Propensity scores: what, why and why not?

Rhian Daniel, Cardiff University


@statnav

Joint workshop — S3RI & Wessex Institute
University of Southampton, 22nd March 2018

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Setting

- We consider settings with a binary exposure A, an outcome Y, and a number of covariates/confounders L.

Setting

- We consider settings with a binary exposure A, an outcome Y, and a number of covariates/confounders L.
- Loosely speaking, we are interested in the effect of A on Y, and in order to estimate this we will account for L in some way.

Setting

- We consider settings with a binary exposure A, an outcome Y, and a number of covariates/confounders L.
- Loosely speaking, we are interested in the effect of A on Y, and in order to estimate this we will account for L in some way.
- We'll introduce propensity-score-based methods for accounting for L.

Setting

- We consider settings with a binary exposure A, an outcome Y, and a number of covariates/confounders L.
- Loosely speaking, we are interested in the effect of A on Y, and in order to estimate this we will account for L in some way.
- We'll introduce propensity-score-based methods for accounting for L.
- In particular we'll discuss the reasons for and the situations in which one might prefer such an approach over 'traditional' regression, as well as pointing out some incorrectly held beliefs on this issue.

Traditional regression

- Traditional regression methods proceed by fitting one model for Y given A and L, eg

$$E(Y|A, L) = \alpha + \beta A + \gamma' L.$$

Traditional regression

- Traditional regression methods proceed by fitting one model for Y given A and L, eg

$$E(Y|A, L) = \alpha + \beta A + \gamma' L.$$

- The field known as causal inference has formalised the causal interpretation of the above:
  - What is the causal quantity (hopefully) being estimated?
  - On what structural assumptions does this rely?
  - What sorts of variables can/should we include in L? Which ones should we omit? (DAGs)

    [Pearl (2009) *Causality*, CUP.]
  - On what parametric assumptions are we additionally relying?

[eg Hernán and Robins (2018) *Causal Inference*, Chapman & Hall.]

Traditional regression

- Traditional regression methods proceed by fitting one model for Y given A and L, eg

$$E(Y|A, L) = \alpha + \beta A + \gamma' L.$$

- The field known as causal inference has formalised the causal interpretation of the above:
  - What is the causal quantity (hopefully) being estimated?
  - On what structural assumptions does this rely?
  - What sorts of variables can/should we include in L? Which ones should we omit? (DAGs)

    [Pearl (2009) *Causality*, CUP.]
  - On what parametric assumptions are we additionally relying?

  [eg Hernán and Robins (2018) *Causal Inference*, Chapman & Hall.]
- We return to these questions in a moment.

WHAT?

Propensity score methods: overview

- Propensity score methods take a different approach.
  [Rosenbaum and Rubin (1983) *Biometrika*, 70(1):41–55.]

Propensity score methods: overview

- Propensity score methods take a different approach.
  [Rosenbaum and Rubin (1983) *Biometrika*, 70(1):41–55.]
- The propensity score is defined as:

$$\pi(L) = E(A|L),$$

the probability of being exposed as a function of covariates L.

Propensity score methods: overview

- Propensity score methods take a different approach.
  [Rosenbaum and Rubin (1983) *Biometrika*, 70(1):41–55.]
- The propensity score is defined as:

$$\pi(L) = E(A|L),$$

  the probability of being exposed as a function of covariates L.
- The key property of the scalar $\pi(L)$ is this: if L is sufficient to adjust for confounding, then so is $\pi(L)$.

Propensity score methods: overview

- Propensity score methods take a different approach.
  [Rosenbaum and Rubin (1983) *Biometrika*, 70(1):41–55.]
- The propensity score is defined as:

$$\pi(L) = E(A|L),$$

  the probability of being exposed as a function of covariates L.
- The key property of the scalar $\pi(L)$ is this: if L is sufficient to adjust for confounding, then so is $\pi(L)$.
- $\pi(\cdot)$ is typically an unknown function. A regression model is often fitted to estimate it, eg

$$\pi(L) = E(A|L) = \text{expit}\left(\nu + \eta' L\right).$$

Propensity score methods: overview

- Propensity score methods take a different approach.
  [Rosenbaum and Rubin (1983) *Biometrika*, 70(1):41–55.]
- The propensity score is defined as:

$$\pi(L) = E(A|L),$$

  the probability of being exposed as a function of covariates L.
- The key property of the scalar $\pi(L)$ is this: if L is sufficient to adjust for confounding, then so is $\pi(L)$.
- $\pi(\cdot)$ is typically an unknown function. A regression model is often fitted to estimate it, eg

$$\pi(L) = E(A|L) = \text{expit}\left(\nu + \eta' L\right).$$

- Propensity score methods use the key property and replace the multivariate L with the scalar $\widehat{\pi}(L)$ in the analysis, eg by fitting

$$E\left(Y|A, L\right) = \theta + \psi A + \phi\widehat{\pi}(L).$$

PS methods: adjusting, stratifying, matching, weighting

- As well as adjusting for the estimated propensity score in a regression model:

$$E(Y|A, L) = \theta + \psi A + \phi\widehat{\pi}(L),$$

since $\widehat{\pi}(L)$ is a scalar, alternative non-model based adjustments are feasible, such as stratifying on $\widehat{\pi}(L)$, matching on $\widehat{\pi}(L)$ and inverse weighting by $\widehat{\pi}(L)$.

PS methods: adjusting, stratifying, matching, weighting

- As well as adjusting for the estimated propensity score in a regression model:

$$E(Y|A, L) = \theta + \psi A + \phi\widehat{\pi}(L),$$

since $\widehat{\pi}(L)$ is a scalar, alternative non-model based adjustments are feasible, such as stratifying on $\widehat{\pi}(L)$, matching on $\widehat{\pi}(L)$ and inverse weighting by $\widehat{\pi}(L)$.

- At least as far as adjusting/stratifying go, we can think of this use of the propensity score as a dimension-reduction approach on L, but one that preserves the confounding adjustment property of the full set L.

WHY (NOT)?

So much controversy and confusion!

- It is natural to ask why, when and according to what measure is one of these approaches (traditional regression vs propensity score methods) preferable to the other.

So much controversy and confusion!

- It is natural to ask why, when and according to what measure is one of these approaches (traditional regression vs propensity score methods) preferable to the other.
- When might this dimension reduction be particularly valuable? Can it also be harmful?

So much controversy and confusion!

- It is natural to ask why, when and according to what measure is one of these approaches (traditional regression vs propensity score methods) preferable to the other.
- When might this dimension reduction be particularly valuable? Can it also be harmful?
- In one approach we have to specify a model for $E(Y|A, L)$ and in the other we have to specify a model for $E(A|L)$. In what situations might one of these be easier to do than the other?

So much controversy and confusion!

- It is natural to ask why, when and according to what measure is one of these approaches (traditional regression vs propensity score methods) preferable to the other.
- When might this dimension reduction be particularly valuable? Can it also be harmful?
- In one approach we have to specify a model for $E(Y|A, L)$ and in the other we have to specify a model for $E(A|L)$. In what situations might one of these be easier to do than the other?
- It is perhaps surprising that so much controversy and confusion have been generated through trying to answer these seemingly simple questions!

So much controversy and confusion!

- It is natural to ask why, when and according to what measure is one of these approaches (traditional regression vs propensity score methods) preferable to the other.
- When might this dimension reduction be particularly valuable? Can it also be harmful?
- In one approach we have to specify a model for $E(Y|A, L)$ and in the other we have to specify a model for $E(A|L)$. In what situations might one of these be easier to do than the other?
- It is perhaps surprising that so much controversy and confusion have been generated through trying to answer these seemingly simple questions!
- Before we can discuss these questions and resolve some of the confusing answers, we need to revisit traditional regression as viewed through a causal inference lens...

Traditional regression through the causal inference lens (1)

eg

$$E(Y|A, L) = g^{-1}(\alpha + \beta A + \gamma' L) \qquad (\dagger)$$

- What is the causal quantity (hopefully) being estimated?

Traditional regression through the causal inference lens (1)

eg

$$E(Y|A, L) = g^{-1}(\alpha + \beta A + \gamma' L) \qquad (\dagger)$$

- What is the causal quantity (hopefully) being estimated?
  — The average treatment effect (ATE) is defined as $E(Y_1 - Y_0)$, where Y_a is the potential outcome that would be seen if A were set to a.

Traditional regression through the causal inference lens (1)

eg

$$E(Y|A, L) = g^{-1}(\alpha + \beta A + \gamma' L) \qquad (\dagger)$$

- What is the causal quantity (hopefully) being estimated?
  — The average treatment effect (ATE) is defined as $E(Y_1 - Y_0)$, where Y_a is the potential outcome that would be seen if A were set to a.
  — Under the assumptions given on the next slide, linear regression (eg $(\dagger)$ with $g(\cdot)$ the identity link) directly targets the ATE's conditional counterpart $E(Y_1 - Y_0|L)$, but this (β) is equal to the ATE if there is no effect modification.

