Selection on Observables

Alessandro Caggia

June 2025

Abstract

1. Conditional Independence

1.1. Idea

- **Selection bias**: those who take the treatment have different potential outcomes than those who do not. The estimation of the treatment effect is confounded by the baseline differences in potential outcomes between treatment and control, eg the treatment group will have higher outcomes not only due to treatment but also due to different characteristics (it can be that they select into treatment or are selected into treatment due to these different characteristics).

1.2. Model

1.2.1. Assumptions

- **Assumption 1 (Unconfoundedness)**: Selection on observables (or unconfoundedness) means that, conditional on observed covariates, the potential outcomes are independent of treatment assignment, that is:

$$(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i \mid X_i.$$

  - Detailed knowledge of the assignment mechanism and we observe all relevant covariates affecting both treatment and outcomes (that is, determining the selection or the self selection). Best case is that treatment is randomized conditional on X, e.g., when assignment depends on a lottery that varies by age or gender, and we control for those (requires great knowledge of the institutional design).

- **Assumption 1b (Conditional Mean Independence)**:

$$\mathbb{E}[Y_{0i} \mid X_i, D_i] = \mathbb{E}[Y_{0i} \mid X_i], \quad \mathbb{E}[Y_{1i} \mid X_i, D_i] = \mathbb{E}[Y_{1i} \mid X_i]$$

  This is weaker than full unconfoundedness. Assumption 1b is implied by assumption 1 but does not imply assumption 1 (think that the equality does not hold for variance)

1.2.2. Identification

- **Under 1 or 1b, the conditional average treatment effect is identified** ad $ATE_X = ATT_X$, that is:

$$\mathbb{E}[Y_{1i} - Y_{0i} \mid X_i] = \mathbb{E}[Y_{1i} - Y_{0i} \mid D_i = 1, X_i]$$

- **Intuition:** Unconfoundedness ensures that treatment is as good as randomly assigned, conditional on X_i. Thus, the treated group is representative of the full population (with the same X_i) in terms of potential outcomes.
- **Trade-off**: The richer the set of covariates X_i (include also lagged outcome variables, as they are often greatly correlated with potential outcomes), the more credible unconfoundedness becomes. But, we also need variation in treatment assignment within X-group in order to estimate the treatment effect for that group (cannot be deterministic at the group level, there would be no variability in T and C for that subgroup eg all treated or all control) But we must exclude any post-treatment variables (variables that are affected by the treatment, you would be controlling for a mechanism).

1.3. Testing unconfoundedness

Is treatment truly uncorrelated to potential outcomes once conditioning?

- the above is **untestable**, as one does not have both the potential outcomes for the two individuals
  - **Do a balancing table on the observable covariates** ... **BUT** what if there are imbalances in unobservable covariates?
  - We can perform **falsification tests using placebo outcomes** (lagged outcomes that you know to be unrelated)... but say you find that there is no effect with respect to placebo outcome well, but what if the imbalance is relevant only with respect to the true outcome?
  - We can perform **pseudo-treatments** to test whether treatment affects outcomes it theoretically shouldn't... but you find that there is no heterogeneity effect of a fake treatment, but what if there is an effect with the true treatment?
- The idea is that if you don't find anything, sure, you cannot conclude assumptions hold, but if you find something, clearly this falsifies your assumptions.

2. Overlap: from ATE(x) to ATE

2.1. ATE(x) and ATE

- Our goal is to estimate ATE, not ATE_x

$$ATE = \mathbb{E}(Y_1 - Y_0) = \mathbb{E}_X\left(\mathbb{E}(Y_1 - Y_0 \mid X)\right) = \mathbb{E}_X(ATE_X)$$

- So basically you take the ATE for every x and then you aggregate by averaging the te thoughout the support (and weight by the covariate distribution).
- So estimating ATE requires observing both control and treated units throughout the support of X.
- **Assumption 2 (Overlap)**: for all $x \in M$

$$0 < \mathbb{P}(D = 1 \mid X = x) < 1$$

  , where M is the support of the covariates. Basically if this is 0 there are no treated units, if thiis is 1 there are no control units.
- Positive probability of observing units both in the control and treatment group for any value of covariates.
- if they are all treated in a group iit does not work (we are not able to estimate an average effect over the population that includes observations with $x = x_0$), or if you have no data on a group it does not work (obv). Maybe you could change the populaitono f interest
- Assumptions 1 and 2, together, are called **strong ignorability**

3. Identification

- Given the ignorability and overlap assumptions, the ATE and the ATT are identified.
- There are two identification approaches:
  1. Based on Conditional Expectations, which leads to estimates using regression.

