ONLINE APPENDIX

The Future Strikes Back: Using Future Treatments To Detect and Reduce Hidden Bias

Felix Elwert
University of Wisconsin-Madison

Fabian T. Pfeffer
University of Michigan

APPENDIX A. GOTTSCHALK'S FUTURE-TREATMENT STRATEGY

A third future treatment estimator was introduced by Gottschalk (1996). Like Mayer (1997), Gottschalk (1996) premises his analysis on the DGP of Figure 2 and derives a future treatment estimator from its covariance structure. Unlike Mayer, Gottschalk explicitly motivates his estimator with an argument that resembles our difference logic: to use the association between F and Y first to measure and then to subtract bias in the association between T and Y.

Definition 4 (Gottschalk's estimator[1]): Gottschalk's estimator for the causal effect of T on Y, b, is given by

$$b_G = b_{YT} - \sigma_{YF.T} = \sigma_{YT} - (\sigma_{YF} - \sigma_{YT}\sigma_{TF}) \,. \qquad \text{(A.1)}$$

This estimator is similar, but not identical, to the Mayer/difference estimator. Whereas the difference estimator subtracts two partial regression coefficients, $b_D = b_{YT.F} - b_{YF.T}$, Gottschalk subtracts a conditional covariance from the unadjusted regression of Y on T.

Like Mayer's (1997) estimator, Gottschalk's estimator is biased when U affects T and F differently, $a \neq d$.

Result A.1 (bias of Gottschalk's [1996] estimator in the best case): Gottschalk's estimator is biased when data are generated by the model in Figure 2,

$$b_G = b + ac\left(1 - \frac{d}{a} + ad\right) = b + B_{OLS}M_G, \qquad \text{(A.2)}$$

But in contrast to Mayer's (1997) estimator, this estimator is not unbiased in the best-case model of Figure 2 when $a = d$.

Corollary A.1: Gottschalk's estimator remains biased when data are generated by the model in Figure 2 and U affects T and F in the same way, $a = d$,

$$b_G = b + a^3c \neq b. \qquad \text{(A.3)}$$

Like the Mayer/difference estimator, but unlike our control estimator, Gottschalk's estimator can increase rather than decrease the bias from unobserved confounding when $a \neq d$. Like the Mayer/difference estimator, Gottschalk's estimator strictly increases bias when a and d have opposite signs. Interestingly, however, unlike Mayer's estimator, Gottschalk's estimator is mostly bias reducing when a and d share the same sign and a is strong or moderately strong. Indeed, for magnitudes of $|a|$ larger than about 0.42 (regardless of the value of d), Gottschalk's estimator is strictly bias-reducing.

[1] Our notation is superficially different from Gottschalk's original notation since we assume standardized variables.

APPENDIX B. FUTURE TREATMENTS AS INSTRUMENTAL VARIABLES

This appendix evaluates the circumstances under which future treatments can, or cannot, serve as instrumental variables (IV). Instrumental variables analysis is a popular strategy for removing bias from unobserved confounding. With a valid IV, F, the causal effect of treatment T on outcome Y in linear DPGs is consistently estimated by the covariance ratio

$$b_{IV} = \frac{\sigma_{FY}}{\sigma_{FT}} . \quad \text{(B.1)}$$

IV analysis in linear models requires two assumptions: (1) the instrumental variable must be associated with T ("relevance"); and (2) the IV must be associated with the outcome only via paths that include the causal effect of the treatment on the outcome ("exclusion") (Brito and Pearl 2002). If both assumptions are met, we say that the instrumental variable is valid.

Future treatments are not valid instrumental variables in any of the DGPs considered in the main body of this paper. The key assumption motivating our future-treatment strategies–that F is a proxy for the unobserved confounder, U–violates the exclusion assumption because it induces an association between F and Y via the open path $F \leftarrow U \rightarrow Y$.

For example, the instrumental variables estimator, using F as instrumental variable, in data generated by Figure 2 would evaluate to

$$b_{IV} = \frac{\sigma_{FY}}{\sigma_{FT}} = \frac{bad+cd}{ad} = b + \frac{c}{a} \neq b . \quad \text{(B.2)}$$

Recalling that all standardized path parameters lie in the interval $(-1, 1)$, it is obvious that the instrumental variables estimator in this case is strictly more biased than the unadjusted OLS estimator because

$$|B_{OLS}| = |ac| < |\frac{a}{c}| = |B_{IV}|, \text{ for all } a, c \neq 0 . \quad \text{(B.3)}$$

Nonetheless, future treatments have previously been used as instrumental variables, when F was assumed *not* to be a proxy for the unobserved confounders U. For example, Duncan et al. (1997) cautiously defend such a scenario for the estimation of causal neighborhood effects. In their application, Y is children's test scores, T is parents' neighborhood environment while living with the child, and F is parents' neighborhood environment after the child has moved out. Their central assumption is that U can be partitioned into two independent components, as shown in Figure B.1: U_1 represents unobserved parenting quality, which affects child test scores and neighborhood choice while the child lives at home; and U_2 represents parent's residential preferences aside from child rearing considerations.

