both groups after applying the weights
  - Or assessing the balance of covariates across both groups after matching after applying the weights

Distance-based matching estimators

- Main issue of propensity-score based matching estimators
  - May be biased if the pscore is misspecified
  - An alternative: **distance**-based matching estimators

Distance-based matching estimators

- Main issue of propensity-score based matching estimators
  - May be biased if the pscore is misspecified
  - An alternative: **distance**-based matching estimators
- Use the Mahalanobis metric $||X_i, X_j||$ to define the distance between two covariate $1 * K$ vectors X_i and X_j:

$$||X_i, X_j|| = (X_i - X_j)' \hat{\Omega}_X^{-1} (X_i - X_j)$$

- where $\hat{\Omega}_X^{-1}$ is the sample covariance matrix of covariates.

Distance-based matching estimators

- Main issue of propensity-score based matching estimators
  - May be biased if the pscore is misspecified
  - An alternative: **distance**-based matching estimators
- Use the Mahalanobis metric $||X_i, X_j||$ to define the distance between two covariate $1 * K$ vectors X_i and X_j:

$$||X_i, X_j|| = (X_i - X_j)' \hat{\Omega}_X^{-1} (X_i - X_j)$$

- where $\hat{\Omega}_X^{-1}$ is the sample covariance matrix of covariates.
- Let $j(i)$ be the index for the closest match defined as: $j(i) = \underset{j:D_j \neq D_i}{\text{argmin}} ||X_i, X_j||$
- The parallel can be done with the simple one-to-one matching sm) or one-to-M matching ($sm - M$) estimators presented previously

Variance of matching estimators

- Main issue: the propensity score has been estimated
- Common practice: **bootstrapping** to recover standard errors
  1. Sample n observations randomly with replacement from the sample at hand to obtain a "bootstrap" data set
  2. Calculate the bootstrap version of the statistic of interest
  3. Repeat steps 1 and 2 a large number of times to obtain an estimate of the bootstrap distribution

Variance of matching estimators

- Main issue: the propensity score has been estimated
- Common practice: **bootstrapping** to recover standard errors
  1. Sample n observations randomly with replacement from the sample at hand to obtain a "bootstrap" data set
  2. Calculate the bootstrap version of the statistic of interest
  3. Repeat steps 1 and 2 a large number of times to obtain an estimate of the bootstrap distribution
- But for some matching estimator (e.g. distance-based ones), **bootstrap is not valid** (Abadie and Imbens, 2006)

A General Variance estimator (Imbens, 2015)

- Idea: use matching (Imbens, 2015)
  - Let $h(i)$ be the closest match for i, **with same treatment**.
  - The conditional variance of the ATE is:

$$\hat{V}(\hat{\rho}_{ATE}) = \frac{1}{N^2}\sum_{i=1}^{N}(1+\frac{K_M(i)}{M})^2\frac{1}{2}(Y_i - Y_{h(i)})^2$$

  - and of the ATT is:

$$\hat{V}(\hat{\rho}_{ATT}) = \frac{1}{N_t^2}\sum_{i=1}^{N}(D_i - (1-D_i)\frac{K_M(i)}{M}]^2\frac{1}{2}(Y_i - Y_{h(i)})^2$$

  - where $K_M(i)$ stands for the number of times unit i is used as a match given that M matches per unit are used.

A General Variance estimator (Imbens, 2015)

- If we assume constant treatment effect and homoskedasticity
  - The variance of the ATT is:

$$\hat{V}(\hat{\rho}_{ATT}) = \frac{1}{N_t^2} \sum_{i=1}^{N} (D_i - (1 - D_i) \frac{K_M(i)}{M}]^2 \frac{1}{2N_t} (Y_i - Y_j(i) - \hat{\rho})^2$$

  - where $K_M(i)$ stands for the number of times unit i is used as a match given that M matches per unit are used
  - and $\hat{\rho}$ is the estimated ATT.

Do matching improve overlap for the non-experimental Lalonde data? (Imbens, 2015)

YES! Distance-based matching allows to balance covariates across groups.

