

Matching

EC 607, Set 7

Edward Rubin
Spring 2020

Prologue

Schedule

Last time

- The conditional independence assumption: $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | X_i$
- Omitted variable bias
- Good vs. bad controls

Today

- Return first round of project proposals.
- Matching estimators (*MHE* 3.2 and Cameron and Trivedi 25.4).

Upcoming

- Admin: Assignment(s)
- No midterm
- Next round of the project proposal

Matching

Matching

The gist

Remember the **conditional independence assumption**[†] in a setting—*i.e.*, treatment is as-good-as random conditional on a known set of covariates?

[†] AKA "selection on observables"

Matching

The gist

Remember the **conditional independence assumption**[†] in a setting—*i.e.*, treatment is as-good-as random conditional on a known set of covariates?

Matching estimators take us at our word.

[†] AKA "selection on observables"

Matching

The gist

Remember the **conditional independence assumption**[†] in a setting—*i.e.*, treatment is as-good-as random conditional on a known set of covariates?

Matching estimators take us at our word.

If we really believe $(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D_i | X_i$, then we can just calculate a bunch of treatment effects conditional on X_i , *i.e.*,

$$\tau(x) = E[Y_{1i} - Y_{0i} \mid X_i = x]$$

[†] AKA "selection on observables"

Matching

The gist

Remember the **conditional independence assumption**[†] in a setting—i.e., treatment is as-good-as random conditional on a known set of covariates?

Matching estimators take us at our word.

If we really believe $(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D_i | X_i$, then we can just calculate a bunch of treatment effects conditional on X_i , i.e.,

$$\tau(x) = E[Y_{1i} - Y_{0i} \mid X_i = x]$$

The idea: Estimate a treatment effect only using observations with (nearly?) identical values of X_i .

[†] AKA "selection on observables"

Matching

The gist

Remember the **conditional independence assumption**[†] in a setting—i.e., treatment is as-good-as random conditional on a known set of covariates?

Matching estimators take us at our word.

If we really believe $(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D_i | X_i$, then we can just calculate a bunch of treatment effects conditional on X_i , i.e.,

$$\tau(x) = E[Y_{1i} - Y_{0i} | X_i = x]$$

The idea: Estimate a treatment effect only using observations with (nearly?) identical values of X_i . The CIA buys us causality within these groups.

[†] AKA "selection on observables"

Matching

Goals

Let's return to **the fundamental problem of causal inference** for a moment.

1. We want/need to know $\tau_i = Y_{1i} - Y_{0i}$.
2. We cannot simultaneously observe *both* Y_{1i} and Y_{0i} .

Matching

Goals

Let's return to **the fundamental problem of causal inference** for a moment.

1. We want/need to know $\tau_i = Y_{1i} - Y_{0i}$.
2. We cannot simultaneously observe *both* Y_{1i} and Y_{0i} .

Most empirical strategies boil to strategies to estimate Y_{0i} for treated individuals—the unobservable counterfactual for the treatment group.

Matching

Goals

Let's return to **the fundamental problem of causal inference** for a moment.

1. We want/need to know $\tau_i = Y_{1i} - Y_{0i}$.
2. We cannot simultaneously observe *both* Y_{1i} and Y_{0i} .

Most empirical strategies boil to strategies to estimate Y_{0i} for treated individuals—the unobservable counterfactual for the treatment group.

Matching is no different.

We match untreated observations to treated observations using X_i , *i.e.*, calculate a $\widehat{Y_{0i}}$ for each Y_{1i} , based upon "matched" untreated individuals.

Matching

More formally

We want to construct a counterfactual for each individual with $\mathbf{D}_i = 1$.

Matching

More formally

We want to construct a counterfactual for each individual with $\mathbf{D}_i = 1$.

The counterfactual for i should only use individuals that match \mathbf{X}_i .

Matching

More formally

We want to construct a counterfactual for each individual with $\mathbf{D}_i = 1$.

The counterfactual for i should only use individuals that match \mathbf{X}_i .

Let there be N_T treated individuals and N_C control individuals. We want

- N_T sets of weights
- with N_C weights in each set

Matching

More formally

We want to construct a counterfactual for each individual with $\mathbf{D}_i = 1$.

The counterfactual for i should only use individuals that match \mathbf{X}_i .

Let there be N_T treated individuals and N_C control individuals. We want

- N_T sets of weights
- with N_C weights in each set: $w_i(j)$ ($i = 1, \dots, N_T; j = 1, \dots, N_C$)

Matching

More formally

We want to construct a counterfactual for each individual with $\mathbf{D}_i = 1$.

The counterfactual for i should only use individuals that match \mathbf{X}_i .

Let there be N_T treated individuals and N_C control individuals. We want

- N_T sets of weights
- with N_C weights in each set: $w_i(j)$ ($i = 1, \dots, N_T; j = 1, \dots, N_C$)

Assume $\sum_j w_i(j) = 1$. Our estimate for the counterfactual of treated i is

$$\widehat{Y_{0i}} = \sum_{j \in (D=0)} w_i(j) Y_j$$

Matching

More formally

If our estimated counterfactual for treated individual i is

$$\widehat{Y}_{0i} = \sum_j w_i(j) Y_j$$

then our estimated treatment effect (for individual i) is

$$\hat{\tau}_i = Y_{1i} - \widehat{Y}_{0i} = Y_{1i} - \sum_j w_i(j) Y_j$$

Matching

More formally

If our estimated counterfactual for treated individual i is

$$\widehat{Y}_{0i} = \sum_j w_i(j) Y_j$$

then our estimated treatment effect (for individual i) is

$$\hat{\tau}_i = Y_{1i} - \widehat{Y}_{0i} = Y_{1i} - \sum_j w_i(j) Y_j$$

\therefore a generic matching estimator for the treatment effect on the treated is

$$\hat{\tau}_M = \frac{1}{N_T} \sum_{i \in (D=1)} (Y_{1i} - \widehat{Y}_{0i}) = \frac{1}{N_T} \sum_{i \in (D=1)} \left(Y_{1i} - \sum_{j \in (D=0)} w_i(j) Y_j \right)$$

Matching

Weight for it[†]

So all we need is those weights and we're done.^{††}

[†]  ^{††} Plus an interesting, policy-relevant setting with valid conditional independence. And data.

Matching

Weight for it[†]

So all we need is those weights and we're done.^{††}

Q Where does one find these handy weights?

[†]  ^{††} Plus an interesting, policy-relevant setting with valid conditional independence. And data.

Matching

Weight for it[†]

So all we need is those weights and we're done.^{††}

Q Where does one find these handy weights?

A You've got options, but you need to choose carefully/responsibly.

E.g., if $w_i(j) = \frac{1}{N_C}$ for all (i, j) , then we're back to a difference in means.

This weighting doesn't abide by our conditional independence assumption.

[†]  ^{††} Plus an interesting, policy-relevant setting with valid conditional independence. And data.

Matching

Weight for it[†]

So all we need is those weights and we're done.^{††}

Q Where does one find these handy weights?

A You've got options, but you need to choose carefully/responsibly.

E.g., if $w_i(j) = \frac{1}{N_C}$ for all (i, j) , then we're back to a difference in means.

This weighting doesn't abide by our conditional independence assumption.

The plan Choose weights $w_i(j)$ that indicate **how close** X_j is to X_i .

[†]  ^{††} Plus an interesting, policy-relevant setting with valid conditional independence. And data.

Matching

Proximity

Our weights $w_i(j)$ should be a measure of **how close** \mathbf{X}_j is to \mathbf{X}_i .

Matching

Proximity

Our weights $w_i(j)$ should be a measure of **how close** \mathbf{X}_j is to \mathbf{X}_i .

If \mathbf{X} is **discrete**, then we can consider equality, i.e., $w_i(j) = \mathbb{I}(\mathbf{X}_i = \mathbf{X}_j)$, scaling as necessary to get $\sum_j w_i(j) = 1$.

Matching

Proximity

Our weights $w_i(j)$ should be a measure of **how close** \mathbf{X}_j is to \mathbf{X}_i .

If \mathbf{X} is **continuous**, then we need *proximity* rather than *equality*.

Matching

Proximity

Our weights $w_i(j)$ should be a measure of **how close** \mathbf{X}_j is to \mathbf{X}_i .

If \mathbf{X} is **continuous**, then we need *proximity* rather than *equality*.

Nearest-neighbor matching chooses the single closest control observation using the Euclidean distance between \mathbf{X}_i and \mathbf{X}_j , i.e.,

$$d_{i,j} = (\mathbf{X}_i - \mathbf{X}_j)' (\mathbf{X}_i - \mathbf{X}_j)$$

Matching

Proximity

Our weights $w_i(j)$ should be a measure of **how close** \mathbf{X}_j is to \mathbf{X}_i .

If \mathbf{X} is **continuous**, then we need *proximity* rather than *equality*.

Nearest-neighbor matching chooses the single closest control observation using the Euclidean distance between \mathbf{X}_i and \mathbf{X}_j , i.e.,

$$d_{i,j} = (\mathbf{X}_i - \mathbf{X}_j)' (\mathbf{X}_i - \mathbf{X}_j)$$

- $\hat{\tau}_i = Y_{1i} - Y_{0j}^i$, where Y_{0j}^i is i 's nearest neighbor in the control group.
- **Estimator:** $\hat{\tau}_M = \frac{1}{N_T} \sum_i \hat{\tau}_i$
- Produces causal estimates if CIA is valid *and* we have sufficient overlap.
- Suffers from arbitrary choices of units.

Matching

Proximity

Our weights $w_i(j)$ should be a measure of **how close** \mathbf{X}_j is to \mathbf{X}_i .

If \mathbf{X} is **continuous**, then we need *proximity* rather than *equality*.

Nearest-neighbor matching with Mahalanobis distance chooses the single closest control using **Mahalanobis** distance between \mathbf{X}_i and \mathbf{X}_j , i.e.,

$$d_{i,j} = (\mathbf{X}_i - \mathbf{X}_j)' \Sigma_X^{-1} (\mathbf{X}_i - \mathbf{X}_j)$$

where Σ_X^{-1} is the covariance matrix of \mathbf{X} .

Matching

Proximity

Our weights $w_i(j)$ should be a measure of **how close** \mathbf{X}_j is to \mathbf{X}_i .

If \mathbf{X} is **continuous**, then we need *proximity* rather than *equality*.

Nearest-neighbor matching with Mahalanobis distance chooses the single closest control using **Mahalanobis** distance between \mathbf{X}_i and \mathbf{X}_j , i.e.,

$$d_{i,j} = (\mathbf{X}_i - \mathbf{X}_j)' \Sigma_X^{-1} (\mathbf{X}_i - \mathbf{X}_j)$$

where Σ_X^{-1} is the covariance matrix of \mathbf{X} .

- **Estimator:** $\hat{\tau}_M = \frac{1}{N_T} \sum_i \hat{\tau}_i$ where $(\hat{\tau}_i = Y_{1i} - Y_{0j}^i)$
- Produces causal estimates if CIA is valid *and* we have sufficient overlap.
- Does not suffer from arbitrary choices of units.

Matching

More neighbors?

Why limit ourselves to a **single** "best" match?

If we're going to let a function/algorithm choose the *nearest* match, can't we also let the function/algorithm choose *how many* matches?

Furthermore, if $N_C \gg N_T$, it we're throwing away *a lot* of information.

We could instead use this information and be more efficient.

Matching

More neighbors!

Kernel matching gives positive weight to all control observations within some **bandwidth** h , with higher weight for closer matches determined by some **kernel function** $K(\cdot)$,

$$w_i(j) = \frac{K\left(\frac{\mathbf{X}_j - \mathbf{X}_i}{h}\right)}{\sum_{j \in (D=0)} K\left(\frac{\mathbf{X}_j - \mathbf{X}_i}{h}\right)}$$

Matching

More neighbors!