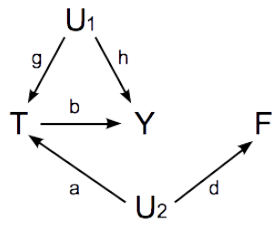


Figure B.1. Model in which future treatments, F, are a valid instrumental variable for the effect $T \to Y$, because the unobservables, U, are suitably partitioned.

If this model is true, then F is indeed a valid instrument for the effect of T on Y, and the instrumental variables estimator evaluates to

$$b_{IV} = \frac{\sigma_{FY}}{\sigma_{FT}} = \frac{abd}{ad} = b. \quad \text{(B.4)}$$

As Duncan et al. (1997) have noted, this model may not be especially robust. Instrumental variables estimation would fail under small modifications of the original model, e.g., if parenting, U_1, is associated with future neighborhood conditions (ibid: p. 249), perhaps because concerned parents move to better neighborhoods, or if parent's neighborhood preferences, U_2, are associated with other unobserved factors, such as parental ability, that also affect child test scores (ibid: p. 230). We capture these scenarios in Figures B.2a and B.2b, in which the instrumental variable estimator evaluates to $b_{IV} = b + \frac{ih}{gi+ad} \neq b$ and $b_{IV} = b + \frac{c}{a} \neq b$, respectively. In both of these more elaborate scenarios, F is not a valid instrumental variable because it is a proxy for one or another unobserved confounder, U_1 or U_2, of T and Y, and hence violates the exclusion condition via the open paths $F \leftarrow U_1 \to Y$ and $F \leftarrow U_2 \to Y$, respectively.

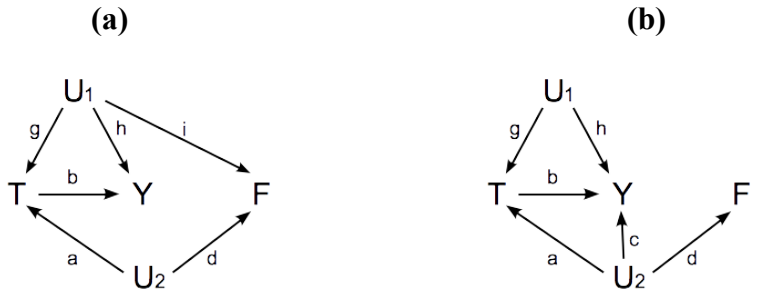


Figure B.2. Two models in which future treatments are not valid instrumental variables for the effect $T \to Y$.

We further note that F also fails as an instrumental variable even if F is not a proxy for unobserved confounders of T and Y, namely in the presence of true state dependence or selection. True state dependence would occur in Duncan et al.'s (1997) scenario if parents develop a taste for the kind of neighborhood they live in (Deluca 2012), as shown in Figure B.3a. In this scenario, the exclusion assumption is violated because F is associated with Y via the open path $F \leftarrow T \leftarrow U_1 \to Y$ (i.e. via a path that does not include the causal effect of T on Y). Consequently, the instrumental variables estimator is biased, $b_{IV} = b + \frac{fgh}{ad+f} \neq b$.

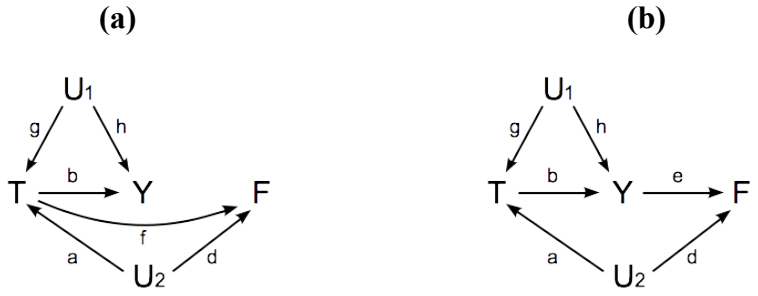


Figure B.3. True state dependence and selection invalidate future treatments as instrumental variables for the effect $T \rightarrow Y$.

Selection would occur if children's test scores affect parents' future residential choice, as shown in Figure B.3b (an admittedly far-fetched proposal, unless, e.g., families relocate in response to children experiencing academic difficulties at a local school). Here, the exclusion condition would be violated because F is directly associated with Y, and the instrumental variable estimator evaluates to

$$b_{IV} = b + \frac{e}{ad+e(b+gh)} \neq b \; . \qquad \text{(B.5)}$$

In a final twist, although true state dependence ($T \rightarrow F$) and selection ($Y \rightarrow F$) invalidate the use of future treatments as instrumental variables, Chan and Kuroki (2010) have shown that descendants of T and Y (which could include future values of the treatment) can sometimes be used to remove unobserved confounding in linear models if true state dependence and selection are suitably mediated in more complicated DGPs. Their results are akin, but not identical, to instrumental variables analysis. To the best of our knowledge, Chan and Kuroki's methodological results have not yet been used in empirical applications.