	Full Sample nor-dif	Matched Sample nor-dif	ratio of nor-dif
Black	2.43	0.00	0.00
Hispanic	-0.05	0.00	-0.00
Age	-0.80	-0.15	0.19
Married	-1.23	-0.28	0.22
Nodegree	0.90	0.25	0.28
Education	-0.68	-0.18	0.26
E'74	-1.57	-0.03	0.02
U'74	1.49	0.02	0.02
E'75	-1.75	-0.07	0.04
U'75	1.19	0.02	0.02

Outline

Subclassification/stratification/blocking estimators

- Starting from the Average Treatment Effect (ATE) and using the law of iterated expectations:

$$E(Y_{1i} - Y_{0i}) = E[[E(Y_{1i}|X_i, D_i = 1) - E(Y_{0i}|X_i, D_i = 1)]]$$

Subclassification/stratification/blocking estimators

- Starting from the Average Treatment Effect (ATE) and using the law of iterated expectations:

$$E(Y_{1i} - Y_{0i}) = E[[E(Y_{1i}|X_i, D_i = 1) - E(Y_{0i}|X_i, D_i = 1)]]$$

- Under CIA, the ATE can be written as:

$$E(Y_{1i} - Y_{0i}) = E[\underbrace{[E(Y_i|X_i, D_i = 1) - E(Y_i|X_i, D_i = 0)]}_{\delta_X}]$$

- and the ATT can be written as:

$$E(Y_{1i} - Y_{0i}|D_i = 1) = E[\underbrace{[E(Y_i|X_i, D_i = 1) - E(Y_i|X_i, D_i = 0)]}_{\delta_X}|D_i = 1]$$

- where δ_X denote the difference in mean outcomes by treatment level **for each value of** X_i.

Subclassification/stratification/blocking estimators

- Stratification = computing differences in mean outcomes between treated and control groups **within small groups (strata/cells) of** X_i
- And then averaging over the set of covariate cells

$$E(Y_{1i} - Y_{0i}) = \sum_x \delta_x P(X_i = x)$$

Subclassification/stratification/blocking estimators

- Stratification = computing differences in mean outcomes between treated and control groups **within small groups (strata/cells) of X_i**
- And then averaging over the set of covariate cells

$$E(Y_{1i} - Y_{0i}) = \sum_x \delta_x P(X_i = x)$$

- Only possible in simple cases
  - **Only discrete** covariates
  - Not too many: in the case of k dummy variables $\Rightarrow 2^k$ different cells: **the curse of dimensionality**
    - Many cells can be empty or do not check the common support
    - To solve this curse of dimensionality: use the propensity score

Subclassification/stratification/blocking estimators in practice (Dehejia and Wahba, 1999; Imbens, 2015)

Subclassification estimators can be implemented in 5 steps:

1. Estimating the propensity score: $p(X_i) = E(D_i|X_i)$

Subclassification/stratification/blocking estimators in practice (Dehejia and Wahba, 1999; Imbens, 2015)

Subclassification estimators can be implemented in 5 steps:

1. Estimating the propensity score: $p(X_i) = E(D_i|X_i)$

2. Checking whether the commun support assumption holds: $0 < p(X_i) < 1$??

Subclassification/stratification/blocking estimators in practice (Dehejia and Wahba, 1999; Imbens, 2015)

Subclassification estimators can be implemented in 5 steps:

1. Estimating the propensity score: $p(X_i) = E(D_i|X_i)$

2. Checking whether the commun support assumption holds: $0 < p(X_i) < 1$??

3. Dividing the sample into strata (blocks):
   - 3.1 Blocking/partitioning of the range of propensity score, the interval $[0, 1]$ into J intervals of the form $[b_{j-1}, b_j)$ where $b_0 = 0$ and $b_J = 1$
   - 3.2 Define block indicator $B_{ij} = 1$ if $b_{j-1} < p(X_i) \leq b_j$

Subclassification/stratification/blocking estimators in practice (Dehejia and Wahba, 1999; Imbens, 2015)

Subclassification estimators can be implemented in 5 steps:

1. Estimating the propensity score: $p(X_i) = E(D_i|X_i)$

2. Checking whether the commun support assumption holds: $0 < p(X_i) < 1$??

3. Dividing the sample into strata (blocks):
   - 3.1 Blocking/partitioning of the range of propensity score, the interval $[0, 1]$ into J intervals of the form $[b_{j-1}, b_j)$ where $b_0 = 0$ and $b_J = 1$
   - 3.2 Define block indicator $B_{ij} = 1$ if $b_{j-1} < p(X_i) \leq b_j$

Subclassification/stratification/blocking estimators in practice (Dehejia and Wahba, 1999; Imbens, 2015)

4 Within each block (stratum) j compute the average treatment effect $\hat{\rho}_j$

4.1 Either difference in means of the outcome $\hat{\tau}_j$ between the treatment and control groups:

$$\hat{\rho}_j = \frac{1}{N_{jt}} \sum_{i:B_{ij}=1} D_i Y_i + \frac{1}{N_{ct}} \sum_{i:B_{ij}=1} (1 - D_i) Y_i$$

4.2 Or estimating $\hat{\rho}_j$ using linear regression with some of the covariates

$$(\hat{alpha}_j, \hat{\rho}_j, \hat{\beta}_j) = \underset{\alpha,\rho,\beta}{\operatorname{argmin}} \sum_{i=1}^{N} B_{ij}(Y_i - \alpha - \rho D_i - X_i'\beta)^2$$

Subclassification/stratification/blocking estimators in practice (Dehejia and Wahba, 1999; Imbens, 2015)

5 Averaging all these effects to estimate the ATE $\hat{\rho}_{ATE}$ or the ATT $\hat{\rho}_{ATT}$:

$$\hat{\rho}_{ATE} = \sum_{j=1}^{J} \frac{N_{jc} + N_{jt}}{N} \hat{\rho}_j$$

$$\hat{\rho}_{ATT} = \sum_{j=1}^{J} \frac{N_{jt}}{N} \hat{\rho}_j$$

The variance of subclassification estimators

- An attractive feature of subclassification estimators
  - Within each block, the variance of the estimator $\hat{\rho}_j$ has the same formula of the Neyman's one in context of randomized experiment:

$$\hat{V}(\hat{\rho}_{ATE,j}) = \frac{s^2_{cj}}{N_{cj}} + \frac{s^2_{tj}}{N_{tj}}$$

- where s^2_{cj} and s^2_{tj} denote respectively the empirical variance of the outcome in the control and the treated group within each block j.

The variance of subclassification estimators

- An attractive feature of subclassification estimators
  - Within each block, the variance of the estimator $\hat{\rho}_j$ has the same formula of the Neyman's one in context of randomized experiment:

$$\hat{V}(\hat{\rho}_{ATE,j}) = \frac{s_{cj}^2}{N_{cj}} + \frac{s_{tj}^2}{N_{tj}}$$

- where s_{cj}^2 and s_{tj}^2 denote respectively the empirical variance of the outcome in the control and the treated group within each block j.

- The variance of the estimator is obtained by adding the within-block variances multiplied by the square of the block proportions:

$$\hat{V}(\hat{\rho}_{ATE}) = \sum_{j=1}^{J} \hat{V}(\hat{\rho}_{ATE,j})(\frac{N_j}{N})^2$$

Outline

Weighting estimators

- We can use directly the pscore to weight the units in order to eliminate biases (Horwitz and Thompson, 1952)
- Once we have estimated $p(X_i)$, we can construct an Inverse Probability Weighting (IPW) estimator of ATE and ATT:

$$\rho^{ht}_{ATE} = \frac{1}{N}\sum_{i=1}^{N}\frac{(D_i Y_i)}{\hat{p}(X_i)} - \frac{1}{N}\sum_{i=1}^{N}\frac{(1-D_i)Y_i}{[1-\hat{p}(X_i)]}$$

and:

$$\rho^{ht}_{ATT} = \frac{1}{N_t}\sum_{i=1}^{N_t} D_i Y_i - \frac{1}{N}\sum_{i=1}^{N}(1-D_i)Y_i\frac{\hat{p}(X_i)}{[1-\hat{p}(X_i)]}$$