2. Based on weights, which leads to matching and propensity score estimators.

ATT Identification

- The ATT can be identified under weaker assumptions. Remember our main problem:

$$\mathbb{E}(Y_{1i} \mid D_i = 1) - \mathbb{E}(Y_{0i} \mid D_i = 0) = \underbrace{\mathbb{E}(Y_{1i} \mid D_i = 1) - \mathbb{E}(Y_{0i} \mid D_i = 1)}_{ATT} + \underbrace{\mathbb{E}(Y_{0i} \mid D_i = 1) - \mathbb{E}(Y_{0i} \mid D_i = 0)}_{\text{Selection Bias}}$$

  We just need the treatemnt to be independent of the outcome Y_0.
- **Assumption 1c**: $\mathbb{E}(Y_0 \mid X, D) = \mathbb{E}(Y_0 \mid X)$. Basically just CMI for Y_0.
- **Assumption 2:** Moreover, in this case, we only need a weaker overlap condition: $\mathbb{P}(D = 1 \mid X) < 1$.
- Idea: we will not need to iterate over all the values of X, but just for those categories where there some treated units. We just need to borrow $\mathbb{E}(Y_0 \mid D = 1, X)$ for values of X in the treatment group, that is, we just need a control analogue for each treated observation. We discard all those groups where we have just control units, hence we do not need $0 < \mathbb{P}(D = 1 \mid X)$, which guarantees a treated analogue for each control observation (that means: for those observations, we have the baseline outcome; we do not know what would have been the outcome under treatment, and we borrow such information from the treated of that same group). For computing the ATE you are basically required to input the missing counterfactual both for members of the T and the C group. That would require full overlap then. Recall also that for ATT you have to recompose the missing coutnerfactual just for Treted units!
- We just care about the treated units in each category, that's why we impose $\mid D = 1$. Then, *within these categories we care about* we borrow information from the control!

$$ATT = \mathbb{E}_X \left[\mathbb{E}(Y \mid X, D = 1) - \mathbb{E}(Y \mid X, D = 0) \mid D = 1\right]$$

- Differently,if you want to identify the ATNT you just need $0 < \mathbb{P}(D = 1 \mid X)$. Need treated units for each group where you have a control (you need to input the counterfactual outcome from the treated), but you don't care if there is a group full of controls (not the control individual you are interested in building a counterfactual). Indeed instead if you consider a group where you have just controls and no treated you would be missing the counterfactual for that group!

4. Estimation

4.1. Nonparametric Estimation of ATE and ATT

Just comptue the emans for each subgroup and then aggregate though a weighted average!

4.2. Estimating m_0 and m_1 by OLS

- Assume linear outcome models:

$$\mu_0(X) = \mathbb{E}(Y_0 \mid X) = \alpha_0 + X\beta_0, \quad \mu_1(X) = \mathbb{E}(Y_1 \mid X) = \alpha_1 + X\beta_1$$

- Estimate by OLS:
  - Regress Y_i on constant and X_i using controls $(D_i = 0)$ get $(\hat{\alpha}_0, \hat{\beta}_0)$
  - Regress Y_i on constant and X_i using treated $(D_i = 1)$ get $(\hat{\alpha}_1, \hat{\beta}_1)$
- Compute:

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

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

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

- $\bar{X}$ = average covariate in full sample; $\bar{X}_1$ = average covariate in treated subsample.

Implementation in Stata

- Procedure:
  - `regress Y X if D==1`
  - `predict Y1hat`
  - `regress Y X if D==0`
  - `predict Y0hat`
  - `means Y1hat Y0hat`
  - `lincom _b[Y1hat] - _b[Y0hat]`
- predict is basically predicitng for everyone, the aberage predicion will be on the aevrage covariates, as above
- Problem: standard errors are incorrect due to first-step estimation uncertainty.
- Correct way: use `teffects ra (Y X) (D), ate` to account for both steps properly via GMM.

Alternative: Pooled Regression with Interaction

- Use a single regression:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 X_i D_i + \varepsilon_i$$

- Then: $\widehat{ATE} = \hat{\beta}_2 + \hat{\beta}_3 \bar{X}$
- Same result as running two separate regressions.
- Don't want to sum coefficients? To get the standard errors for the ATE directly from one regression regress Y_i on constant, X_i, D_i, and $D_i(X_i - \bar{X})$
- Coefficient on D_i is estimated ATE.
- Note: adjust SE for $\bar{X}$ if needed usually small.

Introducing Covariates in OLS

- If controls are binary, a fully saturated model includes all interactions between covariates and with D.
- This is equivalent to a nonparametric model (initial nonparametric estimation via group means).
- **Common practice of linear-additive controls in a non saturatel model is a strong assumption**. "Non-saturated" = not allowing all interactions or group-specific effects. So we assume that the effect of each covariate is constant across groups and doesn't interact with other covariates or treatment (additive and linear).
- If funcional form is worng may misspecify model and bias results.
- If you have continuous covariates and want to flexibly adjust for confounding or allow heterogeneous treatment effects, you should: 1) add linar interaction, 2) add non linear terms.

5. The Problem with Linear Regressions

- Overlap assumption may be hidden in linear models: extrapolation beyond data may occur. Extrapolation beyond the data means that the regression model is used to make predictions in regions of the covariate space (X) where you have little or no data for one treatment group but not the other. Is all about the overlap assumption: if you have a treated in a given covariate region, you also need a control.
- non-parametric methods, Matching / Reweighting if they don't find a valid control counterfactual cannot compute the ATT_x for that covariate. The regression instead keeps the outlier (more infra). The regression uses all observations relying on the linear functional form. 1) li tiene tutti, 2) fitta una forma lineare. The point again is that the OLS is a global tool.

- If covariate distribution differs across treated and control, regression estimates can be very sensitive.
- That's why you should always check covariate distribution before applying regression methods.
- **Rule of thumb:** if covariate means differ between treated and control groups by more than 0.250.5 standard deviations, simple regression may fail to adjust for selection bias. DFF IN COVARIATE MEANS, MEANS UNBALANCE (a priori problem)
  - This indicates poor overlap: treated and control groups differ substantially in observed characteristics.
  - In such cases, regression relies on extrapolation and may produce biased estimates, especially under model misspecification.
  - Instead, consider matching, trimming, or nonparametric methods that restrict to common support and reduce reliance on functional form assumptions.