Traditional regression through the causal inference lens (1)

eg

$$E(Y|A, L) = g^{-1}(\alpha + \beta A + \gamma' L) \qquad (\dagger)$$

- What is the causal quantity (hopefully) being estimated?
  — The average treatment effect (ATE) is defined as $E(Y_1 - Y_0)$, where Y_a is the potential outcome that would be seen if A were set to a.
  — Under the assumptions given on the next slide, linear regression (eg $(\dagger)$ with $g(\cdot)$ the identity link) directly targets the ATE's conditional counterpart $E(Y_1 - Y_0|L)$, but this (β) is equal to the ATE if there is no effect modification.
  — Otherwise, the ATE can be obtained via a small additional step. Eg for $(\dagger)$, and under the assumptions on the next slide,

$$E(Y_1 - Y_0) = E\left\{g^{-1}(\alpha + \beta + \gamma' L) - g^{-1}(\alpha + \gamma' L)\right\}.$$

Traditional regression through the causal inference lens (1)

eg

$$E(Y|A,L) = g^{-1}(\alpha + \beta A + \gamma' L) \qquad (\dagger)$$

- What is the causal quantity (hopefully) being estimated?
  — The average treatment effect (ATE) is defined as $E(Y_1 - Y_0)$, where Y_a is the potential outcome that would be seen if A were set to a.
  — Under the assumptions given on the next slide, linear regression (eg (†) with $g(\cdot)$ the identity link) directly targets the ATE's conditional counterpart $E(Y_1 - Y_0|L)$, but this (β) is equal to the ATE if there is no effect modification.
  — Otherwise, the ATE can be obtained via a small additional step. Eg for (†), and under the assumptions on the next slide,

$$E(Y_1 - Y_0) = E\left\{g^{-1}(\alpha + \beta + \gamma' L) - g^{-1}(\alpha + \gamma' L)\right\}.$$

  — NB propensity score methods also target the ATE, or can be made to via a simple step as above.

Traditional regression through the causal inference lens (2)

eg

$$E(Y|A, L) = g^{-1}(\alpha + \beta A + \gamma' L) \qquad (\dagger)$$

- On what structural assumptions does this causal interpretation rely?
- What sorts of variables can/should we include in L? Which ones should we omit?
- On what parametric assumptions are we additionally relying?

Traditional regression through the causal inference lens (2)

eg

$$E(Y|A, L) = g^{-1}(\alpha + \beta A + \gamma' L) \qquad (\dagger)$$

- On what structural assumptions does this causal interpretation rely?
  — The key one is conditional exchangeability, $Y_a \perp\!\!\!\perp A|L,\ a = 0, 1$.

- What sorts of variables can/should we include in L? Which ones should we omit?

- On what parametric assumptions are we additionally relying?

Traditional regression through the causal inference lens (2)

eg

$$E(Y|A, L) = g^{-1}(\alpha + \beta A + \gamma' L) \qquad (\dagger)$$

- On what structural assumptions does this causal interpretation rely?
  — The key one is conditional exchangeability, $Y_a \perp\!\!\!\perp A|L$, $a = 0, 1$.
  — Also no interference, consistency, positivity. (More details at the end.)

  [eg Hernán and Robins (2018) *Causal Inference*, Chapman & Hall.]
- What sorts of variables can/should we include in L? Which ones should we omit?
- On what parametric assumptions are we additionally relying?

Traditional regression through the causal inference lens (2)

eg

$$E(Y|A, L) = g^{-1}(\alpha + \beta A + \gamma' L) \qquad (\dagger)$$

- On what structural assumptions does this causal interpretation rely?
  — The key one is conditional exchangeability, $Y_a \perp\!\!\!\perp A|L$, $a = 0, 1$.
  — Also no interference, consistency, positivity. (More details at the end.)

  [eg Hernán and Robins (2018) *Causal Inference*, Chapman & Hall.]
- What sorts of variables can/should we include in L? Which ones should we omit?
  — To avoid bias, we need to choose L so that conditional exchangeability holds. For efficiency, other variables may be useful/detrimental without affecting conditional exchangeability. See next slides.
- On what parametric assumptions are we additionally relying?

Traditional regression through the causal inference lens (2)

eg

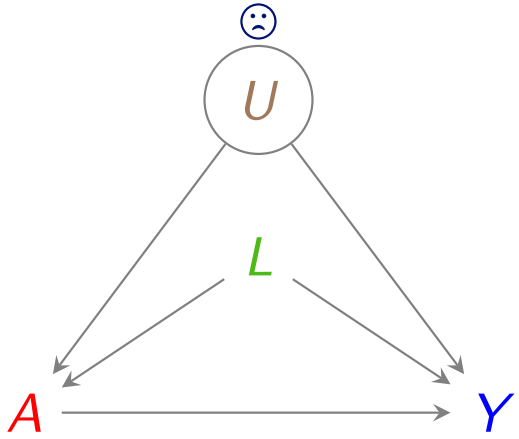
$$E(Y|A, L) = g^{-1}(\alpha + \beta A + \gamma' L) \qquad (\dagger)$$

- On what structural assumptions does this causal interpretation rely?
  — The key one is conditional exchangeability, $Y_a \perp\!\!\!\perp A|L$, $a = 0, 1$.
  — Also no interference, consistency, positivity. (More details at the end.)

  [eg Hernán and Robins (2018) *Causal Inference*, Chapman & Hall.]
- What sorts of variables can/should we include in L? Which ones should we omit?
  — To avoid bias, we need to choose L so that conditional exchangeability holds. For efficiency, other variables may be useful/detrimental without affecting conditional exchangeability. See next slides.
- On what parametric assumptions are we additionally relying?
  — The regression model needs to have the correct functional form, eg if ${L_1}^2$ is wrongly omitted from $(\dagger)$ then our estimator of the ATE will be biased.

Trad reg: which variables should be included in L? (1)

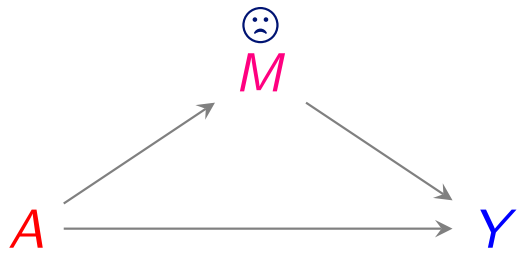
Pearl's work has taught us...



- Assuming conditional exchangeability given L means there can exist no unmeasured common causes U of A and Y.

Trad reg: which variables should be included in L? (1)

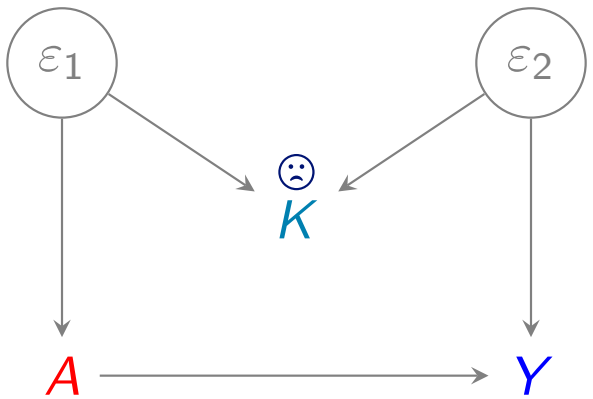
Pearl's work has taught us...



- Assuming conditional exchangeability given L means there can exist no unmeasured common causes U of A and Y.
- And also it means that L should not include any consequences M of exposure...

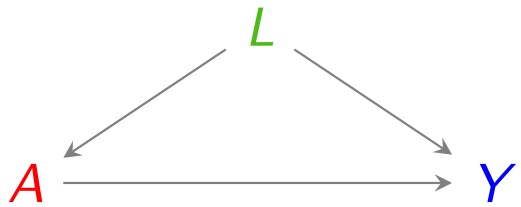
Trad reg: which variables should be included in L? (1)

Pearl's work has taught us...



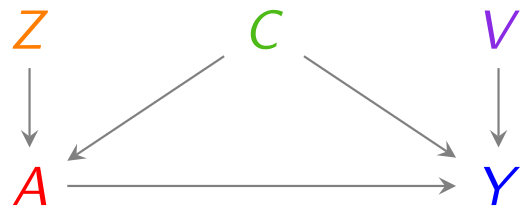
- Assuming conditional exchangeability given L means there can exist no unmeasured common causes U of A and Y.
- And also it means that L should not include any consequences M of exposure...
- ...nor any 'unresolved' colliders K.

Trad reg: which variables should be included in L? (2)



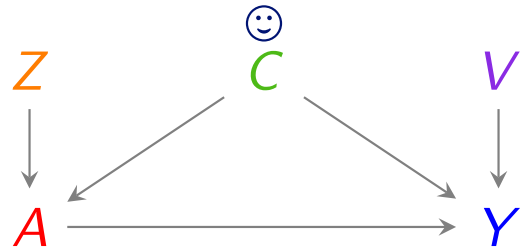
- We can explicitly split L...

Trad reg: which variables should be included in L? (2)



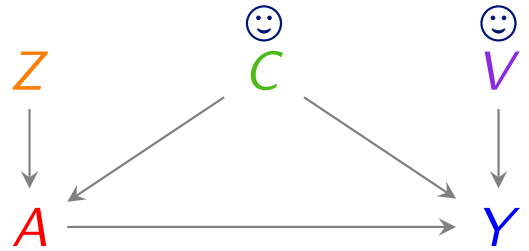
- We can explicitly split L...
- ...into confounders C, instruments Z and risk factors V.