APPENDIX C. PROOFS OF RESULTS 11 AND 12

The temporal order of variables is $(O, V, \boldsymbol{Q}, \boldsymbol{X}) \prec T \prec Y \prec F$; the ordering of the variables in $(O, U, \boldsymbol{Q}, \boldsymbol{X})$ is irrelevant. As is usual in working with graphical causal models, we use the rules of d-separation and assume faithfulness (Pearl 2009).

Proof of Result 11:
We first show that, under Assumption 1, unobserved confounding between T and Y conditional on $\boldsymbol{X}$ implies $F \neg\perp Y|(T, \boldsymbol{X})$.

1. Confounding between T and Y conditional on $\boldsymbol{X}$ (which may b empty) implies a d-connected path, $\pi_{TY}(O)$, from T to Y via an unobserved parent of T, $O \notin \boldsymbol{X}$: $T \leftarrow O$ ... $\rightarrow Y$. O is a non-collider on $\pi_{TY}(O)$, and all variables in $\boldsymbol{X}$ that are on $\pi_{TY}(O)$, if any, are colliders on $\pi_{TY}(O)$.

2. Now show that unobserved confounding between T and Y along the path $\pi_{TY}(O)$ implies the existence of a path between F and Y, π_{FY} that is d-connected conditional on $(T, \boldsymbol{X})$. We distinguish two cases:

2.1. Suppose that O is V. Then Assumption 1 and confounding along the path $\pi_{TY}(O)$ imply the existence of a path from F to Y via O, $\pi_{FY}(O)$: $F \leftarrow \cdots O$ ... $\rightarrow Y$, which is d-connected conditional on $(T, \boldsymbol{X})$ because (a) the path segment $F \leftarrow \cdots O$ of $\pi_{FY}(O)$ is d-connected by Assumption 1; (b) the path segment O ... $\rightarrow Y$ of $\pi_{FY}(O)$ is d-connected because it is a path segment of $\pi_{TY}(O)$, which is d-connected by premise above; and (c) the entire path $\pi_{FY}(O)$ is d-connected because the variable $O \notin \boldsymbol{X}$ that connects its two segments is either an unconditioned non-collider on $\pi_{FY}(O)$, or a collider on $\pi_{FY}(O)$ whose descendant T is conditioned.

2.2. Suppose that O is not V. Then Assumption 1 and confounding along $\pi_{TY}(O)$ imply the existence of a path from F to Y via V, $\pi_{FY}(V)$: $F \leftarrow \cdots V \rightarrow T \leftarrow O$ ... $\rightarrow Y$, which is d-connected conditional on $(T, \boldsymbol{X})$ because (a) the segment $F \leftarrow \cdots V \rightarrow T$ of $\pi_{FY}(V)$ is d-connected by Assumption 1, (b) the segment $T \leftarrow O$ ... $\rightarrow Y$ of $\pi_{FY}(V)$ is d-connected because it is the d-connected confounding path $\pi_{TY}(O)$, and the entire path of $\pi_{FY}(V)$ is open because we condition on the collider T that connects its two segments.

2.3. Since two variables that are d-connected via at least one path are statistically associated under the usual conditions (Verma and Pearl 1988), d-connection of either $\pi_{FY}(O)$ or $\pi_{FY}(V)$ conditional on $(T, \boldsymbol{X})$ implies $F \neg\perp Y|(T, \boldsymbol{X})$

3. By contraposition, conditional independence $F \perp Y|(T, \boldsymbol{X})$ implies the absence of confounding. ■

Proof of Result 12:
Since Assumption 2 implies Assumption 1, the first part of Result 12 is given by Result 11. To prove the second part, we need to show that $F \neg\perp Y|(T, \mathbf{X})$ implies the presence of unobserved confounding between T and Y, conditional on $\boldsymbol{X}$.

1. $F \neg\perp Y|(T, \boldsymbol{X})$ implies the existence of a path between F and Y, π_{YF} that is d-connected conditional on T and $\boldsymbol{X}$.

1.1. There are four possible types of paths between F and Y, depending on whether they start with an arrow into or out of Y and end with an arrow into or out of F.

Paths π_{YF} of the type $Y \rightarrow \cdots \rightarrow F$ that do not contain colliders are ruled out by Assumption 3. Such paths containing colliders, C, are d-separated because the collider is unconditioned, $C \notin (T, \boldsymbol{X})$, as $\boldsymbol{X} \prec T \prec Y \prec C$.