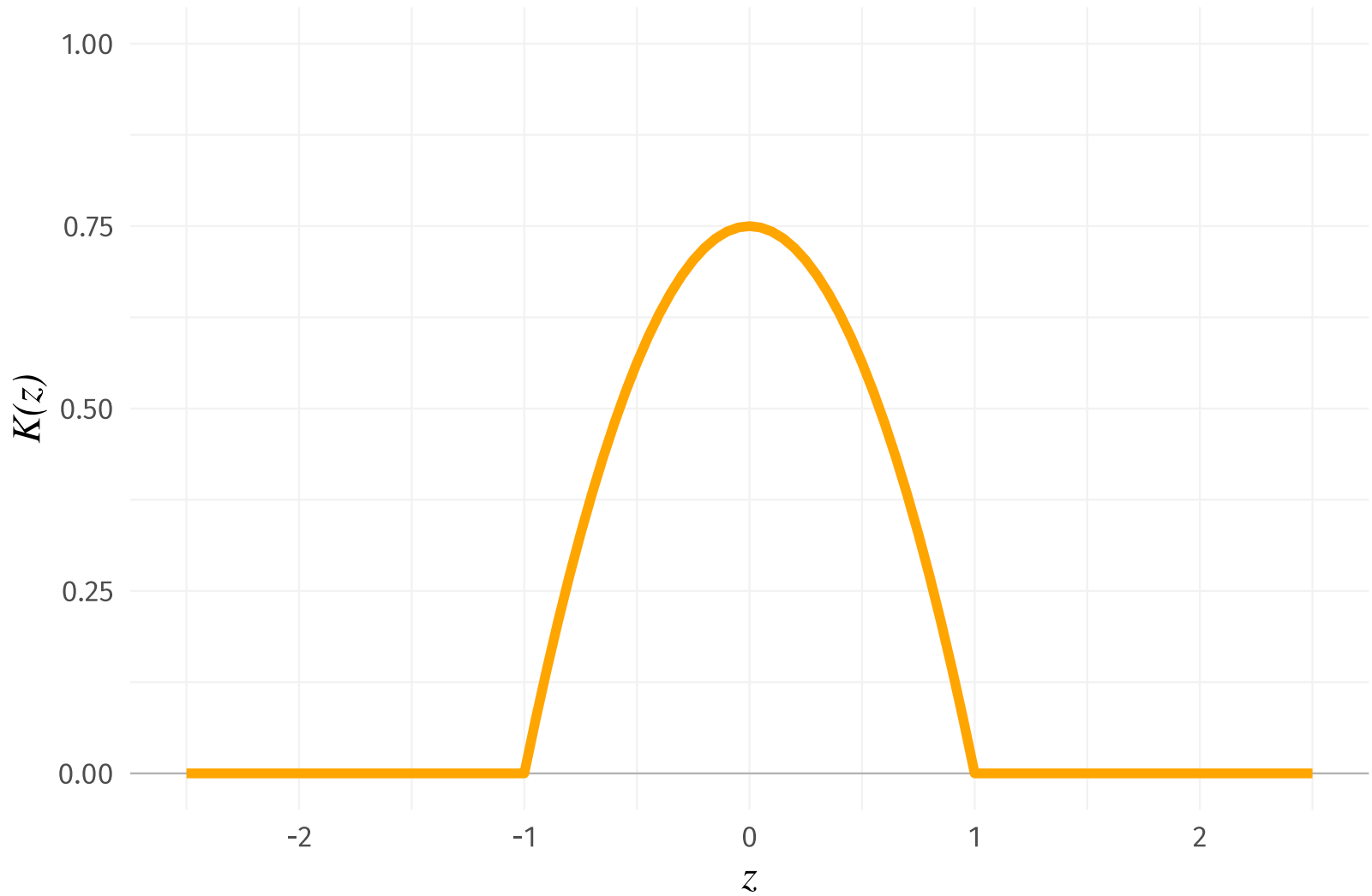
Kernel matching gives positive weight to all control observations within some **bandwidth** h , with higher weight for closer matches determined by some **kernel function** $K(\cdot)$,

$$w_i(j) = \frac{K\left(\frac{\mathbf{X}_j - \mathbf{X}_i}{h}\right)}{\sum_{j \in (D=0)} K\left(\frac{\mathbf{X}_j - \mathbf{X}_i}{h}\right)}$$

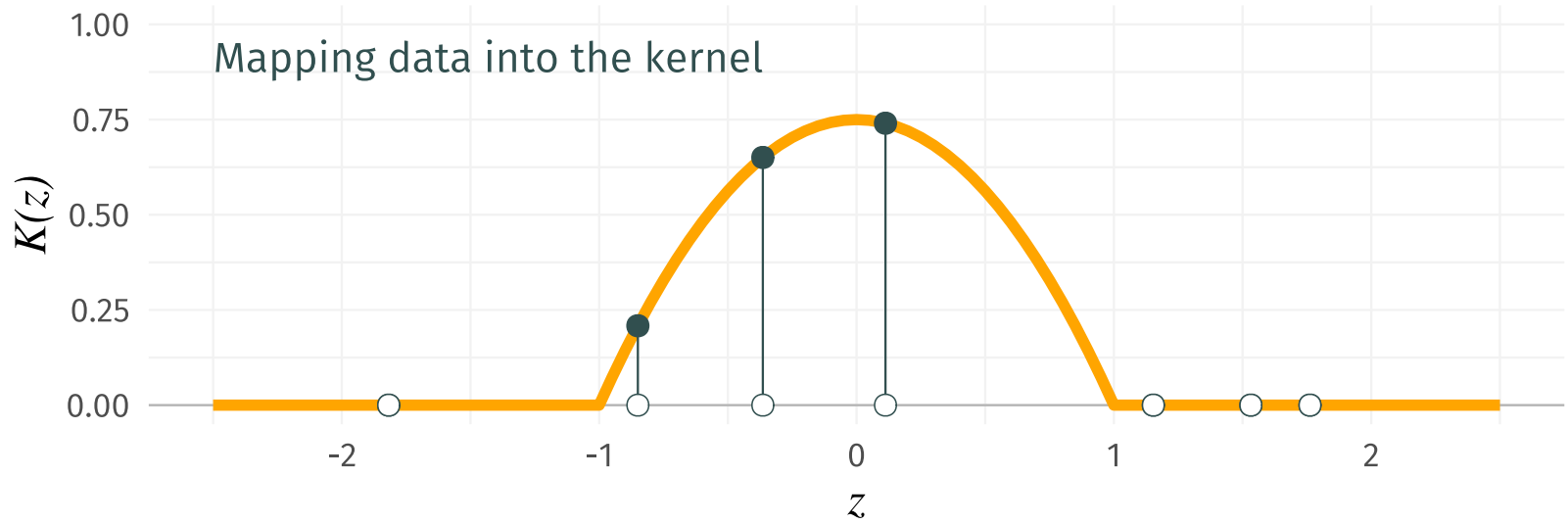
Example The *Epanechnikov kernel* is defined as

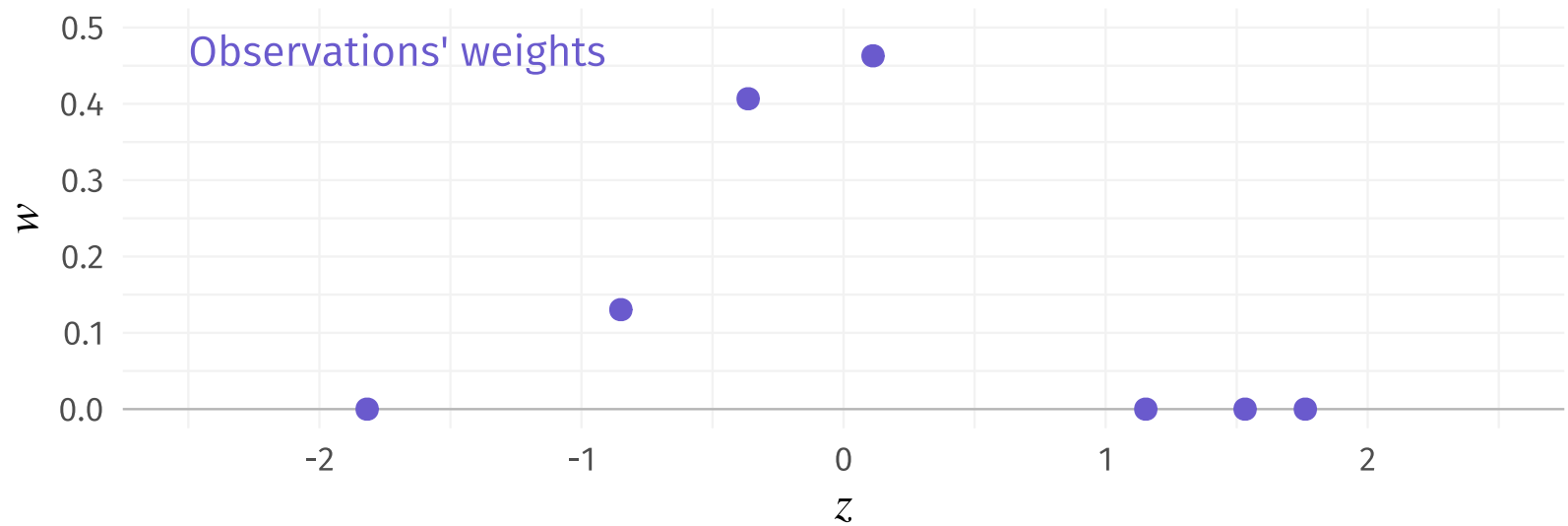
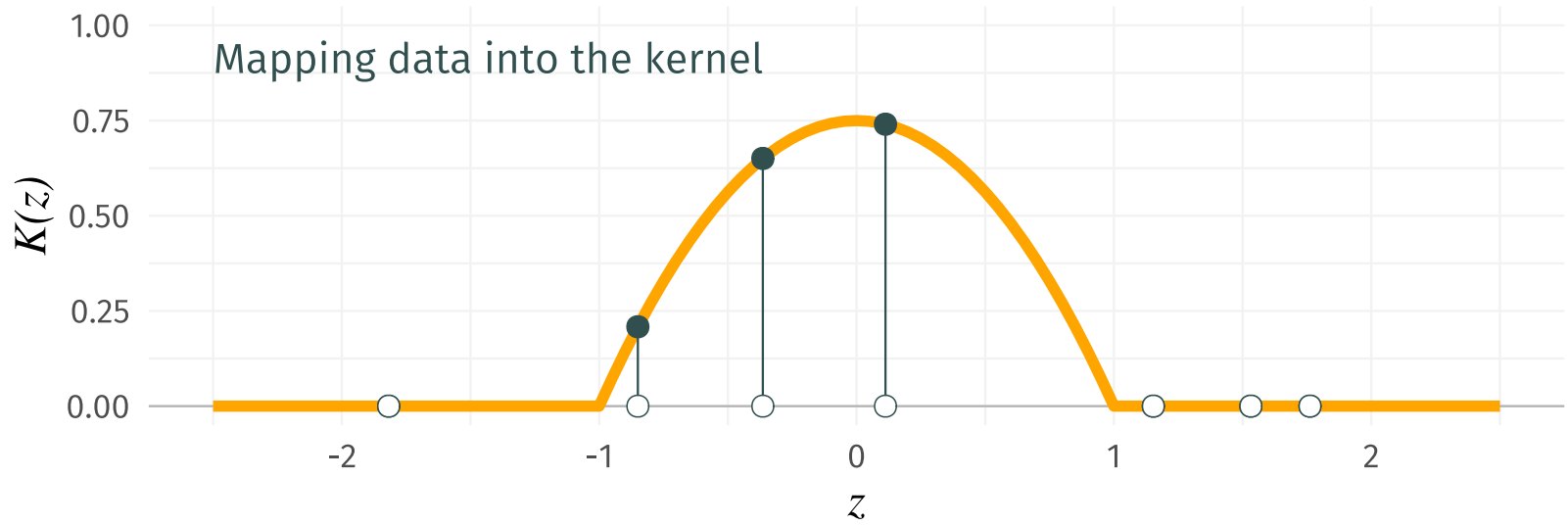
$$K(z) = \frac{3}{4}(1 - z^2) \times \mathbb{I}(|z| < 1)$$