Trad reg: which variables should be included in L? (2)



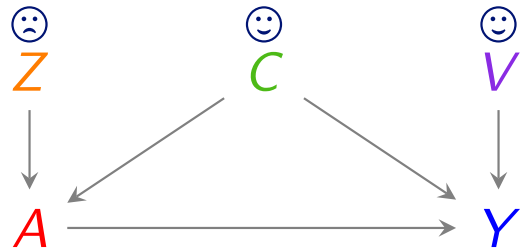
- We can explicitly split L...
- ...into confounders C, instruments Z and risk factors V.
- Adjustment for C is necessary for consistent estimation

Trad reg: which variables should be included in L? (2)



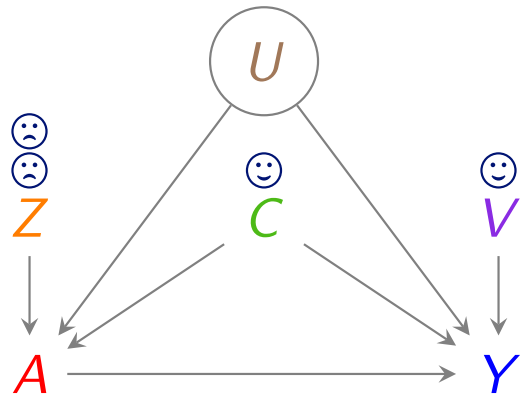
- We can explicitly split L...
- ...into confounders C, instruments Z and risk factors V.
- Adjustment for C is necessary for consistent estimation, adjustment for V is helpful for precision

Trad reg: which variables should be included in L? (2)



- We can explicitly split L...
- ...into confounders C, instruments Z and risk factors V.
- Adjustment for C is necessary for consistent estimation, adjustment for V is helpful for precision, but adjustment for Z is detrimental for precision.

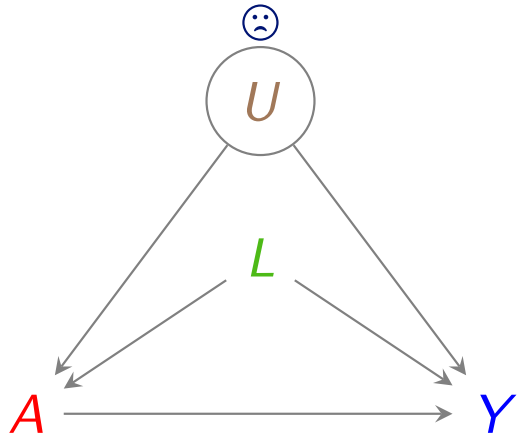
Trad reg: which variables should be included in L? (2)



- We can explicitly split L...
- ...into confounders C, instruments Z and risk factors V.
- Adjustment for C is necessary for consistent estimation, adjustment for V is helpful for precision, but adjustment for Z is detrimental for precision.
- Adjustment for Z also amplifies bias due to any unmeasured confounders U.

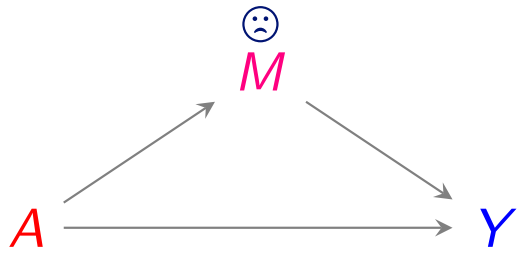
[Wooldridge (2016) *Research in Economics*, 70(2):232–7.]

What about when using propensity score methods? (1)



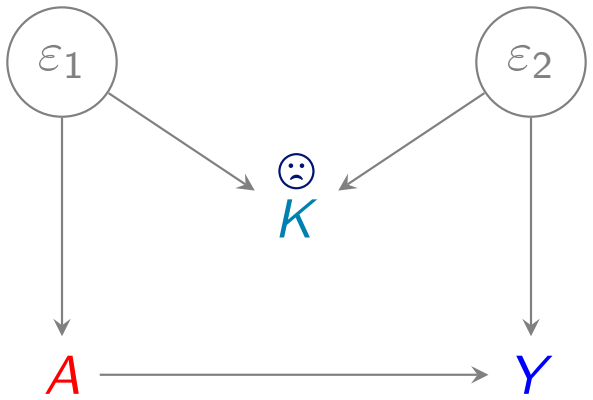
- The story is the same when considering methods based on the propensity score: unmeasured confounders U are bad news. Anecdotally, this is not always appreciated.

What about when using propensity score methods? (1)



- The story is the same when considering methods based on the propensity score: unmeasured confounders U are bad news. Anecdotally, this is not always appreciated.
- L still can't include any consequence M of exposure...

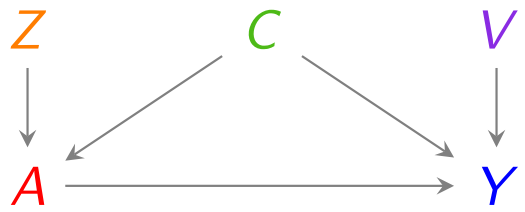
What about when using propensity score methods? (1)



- The story is the same when considering methods based on the propensity score: unmeasured confounders U are bad news. Anecdotally, this is not always appreciated.
- L still can't include any consequence M of exposure...
- ...nor any 'unresolved' colliders K. Rubin refuses to acknowledge this!

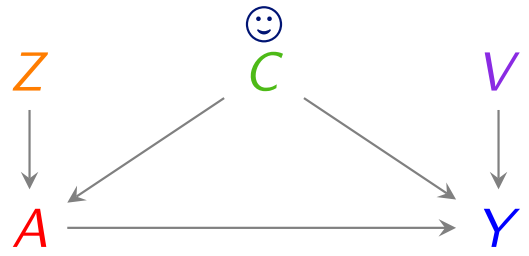
[Rubin (2007) *Statistics in Medicine*, 26(1): 20–36 and the letters that followed by Shrier, Rubin, Pearl, Sjölander.]

What about when using propensity score methods? (2)



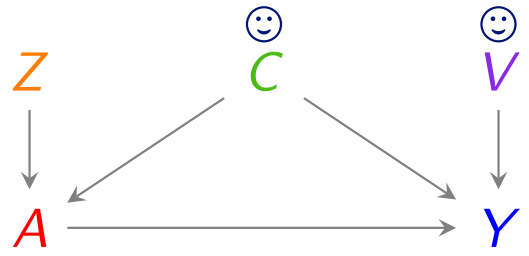
- What about the choice between C, V and Z when using a propensity score approach? One might naïvely imagine that the arguments change when we model $E(A|L)$ instead of $E(Y|A, L)$.

What about when using propensity score methods? (2)



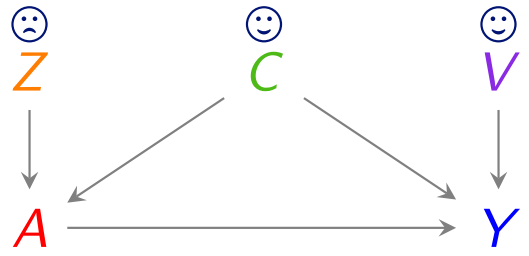
- What about the choice between C, V and Z when using a propensity score approach? One might naïvely imagine that the arguments change when we model $E(A|L)$ instead of $E(Y|A, L)$.
- But no. Including C in the PS model is necessary for consistent estimation

What about when using propensity score methods? (2)



- What about the choice between C, V and Z when using a propensity score approach? One might naïvely imagine that the arguments change when we model $E(A|L)$ instead of $E(Y|A, L)$.
- But no. Including C in the PS model is necessary for consistent estimation, including V is helpful for precision (even though its true coefficient is zero)

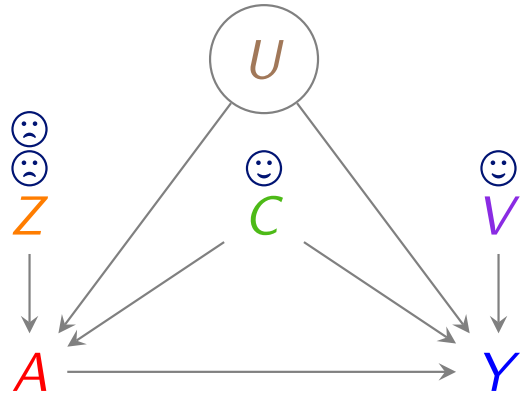
What about when using propensity score methods? (2)



- What about the choice between C, V and Z when using a propensity score approach? One might naïvely imagine that the arguments change when we model $E(A|L)$ instead of $E(Y|A, L)$.
- But no. Including C in the PS model is necessary for consistent estimation, including V is helpful for precision (even though its true coefficient is zero), and including Z is detrimental for precision.

  [Brookhart et al (2006) *AJE* 163:1149–56.]

What about when using propensity score methods? (2)



- What about the choice between C, V and Z when using a propensity score approach? One might naïvely imagine that the arguments change when we model $E(A|L)$ instead of $E(Y|A, L)$.
- But no. Including C in the PS model is necessary for consistent estimation, including V is helpful for precision (even though its true coefficient is zero), and including Z is detrimental for precision.

  [Brookhart et al (2006) *AJE* 163:1149–56.]
- Again, including Z in the PS model amplifies bias due to any unmeasured confounders U.

  [Pearl (2011) *AJE*, 174:1223–7.]

What then is the difference?

- So far, we have noted that traditional regression and propensity score approaches demand that we collect the same set of covariates L, namely those given which conditional exchangeability holds (for bias reduction) plus any variables that predict the outcome (for precision). Predictors of exposure (only) are harmful (precision, bias amplification) for both approaches.

What then is the difference?

- So far, we have noted that traditional regression and propensity score approaches demand that we collect the same set of covariates L, namely those given which conditional exchangeability holds (for bias reduction) plus any variables that predict the outcome (for precision). Predictors of exposure (only) are harmful (precision, bias amplification) for both approaches.
- The choice then is between, for traditional regression, specifying a parametric model for $E(Y|A, L)$, and, for propensity score methods, specifying a parametric model for $E(A|L)$.