Paths π_{YF} of the types $Y \leftarrow \cdots \leftarrow F$ and $Y \rightarrow \cdots \leftarrow F$ are d-separated because they contain an unconditioned collider, $C \notin (T, \boldsymbol{X})$, as $\boldsymbol{X} \prec T \prec F \prec C$.

Therefore, for π_{YF} to be d-connected it must contain arrows into Y and F, $Y \leftarrow \cdots \rightarrow F$.

1.2. For π_{YF} to be d-connected conditional on $(T, \boldsymbol{X})$ it must end with an arrow from an unobserved variable, R, into F, $R \rightarrow F$, because paths π_{YF} that end with $\boldsymbol{X} \rightarrow F$ or $T \rightarrow F$ would be d-separated conditional on $(T, \boldsymbol{X})$, and paths P_{YF} that end with $Y \rightarrow F$ are ruled out by Assumption 3. Clearly, R is a non-collider on π_{YF}.

2. By assumption 2, $R \in \boldsymbol{Q}$, which implies the d-connected path π_{TF}, $T \leftarrow R \rightarrow F$.

3. It follows that there exists a non-causal path π_{TY}, $T \leftarrow R \ldots \rightarrow Y$, which is d-connected conditional on $\boldsymbol{X}$ since its segment $R \ldots \rightarrow Y$ is a segment on π_{YF}, which is d-connected conditional on $(T, \boldsymbol{X})$, and R is an unconditioned non-collider on π_{TY}.

4. Since two variables that are d-connected via at least one path are statistically associated under the usual conditions (Verma and Pearl 1988), the d-connected path π_{TY} represents unobserved confounding between T and Y conditional on $\boldsymbol{X}$. ■

APPENDIX D. REPLICATION OF MAYER (1997)

Table D.1. Descriptives
Means (standard deviations in parentheses); weighted

	Main Sample		Analytic Sample	
	Mayer (1997)	Replication	Mayer (1997)	Replication
Years of Education	12.793	12.838	(a)	12.886
	(1.940)	(1.928)	(a)	(1.957)
Log family income	10.687	11.840	(a)	11.938
	(0.572)	(0.447)	(a)	(0.357)
Log family size	1.647	1.576	(a)	1.609
	(0.331)	(0.331)	(a)	(0.338)
Parent is black	0.141	0.151	(a)	0.170
	(0.347)	(0.358)	(a)	(0.376)
Parent's age	40.127	40.691	(a)	40.899
	(6.163)	(5.908)	(a)	(5.646)
Parent's years of education	12.590	12.593	(a)	12.663
	(2.722)	(2.768)	(a)	(2.859)
Child is a boy	0.481	0.494	(a)	0.717
	(0.498)	(0.500)	(a)	(0.450)
Observations	3,275	3,357	1,853	1,513

Estimates as reported in Mayer (1997), Table A.2 (pp.162-163); (a) Estimates not directly reported in Mayer (1997)

Table D.2. Full Regression Results
OLS coefficient estimates (standard errors in parentheses); weighted

	Main Sample Unstandardized Coefficients		Analytic Sample Standardized Coefficients	
	Mayer (1997)	Replication	Mayer (1997)	Replication
Log family income	0.784	0.749	0.186	0.185
	(0.065)	(0.074)	(a)	(0.038)
Log family size	-0.714	-0.700	(a)	-0.157
	(0.091)	(0.092)	(a)	(0.025)
Parent is black	0.257	0.276	(a)	0.031
	(0.091)	(0.088)	(a)	(0.034)
Parent's age	0.023	0.033	(a)	0.115
	(0.005)	(0.005)	(a)	(0.025)
Parent's years of education	0.235	0.293	(a)	0.476
	(0.013)	(0.011)	(a)	(0.026)
Child is a boy	-0.032	-0.181	(a)	0.008
	(0.059)	(0.057)	(a)	(0.026)
Constant	1.651	0.082	(a)	-0.042
	(0.652)	(0.824)	(a)	(0.030)
N	3,275	3,357	1,853	1,513
R^2	0.265	0.274	(a)	0.301

Estimates as reported in Mayer (1997), Table B.6 (p.174) for main sample, Table 5.3 (p. 92) for analytic sample; (a) Estimates not directly reported in Mayer (1997)