The Epanechnikov kernel $K(z) = \frac{3}{4}(1 - z^2) \times \mathbb{I}(|z| < 1)$

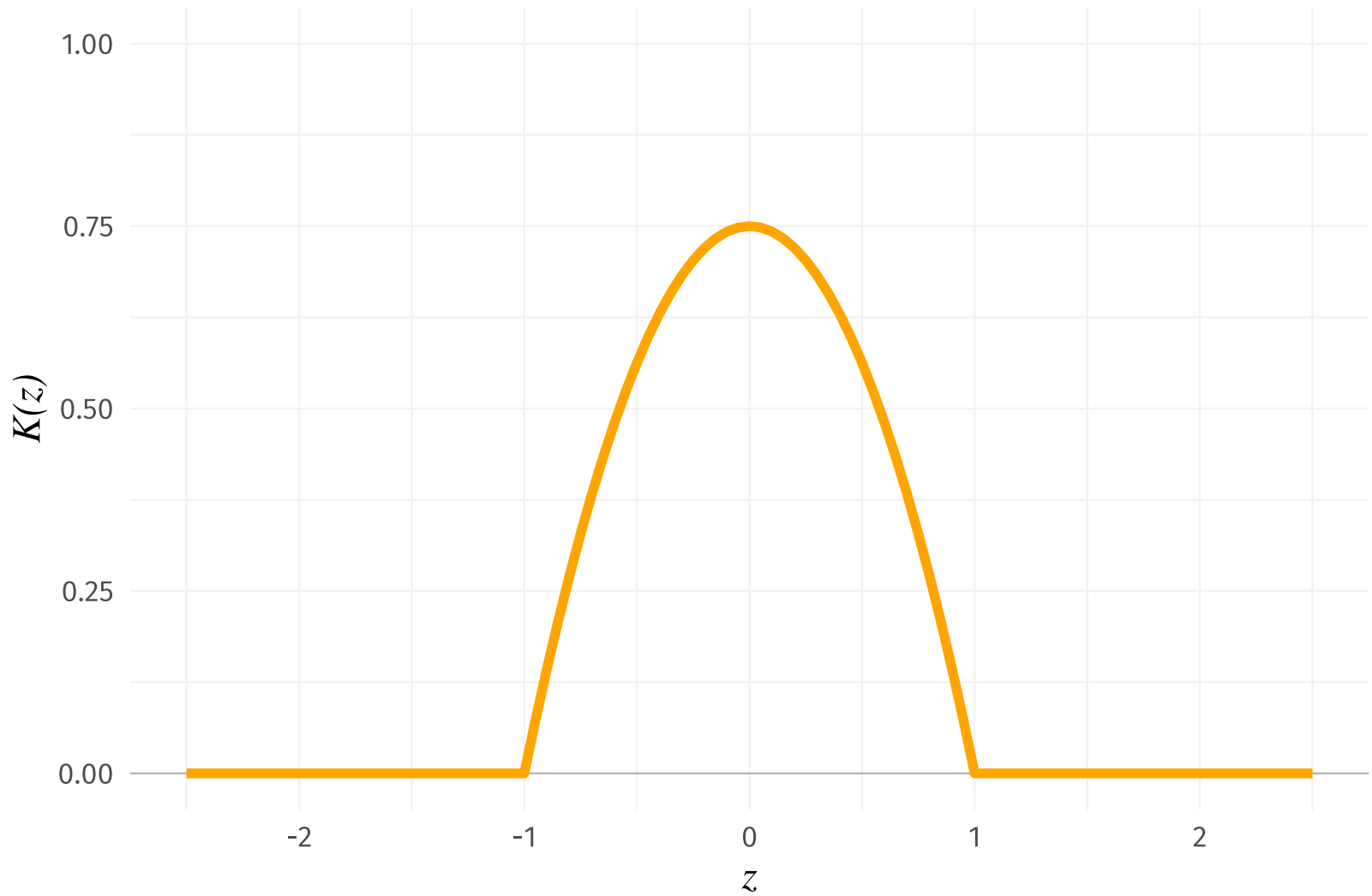




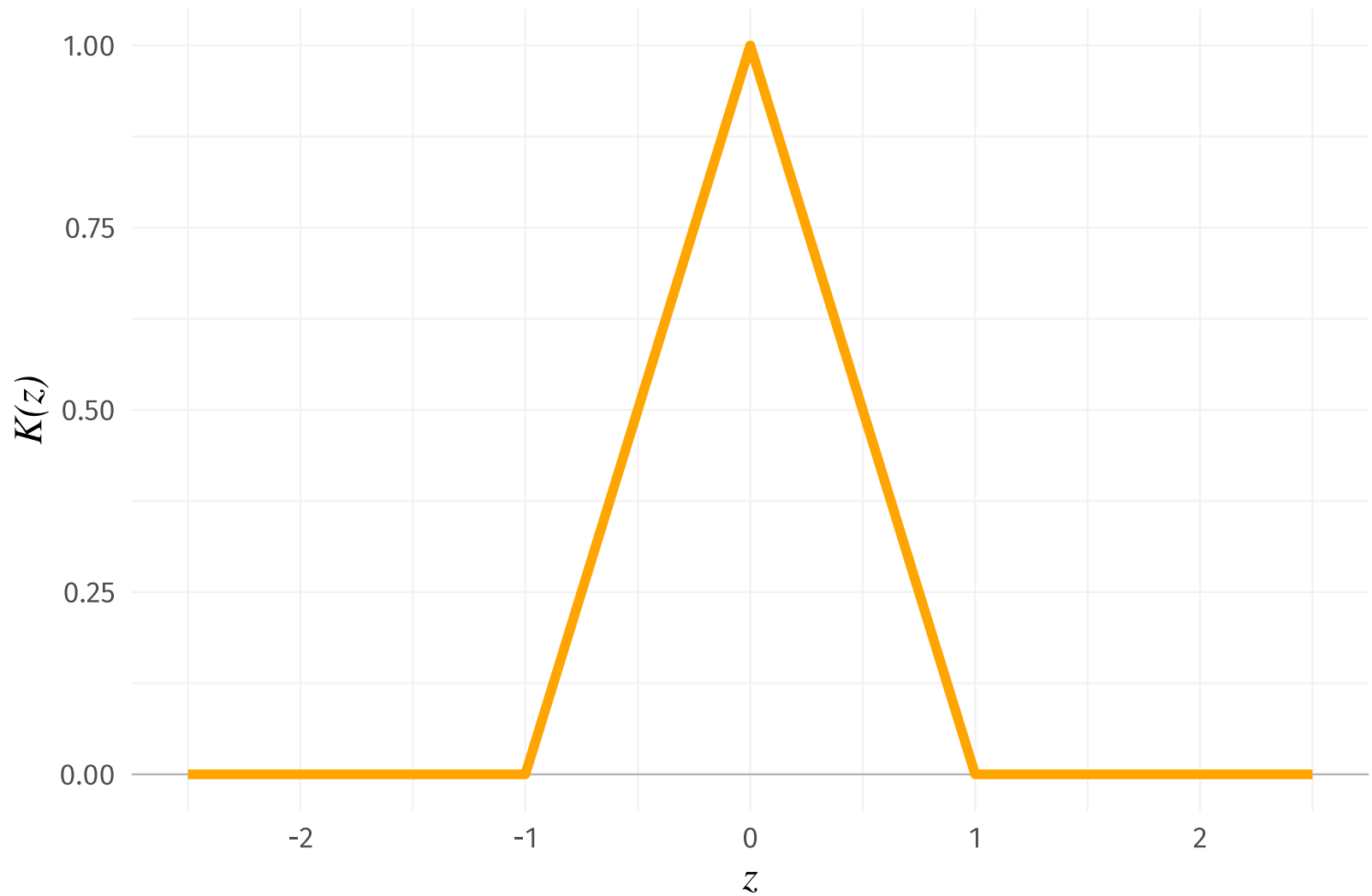




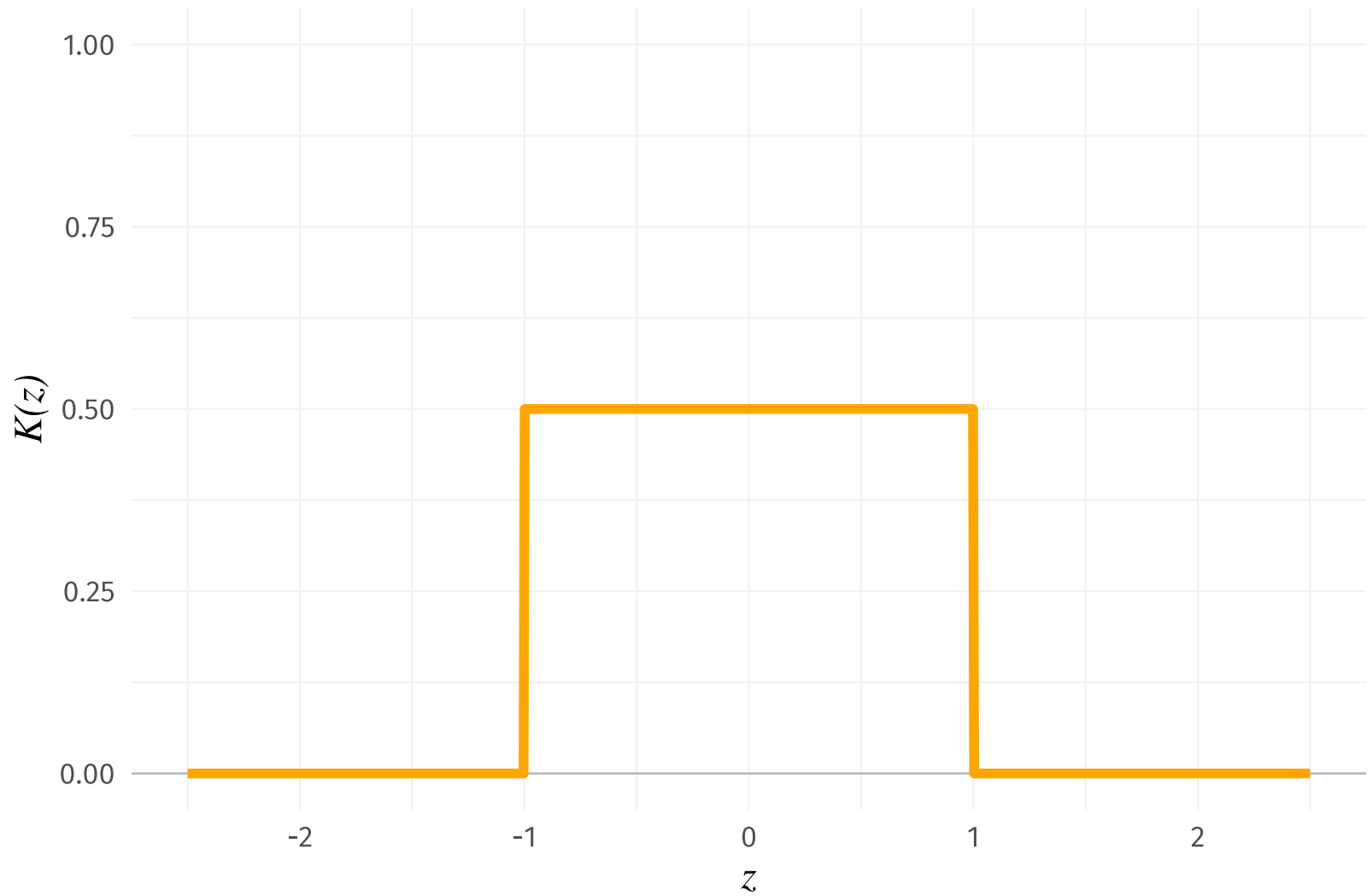
The Epanechnikov kernel $K(z) = \frac{3}{4}(1 - z^2) \times \mathbb{I}(|z| < 1)$