What then is the difference?

- So far, we have noted that traditional regression and propensity score approaches demand that we collect the same set of covariates L, namely those given which conditional exchangeability holds (for bias reduction) plus any variables that predict the outcome (for precision). Predictors of exposure (only) are harmful (precision, bias amplification) for both approaches.
- The choice then is between, for traditional regression, specifying a parametric model for $E(Y|A, L)$, and, for propensity score methods, specifying a parametric model for $E(A|L)$.
- If we misspecify the functional form of $E(Y|A, L)$, the traditional regression estimator (of the ATE) will be biased. If we misspecify the functional form of $E(A|L)$, the propensity score estimators (of the ATE) will be biased.

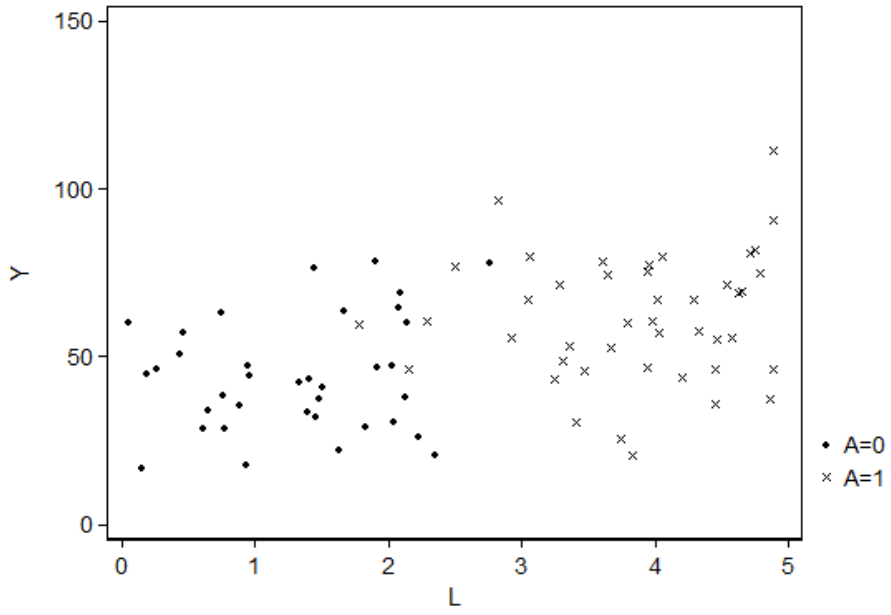
What then is the difference?

- So far, we have noted that traditional regression and propensity score approaches demand that we collect the same set of covariates L, namely those given which conditional exchangeability holds (for bias reduction) plus any variables that predict the outcome (for precision). Predictors of exposure (only) are harmful (precision, bias amplification) for both approaches.
- The choice then is between, for traditional regression, specifying a parametric model for $E(Y|A, L)$, and, for propensity score methods, specifying a parametric model for $E(A|L)$.
- If we misspecify the functional form of $E(Y|A, L)$, the traditional regression estimator (of the ATE) will be biased. If we misspecify the functional form of $E(A|L)$, the propensity score estimators (of the ATE) will be biased.
- The dimension of L is the same in both, so aren't both jobs equally hard, and the consequences of mistakes equally severe?

What then is the difference?

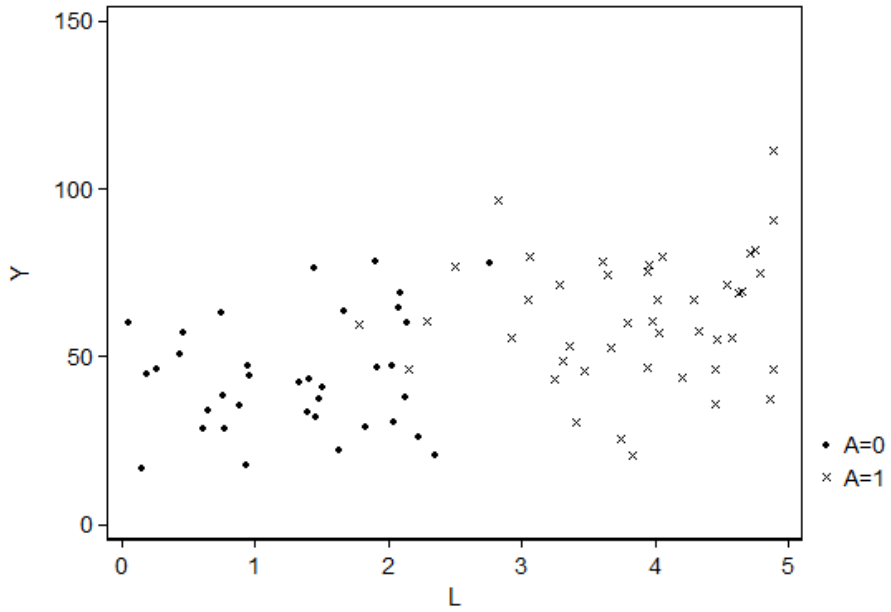
- So far, we have noted that traditional regression and propensity score approaches demand that we collect the same set of covariates L, namely those given which conditional exchangeability holds (for bias reduction) plus any variables that predict the outcome (for precision). Predictors of exposure (only) are harmful (precision, bias amplification) for both approaches.
- The choice then is between, for traditional regression, specifying a parametric model for $E(Y|A, L)$, and, for propensity score methods, specifying a parametric model for $E(A|L)$.
- If we misspecify the functional form of $E(Y|A, L)$, the traditional regression estimator (of the ATE) will be biased. If we misspecify the functional form of $E(A|L)$, the propensity score estimators (of the ATE) will be biased.
- The dimension of L is the same in both, so aren't both jobs equally hard, and the consequences of mistakes equally severe? Well, no...

Poor overlap (1)



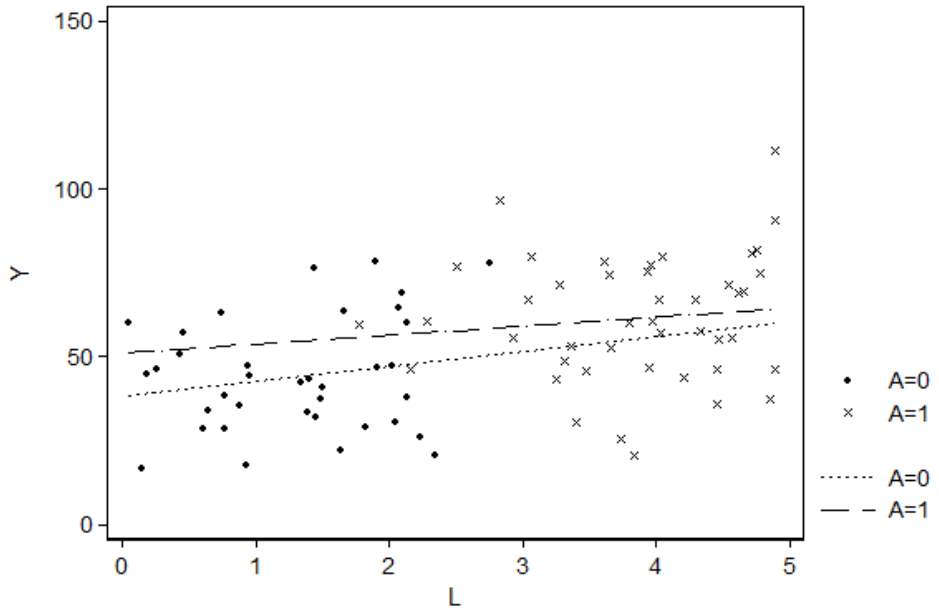
— An important and common situation is one in which there is little overlap between the L-values of the exposed and unexposed groups.

Poor overlap (1)



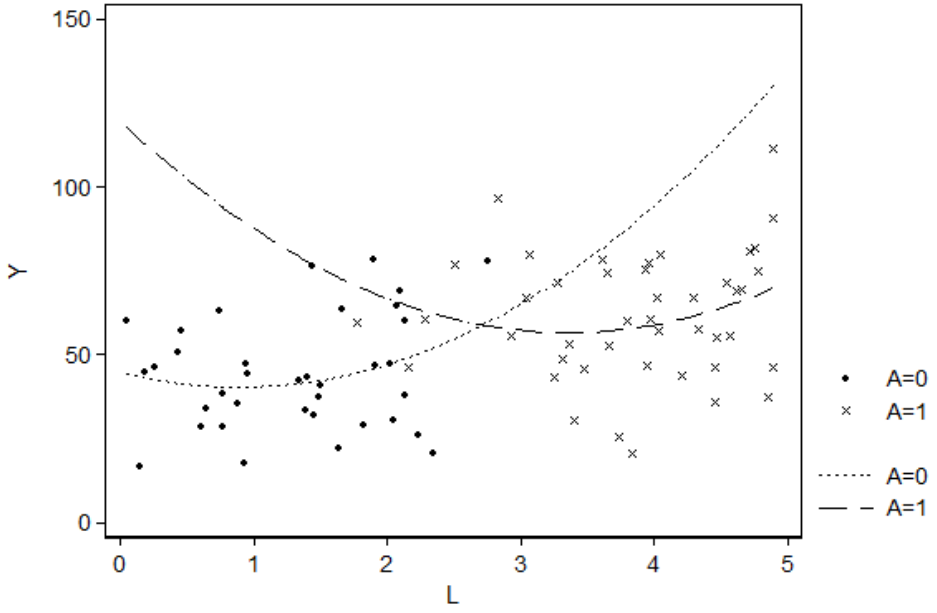
— An important and common situation is one in which there is little overlap between the L-values of the exposed and unexposed groups.
— Consider what then happens to the traditional regression estimator, which relies on correctly specifying the functional form of $E(Y|A, L)$.