5.1. Matching estimator

Built with a focus on overlap. For every treatment variable, you try to find in the sample the control observations that are closest to the treated unit in terms of covariates. Multiple ways to define the closest.

- exact
- One dimension distance or weighted average of multiple dimensions distances (but how to asisgne weigths? imagine the acse of two cobariates age adn educ. a 1 yr diffeerence in age is lss relevant than a 1 yr difference in educ (reflected in sd). SO divide by sd!)

Matching Estimator

- Matching estimators rely on the **Unconfoundedness** (Ignorability) assumption:

$$D \perp (Y_1, Y_0) \mid X$$

- Under this, the conditional average treatment effect is identified as:

$$ATE_X = \mathbb{E}(Y \mid X, D = 1) - \mathbb{E}(Y \mid X, D = 0)$$

- The matching estimator compares treated and control units with identical or similar covariates X.

Fully Saturated Regression vs. Exact Matching (saturated = dummies)

- In practice, we can run a **fully saturated regression** of Y on D with controls X and $D \times X$ interaction terms. The interaction coefficients identify:

$$\mathbb{E}(Y \mid X, D = 1) - \mathbb{E}(Y \mid X, D = 0)$$

- Instead of a fully saturated regression, we can, equivalently, do **Exact Matching**:
  - Stratify the data into cells for each unique combination of the covariates X.
  - For each cell, compute the conditional treatment effect ATE_X as:

$$\widehat{ATE}_x = \bar{Y}_{T,x} - \bar{Y}_{C,x}$$

  - To obtain the overall ATE, average these cell-level effects using weights based on the distribution of X:

$$\widehat{ATE} = \sum_x \widehat{ATE}_x \cdot \mathbb{P}(X = x)$$

Regression vs. Matching

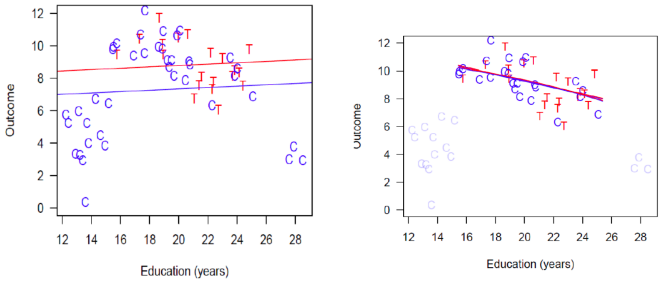
- Both regression and matching rely on the **selection on observables** (ignorability) assumption:

$$(Y_1, Y_0) \perp D \mid X$$

- **Regression** can be viewed as a form of pseudo-matching that relies on functional form assumptions (e.g., linearity):
  - It estimates conditional expectations $\mathbb{E}(Y \mid X, D)$ using a parametric model (e.g., linear regression).
  - Requires correct specification of the functional form (e.g., linearity, additivity).
  - Sensitive to extrapolation, especially when X distributions differ across groups. it uses the full support even if you have T observations just on a limited subset of it. to do such predictions it leverages the linear structure assumption.

$$Y_i = \alpha + \beta D_i + X_i'\gamma + D_i \cdot X_i'\delta + \varepsilon_i$$

  - issue 1) there is no data *& 2) assumption on parametric form is wrong*. were it correct it would be fine.
  - check: try to change the functional form specification and see if the ATE remains
  - you have to think about how the line is built (all the other things that are still on below the treatment dummy! that's why it has a different shape)



(a) Regression in unmatched sample **(b)** Regression in matched sample

About the graph: clearly, the overlap assumption does not hold. See how the gap between the two lines is the treatment effect. **Key point:** Regression can **mask the relevance of the overlap condition**. It produces estimates even in regions with no treated or control observations (extrapolation). **It uses all the observations through assuming some generlaizabel relaitonship!**.

5.1.1. Example

- Suppose:
  - Treated units have $X \in [5, 10]$
  - Control units have $X \in [0, 5]$
- To estimate the counterfactual for a treated unit with $X = 8$, we need:

$$\mathbb{E}(Y \mid X = 8, D = 0)$$

- However:
  - There is **no data** for $D = 0$ at $X = 8$
  - Regression uses the model (e.g., a linear function) to **extrapolate**
  - This estimate is **not identified from the data** and depends on the **correctness of the functional form**
- so you have a treated at 8, no controls at 8. So the reg extrapolates from $[0, 5]$ controls the functional form of the relationship and builds a counterfactual. Then the ate is the difference between the lines.
- If the regression is **not fully saturated** in $X \times D$ interactions:
  - It imposes constant treatment effects across X.
  - The resulting ATE estimate is a weighted average with weights not equal to $\mathbb{P}(D = 1 \mid X)$, hence **not equal to the average of** ATE_X (see AP for understanding

the weights).[1]⇒ IPW **downweights** the overrepresented and **upweights** the underrepresented units to reconstruct what *random assignment* would look like.
    - works as a post-stratification. issue: rare units (noisy, outlliers) get large weigths (overweights outliers,).
- **Matching** does not rely on a functional form, but:
    - many observations are unmateched and discarder
    - May be less statistically efficient in small samples.
    - Explicitly enforces comparison of similar observations.