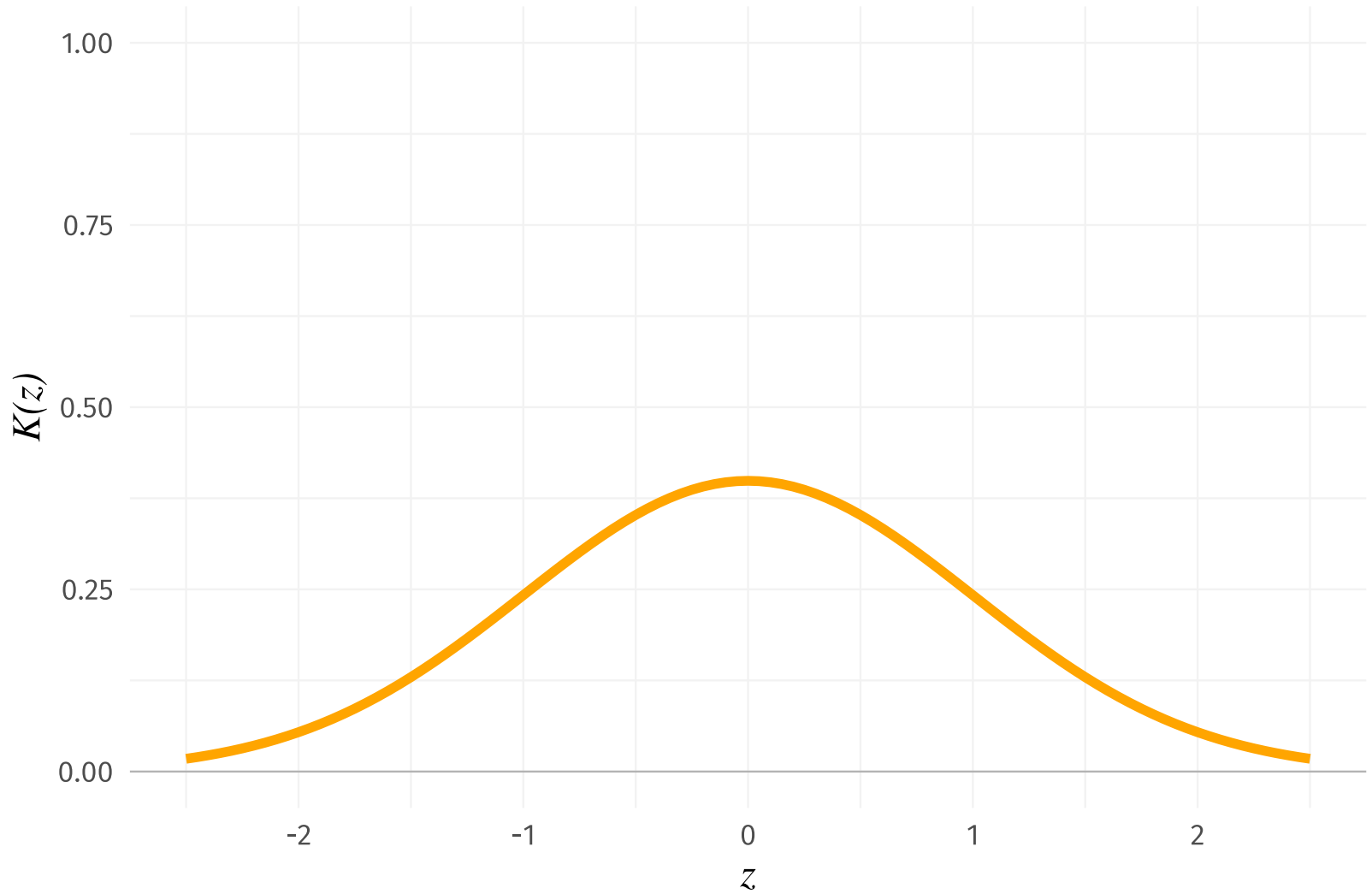
The Triangle kernel $K(z) = (1 - |z|) \times \mathbb{I}(|z| < 1)$



The Uniform kernel $K(z) = \frac{1}{2} \times \mathbb{I}(|z| < 1)$



The Gaussian kernel $K(z) = (2\pi)^{-1/2} \exp(-z^2/2)$



Kernels

Aside

Kernel functions are good for more than just matching.

You will most commonly see/use them smoothing out densities—providing a smooth, moving-window average.

Kernels

Aside

Kernel functions are good for more than just matching.

You will most commonly see/use them smoothing out densities—providing a smooth, moving-window average.

E.g., R's (`ggplot2`'s) smooth, density-plotting function `geom_density()`.

`geom_density()` defaults to `kernel = "gaussian"`, but you can specify many other kernel functions (including `"epanechnikov"`).

Kernels

Aside

Kernel functions are good for more than just matching.

You will most commonly see/use them smoothing out densities—providing a smooth, moving-window average.

E.g., R's (`ggplot2`'s) smooth, density-plotting function `geom_density()`.

`geom_density()` defaults to `kernel = "gaussian"`, but you can specify many other kernel functions (including `"epanechnikov"`).

You can also change the `bandwidth` argument. The default is a bandwidth-choosing function called `bw.nrd0()`.

Matching

Adding neighbors

As we add more neighbors—either moving from 1 to $n > 1$ or increasing our bandwidth—we potentially increase the efficiency of our estimator.

Matching

Adding neighbors

As we add more neighbors—either moving from 1 to $n > 1$ or increasing our bandwidth—we potentially increase the efficiency of our estimator.

We need to **be careful not to add too many controls** for each treated i .

Matching

Adding neighbors

As we add more neighbors—either moving from 1 to $n > 1$ or increasing our bandwidth—we potentially increase the efficiency of our estimator.

We need to **be careful not to add too many controls** for each treated i .

CIA requires that we're actually conditioning on the observables—it does not allow us to take a simple average across all control observations.

Matching

The curse of dimensionality[†]

It turns out kernel- and bandwidth-selection are not our biggest enemies.

[†] I'm not sure if this is a title for Harry Potter or Indiana Jones... crossover anyone?

Matching

The curse of dimensionality[†]

It turns out kernel- and bandwidth-selection are not our biggest enemies.

As the dimension of \mathbf{X} expands (matching on more variables), it becomes **harder and harder to find a nice, close control** for each treated unit.

[†] I'm not sure if this is a title for Harry Potter or Indiana Jones... crossover anyone?

Matching

The curse of dimensionality[†]

It turns out kernel- and bandwidth-selection are not our biggest enemies.

As the dimension of \mathbf{X} expands (matching on more variables), it becomes **harder and harder to find a nice, close control** for each treated unit.

We need a way to shrink the dimensionality of \mathbf{X} .

[†] I'm not sure if this is a title for Harry Potter or Indiana Jones... crossover anyone?

Propensity-score methods

Propensity-score methods

Setup

Let's begin with two assumptions—one old and one new.

1. **Conditional independence:** $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | X_i$
2. **Overlap:** $0 < \Pr(D_i = 1 | X_i) < 1$

Propensity-score methods

Setup

Let's begin with two assumptions—one old and one new.

1. **Conditional independence:** $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | X_i$
2. **Overlap:** $0 < \Pr(D_i = 1 | X_i) < 1$

We can estimate an average treatment effect by conditioning on X_i .

Propensity-score methods

Setup

Let's begin with two assumptions—one old and one new.

1. **Conditional independence:** $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | X_i$
2. **Overlap:** $0 < \Pr(D_i = 1 | X_i) < 1$

We can estimate an average treatment effect by conditioning on X_i .

However, overlap may fail if the dimensions of X are large and N is finite.

Propensity-score methods

Setup

Let's begin with two assumptions—one old and one new.

1. **Conditional independence:** $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | X_i$
2. **Overlap:** $0 < \Pr(D_i = 1 | X_i) < 1$

We can estimate an average treatment effect by conditioning on X_i .

However, overlap may fail if the dimensions of X are large and N is finite.

Propensity scores provide a solution to this mess.

Propensity-score methods

The magic

It turns out that if $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | X_i$, then we actually only need to match/condition on $p(X_i) = E[D_i | X_i]$.

Propensity-score methods

The magic

It turns out that if $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | X_i$, then we actually only need to match/condition on $p(X_i) = E[D_i | X_i]$.

$p(X_i)$ is the **propensity score**

Propensity-score methods

The magic

It turns out that if $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | X_i$, then we actually only need to match/condition on $p(X_i) = E[D_i | X_i]$.

$p(X_i)$ is the **propensity score**, the probability of treatment given X_i .

Propensity-score methods

The magic

It turns out that if $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | X_i$, then we actually only need to match/condition on $p(X_i) = E[D_i | X_i]$.

$p(X_i)$ is the **propensity score**, the probability of treatment given X_i .

Propensity-score theorem If $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | X_i$, then $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | p(X_i)$.

Propensity-score methods

The magic

It turns out that if $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | X_i$, then we actually only need to match/condition on $p(X_i) = E[D_i | X_i]$.

$p(X_i)$ is the **propensity score**, the probability of treatment given X_i .

Propensity-score theorem If $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | X_i$, then $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | p(X_i)$.

This theorem extends our CIA to a one-dimensional score, avoiding the curse of dimensionality.

Propensity-score methods

Theorem If $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | X_i$, then $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | p(X_i)$.

Proof

Propensity-score methods

Theorem If $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | X_i$, then $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | p(X_i)$.

Proof

To prove this theorem, we will show $\Pr(D_i = 1 \mid Y_{0i}, Y_{1i}, p(X_i)) = p(X_i)$,
i.e., D_i is independent of (Y_{0i}, Y_{1i}) after conditioning on $p(X_i)$.

Propensity-score methods

Theorem If $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | X_i$, then $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | p(X_i)$.

Proof

$$\Pr \left[D_i = 1 \mid Y_{0i}, Y_{1i}, p(X_i) \right]$$

Propensity-score methods

Theorem If $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | X_i$, then $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | p(X_i)$.

Proof

$$\begin{aligned} \Pr \left[D_i = 1 \mid Y_{0i}, Y_{1i}, p(X_i) \right] \\ = E \left[D_i \mid Y_{0i}, Y_{1i}, p(X_i) \right] \end{aligned}$$

Propensity-score methods

Theorem If $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | X_i$, then $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | p(X_i)$.

Proof

$$\begin{aligned} \Pr[D_i = 1 | Y_{0i}, Y_{1i}, p(X_i)] \\ &= E[D_i | Y_{0i}, Y_{1i}, p(X_i)] \\ &= E\left[E(D_i | Y_{0i}, Y_{1i}, p(X_i), X_i) | Y_{0i}, Y_{1i}, p(X_i)\right] \end{aligned}$$

Propensity-score methods

Theorem If $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | X_i$, then $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | p(X_i)$.