Poor overlap (2)



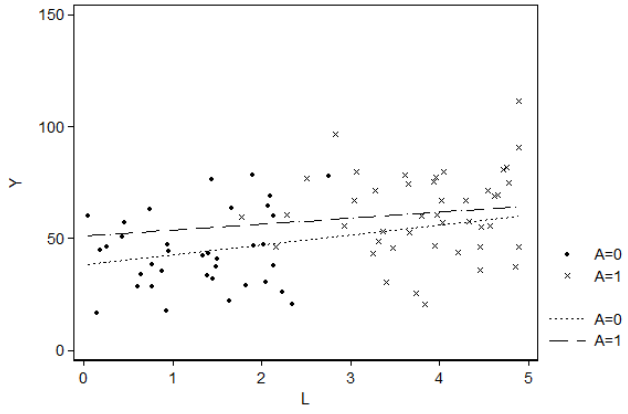
There is little information in the data to choose between this linear model . . .

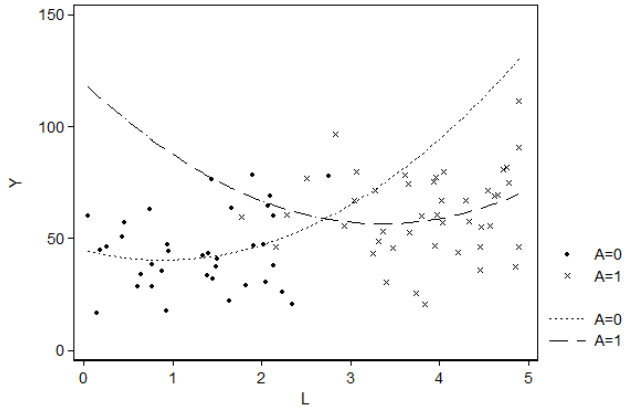
Poor overlap (3)



. . . and this quadratic model, even though the estimated ATEs are v different.

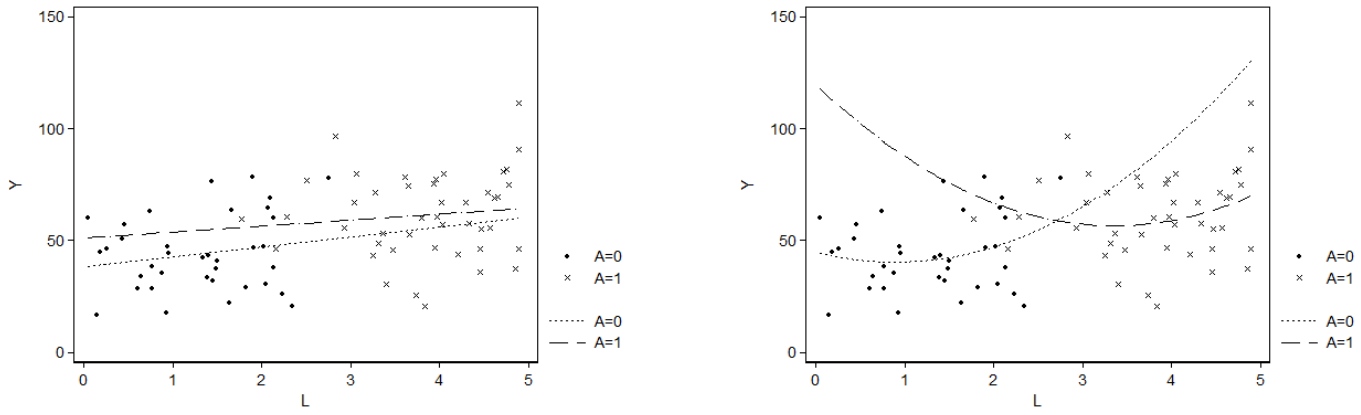
Poor overlap (4)





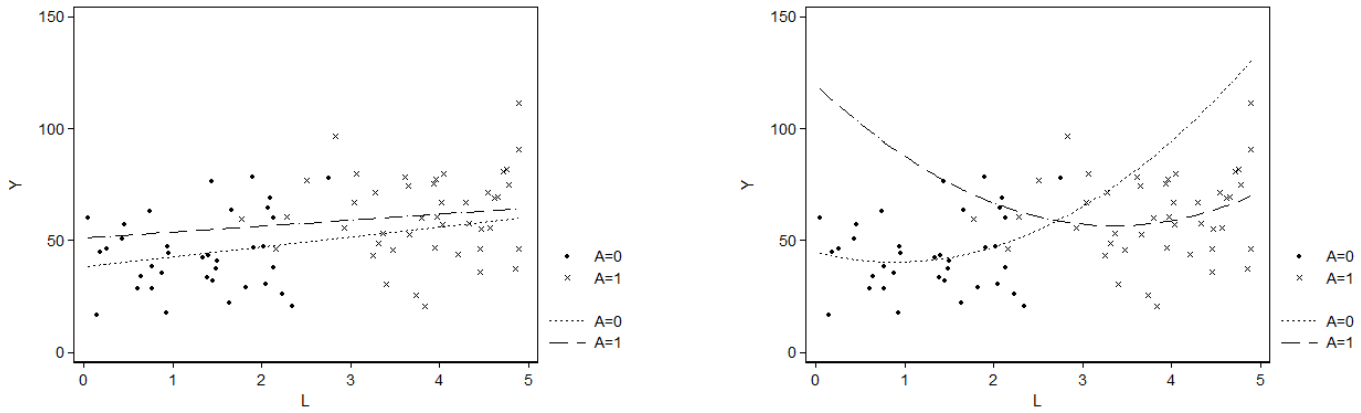
- Both models fit the data almost equally well, but lead to different estimates of the ATE.

Poor overlap (4)



- Both models fit the data almost equally well, but lead to different estimates of the ATE.
- We can only check how well the model fits the observed data, but, whenever there is poor overlap, the traditional regression estimator relies on the extrapolation of these fitted relationships to regions where there is little data to support the model choice.

Poor overlap (4)

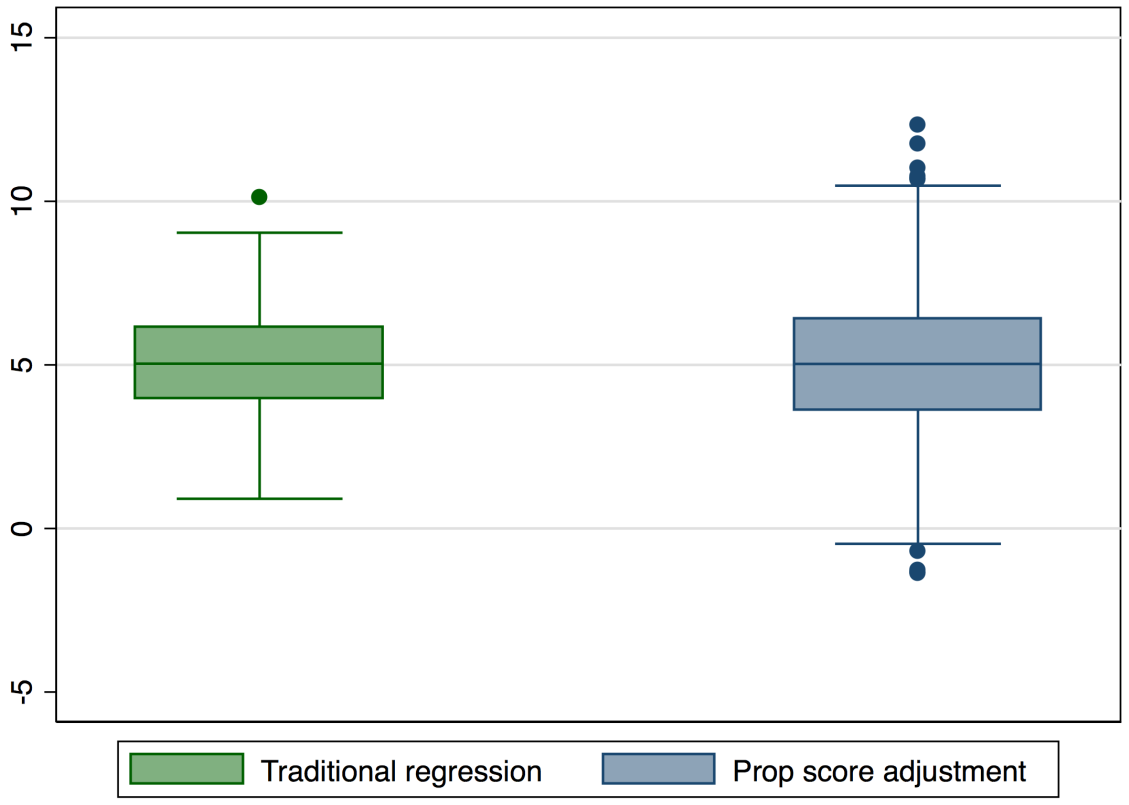


- Both models fit the data almost equally well, but lead to different estimates of the ATE.
- We can only check how well the model fits the observed data, but, whenever there is poor overlap, the traditional regression estimator relies on the extrapolation of these fitted relationships to regions where there is little data to support the model choice.
- Regression methods flag this only very mildly (via slightly increased estimated SEs) and proceed to give 'precise' estimates based on extrapolations.