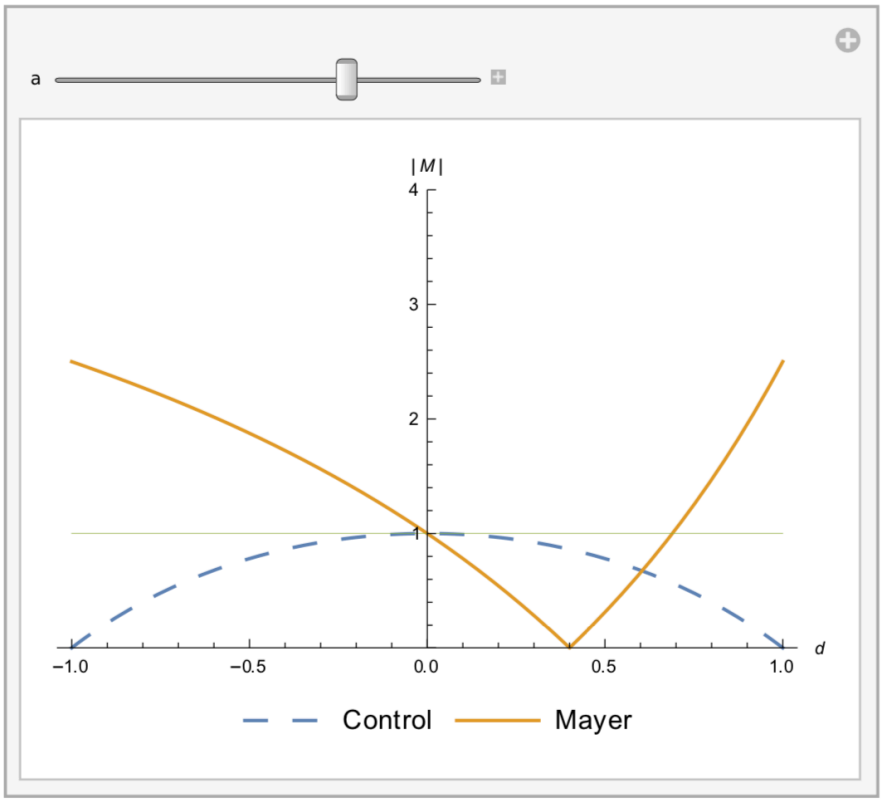
APPENDIX E. ANIMATION OF ANALYTICAL RESULTS

Mathematica code to create interactive animations of our analytical results is available in the replication package to this article (http:// doi.org/10.3886/E104060V1). Here, we provide the code and still images. Using the Mathematica version enables readers to change the sliders to explore the graphs further.

E.1 BEST-CASE SCENARIO: ABSOLUTE BIAS FACTORS: CONTROL, MAYER

Figure 3 in the paper

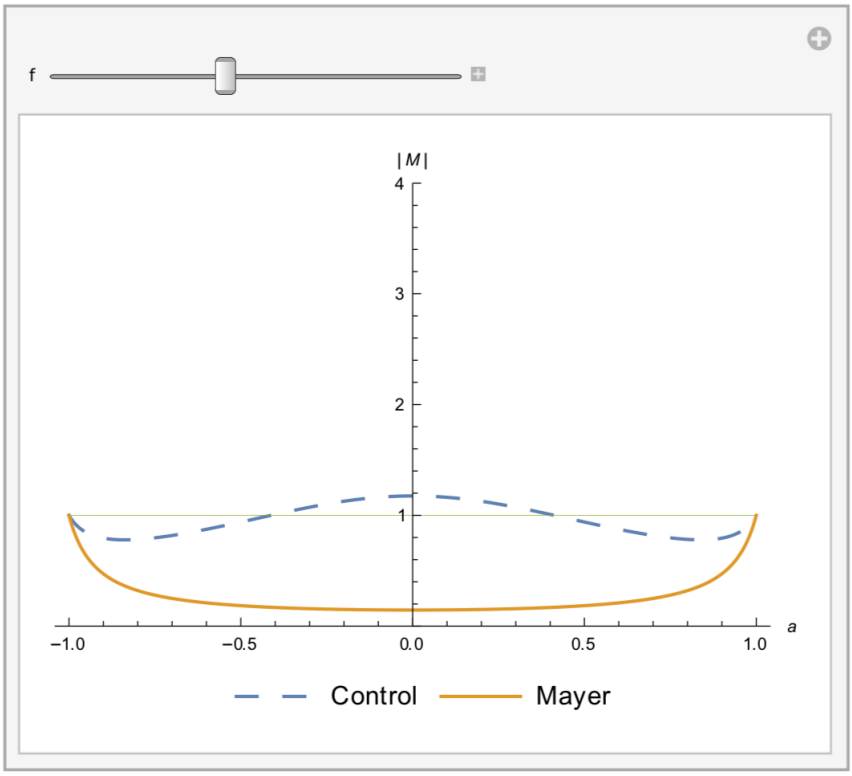
```
Manipulate[Plot[{Abs[(1 - d^2)/(1 - a^2 d^2)], Abs[((a - d))/(a - a^2 d)], 1},
  {d, -1, 1}, PlotRange → {0, 4}, PlotStyle → {Dashing[Large], Dashing[None], Thin},
  AxesLabel → {d, Abs[M]}, PlotLegends → Placed[{"Control", "Mayer"}, Bottom]], {a, -1, 1}]
```