Proof

$$\begin{aligned} \Pr[D_i = 1 | Y_{0i}, Y_{1i}, p(X_i)] \\ &= E[D_i | Y_{0i}, Y_{1i}, p(X_i)] \\ &= E\left[E(D_i | Y_{0i}, Y_{1i}, p(X_i), X_i) | Y_{0i}, Y_{1i}, p(X_i)\right] \\ &= E\left[E(D_i | Y_{0i}, Y_{1i}, X_i) | Y_{0i}, Y_{1i}, p(X_i)\right] \end{aligned}$$

Propensity-score methods

Theorem If $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | X_i$, then $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | p(X_i)$.

Proof

$$\Pr\left[D_i = 1 \mid Y_{0i}, Y_{1i}, p(X_i)\right] = \dots = E\left[E\left(D_i \mid Y_{0i}, Y_{1i}, X_i\right) \mid Y_{0i}, Y_{1i}, p(X_i)\right]$$

Propensity-score methods

Theorem If $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | X_i$, then $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | p(X_i)$.

Proof

$$\begin{aligned}\Pr\left[D_i = 1 \mid Y_{0i}, Y_{1i}, p(X_i)\right] &= \dots = E\left[E\left(D_i \mid Y_{0i}, Y_{1i}, X_i\right) \mid Y_{0i}, Y_{1i}, p(X_i)\right] \\ &= E\left[E\left(D_i \mid X_i\right) \mid Y_{0i}, Y_{1i}, p(X_i)\right]\end{aligned}$$

Propensity-score methods

Theorem If $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | X_i$, then $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | p(X_i)$.

Proof

$$\begin{aligned}\Pr\left[D_i = 1 \mid Y_{0i}, Y_{1i}, p(X_i)\right] &= \dots = E\left[E\left(D_i \mid Y_{0i}, Y_{1i}, X_i\right) \mid Y_{0i}, Y_{1i}, p(X_i)\right] \\ &= E\left[E\left(D_i \mid X_i\right) \mid Y_{0i}, Y_{1i}, p(X_i)\right] \\ &= E\left[p(X_i) \mid Y_{0i}, Y_{1i}, p(X_i)\right]\end{aligned}$$

Propensity-score methods

Theorem If $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | X_i$, then $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | p(X_i)$.

Proof

$$\begin{aligned}\Pr\left[D_i = 1 \mid Y_{0i}, Y_{1i}, p(X_i)\right] &= \dots = E\left[E\left(D_i \mid Y_{0i}, Y_{1i}, X_i\right) \mid Y_{0i}, Y_{1i}, p(X_i)\right] \\ &= E\left[E\left(D_i \mid X_i\right) \mid Y_{0i}, Y_{1i}, p(X_i)\right] \\ &= E\left[p(X_i) \mid Y_{0i}, Y_{1i}, p(X_i)\right] \\ &= p(X_i)\end{aligned}$$

Propensity-score methods

Theorem If $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | X_i$, then $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | p(X_i)$.

Proof

$$\begin{aligned}\Pr\left[D_i = 1 \mid Y_{0i}, Y_{1i}, p(X_i)\right] &= \dots = E\left[E\left(D_i \mid Y_{0i}, Y_{1i}, X_i\right) \mid Y_{0i}, Y_{1i}, p(X_i)\right] \\ &= E\left[E\left(D_i \mid X_i\right) \mid Y_{0i}, Y_{1i}, p(X_i)\right] \\ &= E\left[p(X_i) \mid Y_{0i}, Y_{1i}, p(X_i)\right] \\ &= p(X_i)\end{aligned}$$

$$\therefore (Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | X_i \implies (Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i | p(X_i) \quad \checkmark$$

Propensity-score methods

Intuition

Q What's going on here?

\mathbf{X}_i carries way more information than $p(\mathbf{X}_i)$, so how can we still get conditional independence of treatment by only conditioning on $p(\mathbf{X}_i)$?

Propensity-score methods

Intuition

Q What's going on here?

\mathbf{X}_i carries way more information than $p(\mathbf{X}_i)$, so how can we still get conditional independence of treatment by only conditioning on $p(\mathbf{X}_i)$?

A₁ Conditional independence of treatment isn't about extracting all of the information possible from \mathbf{X}_i . We actually only care about creating a situation in which \mathbf{D}_i | something is independent of $(\mathbf{Y}_{0i}, \mathbf{Y}_{1i})$.

Propensity-score methods

Intuition

Q What's going on here?

\mathbf{X}_i carries way more information than $p(\mathbf{X}_i)$, so how can we still get conditional independence of treatment by only conditioning on $p(\mathbf{X}_i)$?

A₁ Conditional independence of treatment isn't about extracting all of the information possible from \mathbf{X}_i . We actually only care about creating a situation in which \mathbf{D}_i | something is independent of $(\mathbf{Y}_{0i}, \mathbf{Y}_{1i})$.

A₂ Back to our main concern: **selection bias**. People select into treatment. If \mathbf{X} says two people were equally likely to be treated, and if \mathbf{X}_i explains all of selection (CIA), then there cannot be selection between these two people.

Propensity-score methods

Estimation

So where do propensity scores come from?

Propensity-score methods

Estimation

So where do propensity scores come from?

We estimate them—and there are a lot of ways to do that.

1. Flexible (*i.e.*, interactions) logit specification
2. Kernel regression (remember kernel functions?)
3. Many others—machine learning, series-logit estimator, *etc.*

Propensity-score methods

Estimation

So where do propensity scores come from?

We estimate them—and there are a lot of ways to do that.

1. Flexible (*i.e.*, interactions) logit specification
2. Kernel regression (remember kernel functions?)
3. Many others—machine learning, series-logit estimator, *etc.*

Q Can we just use plain OLS (linear probability model)?

Propensity-score methods

Estimation

So where do propensity scores come from?

We estimate them—and there are a lot of ways to do that.

1. Flexible (*i.e.*, interactions) logit specification
2. Kernel regression (remember kernel functions?)
3. Many others—machine learning, series-logit estimator, *etc.*

Q Can we just use plain OLS (linear probability model)?

A Sort of. Think about FWL. This route is going to be the same as a regression conditioning on \mathbf{X}_i .

Propensity-score methods

Estimation

From *MHE* (p. 83)

Question

A big question here is how to best model and estimate $p(\mathbf{X}_i)$...

Answer

The answer to this is inherently application-specific. A growing empirical literature suggests that a logit model for the propensity score with a few polynomial terms in continuous covariates works well in practice...

Propensity-score methods

Application

So you have some estimated propensity scores $\hat{p}(\mathbf{X}_i)$. What next?

Propensity-score methods

Application

So you have some estimated propensity scores $\hat{p}(\mathbf{X}_i)$. What next?

Option 1 Conditioning via regression

Propensity-score methods

Application

So you have some estimated propensity scores $\hat{p}(\mathbf{X}_i)$. What next?

Option 1 Conditioning via regression

Option 1a Use a **regression to condition** on $p(\mathbf{X}_i)$, i.e.,

$$Y_i = \alpha + \delta D_i + \beta p(\mathbf{X}_i) + u_i \quad (1a)$$

Propensity-score methods

Application

So you have some estimated propensity scores $\hat{p}(\mathbf{X}_i)$. What next?

Option 1 Conditioning via regression

Option 1a Use a **regression to condition** on $p(\mathbf{X}_i)$, i.e.,

$$Y_i = \alpha + \delta D_i + \beta p(\mathbf{X}_i) + u_i \quad (1a)$$

Option 1b If we think treatment effects are heterogeneous and may covary with \mathbf{X} , then we might want to also **interact** treatment with $p(\mathbf{X}_i)$, i.e.,

$$Y_i = \alpha + \delta_1 D_i + \delta_2 D_i p(\mathbf{X}_i) + \beta p(\mathbf{X}_i) + u_i \quad (1b)$$

Propensity-score methods

Heterogeneity with regression

Let's think a bit more about heterogeneous treatment effects in this setting.

$$\begin{aligned}Y_{0i} &= \alpha + \beta X_i + u_i \\Y_{1i} &= Y_{0i} + \delta_1 + \delta_2 X_i\end{aligned}$$

i.e., the treatment effect depends upon X_i .

Propensity-score methods

Heterogeneity with regression

Let's think a bit more about heterogeneous treatment effects in this setting.

$$\begin{aligned}Y_{0i} &= \alpha + \beta X_i + u_i \\Y_{1i} &= Y_{0i} + \delta_1 + \delta_2 X_i\end{aligned}$$

i.e., the treatment effect depends upon X_i .

$$Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$$

Propensity-score methods

Heterogeneity with regression

Let's think a bit more about heterogeneous treatment effects in this setting.

$$\begin{aligned}Y_{0i} &= \alpha + \beta X_i + u_i \\Y_{1i} &= Y_{0i} + \delta_1 + \delta_2 X_i\end{aligned}$$

i.e., the treatment effect depends upon X_i .

$$\begin{aligned}Y_i &= D_i Y_{1i} + (1 - D_i) Y_{0i} \\&= D_i \left(Y_{0i} + \delta_1 + \delta_2 X_i \right) + (1 - D_i) Y_{0i}\end{aligned}$$

Propensity-score methods

Heterogeneity with regression

Let's think a bit more about heterogeneous treatment effects in this setting.

$$\begin{aligned}Y_{0i} &= \alpha + \beta X_i + u_i \\Y_{1i} &= Y_{0i} + \delta_1 + \delta_2 X_i\end{aligned}$$

i.e., the treatment effect depends upon X_i .

$$\begin{aligned}Y_i &= D_i Y_{1i} + (1 - D_i) Y_{0i} \\&= D_i \left(Y_{0i} + \delta_1 + \delta_2 X_i \right) + (1 - D_i) Y_{0i} \\&= Y_{0i} + \delta_1 D_i + \delta_2 D_i X_i\end{aligned}$$

Propensity-score methods

Heterogeneity with regression

Let's think a bit more about heterogeneous treatment effects in this setting.

$$\begin{aligned}Y_{0i} &= \alpha + \beta X_i + u_i \\Y_{1i} &= Y_{0i} + \delta_1 + \delta_2 X_i\end{aligned}$$

i.e., the treatment effect depends upon X_i .

$$\begin{aligned}Y_i &= D_i Y_{1i} + (1 - D_i) Y_{0i} \\&= D_i \left(Y_{0i} + \delta_1 + \delta_2 X_i \right) + (1 - D_i) Y_{0i} \\&= Y_{0i} + \delta_1 D_i + \delta_2 D_i X_i \\&= \alpha + \delta_1 D_i + \delta_2 D_i X_i + \beta X_i + u_i\end{aligned}$$

Propensity-score methods

Heterogeneity

This final equation

$$Y_i = \alpha + \delta_1 D_i + \delta_2 D_i X_i + \beta X_i + u_i$$

Propensity-score methods

Heterogeneity

This final equation

$$Y_i = \alpha + \delta_1 D_i + \delta_2 D_i X_i + \beta X_i + u_i$$

suggests that we want $p(X_i)$ and $D_i p(X_i)$, i.e.,

$$Y_i = \alpha + \delta_1 D_i + \delta_2 D_i p(X_i) + \beta p(X_i) + u_i \quad (1b)$$

Propensity-score methods

Heterogeneity

This final equation

$$Y_i = \alpha + \delta_1 D_i + \delta_2 D_i X_i + \beta X_i + u_i$$

suggests that we want $p(X_i)$ and $D_i p(X_i)$, i.e.,

$$Y_i = \alpha + \delta_1 D_i + \delta_2 D_i p(X_i) + \beta p(X_i) + u_i \quad (1b)$$

which yields

1. a **group-specific treatment effect** $\delta_1 + \delta_2 X_i$ for each X_i
2. an **average treatment effect** $\delta_1 + \delta_2 \bar{p}(X_i)$

Propensity-score methods

More flexibility

We motivated propensity scores with a desire to reduce dimensionality and estimate/choose/assume fewer parameters.