A very simple simulation

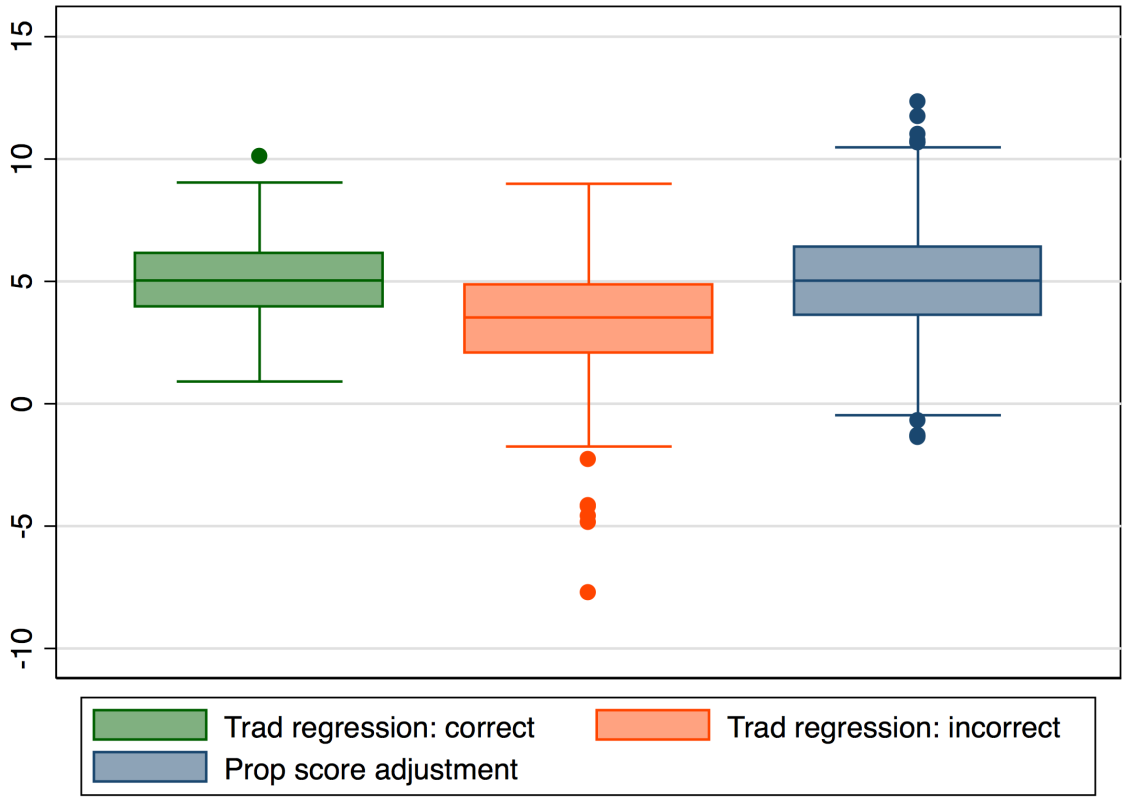
```
gen b_reg_c=.
gen b_reg_w=.
gen b_ps=.

forvalues i=1(1)500 {
qui gen C=.7*(rnormal())^2 if _n<=100
qui replace C=C+rnormal() if _n<=100

qui gen X=runiform()<1/(1+exp(-(5*(C-.7)))) if _n<=100
qui gen Y=C+1.5*C^2+5*X+5*rnormal() if _n<=100

qui logit X C if _n<=100
qui predict ps if _n<=100

qui gen Csq=C^2 if _n<=100
qui reg Y C Csq X if _n<=100
qui replace b_reg_c=_b[X] in `i'

qui gen CX=C*X if _n<=100
qui summ C if _n<=100
local m=r(mean)
qui reg Y C X CX if _n<=100
qui replace b_reg_w=_b[X]+_b[CX]*`m' in `i'

qui reg Y ps X if _n<=100
qui replace b_ps=_b[X] in `i'

keep b_*
}
```

Simulations to demonstrate this (true value of ATE = 5)



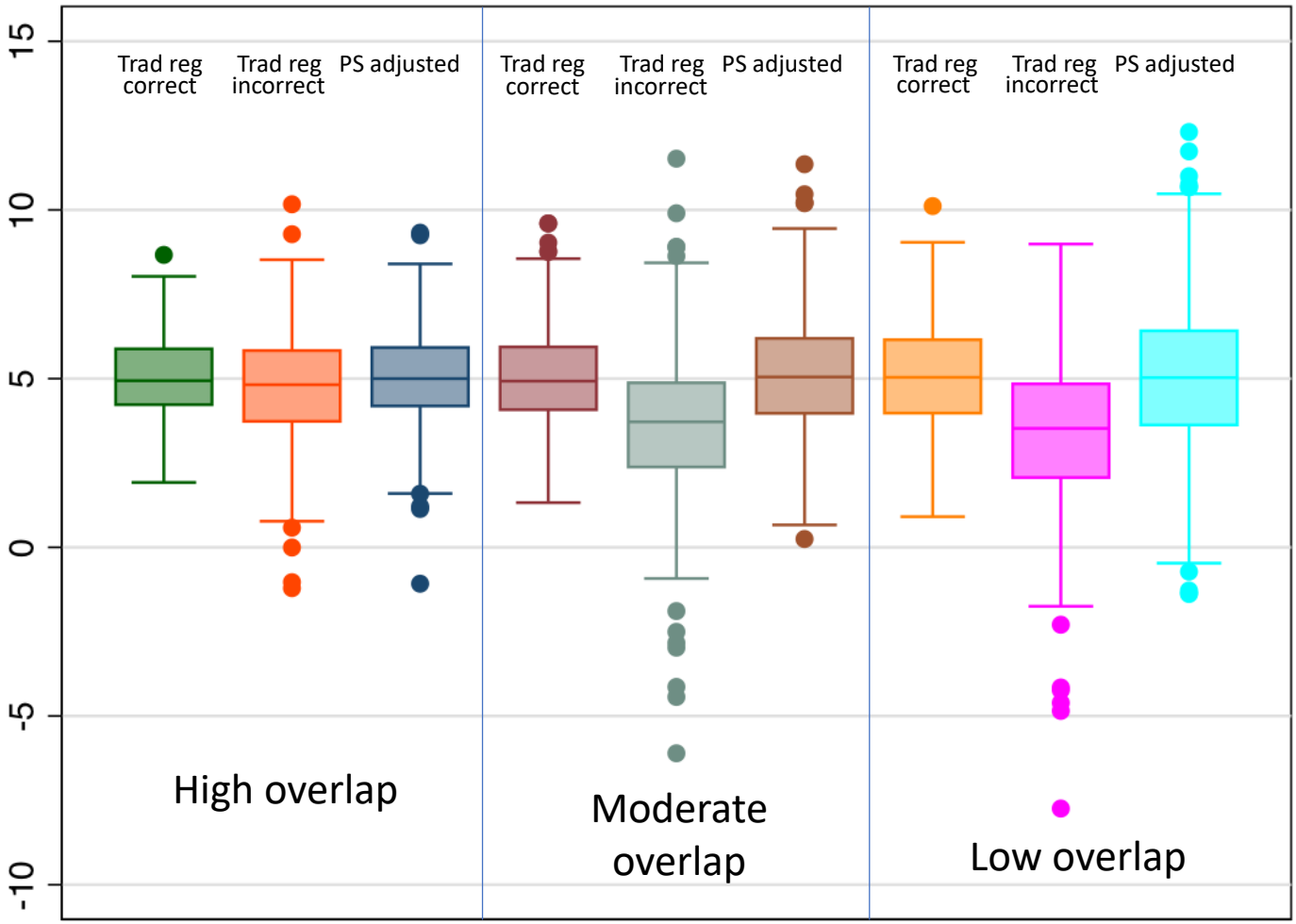
— When the model for $E(Y|A, L)$ is correctly specified, all is well, and, indeed, traditional regression is more efficient than adjustment for the estimated propensity score, as theory dictates.

Simulations to demonstrate this (true value of ATE = 5)



— But more relevant is the performance of traditional regression with an incorrectly-specified model: it is biased.

Simulations to demonstrate this (true value of ATE = 5)



Some remarks (1)

- As the overlap decreases, traditional regression based on a correctly-specified model stays unbiased and efficient, but the bias in the corresponding estimator with an incorrectly-specified model gets worse.

Some remarks (1)

- As the overlap decreases, traditional regression based on a correctly-specified model stays unbiased and efficient, but the bias in the corresponding estimator with an incorrectly-specified model gets worse.
- As the overlap decreases, we are increasingly likely to choose the wrong model, so low overlap stings traditional regression twice.

Some remarks (1)

- As the overlap decreases, traditional regression based on a correctly-specified model stays unbiased and efficient, but the bias in the corresponding estimator with an incorrectly-specified model gets worse.
- As the overlap decreases, we are increasingly likely to choose the wrong model, so low overlap stings traditional regression twice.
- In contrast, there is no reason to believe that low overlap increases the risk of getting the PS model wrong.

Some remarks (2)

- Note how the (correctly-specified) PS adjusted estimator gets less precise as the overlap decreases.

Some remarks (2)

- Note how the (correctly-specified) PS adjusted estimator gets less precise as the overlap decreases.
- This is because (for linear regression in the absence of effect modification), the propensity score adjusted estimator of the ATE is:

$$\hat{\psi} = \frac{\sum_{i=1}^{n} \{A_i - \widehat{\pi}(L_i)\} \left\{ Y_i - \widehat{\theta} - \widehat{\phi}\widehat{\pi}(L_i) \right\}}{\sum_{i=1}^{n} \{A_i - \widehat{\pi}(L_i)\} A_i}$$

Note how regions of low overlap are down-weighted.