6. Matching Techniques

6.1. Exact Matching on Covariates

- If we have one discrete covariate we can match each treatment to a unique control based on the values of that covariate (assuming exists a control for erach treated).
- You will get a set of N_T matched pairs (size of the treatment group).
- We can get an unbiased estimator for the ATT(X) for each match

$$\hat{\rho}_i^{\text{match}} = Y_i - Y_{m_i^c}$$

wh yis this the att_x? you ahve the outcome for the treated minus the outcome that the treated would have been had he been in the control (counterfactual).
- And recover the ATT as:

$$\hat{\rho}^{\text{match}} = \frac{1}{N_T} \sum_{i:D_i=1} \hat{\rho}_i^{\text{match}}$$

- issue: With a few discrete X we can do exact matching but:
    - But, with K binary variables, the numbers of cells is 2^K . If we have more cells than sample size, we will not find control and treated observations in each cell.
    - With continuous variables, it is not possible to do 1-to-1 matching ⇒ see next sections

6.2. Multivariate Distance Matching

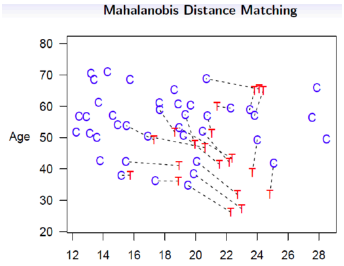
- Use some metric distance to define “close” matches on many covariates.
- Distance between the vector of characteristics of treated and control units:

$$\text{Distance}(X_T, X_C) = \sqrt{(X_T - X_C)' S^{-1} (X_T - X_C)}$$

- **Mahalanobis matching:** S is the covariance matrix of X.
    - If covariates have different units (e.g., income in $1000s vs. age in years), Euclidean distance would over-weight large-scale variables (tribiall ythe difference is larger). Mahalanobis rescales each variable by its variance
    - Mahalanobis accounts for correlation. If two variables are highly correlated, the "effective dimensionality" is lower (why penalizing twice the same stuff?)

  Euclidean matching: S is the identity matrix.
- Match each treated unit to the nearest control unit using this distance. Still, matching techniques above always find a match for every treated unit, regardless of how far the closest control is. Drop control units that are not used.

[1] In Inverse Probability Weighting (IPW), we instead reweight observations using the inverse of the estimated propensity score, $\mathbb{P}(D = 1 \mid X)$, to simulate a pseudo-population where treatment is randomly assigned. Units with high treatment probability get lower weight, and those with low probability get higher weight, balancing covariates across treatment status to consistently estimate the ATE. (more infra)



6.3. Propensity score matching

6.3.1. Definition and properties

- Using the propensity score allows to reduce the large dimensionality of the matching problem to only one dimension.
- **Propensity Score p(x)** : conditional probability of being in the treated group given pre-treatment variables:

$$p(X) \equiv \Pr(D = 1 \mid X) = \mathbb{E}(D \mid X)$$

- Instead of matching on the numerous covariates X, we can match on a single variable: the propensity score.

Ignorability conditional on Propensity Score: If $D \perp (Y_1, Y_0) \mid X \Rightarrow D \perp (Y_1, Y_0) \mid p(X)$. If we have ignorability conditioning on X, we also have it by conditioning only on $p(X)$. Basically, we are showing that a **single fucnion of multiple x** is just as good as multiple x used individually.

6.3.2. Identification of the PS

- We proved that given ignorability conditional on X, we have:

$$\Pr(D = 1 \mid Y_1, Y_0, p(X)) = \Pr(D = 1 \mid p(X))$$

  Treatment is independent of potential outcomes conditional on the pscore.
  Note this is untestable bc you do not have the two outocmes for both idnivduals.
- We can then define:

$$\text{ATE}_{p(X)} = \mathbb{E}(Y_{1i} - Y_{0i} \mid p(X_i))$$

- This parameter is identified since:

$$\begin{aligned}\mathbb{E}(Y_{1i} - Y_{0i} \mid p(X_i)) &= \mathbb{E}(Y_{1i} \mid p(X_i)) - \mathbb{E}(Y_{0i} \mid p(X_i)) \\ &= \mathbb{E}(Y_{1i} \mid p(X_i), D_i = 1) - \mathbb{E}(Y_{0i} \mid p(X_i), D_i = 0) \quad \text{(art: conditioning has no effect)} \\ &= \mathbb{E}(Y_i \mid p(X_i), D_i = 1) - \mathbb{E}(Y_i \mid p(X_i), D_i = 0)\end{aligned}$$

- Which we observe from the data.

For the ATT:

$$\begin{aligned}\mathbb{E}(\text{ATE}_{p(X)} \mid D_i = 1) &= \mathbb{E}\left[\mathbb{E}(Y_{1i} - Y_{0i} \mid p(X_i)) \mid D_i = 1\right] \quad \text{(by definition)} \\ &= \mathbb{E}\left[\mathbb{E}(Y_{1i} - Y_{0i} \mid p(X_i), D_i = 1) \mid D_i = 1\right] \quad \text{(by ignorability)} \\ &= \mathbb{E}(Y_{1i} - Y_{0i} \mid D_i = 1) \quad \text{(LIE)} \\ &= \text{ATT}\end{aligned}$$

As always, you care just about treating and reconstructing just their outcomes. I am not writing here again such discussion, but as in the general case here we are reconstructing the counterfactual by using control data in the same stratum!