Adding $p(\mathbf{X}_i)$ and $\mathbf{D}_i p(\mathbf{X}_i)$ as covariates in a linear regression doesn't quite exhaust our potential for flexible/nonparametric estimation.

Propensity-score methods

Blocking

Option 2 Block (stratify) on propensity scores.

Propensity-score methods

Blocking

Option 2 Block (stratify) on propensity scores.

1. Divide the range of $\hat{p}(\mathbf{X}_i)$ into K blocks (e.g., 0.05-wide blocks).
2. Place each observation into a block via its $\hat{p}(\mathbf{X}_i)$.
3. Calculate $\hat{\tau}_k$ for each block via difference in means.
4. Average the $\hat{\tau}_k$ using their shares of the sample, i.e.,

$$\hat{\tau}_{\text{Block}} = \sum_{k=1}^K \hat{\tau}_k \frac{N_{1k} + N_{0k}}{N}$$

Propensity-score methods

Blocking

Option 2 Block (stratify) on propensity scores.

1. Divide the range of $\hat{p}(\mathbf{X}_i)$ into K blocks (e.g., 0.05-wide blocks).
2. Place each observation into a block via its $\hat{p}(\mathbf{X}_i)$.
3. Calculate $\hat{\tau}_k$ for each block via difference in means.
4. Average the $\hat{\tau}_k$ using their shares of the sample, i.e.,

$$\hat{\tau}_{\text{Block}} = \sum_{k=1}^K \hat{\tau}_k \frac{N_{1k} + N_{0k}}{N}$$

Note Blocking is similar to NN/kernel matching using $p(\mathbf{X}_i)$ as distance.

Propensity-score methods

Choosing blocks

Blocking on propensity scores requires defining defining blocks.

One common route involves some iteration.

1. **Choose blocks.**
2. Check the **balance of the covariates** within each block.[†]
 - If covariates are **not balanced**, then split your blocks and repeat.
 - If covariates are **balanced**, then stop.

[†] Keep multiple-hypothesis testing in mind. With many covariates and many blocks, you are bound to find statistically significant relationships—even if you are balanced in truth.

Propensity-score methods

Overlap

Blocking emphasizes our overlap assumption, *i.e.*, $0 < \Pr(\mathbf{D}_i | \mathbf{X}_i) < 1$.

If a block contains zero treated/control units, we cannot calculate $\hat{\tau}_k$.

Propensity-score methods

Overlap

Blocking emphasizes our overlap assumption, *i.e.*, $0 < \Pr(\mathbf{D}_i | \mathbf{X}_i) < 1$.

If a block contains zero treated/control units, we cannot calculate $\hat{\tau}_k$.

Caution Logit can hide violations—it forces $0 < \hat{p}(\mathbf{X}_i) < 1$.

Propensity-score methods

Overlap

Blocking emphasizes our overlap assumption, *i.e.*, $0 < \Pr(\mathbf{D}_i | \mathbf{X}_i) < 1$.

If a block contains zero treated/control units, we cannot calculate $\hat{\tau}_k$.

Caution Logit can hide violations—it forces $0 < \hat{p}(\mathbf{X}_i) < 1$.

Common practice Empirically enforce overlap:

- Drop control units with $\hat{p}(\mathbf{X}_i)$ below the minimum propensity score in the treatment group.
- Drop treated units with $\hat{p}(\mathbf{X}_i)$ above the maximum propensity score in the control group.

Propensity-score methods

Weighting

Option 3 Weight observations by the inverse propensity score.

Propensity-score methods

Weighting

Option 3 Weight observations by the inverse propensity score.

Q How does weighting by $1/\hat{p}(\mathbf{X}_i)$ make sense?

Propensity-score methods

Weighting

Option 3 Weight observations by the inverse propensity score.

Q How does weighting by $1/\hat{p}(\mathbf{X}_i)$ make sense?

A Consider our old (likely biased) friend the difference in means, *i.e.*,

$$\hat{\tau}_{\text{Diff}} = \bar{Y}_T - \bar{Y}_C = \frac{\sum_i D_i Y_i}{\sum_i D_i} - \frac{\sum_i (1 - D_i) Y_i}{\sum_i (1 - D_i)}$$

Propensity-score methods

Weighting

Option 3 Weight observations by the inverse propensity score.

Q How does weighting by $1/\hat{p}(\mathbf{X}_i)$ make sense?

A Consider our old (likely biased) friend the difference in means, *i.e.*,

$$\hat{\tau}_{\text{Diff}} = \bar{Y}_T - \bar{Y}_C = \frac{\sum_i D_i Y_i}{\sum_i D_i} - \frac{\sum_i (1 - D_i) Y_i}{\sum_i (1 - D_i)}$$

which we've discussed is biased due to selection into treatment, *i.e.*,

$$E[Y_{0i} | D_i = 1] \neq E[Y_{0i}]$$

Propensity-score methods

Weighting, justified

Suppose we know $p(\mathbf{X}_i)$ and we weight each **treated** individual by $1/p(\mathbf{X}_i)$

Propensity-score methods

Weighting, justified

Suppose we know $p(\mathbf{X}_i)$ and we weight each **treated** individual by $1/p(\mathbf{X}_i)$

$$E \left[\frac{D_i Y_i}{p(\mathbf{X}_i)} \right]$$

Propensity-score methods

Weighting, justified

Suppose we know $p(\mathbf{X}_i)$ and we weight each **treated** individual by $1/p(\mathbf{X}_i)$

$$E\left[\frac{D_i Y_i}{p(\mathbf{X}_i)}\right] = E\left[\frac{D_i (D_i Y_{1i} + (1 - D_i) Y_{0i})}{p(\mathbf{X}_i)}\right]$$

Propensity-score methods

Weighting, justified

Suppose we know $p(\mathbf{X}_i)$ and we weight each **treated** individual by $1/p(\mathbf{X}_i)$

$$E\left[\frac{D_i Y_i}{p(\mathbf{X}_i)}\right] = E\left[\frac{D_i (D_i Y_{1i} + (1 - D_i) Y_{0i})}{p(\mathbf{X}_i)}\right] = E\left[\frac{D_i Y_{1i}}{p(\mathbf{X}_i)}\right]$$

Propensity-score methods

Weighting, justified

Suppose we know $p(\mathbf{X}_i)$ and we weight each **treated** individual by $1/p(\mathbf{X}_i)$

$$\begin{aligned} E\left[\frac{D_i Y_i}{p(\mathbf{X}_i)}\right] &= E\left[\frac{D_i (D_i Y_{1i} + (1 - D_i) Y_{0i})}{p(\mathbf{X}_i)}\right] = E\left[\frac{D_i Y_{1i}}{p(\mathbf{X}_i)}\right] \\ &= E\left(E\left[\frac{D_i Y_{1i}}{p(\mathbf{X}_i)} \mid \mathbf{X}_i\right]\right) \end{aligned}$$

Propensity-score methods

Weighting, justified

Suppose we know $p(\mathbf{X}_i)$ and we weight each **treated** individual by $1/p(\mathbf{X}_i)$

$$\begin{aligned} E\left[\frac{D_i Y_i}{p(\mathbf{X}_i)}\right] &= E\left[\frac{D_i (D_i Y_{1i} + (1 - D_i) Y_{0i})}{p(\mathbf{X}_i)}\right] = E\left[\frac{D_i Y_{1i}}{p(\mathbf{X}_i)}\right] \\ &= E\left(E\left[\frac{D_i Y_{1i}}{p(\mathbf{X}_i)} \mid \mathbf{X}_i\right]\right) = E\left(\frac{E[D_i \mid \mathbf{X}_i] E[Y_{1i} \mid \mathbf{X}_i]}{p(\mathbf{X}_i)}\right) \end{aligned}$$

Propensity-score methods

Weighting, justified

Suppose we know $p(\mathbf{X}_i)$ and we weight each **treated** individual by $1/p(\mathbf{X}_i)$

$$E\left[\frac{D_i Y_i}{p(\mathbf{X}_i)}\right] = E\left[\frac{D_i (D_i Y_{1i} + (1 - D_i) Y_{0i})}{p(\mathbf{X}_i)}\right] = E\left[\frac{D_i Y_{1i}}{p(\mathbf{X}_i)}\right]$$

$$= E\left(E\left[\frac{D_i Y_{1i}}{p(\mathbf{X}_i)} \mid \mathbf{X}_i\right]\right) = E\left(\frac{E[D_i \mid \mathbf{X}_i] E[Y_{1i} \mid \mathbf{X}_i]}{p(\mathbf{X}_i)}\right)$$

$$= E\left(\frac{p(\mathbf{X}_i) E[Y_{1i} \mid \mathbf{X}_i]}{p(\mathbf{X}_i)}\right)$$

Propensity-score methods

Weighting, justified

Suppose we know $p(\mathbf{X}_i)$ and we weight each **treated** individual by $1/p(\mathbf{X}_i)$

$$E\left[\frac{D_i Y_i}{p(\mathbf{X}_i)}\right] = E\left[\frac{D_i (D_i Y_{1i} + (1 - D_i) Y_{0i})}{p(\mathbf{X}_i)}\right] = E\left[\frac{D_i Y_{1i}}{p(\mathbf{X}_i)}\right]$$

$$= E\left(E\left[\frac{D_i Y_{1i}}{p(\mathbf{X}_i)} \mid \mathbf{X}_i\right]\right) = E\left(\frac{E[D_i \mid \mathbf{X}_i] E[Y_{1i} \mid \mathbf{X}_i]}{p(\mathbf{X}_i)}\right)$$

$$= E\left(\frac{p(\mathbf{X}_i) E[Y_{1i} \mid \mathbf{X}_i]}{p(\mathbf{X}_i)}\right) = E\left(E[Y_{1i} \mid \mathbf{X}_i]\right)$$

Propensity-score methods

Weighting, justified

Suppose we know $p(\mathbf{X}_i)$ and we weight each **treated** individual by $1/p(\mathbf{X}_i)$

$$\begin{aligned} E\left[\frac{D_i Y_i}{p(\mathbf{X}_i)}\right] &= E\left[\frac{D_i (D_i Y_{1i} + (1 - D_i) Y_{0i})}{p(\mathbf{X}_i)}\right] = E\left[\frac{D_i Y_{1i}}{p(\mathbf{X}_i)}\right] \\ &= E\left(E\left[\frac{D_i Y_{1i}}{p(\mathbf{X}_i)} \mid \mathbf{X}_i\right]\right) = E\left(\frac{E[D_i \mid \mathbf{X}_i] E[Y_{1i} \mid \mathbf{X}_i]}{p(\mathbf{X}_i)}\right) \\ &= E\left(\frac{p(\mathbf{X}_i) E[Y_{1i} \mid \mathbf{X}_i]}{p(\mathbf{X}_i)}\right) = E\left(E[Y_{1i} \mid \mathbf{X}_i]\right) = E[Y_{1i}] \end{aligned}$$