Some remarks (2)

- Note how the (correctly-specified) PS adjusted estimator gets less precise as the overlap decreases.
- This is because (for linear regression in the absence of effect modification), the propensity score adjusted estimator of the ATE is:

$$\hat{\psi} = \frac{\sum_{i=1}^{n} \{A_i - \widehat{\pi}(L_i)\} \left\{ Y_i - \widehat{\theta} - \widehat{\phi}\widehat{\pi}(L_i) \right\}}{\sum_{i=1}^{n} \{A_i - \widehat{\pi}(L_i)\} A_i}$$

  Note how regions of low overlap are down-weighted.
- As the overlap decreases, the PS-adjusted estimator more honestly reflects the loss of information regarding the treatment effect.

Some remarks (2)

- Note how the (correctly-specified) PS adjusted estimator gets less precise as the overlap decreases.
- This is because (for linear regression in the absence of effect modification), the propensity score adjusted estimator of the ATE is:

$$\hat{\psi} = \frac{\sum_{i=1}^{n} \{A_i - \widehat{\pi}(L_i)\} \left\{ Y_i - \widehat{\theta} - \widehat{\phi}\widehat{\pi}(L_i) \right\}}{\sum_{i=1}^{n} \{A_i - \widehat{\pi}(L_i)\} A_i}$$

  Note how regions of low overlap are down-weighted.
- As the overlap decreases, the PS-adjusted estimator more honestly reflects the loss of information regarding the treatment effect.
- The estimated SEs for other propensity score methods also become increasingly large as overlap decreases and the poor overlap is flagged up by small strata, no available matches, extreme weights.

Another issue: finite sample bias (1)

- The parameters of regression models are typically estimated by maximum likelihood.

Another issue: finite sample bias (1)

- The parameters of regression models are typically estimated by maximum likelihood.
- ML estimators are, in general, only asymptotically unbiased.

Another issue: finite sample bias (1)

- The parameters of regression models are typically estimated by maximum likelihood.
- ML estimators are, in general, only asymptotically unbiased.
- For logistic and Cox regression, ML parameter estimators can be noticeably biased in small samples.

Another issue: finite sample bias (1)

- The parameters of regression models are typically estimated by maximum likelihood.
- ML estimators are, in general, only asymptotically unbiased.
- For logistic and Cox regression, ML parameter estimators can be noticeably biased in small samples.
- In particular, bias increases as the number of events per parameter decreases.

Another issue: finite sample bias (1)

- The parameters of regression models are typically estimated by maximum likelihood.
- ML estimators are, in general, only asymptotically unbiased.
- For logistic and Cox regression, ML parameter estimators can be noticeably biased in small samples.
- In particular, bias increases as the number of events per parameter decreases.
- The more confounders L we adjust for, the larger this finite-sample bias.

Another issue: finite sample bias (1)

- The parameters of regression models are typically estimated by maximum likelihood.
- ML estimators are, in general, only asymptotically unbiased.
- For logistic and Cox regression, ML parameter estimators can be noticeably biased in small samples.
- In particular, bias increases as the number of events per parameter decreases.
- The more confounders L we adjust for, the larger this finite-sample bias.
- Rule of thumb: "need 10 or more events per parameter being estimated".

[Peduzzi et al (1995) *J Clin Epidemiol*.]

Another issue: finite sample bias (2)

- Why is this more of a problem for traditional regression than PS methods?

Another issue: finite sample bias (2)

- Why is this more of a problem for traditional regression than PS methods?
- Often outcomes are rare but exposures are not, so the events per parameter can be far lower for $E(Y|A, L)$ than for $E(A|L)$.

Another issue: finite sample bias (2)

- Why is this more of a problem for traditional regression than PS methods?
- Often outcomes are rare but exposures are not, so the events per parameter can be far lower for $E(Y|A, L)$ than for $E(A|L)$.
- More importantly, the finite sample bias affects individual parameter estimators but not the estimator of the conditional expectation (the predictions).

Another issue: finite sample bias (2)

- Why is this more of a problem for traditional regression than PS methods?
- Often outcomes are rare but exposures are not, so the events per parameter can be far lower for $E(Y|A, L)$ than for $E(A|L)$.
- More importantly, the finite sample bias affects individual parameter estimators but not the estimator of the conditional expectation (the predictions).
- So, for both reasons, $\widehat{\pi}(L)$ can be estimated without bias while the individual parameter estimates for the parameters of $E(Y|A, L)$ are severely biased.

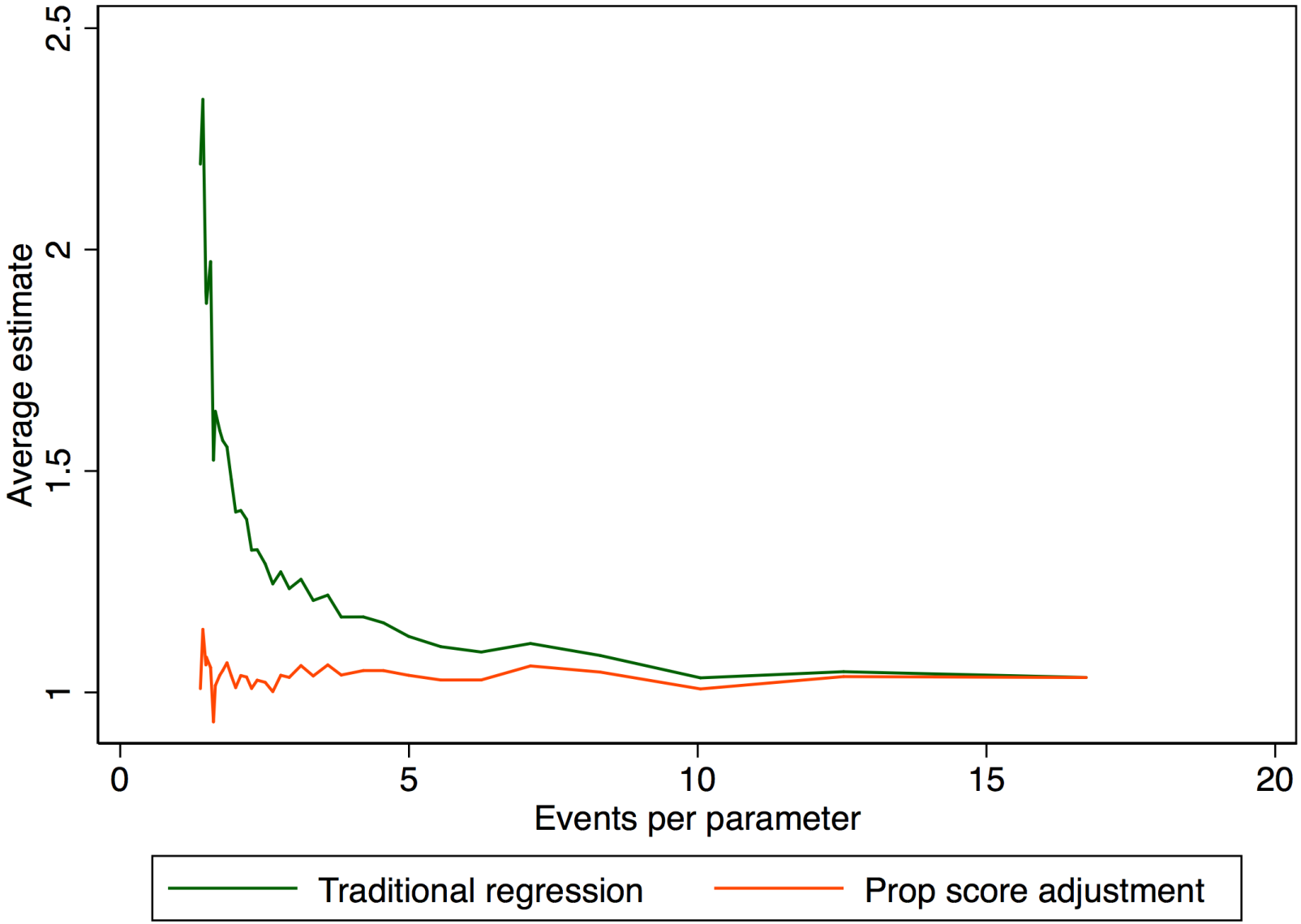
Another issue: finite sample bias (2)

- Why is this more of a problem for traditional regression than PS methods?
- Often outcomes are rare but exposures are not, so the events per parameter can be far lower for $E(Y|A, L)$ than for $E(A|L)$.
- More importantly, the finite sample bias affects individual parameter estimators but not the estimator of the conditional expectation (the predictions).
- So, for both reasons, $\widehat{\pi}(L)$ can be estimated without bias while the individual parameter estimates for the parameters of $E(Y|A, L)$ are severely biased.
- This is so even in the absence of any model misspecification.