Weighting estimators

- In practice, we normalize weights placed on treated and control units so that the sum of weights normalize to one for each group
- The normalized IPW estimators are:

$$\rho_{ATE}^{ht} = \sum_{i=1}^{N} \frac{\frac{(D_i Y_i)}{\hat{p}(X_i)}}{\sum_{j=1}^{N} \frac{1}{\hat{p}(X_i)}} - \sum_{i=1}^{N} \frac{\frac{(1-D_i) Y_i}{[1-\hat{p}(X_i)]}}{\sum_{j=1}^{N} \frac{1}{[1-\hat{p}(X_i)]}}$$

And:

$$\rho_{ATT}^{ht} = \frac{1}{N_t} \sum_{i=1}^{N_t} D_i Y_i - \sum_{i=1}^{N} \frac{(1-D_i) Y_i \frac{\hat{p}(X_i)}{[1-\hat{p}(X_i)]}}{\sum_{j=1}^{N} \frac{\hat{p}(X_i)}{[1-\hat{p}(X_i)]}}$$

Weighting estimators

- An attractive estimator at first sight
  - Easy to implement once the pscore has been estimated

Weighting estimators

- An attractive estimator at first sight
  - Easy to implement once the pscore has been estimated
- BUT not recommended
  - Strongly sensitive to low/large values of the pscore
  - Sensitive to the specification of the pscore

Do subclassification or matching estimators overcome the Lalonde's critique (Dehejia and Wahba, 1999)?

YES! These two estimators find ATT effects consistent with those found from experimental data

Table 3. Estimated Training Effects for the NSW Male Participants Using Comparison Groups from PSID and CPS-SSA

	NSW Earnings Less Comparison Group Earnings		NSW Treatment Earnings Less Comparison Group Earnings, Conditional On The Estimated Propensity Score					
			Quadratic in Score[b]	Stratifying on the Score			Matching on the Score	
	(1) Unadjusted	(2) Adjusted[a]	(3)	(4) Un-adjusted	(5) Adjusted[a]	(6) Obs.[g]	(7) Un-adjusted	(8) Adjusted[f]
NSW	1,794 (633)	1,672 (638)						
PSID-1[c]	-15,205 (1154)	731 (886)	294 (1389)	1,608 (1571)	1,494 (1581)	1,255	1,691 (2209)	1,473 (809)
PSID-2[d]	-3,647 (959)	683 (1028)	496 (1193)	2,220 (1768)	2,235 (1793)	389	1,455 (2303)	1,480 (808)
PSID-3[d]	1,069 (899)	825 (1104)	647 (1383)	2,321 (1994)	1,870 (2002)	247	2,120 (2335)	1,549 (826)
CPS-1[e]	-8,498 (712)	972 (550)	1117 (747)	1,713 (1115)	1,774 (1152)	4,117	1,582 (1069)	1,616 (751)
CPS-2[e]	-3,822 (670)	790 (658)	505 (847)	1,543 (1461)	1,622 (1346)	1493	1,788 (1205)	1,563 (753)
CPS-3[e]	-635 (657)	1,326 (798)	556 (951)	1,252 (1617)	2,219 (2082)	514	587 (1496)	662 (776)

BUT regressions can do the same using pscore for **sample selection** (Angrist and Pischke, 2008)

Picking only observations for which $0.1 < p(\hat{X}) < 0.9$ (Crump et al., 2009) strongly improve overlap and allows to obtain estimates from regressions close to the ATT obtained from experimental data

Table 3.3.3: Regression estimates of NSW training effects using alternate controls

Specification	Full Samples			P-Score Screened Samples	
	NSW	CPS-1	CPS-3	CPS-1	CPS-3
	(1)	(2)	(3)	(4)	(5)
Raw Difference	1,794 (633)	-8,498 (712)	-635 (657)		
Demographic controls	1,670 (639)	-3,437 (710)	771 (837)	-3,361 (811) [139/497]	890 (884) [154/154]
1975 Earnings	1,750 (632)	-78 (537)	-91 (641)	no obs. [0/0]	166 (644) [183/427]
Demographics, 1975 Earnings	1,636 (638)	623 (558)	1,010 (822)	1,201 (722) [149/357]	1,050 (861) [157/162]
Demographics, 1974 and 1975 Earnings	1,676 (639)	794 (548)	1,369 (809)	1,362 (708) [151/352]	649 (853) [147/157]