Propensity-score methods

Weighting, justified

Suppose we know $p(\mathbf{X}_i)$ and we weight each **treated** individual by $1/p(\mathbf{X}_i)$

$$\begin{aligned} E\left[\frac{D_i Y_i}{p(\mathbf{X}_i)}\right] &= E\left[\frac{D_i (D_i Y_{1i} + (1 - D_i) Y_{0i})}{p(\mathbf{X}_i)}\right] = E\left[\frac{D_i Y_{1i}}{p(\mathbf{X}_i)}\right] \\ &= E\left(E\left[\frac{D_i Y_{1i}}{p(\mathbf{X}_i)} \mid \mathbf{X}_i\right]\right) = E\left(\frac{E[D_i \mid \mathbf{X}_i] E[Y_{1i} \mid \mathbf{X}_i]}{p(\mathbf{X}_i)}\right) \\ &= E\left(\frac{p(\mathbf{X}_i) E[Y_{1i} \mid \mathbf{X}_i]}{p(\mathbf{X}_i)}\right) = E\left(E[Y_{1i} \mid \mathbf{X}_i]\right) = E[Y_{1i}] \end{aligned}$$

Similarly, weighting **control** individuals by $1/(1 - p(\mathbf{X}_i))$ yields

$$E\left[\frac{(1 - D_i) Y_i}{1 - p(\mathbf{X}_i)}\right] = E[Y_{0i}]$$

Propensity-score methods

Weighting: The estimator

Thus, we can estimate an unbiased treatment effect via

$$\hat{\tau}_{p\text{Weight}} = \frac{1}{N} \sum_{i=1}^N \left[\frac{D_i Y_i}{p(\mathbf{X}_i)} - \frac{(1 - D_i) Y_i}{1 - p(\mathbf{X}_i)} \right]$$

Propensity-score methods

Weighting: The estimator

Thus, we can estimate an unbiased treatment effect via

$$\hat{\tau}_{p\text{Weight}} = \frac{1}{N} \sum_{i=1}^N \left[\frac{D_i Y_i}{p(\mathbf{X}_i)} - \frac{(1 - D_i) Y_i}{1 - p(\mathbf{X}_i)} \right]$$

Intuition We're trying to overcome selection bias, *i.e.*, treated individuals were more likely to be treated as a function of \mathbf{X}_i —producing higher $p(\mathbf{X}_i)$.

Propensity-score methods

Weighting: The estimator

Thus, we can estimate an unbiased treatment effect via

$$\hat{\tau}_{p\text{Weight}} = \frac{1}{N} \sum_{i=1}^N \left[\frac{D_i Y_i}{p(\mathbf{X}_i)} - \frac{(1 - D_i) Y_i}{1 - p(\mathbf{X}_i)} \right]$$

Intuition We're trying to overcome selection bias, *i.e.*, treated individuals were more likely to be treated as a function of \mathbf{X}_i —producing higher $p(\mathbf{X}_i)$.

We want to get back to *as-good-as random* variation in treatment.

So we upweight **(1) treated** individuals with low $p(\mathbf{X}_i)$ and **(2) control** observations with high $p(\mathbf{X}_i)$.

Propensity-score methods

Weighting: The example

Suppose for some individual i , $p(\mathbf{X}_i) = 0.80$.

Propensity-score methods

Weighting: The example

Suppose for some individual i , $p(\mathbf{X}_i) = 0.80$.

This propensity score says someone with this set of \mathbf{X}_i was four-times more likely to be **treated** than **control**.

Propensity-score methods

Weighting: The example

Suppose for some individual i , $p(\mathbf{X}_i) = 0.80$.

This propensity score says someone with this set of \mathbf{X}_i was four-times more likely to be **treated** than **control**.

Our weights fix this imbalance for each \mathbf{X}_i .

Propensity-score methods

Weighting: The example

Suppose for some individual i , $p(\mathbf{X}_i) = 0.80$.

This propensity score says someone with this set of \mathbf{X}_i was four-times more likely to be **treated** than **control**.

Our weights fix this imbalance for each \mathbf{X}_i .

- If i is **treated**, then her weight is $1/p(\mathbf{X}_i) = 1/0.80 = 1.25$

Propensity-score methods

Weighting: The example

Suppose for some individual i , $p(\mathbf{X}_i) = 0.80$.

This propensity score says someone with this set of \mathbf{X}_i was four-times more likely to be **treated** than **control**.

Our weights fix this imbalance for each \mathbf{X}_i .

- If i is **treated**, then her weight is $1/p(\mathbf{X}_i) = 1/0.80 = 1.25$
- If i is **control**, then her weight is $1/(1 - p(\mathbf{X}_i)) = 1/(1 - 0.80) = 5$

Propensity-score methods

Weighting: The example

Suppose for some individual i , $p(\mathbf{X}_i) = 0.80$.

This propensity score says someone with this set of \mathbf{X}_i was four-times more likely to be **treated** than **control**.

Our weights fix this imbalance for each \mathbf{X}_i .

- If i is **treated**, then her weight is $1/p(\mathbf{X}_i) = 1/0.80 = 1.25$
- If i is **control**, then her weight is $1/(1 - p(\mathbf{X}_i)) = 1/(1 - 0.80) = 5$

And guess what $5/1.25$ is...

Propensity-score methods

Weighting: The example

Suppose for some individual i , $p(\mathbf{X}_i) = 0.80$.

This propensity score says someone with this set of \mathbf{X}_i was four-times more likely to be **treated** than **control**.

Our weights fix this imbalance for each \mathbf{X}_i .

- If i is **treated**, then her weight is $1/p(\mathbf{X}_i) = 1/0.80 = 1.25$
- If i is **control**, then her weight is $1/(1 - p(\mathbf{X}_i)) = 1/(1 - 0.80) = 5$

And guess what $5/1.25$ is... 4!

Propensity-score methods

Weighting: The example

Suppose for some individual i , $p(\mathbf{X}_i) = 0.80$.

This propensity score says someone with this set of \mathbf{X}_i was four-times more likely to be **treated** than **control**.

Our weights fix this imbalance for each \mathbf{X}_i .

- If i is **treated**, then her weight is $1/p(\mathbf{X}_i) = 1/0.80 = 1.25$
- If i is **control**, then her weight is $1/(1 - p(\mathbf{X}_i)) = 1/(1 - 0.80) = 5$

And guess what $5/1.25$ is... 4! This weighting scheme gets us back to equal representation for each set of \mathbf{X}_i .

Propensity-score methods

Weighting: Last issue

Practical issue Nothing guarantees $\sum_i \hat{p}(\mathbf{X}_i) = 1$.

Propensity-score methods

Weighting: Last issue

Practical issue Nothing guarantees $\sum_i \hat{p}(\mathbf{X}_i) = 1$.

Solution Normalize weights by their total sum.

Propensity-score methods

Weighting: Last issue

Practical issue Nothing guarantees $\sum_i \hat{p}(\mathbf{X}_i) = 1$.

Solution Normalize weights by their total sum.

Applying the normalized (and estimated) propensity scores

$$\hat{\tau}_{p\text{Weight}} = \sum_{i=1}^N \frac{\frac{D_i Y_i}{\hat{p}(\mathbf{X}_i)}}{\sum_i \frac{D_i}{\hat{p}(\mathbf{X}_i)}} - \sum_{i=1}^N \frac{\frac{(1 - D_i) Y_i}{1 - \hat{p}(\mathbf{X}_i)}}{\sum_i \frac{(1 - D_i)}{1 - \hat{p}(\mathbf{X}_i)}}$$

Propensity-score methods

Weighting: Last issue

Practical issue Nothing guarantees $\sum_i \hat{p}(\mathbf{X}_i) = 1$.

Solution Normalize weights by their total sum.

Applying the normalized (and estimated) propensity scores

$$\hat{\tau}_{p\text{Weight}} = \sum_{i=1}^N \frac{\frac{D_i Y_i}{\hat{p}(\mathbf{X}_i)}}{\sum_i \frac{D_i}{\hat{p}(\mathbf{X}_i)}} - \sum_{i=1}^N \frac{\frac{(1 - D_i) Y_i}{1 - \hat{p}(\mathbf{X}_i)}}{\sum_i \frac{(1 - D_i)}{1 - \hat{p}(\mathbf{X}_i)}}$$

Hirano, Imbens, and Ridder (2003) suggests this estimator is efficient.

Propensity-score methods

Why choose one?

There's nothing special about weighted averages—regression can weight.

Thus, a **regression-based estimate**

$$Y_i = \alpha + X_i\beta + \tau D_i + u_i$$

Propensity-score methods

Why choose one?

There's nothing special about weighted averages—regression can weight.

Thus, a **regression-based estimate**

$$Y_i = \alpha + X_i\beta + \tau D_i + u_i$$

with **weights**

$$w_i = \sqrt{\frac{D_i}{\hat{p}(X_i)} + \frac{(1 - D_i)}{1 - \hat{p}(X_i)}}$$

Propensity-score methods

Why choose one?

There's nothing special about weighted averages—regression can weight.

Thus, a **regression-based estimate**

$$Y_i = \alpha + X_i\beta + \tau D_i + u_i$$

with **weights**

$$w_i = \sqrt{\frac{D_i}{\hat{p}(X_i)} + \frac{(1 - D_i)}{1 - \hat{p}(X_i)}}$$

offers a *doubly robust* property—you have two chances to be right: $p(X_i)$ or the regression specification.

Propensity-score methods

Why choose one? Part two

An alternative, doubly robust method combines propensity-score blocking with regression.

Propensity-score methods

Why choose one? Part two

An alternative, doubly robust method combines propensity-score blocking with regression.

Step 1 For each block k , we run the regression

$$Y_i = \alpha_k + X_i\beta_k + \tau_k D_i + u_i$$

Propensity-score methods

Why choose one? Part two

An alternative, doubly robust method combines propensity-score blocking with regression.

Step 1 For each block k , we run the regression

$$Y_i = \alpha_k + \mathbf{X}_i\beta_k + \tau_k\mathbf{D}_i + u_i$$

Step 2 Aggregate block-level treatment-effect estimates

$$\hat{\tau} = \sum_{k=1}^K \hat{\tau}_k \frac{N_{1k} + N_{0k}}{N}$$

Propensity-score methods

Major requirements

Don't get (too) caught up in the bells and whistles.

We still have two **major** requirements for any of these methods to work?

Propensity-score methods

Major requirements

Don't get (too) caught up in the bells and whistles.

We still have two **major** requirements for any of these methods to work?

1. Is the **conditional-independence assumption** true?

Propensity-score methods

Major requirements

Don't get (too) caught up in the bells and whistles.

We still have two **major** requirements for any of these methods to work?

1. Is the **conditional-independence assumption** true?
2. Do we have **overlap** between treatment and control units.

Propensity-score methods

Major requirements

Don't get (too) caught up in the bells and whistles.

We still have two **major** requirements for any of these methods to work?