6.3.3. Implementation of Matching based on the pscore

- Estimate Propensity Score using a probability model (eg parametric = probit)
- Estimate treatment effect:
    - We would like to match treated and controls with identical (estimated) probability of being treated.
    - Since p(X) is continuous, we will not find treated and control units with identical values!

  - To match on the pscore, there are several alternatives to exact matching: stratification, nearest neighbor, radius, kernel.
  - Compute the treatment effect for each value of the estimated propensity score (way of sayign that you compute treatmeent effect for each mathced couple)
  - Average over all the estimated $ATE_{p(X)}$.

I. Estimation of the Propensity Score

- Estimate the pscore using any standard probability model (logit, probit: we want the value to be bewteen 0 and 1).
- If we use a Logit model, allowing for linear and higher-order terms of covariates:

$$\Pr(D = 1 \mid X) = \frac{e^{\gamma h(X_i)}}{1 + e^{\gamma h(X_i)}}$$

- the equation on the right is the structural form of the probit.
- γ is a scale parameter (speed of the transition)
- $h(X_i)$ is what makes the formula a function of you predictors

$$h(X_i) = (x_{1i}, x_{2i}, x_{1i}^2, x_{2i}^2, x_{1i}x_{2i})$$

- But, we would fall in the same dimensionality problem as before. The formula above explodes quickly. You may have more variavles than observaitons if yoy add all variables and interactions in otder to perfectly model the propesity score. Note that each vairable in the fucniton above owul dget a different parameter to be estianated!
- The point is that here we do not need a fully saturated model.
- Truly, you care just about the pscore balancing property so that T and C group are as similar as possible:

$$\Pr(D = 1 \mid X, p(X)) = \Pr(D = 1 \mid p(X))$$

No need to find the best prediction for the p score!. We just need to include in the probability model a parsimonious set of variables that gives us an estimate of the propensity score satisfying the balancing property

- Note that you cannot test the ignorability condition above. You can test that fiven the pscore the covaraietes are balances but this is 1) just in sample, 2) know nothing about balance in unobervables!

6.3.4. Stratification on the Pscore combined with regression: to be 100% sure

- Combine the stratification method with regressions.
- Since we use an estimated pscore, some correlations within stratum between treatment indicator and covariates may remain in finite samples.
- Regression adjustment within blocks can improve precision and reduce any remaining conditional bias within block.
- We run, within each stratum, an OLS regression:

$$Y_{is} = \beta_{0s} + X_s\beta_s + \rho_s D_s + \varepsilon_{is}$$

Note: you could have done pooled regression with stratum dummies and interactions.

6.3.5. Estimation of ATT using pscore: alternative methods

- Up to now exact matching on the propensity score, but there may be no exact match.
- But, we can match each treated unit with the nearest control in terms of the pscore.
- Three alternative methods:
  - **Nearest neighbor matching**: for each treated unit, find the nearest control unit in terms of pscore
  - **Radius (caliper) matching**: for each treated unit, find all the control units whose score differs by less than a chosen tolerance r (it tries to avoid "bad" matches, does not go too far). Then do average of matched control units.
  - **Kernel matching**: each treated unit is matched with a weighted average of all control units, with weights inversely proportional to the distance between the scores (uniform kernel: all have same weight, triangular kernel: pyramidal).
    see how in radius and kernel you have more controls per treated uni!

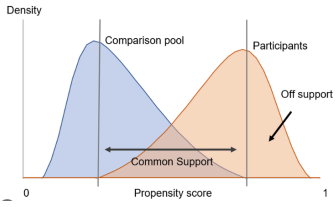
6.3.6. SE with the ps

- You have a difference in outcomes for the matched treated and control groups. Now you have to find Standard Errors!
- **Bootstrap** is a resampling method used to estimate standard errors by repeatedly drawing samples *with replacement* from the observed data. Synthetic way of generating sampling variability (Neyman style of inference). You have just a ssample of 1000 (representative), and you want to replicate what could be happening in the population. In ols gives same SEs of OLS (no formal proof, but empiric).
- Standard Errors are often obtained using bootstrap resampling methods. It is important to estimate the two stages (the probability model for the pscore and the ATT) simultaneously, not in two steps. Resample symultaneously both stages:
  1. Estimation of the propensity score model.
  2. Estimation of the ATT based on matched samples (eg simple differenc ein mean as above).

  Estimating them in two steps (fixing pscore) underestimates variability (obv), also the propensity score depends on the smapling variability.
- **Important warning:** The bootstrap is valid for smooth matching methods (e.g., kernel), but **not valid for nearest neighbor matching**. kernel matching is continous, Nearest eigh is dicreete = we have a jump in matches, all units will be matched differnly.
  note: you can use bootstrapping also for power computation

6.3.7. Common Support / Overlap

- Very imporant, it may be naturally satisfied by the type of matching i do. If you think about it still overlap is a great issue bc you could end up matching units far away (is absolutely not an issue if you do exact matching).
- Do histograms of estimated pscores for control and treated groups with bins equal to pscore strata
- Theoretical assumption: common support is about the existence of comparable treated and control units for each covariate profile.
- Estimated common support is the intersection of the predicted propensity score for treated and control observations.
- Empirical enforcement: we often trim tails to avoid poor matches or extrapolation when overlap is weak or violated. Drop the control obs. with estimated pscore lower to the min pscore for the treated, and the treated obs. with estimated pscore higher than the max for the controls.