E.2 BIAS MULTIPLIERS OF THE CONTROL AND MAYER/DIFFERENCE ESTIMATORS FOR a=d

Table 1 & Result 8 in the paper

```
Manipulate[
 Plot[{Abs[(1 - f - f^2 - a^2 f - a^2)/(1 - (f + a^2)^2)], Abs[(-f - f^2 - a^2 f)/(1 - (f + a^2)^2)], 1},
  {a, -1, 1}, PlotRange → {0, 4}, PlotStyle → {Dashing[Large], Dashing[None], Thin},
  AxesLabel → {a, Abs[M]}, PlotLegends → Placed[{"Control", "Mayer"}, Bottom]], {f, -1, 1}]
```

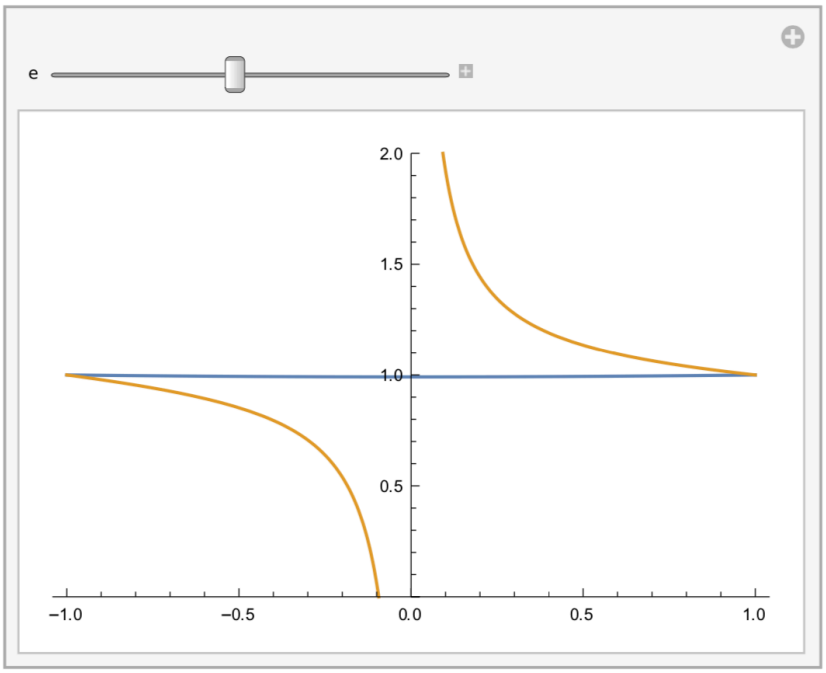


E.3 PURE SELECTION (NO CONFOUNDING)

Equations 17 & 18 in the paper

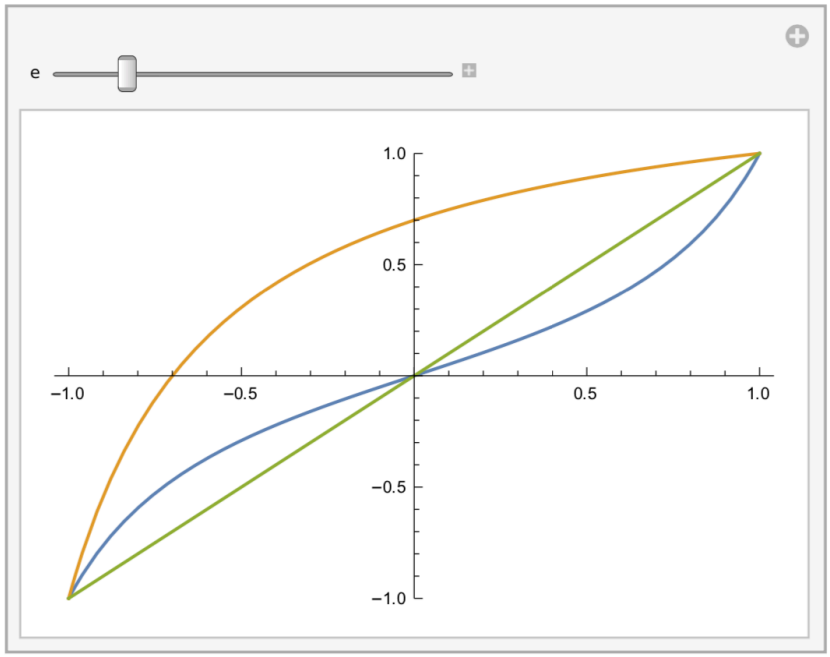
Bias Factor Control vs Mayer/Difference (note: OLS bias is 0) as function of selection, e.

```
Manipulate[Plot[{(1 - e e)/(1 - e e b b), (b - e)/(b - b b e)}, {b, -1, 1}, PlotRange → {0, 2}], {e, -1, 1}]
```