Another very quick simulation

```
forvalues j=1(1)50 {
qui gen epp_‘j’=.
qui gen b_reg_‘j’=.
qui gen b_ps_‘j’=.
forvalues i=1(1)100 {
qui gen X=runiform()<0.5 if _n<=100
forvalues k=1(1)‘j’ {
qui gen C‘k’=runiform()<0.5 if _n<=100
}
qui gen Y=runiform()<1/(1+exp(-(-0.5+X))) if _n<=100
cap qui logit Y X C* if _n<=100, asis
if _rc==0 {
qui replace b_reg_‘j’=_b[X] in ‘i’
}
qui count if Y==1 & _n<=100
qui replace epp_‘j’=r(N)/(2+‘j’) in ‘i’
cap qui logit X C* if _n<=100, asis
if _rc==0 {
qui predict ps if _n<=100
cap qui logit Y X ps if _n<=100
}
if _rc==0 {
qui replace b_ps_‘j’=_b[X] in ‘i’
}
keep b_* epp_*
}
}
```

Simulations to demonstrate this (true value = 1)



Remark

- It may seem that we are 'getting something for nothing' on the previous slide.

Remark

- It may seem that we are 'getting something for nothing' on the previous slide.
- It's reassuring to note that there is a price to pay.

Remark

- It may seem that we are 'getting something for nothing' on the previous slide.
- It's reassuring to note that there is a price to pay.
- When adjusting for $\widehat{\pi}(L)$ instead of L, we stop learning anything about the individual 'effects' of variables in L.

Remark

- It may seem that we are 'getting something for nothing' on the previous slide.
- It's reassuring to note that there is a price to pay.
- When adjusting for $\widehat{\pi}(L)$ instead of L, we stop learning anything about the individual 'effects' of variables in L.
- Scientifically, this is usually a price we are happy to pay for the reward of removing the finite sample bias in the parameter of interest.

Other criticisms

Stephen Senn deserves a slide to himself ☺

- Senn et al (Statistics in Medicine, 2007) were critical of propensity score methods, showing that they are always less efficient than traditional regression methods.

Stephen Senn deserves a slide to himself ☺

- Senn et al (Statistics in Medicine, 2007) were critical of propensity score methods, showing that they are always less efficient than traditional regression methods.
- They considered only linear regression, hence missed the issue of finite sample bias.

Stephen Senn deserves a slide to himself ☺

- Senn et al (Statistics in Medicine, 2007) were critical of propensity score methods, showing that they are always less efficient than traditional regression methods.
- They considered only linear regression, hence missed the issue of finite sample bias.
- They also only considered correctly-specified models, hence poor overlap and misspecification were not considered.

Stephen Senn deserves a slide to himself ☺

- Senn et al (Statistics in Medicine, 2007) were critical of propensity score methods, showing that they are always less efficient than traditional regression methods.
- They considered only linear regression, hence missed the issue of finite sample bias.
- They also only considered correctly-specified models, hence poor overlap and misspecification were not considered.
- Stephen Senn (and many others) stress that estimated standard errors should always be adjusted when using a propensity score approach to allow for the fact that the propensity scores were estimated, not known.

Another controversy

Bayesian propensity score

- Another interesting debate is whether or not Bayesians are allowed even to mention the concept of the propensity score!

Bayesian propensity score

- Another interesting debate is whether or not Bayesians are allowed even to mention the concept of the propensity score!
- Logically, it seems that the propensity score is irrelevant to a Bayesian, even when it is known. And yet, many so-called Bayesian propensity score approaches exist...

Bayesian propensity score

- Another interesting debate is whether or not Bayesians are allowed even to mention the concept of the propensity score!
- Logically, it seems that the propensity score is irrelevant to a Bayesian, even when it is known. And yet, many so-called Bayesian propensity score approaches exist...
- See Robins et al *Biometrics* 71(2): 296–9 for an excellent and amusing account.

Finally. . .

Double robust methods, machine learning, longitudinal data

- Arguably the most valuable use of the propensity score is not instead of traditional regression but in addition to it.

Double robust methods, machine learning, longitudinal data

- Arguably the most valuable use of the propensity score is not instead of traditional regression but in addition to it.
- Many, many methods exist that combine the two working models under a so-called double robust approach.

Double robust methods, machine learning, longitudinal data

- Arguably the most valuable use of the propensity score is not instead of traditional regression but in addition to it.
- Many, many methods exist that combine the two working models under a so-called double robust approach.
- As well as being appealing in the sense of having two bites at the cherry, double robust estimators also have many other nice properties such as faster convergence and more tangible analytical inferences.

Double robust methods, machine learning, longitudinal data

- Arguably the most valuable use of the propensity score is not instead of traditional regression but in addition to it.
- Many, many methods exist that combine the two working models under a so-called double robust approach.
- As well as being appealing in the sense of having two bites at the cherry, double robust estimators also have many other nice properties such as faster convergence and more tangible analytical inferences.
- This means that machine learning methods are often made more feasible when used in conjunction with double robust estimators.

Double robust methods, machine learning, longitudinal data

- Arguably the most valuable use of the propensity score is not instead of traditional regression but in addition to it.
- Many, many methods exist that combine the two working models under a so-called double robust approach.
- As well as being appealing in the sense of having two bites at the cherry, double robust estimators also have many other nice properties such as faster convergence and more tangible analytical inferences.
- This means that machine learning methods are often made more feasible when used in conjunction with double robust estimators.
- Another area related to propensity scores is when making causal inferences from longitudinal data in the presence of time-dependent confounders affected by previous exposures.

In summary

PS: why and why not?

The bad news:

- PS methods are not a solution to unmeasured confounding.

PS: why and why not?

The bad news:

- PS methods are not a solution to unmeasured confounding.
- PS methods are not a solution to collider stratification bias.

PS: why and why not?

The bad news:

- PS methods are not a solution to unmeasured confounding.
- PS methods are not a solution to collider stratification bias.
- PS methods are typically less efficient than traditional regression methods when all models are correctly specified.

PS: why and why not?

The bad news:

- PS methods are not a solution to unmeasured confounding.
- PS methods are not a solution to collider stratification bias.
- PS methods are typically less efficient than traditional regression methods when all models are correctly specified.
- PS methods are logically off limits if you're a Bayesian.

PS: why and why not?

The bad news:

- PS methods are not a solution to unmeasured confounding.
- PS methods are not a solution to collider stratification bias.
- PS methods are typically less efficient than traditional regression methods when all models are correctly specified.
- PS methods are logically off limits if you're a Bayesian.

The good news:

PS: why and why not?

The bad news:

- PS methods are not a solution to unmeasured confounding.
- PS methods are not a solution to collider stratification bias.
- PS methods are typically less efficient than traditional regression methods when all models are correctly specified.
- PS methods are logically off limits if you're a Bayesian.

The good news:

- PS methods can give less biased and more honest inference in situations when there is poor overlap.

PS: why and why not?

The bad news:

- PS methods are not a solution to unmeasured confounding.
- PS methods are not a solution to collider stratification bias.
- PS methods are typically less efficient than traditional regression methods when all models are correctly specified.
- PS methods are logically off limits if you're a Bayesian.

The good news:

- PS methods can give less biased and more honest inference in situations when there is poor overlap.
- PS methods suffer far less from finite sample bias in non-linear regression models when the number of confounders is large.

PS: why and why not?

The bad news:

- PS methods are not a solution to unmeasured confounding.
- PS methods are not a solution to collider stratification bias.
- PS methods are typically less efficient than traditional regression methods when all models are correctly specified.
- PS methods are logically off limits if you're a Bayesian.

The good news:

- PS methods can give less biased and more honest inference in situations when there is poor overlap.
- PS methods suffer far less from finite sample bias in non-linear regression models when the number of confounders is large.
- PS methods, particularly adjustment and weighting, extend well to more complex settings.

PS: why and why not?

The bad news:

- PS methods are not a solution to unmeasured confounding.
- PS methods are not a solution to collider stratification bias.
- PS methods are typically less efficient than traditional regression methods when all models are correctly specified.
- PS methods are logically off limits if you're a Bayesian.

The good news:

- PS methods can give less biased and more honest inference in situations when there is poor overlap.
- PS methods suffer far less from finite sample bias in non-linear regression models when the number of confounders is large.
- PS methods, particularly adjustment and weighting, extend well to more complex settings.
- The two approaches can be combined leading to, in some cases, the best of both worlds.

Extra slides

- No interference: The potential outcome for individual i doesn't depend on the hypothetical level that the exposure for individual $j \neq i$ is set to. This means that a single index a on the potential outcome Y_a is sufficient.
- Consistency: $Y_a = Y$ for any individual for whom $A = a$.
- Positivity: For all l and for any $\varepsilon > 0$ such that the density $f_L(l)$ of L evaluated at l is greater than ε, there exists a $\delta > 0$ such that $\delta < \pi(l) < 1 - \delta$.
- Conditional exchangeability: $Y_a \perp\!\!\!\perp A|L, \ a = 0, 1$.

Suppose the regression model is

$$E\left(Y|A,L\right)=g^{-1}\left\{m(A,L)\right\} \qquad (\dagger)$$

where $m(A,L)$ is the linear predictor (and may contain product terms between A and elements in L). Then,

$$\begin{aligned}
E\left(Y_a\right) &= E\left\{E\left(Y_a|L\right)\right\} \\
&= E\left\{E\left(Y_a|A=a,L\right)\right\} \quad \text{(by cond. exch.)} \\
&= E\left\{E\left(Y|A=a,L\right)\right\} \quad \text{(by consistency*)} \\
&= E\left[g^{-1}\left\{m(a,L)\right\}\right] \quad \text{(by } (\dagger)\text{)}
\end{aligned}$$

*positivity ensures that this conditional expectation exists.
Thus,

$$E\left(Y_1-Y_0\right)=E\left[g^{-1}\left\{m(1,L)\right\}-g^{-1}\left\{m(0,L)\right\}\right]$$