- Check: drop the 1% lower and upper tails of the distribution. Repeat the estimation of ATT in the common support and check whether results are sensitive or not.

- Also check for balance on covariates once restrict to obs. in the common support :)

Further Considerations:

- Trimming the sample alters the estimand: it is no longer the ATT or ATE for the full population, but for the subset within common support. Trimming sacrifices external validity (results no longer generalize to the full sample), but improves internal validity by ensuring more comparable treated and control groups.
- We trim observations with propensity scores close to 1 or 0, because it is difficult to find comparable units in the opposite treatment arm with similar scores.
- Keeping such observations would force extreme extrapolations, reducing credibility of the estimated effect.
- Imbens and Rubin (2015) propose an automatic trimming rule: drop all observations with estimated propensity score outside a pre-defined treshold.

7. Weighting techniques: Inverse Probability Weighting

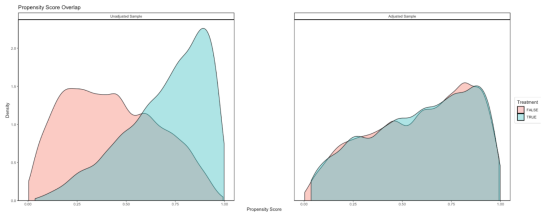
- Instead of using matching estimators, we can estimate treatment effects using **Inverse Probability Weighting (IPW)**, first we will see how the weighti works and do a simple difference in means, in the last few liens we well integrate this in a regression framework.
- Each observation is weighted by the inverse of the estimated propensity score in the regression: The **Inverse Probability Weights** are defined as:

$$w_i = \begin{cases} \dfrac{1}{p(X_i)} & \text{if } D_i = 1 \quad \text{(treated)} \\ \dfrac{1}{1 - p(X_i)} & \text{if } D_i = 0 \quad \text{(control)} \end{cases}$$

Rationale:

  - $\hat{e}(X_i)$ is the estimated propensity score, i.e., the probability of receiving treatment given covariates: $\hat{e}(X_i) = \mathbb{P}(D_i = 1 \mid X_i)$.
  - Treated units with a high $\hat{e}(X_i)$ (i.e., more likely to be treated) are *downweighted*.
  - Control units with a low $\hat{e}(X_i)$ (i.e., unlikely to be treated, these are common) are *downweighted*. Control units with a high $\hat{e}(X_i)$ (i.e., likely to be treated, these are rare) are *upweighted*.

- This creates a *pseudo-population* in which the distribution of covariates is similar across treated and untreated units:
  - Covariate imbalance is corrected at the population level.
  - Individuals with a high probability of treatment (based on covariates) are downweighted.
  - This reweighting restores balance: treatment becomes independent of covariates.
  - Everyone is equally likely to be treated! You have a treatemtn across balanced obervations
- Conceptually, IPW is very similar to *selection on observables*:
  - Selection on observables addresses bias via **conditioning** (e.g., regression).
  - IPW achieves balance via **reweighting**.
  - Once balanced, there's no need to control for covariates in the outcome model.
  - This avoids bias from functional form misspecification in OLS.
- **Advantages of IPW over Propensity Score Matching (PSM):**
  1. PSM is (usually) non-smooth, which complicates inference (e.g., bootstrapping).
  2. IPW can be extended to *doubly robust* estimators (AIPW).
  3. PSM introduces additional bias due to imperfect matches.
  4. PSM does not guarantee full covariate balance.
  5. PSM discards unmatched observations, making it less efficient.
- **Limitations of IPW relative to PSM:**
  1. PSM does not require correct model specification for the propensity score or the outcome.
  2. PSM restricts attention to the region of common support, avoiding extrapolation.
  3. IPW may assign very large weights to units with extreme propensity scores, increasing variance and sensitivity to misspecification.
- **Intuition: self-selection into treatment**
  - Individuals self-select into treatment based on observable characteristics.
  - These characteristics may correlate with potential outcomes.
  - IPW corrects for this by downweighting overrepresented strata and upweighting underrepresented ones.
  - This rebalances the sample, leadign to similar "synthetic" individuals in both T and C groups, allowing valid estimation of treatment effects.



- The idea originates from Horvitz and Thompson (1952), and allows identification of counterfactual means via:

$$E\left(\frac{Y_i D_i}{p(X_i)}\right) = E(Y_{1i}), \quad E\left(\frac{Y_i(1 - D_i)}{1 - p(X_i)}\right) = E(Y_{0i})$$