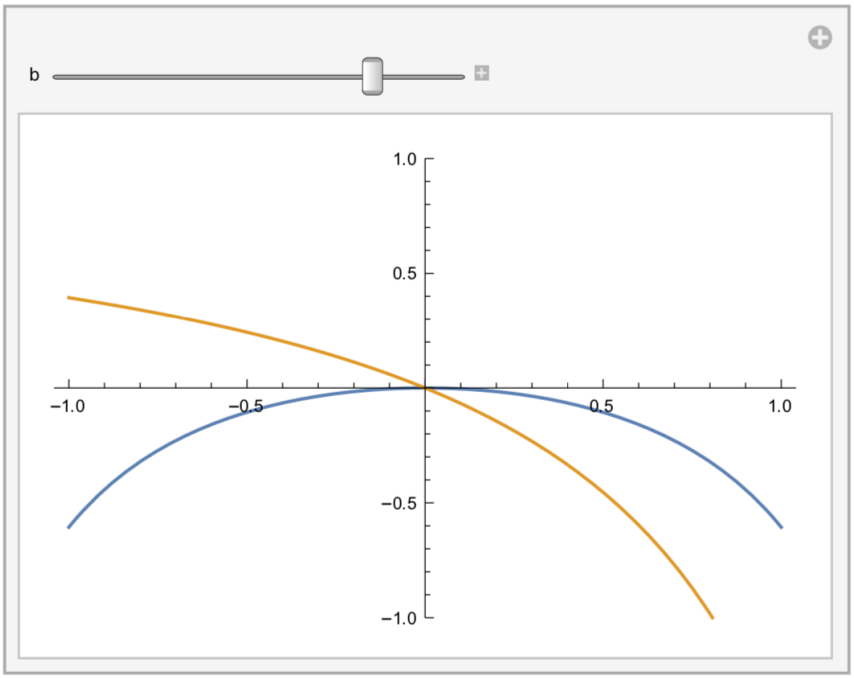
Estimates of Control and Mayer Estimators as function of selection, e

```
Manipulate[
  Plot[{b (1 - e e)/(1 - e e b b), b (b - e)/(b - b b e), b}, {b, -1, 1}, PlotRange → {-1, 1}], {e, -1, 1}]
```



Bias in Control and Mayer Estimators as function of treatment eect, b

```
Manipulate[
  Plot[{b (1 - e e) / (1 - e e b b) - b,  b (b - e) / (b - b b e) - b}, {e, -1, 1}, PlotRange → {-1, 1}], {b, -1, 1}]
```



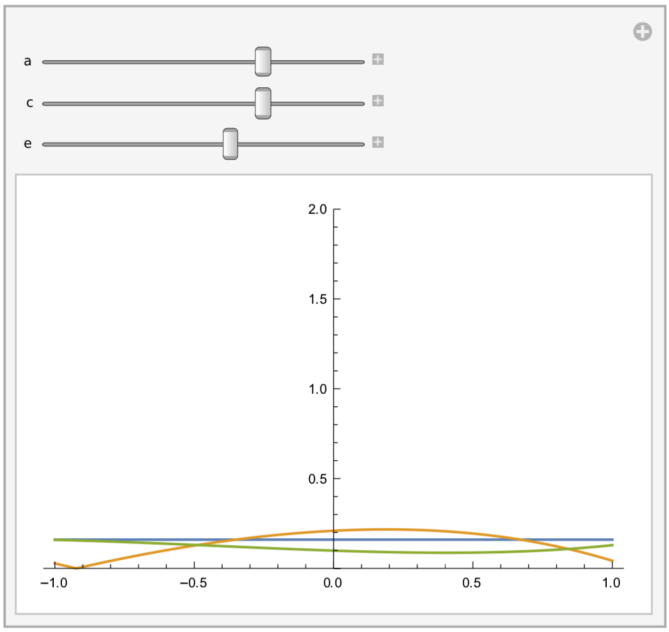
Bias of Control and Mayer Estimators with a=d

```
Manipulate[
  Plot[{b (1 - e e) / (1 - e e b b),  b (b - e) / (b - b b e) , b}, {b, -1, 1}, PlotRange → {-1, 1}], {e, -1, 1}]
```