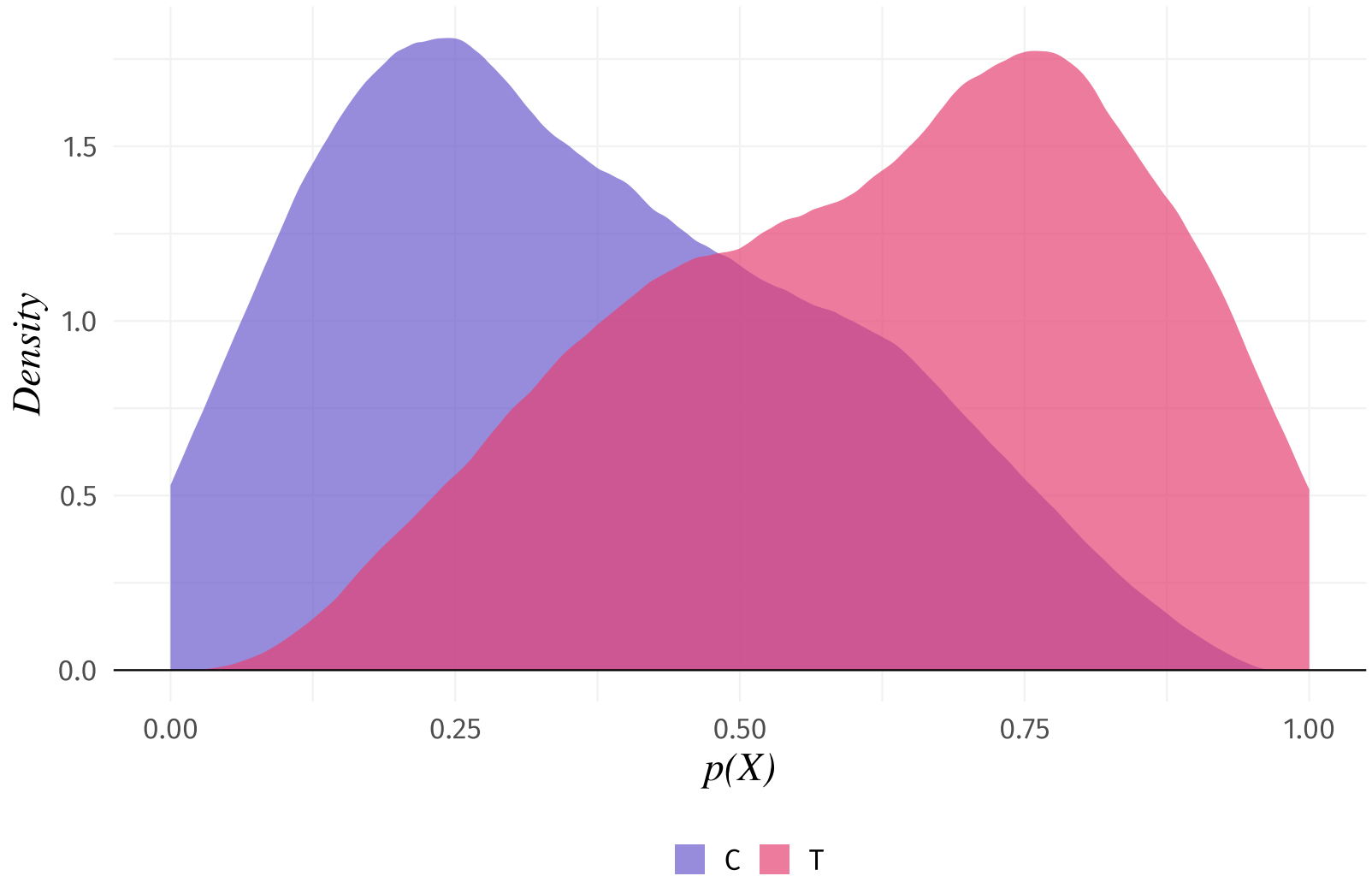
1. Is the **conditional-independence assumption** true?
2. Do we have **overlap** between treatment and control units.

We can look for evidence of (2) in the data—particularly if we're using propensity-score methods.[†]

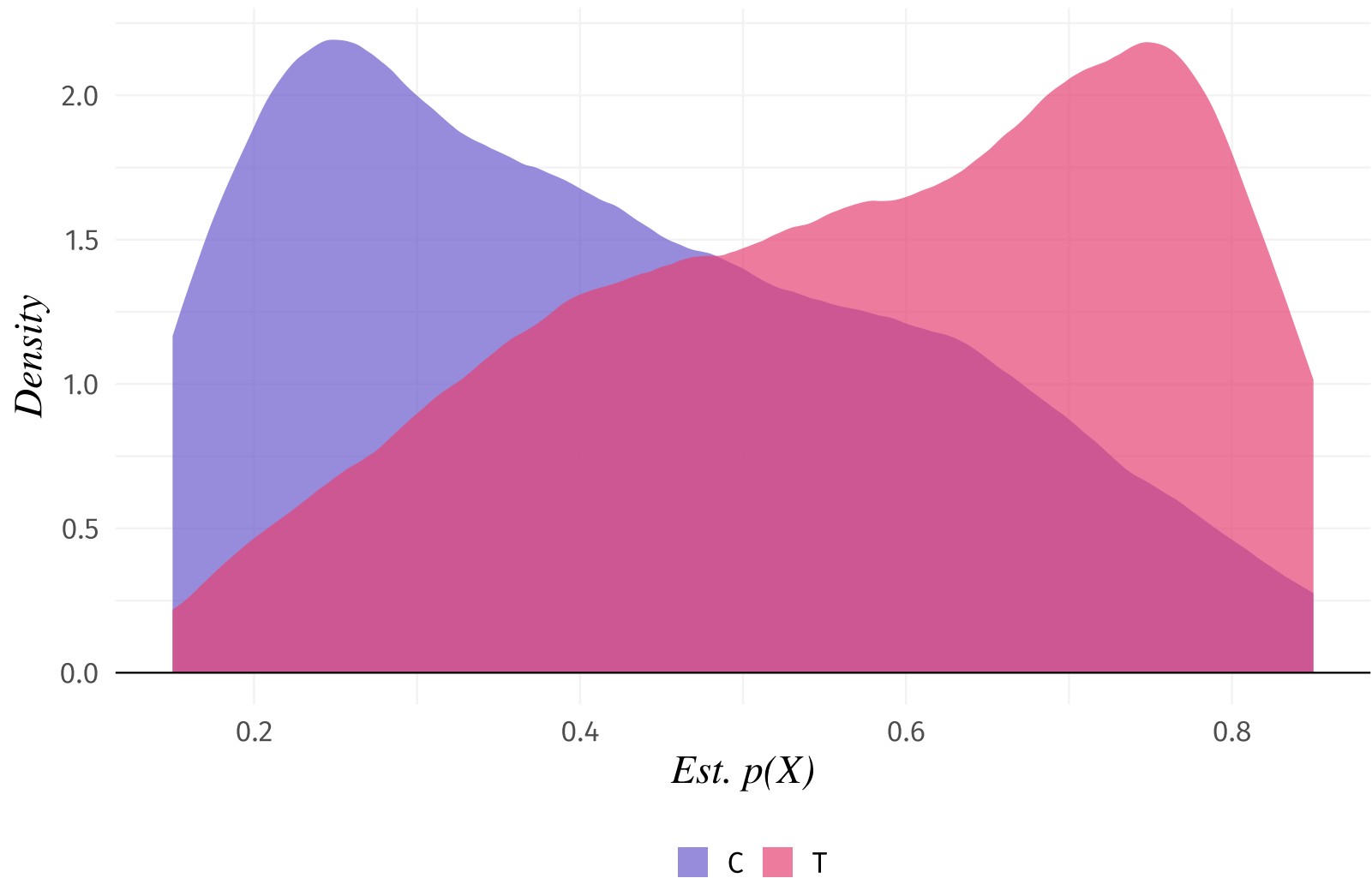
How? Plot the distributions of $p(\mathbf{X}_i)$ for **T** and **C**.

[†] Checking for overlap in \mathbf{X} -space, can be tough as the dimensions of \mathbf{X} expand.

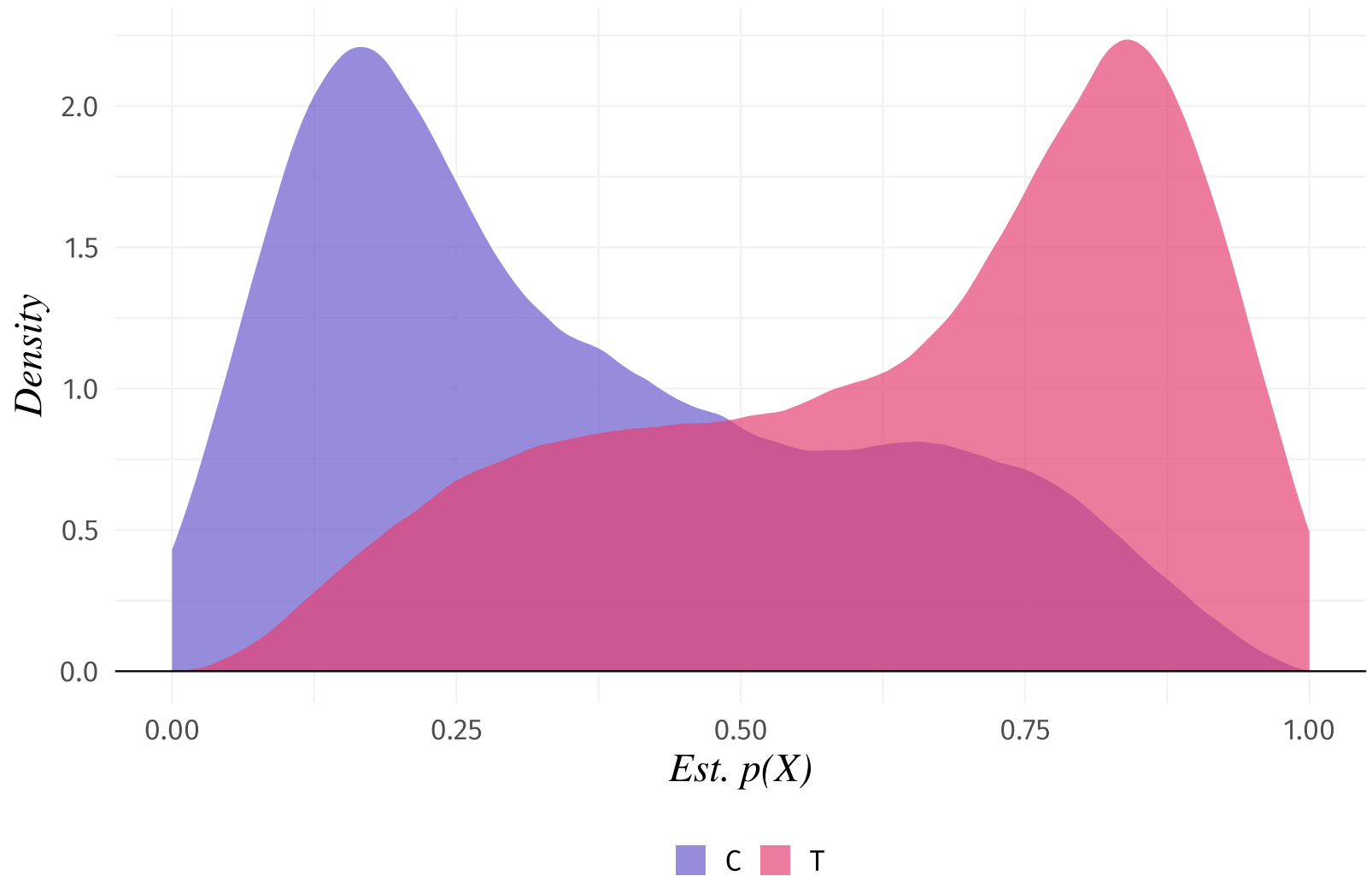
Missing overlap in $p(\mathbf{X}_i)$



Authentic (enforced) overlap in $p(\mathbf{X}_i)$



Logit-based $\hat{p}(\mathbf{X}_i)$ hiding some of the missing overlap in $p(\mathbf{X}_i)$



Overlap in one dimension does not guarantee in two dimensions.

Note Shading denotes **share of treatment**: **white** =0% and **pink**=100%.



Table of contents

Admin

1. Schedule
2. Follow up

General matching

1. The gist
2. Goals
3. Generic matching
4. Weights
 - Discrete \mathbf{X}
 - Nearest neighbor, Euclidean
 - Nearest neighbor, Mahalanobis
 - Kernel matching

Propensity-score methods

1. Setup
2. Propensity-score theorem
 - The magic
 - The proof
 - Intuition
3. Estimation
4. Application
 - Regression
 - Heterogeneity
 - Blocking on $p(\mathbf{X}_i)$
 - Weighting with $p(\mathbf{X}_i)$
 - Doubly robust methods
5. Overlap plots