  Very interesitng: we are saying that the observed outcomes of the treatment units reweighted using inverse probability of treatment are equal to the Average potential outcome if everyone were treated (wow, you cannot observe it for every i). We reweight treated outcomes so they represent what would happen if the entire population were treated. See the figure above: if you reweight the treatment it has a distribtion that is eaual to the one of the control reweighed. the distribution of covariaets in the weighted T is equal to the distribution of covariates in the weighted C. The fact that the treated distribution is equal to the control distribution is due to the fact that: in the population, there is a given distribution over the covariates, and this distribution is $f(x) = \mathbb{P}(X_i = x)$. then, treated observations and control observations are weighted in such a way as to replicate this distribution $f_X(x)$.. THE IDEA IS: NOW THE TREATED OUTCOME IS REPRESENTATIVE OF THE FULL POPULATION AND TAS WELL THE CONTROL OUTCOME IS REPRESENTATIVE OF THE FULL POPULAITON.
- Thus you can identify:

$$ATE = E(Y_{1i} - Y_{0i})$$

7.1. Weighted regression

IPW can be combined with regression adjustment to form a **doubly robust** estimator:

- simply you do an ols where observations are weighted of Yi on Di + can add other controls.
- The estimator is consistent if either (at least one of the two) the propensity score model or the regression model is correctly specified (eg linear form with no omitted covariates).
- And one can prove that this ATE is equal to the coefficient on the treatment on the weighted regression of Yi on Di.
- In Stata, the syntax is: `teffects ipwra (Y X1 X2 X3) (D X1 X2 X3 X4)`
- *Note: We typically include more variables in the selection equation (pscore prediction) than in the outcome equation to improve covariate balancing without overfitting the outcome model.*

8. AP. Angrist, Joshua (1998) Estimating the Labor Market Impact on Voluntary Military Service Using Social Security Data on Military Applicants.

8.1. Design and CIA

- **Goal** is to measure effect of voluntary military service on labor earnings.
- **Ideal experiemnt**: take people at random and send them to war. ATT $= \mathbb{E}(Y_{1i} - Y_{0i} \mid D_i = 1)$: tells us whether, on average, veterans benefited or not from military service.
- **Issue 1**: Veterans are both self-selected and screened by the military (selection bias).
- **Solution (I):** restrict the sample to people who applied to the military. Loss in terms of external validity BUT no selection bias.
  - Treatment Group: enlisted applicants (veterans)
  - Control Group: non-enlisted applicants (subgroup of non-veterans).
- **Issue 2**: we may say that those that did not pass the screen are intrinsically different from those that did. These differences may lead them to have lower average earnings today.
- **Solution 2** Here's the twist.we can use selection on observables because we exactly know the criteria used by the military to do the screening (age, schooling, and test scores), and in our data we observe such covariates. **Great treatment knowledge** allows us to have CIA by great selection on observables. **CIA**: veteran status is independent of potential outcomes after controlling for age, schooling, and test scores.

$$(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i \mid X_i.$$

so we are matching people in the treatment and the control group that have the same covariates! Similar in observable characteristics to the treated group (enlisted applicants) but were not selected or chose not to enlist (unlucky)!

- We are interested in the ATT, so the overlap assumption is

$$\mathbb{P}(D = 1 \mid X) < 1$$

8.2. Data

- A fully-saturated model would require to include all combinations of covariates.
- Administrative data from the US military (application records) and earnings data from the Social Security Administration.
- Earnings data available at aggregate cell-level: 8,760 cells defined by race, year of birth, schooling level, year of application, score group, and veteran status.
- Further restriction on number of observations per cell (>25): 5,654 cells.

8.3. Matching

- **Construct covariates-cells** based on: race, application year, schooling at application, test score group in qualification test by the army, and year of birth.
- *Note that Angrist also matches on Year of Birth, which was not correlated to the probability of serving, but as it affects earnings, including it improves efficiency (Year of birth affects a lof the outcone, conditioning on it makes comparisons within year groups more precise).*
- The Standard Matching estimator with discrete X is:

$$ATT = \mathbb{E}[Y_{1i} - Y_{0i} \mid D_i = 1] = \sum_k \text{ATE}_X \cdot \mathbb{P}(X_i = x \mid D_i = 1)$$

note how the summation is across cells. Note the probability is on the distribution of covariates among treated, so gives the ATT.
- Formula above, in words: Angrist replaces ATEx by the sample veterannonveteran earnings difference for each combination of covariates ($\bar{y}_{1k} - \bar{y}_{0k}$). Recall that the simple difference in means is the ATT + selection bias (recall lec 1), and that with CIA ATE = ATT. Then he comabines in a weighted average using the empirical distribution of covariates among veterans at values where the difference in mean outcomes are defined. Idea: the ATEx is estiametd on a cell as the duifference beteen the outocme of the teated and the counteractual outcome of the control. than this is weighted considering how relevant that cell is.

8.4. Regression

- Regression:

$$\bar{y}_{Dk} = \beta_k + \alpha D + \bar{\varepsilon}_{Dk}$$

- α is the veteran effect, and β_k is a cell specigic effect (combination of x).
- Grouped data: weighted least squares using population cell counts as weights, same estimates as micro data weighted by inverse sampling rates (see AP), that is:

$$y_i = \sum_k d_{ik}\beta_k + \alpha D_i + \varepsilon_i$$

- where d_{ik} indicates whether individual i belongs the X-cell k ($X_i = x_k$). There is only one ik combinaiton different from 0! (elegant way to match individual to group)
- Saturated in X, but does not include the interaction XD. I think that if you add interactions in the reg you get the right ATT, this bc the interactions will absorb the heterogeneous TEs

Implicit Weights in Regression (vs. Matching)

Key idea: Regression and matching differ in how they implicitly weight the conditional treatment effects ATE_X.
Wait we had said that fully saturated reg is equivalent to exact matching, so what's the issue. It is equivalent as long as TE is homogeneous. Already above when we said we had no difference there were different weights, but ATE was the same because TE was homogeneous. Now we have heterogeneous TE ATE and different weights → different ATE!

- Regression weights ATE_X by the conditional variance of treatment D_i given covariates X_i: this reflects how much variation in treatment exists within each covariate group. SO THIS IS A T-C VARIANCE WEIGTHED AVERAGE OF THE ATT.
- Matching weights ATE_X by the conditional probability of treatment at each X_i. THIS IS THE ATT.

Regression weighting:

- Consider a saturated regression of Y_i on D_i and X_i (no interaction terms), and let δ_R be the coefficient on D_i (simplifies the analysis to a single coefficien δ).
- Then:

$$\delta_R = \frac{\text{Cov}(Y_i, \tilde{D}_i)}{\text{Var}(\tilde{D}_i)} = \frac{\mathbb{E}\left[(Y_i - \mathbb{E}[Y_i])(D_i - \mathbb{E}[D_i \mid X_i])\right]}{\mathbb{E}\left[(D_i - \mathbb{E}[D_i \mid X_i])^2\right]}$$

$$= \frac{\mathbb{E}\left[\sigma_D^2(X_i) \cdot ATE_X\right]}{\mathbb{E}\left[\sigma_D^2(X_i)\right]}$$

where $\sigma_D^2(X_i) = \text{Var}(D_i \mid X_i) = \mathbb{E}\left[(D_i - \mathbb{E}[D_i \mid X_i])^2 \mid X_i\right]$.

- Thus, regression produces a **variance-weighted average** of ATE_X.

Matching weighting:

$$ATT = \sum_X ATE_X \cdot \Pr(Xi = X \mid Di = 1)$$

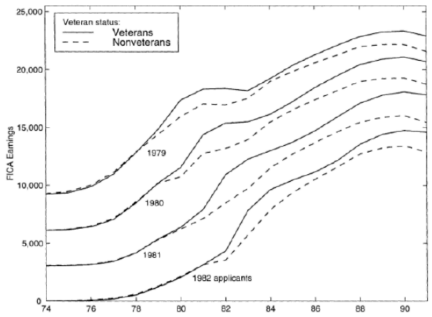
look this is the same fomrula as before. Recall ATT $= \mathbb{E}[Y_1 - Y_0 \mid D = 1]$, and that ATT $\approx \frac{1}{N_1} \sum_{i:D_i=1} ATE_{X_i}$. Now the unit is not individuals but groups, and each group has diff n of individuals ⇒ we weight for a the ration of n of the group / the total treated. Notice how the aggreagte does not show thsi but this is an average ATT $= \frac{1}{N_1} \sum_{i:D_i=1} (Y_{1i} - Y_{0i})$ just on the treated!

Summary of Differences:

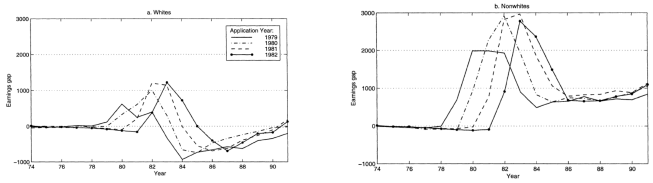
- Matching gives more weight to cells where the treatment probability is high (i.e., treated observations are more concentrated).
- Regression gives more weight to cells where treatment assignment is more variable (i.e., $\Pr(D = 1 \mid X)$ is near 0.5), maximizing variance. This is standard for OLS: ols minizmies errors. Where you haev more C adn T you have higher conditional variance, so more error to be minimzied. OLS will care a lot about these bins!

8.5. Results

- Pre-trends are fine etween T and C (we have their earnings before and after entering the military)
- Graphical analysis; seems veterans are beter off



- BUT, bveterans on aberage hhave hhigehr test scores and are mroe likey to be college graduate (slextion on observables!)
- Regression with controls and matching show that simple comparison in means overestimate the veteran effect (a simple regression of earnings on veteran status is affected by positive selection bias).



(a) White applicants **(b)** Non-white applicants

Figure 2. Earnings by veteran status and race

the desctipiton of the figure is as clear as you see

- I Regression and Matching estimates are almost identical until 1984.
- After 1984 regression gives larger estimates than matching.
- The matching estimator gives the largest weight to the covariate-specific ATE for men with high probability of serving.
- The regression estimator gives more weight to covariate-specific estimates where the probability of military service is close to 0.5 (where variance is maximized).
- This leads to a higher treatment effect by regressions than the one found by matching estimators if those with high probability of service ex-ante, have lower treatment effect as suggested in Figure 4 (coming next).

8.6. problems

- A large fraction of the applicants who do not enlist, appear to qualify for enlistment in any case. Issue: after controlling for observed characteristics X (like age, education, test score), treatment assignment D (enlistment) should beis as good as random.Many non-enlisted applicants appear "qualified" based on observable X (they pass the criteria). BUT THEY AR ENOT TREATED! This implies selection into treatment is not fully explained by X
- That is why, the paper also uses an IV strategy that achieves identification based on a different set of assumptions.
- Angrist claims that the fact that he obtains qualitatively similar conclusions with the three methods gives further credibility to the results. Do you agree? Yes triangularization helps (resutls less likely to be spurious due to model-specific issues.)