E.4 SELECTION WITH CONFOUNDING

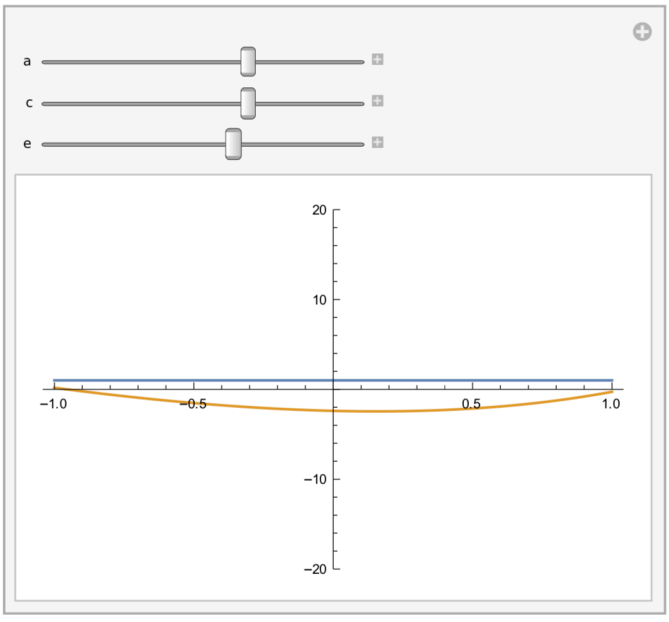
OLS, Mayer/Difference, Control methods bias (absolute bias), a=d

```
Manipulate[Plot[{Abs[a c], Abs[(b (1 - a^2) - e)/((1 - a^2) - e (b - a c)) - b], Abs[(b + a c - (a^2 b + a c + e) (a^2 + b e + a c e))/(1 - (a^2 + b e + a c e)^2) - b]},
 {b, -1, 1}, PlotRange → {0, 2}], {a, -1, 1}, {c, -1, 1}, {e, -1, 1}]
```



Bias factor of difference estimator

```
Manipulate[Plot[{1, ((1 - b^2 + a b c) e)/(a c (-1 + a^2 + b e - a c e))}, {b, -1, 1}, PlotRange → {-20, 20}],
 {a, -1, 1}, {c, -1, 1}, {e, -1, 1}]
```

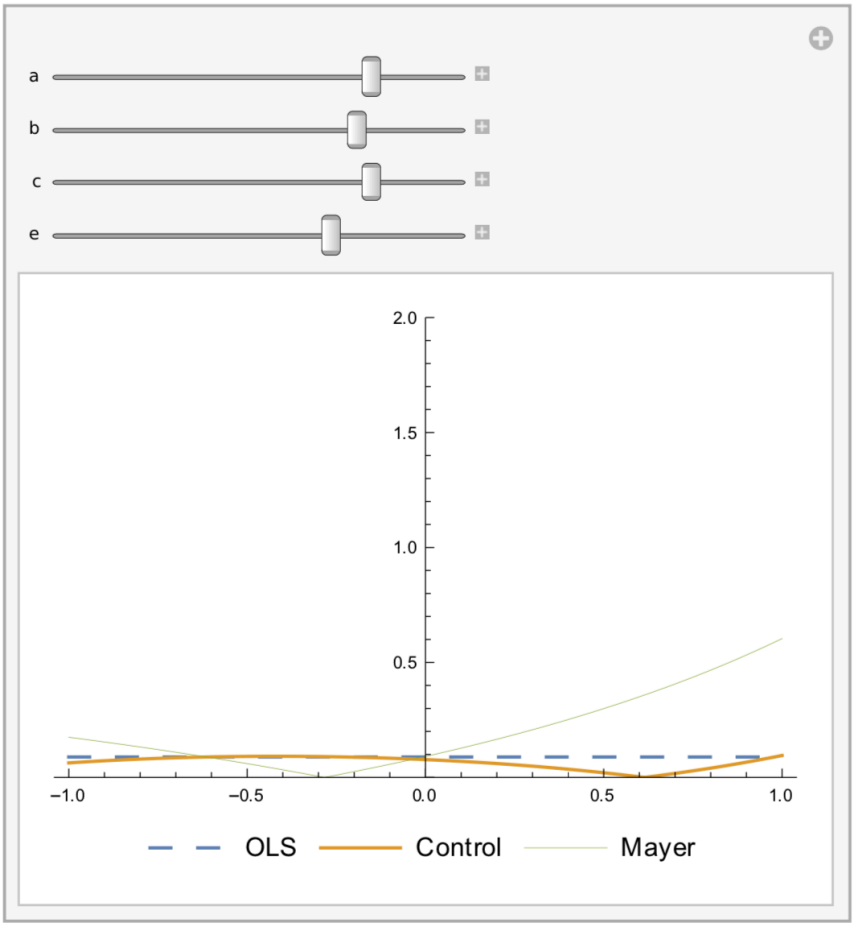


E.5 SELECTION WITH A≠D

Table 2, Results 9&10

Graph absolute bias, OLS, Control Difference

```
Manipulate[
 Plot[{Abs[a c], Abs[(b + a c - (e + d c + a d b) (b e + a c e + a d))/(1 - (b e + a c e + a d)^2 - b]}, Abs[(b + a c - (b a d + e + c d))/(1 - (b e + a c e + a d)) - b]},
  {d, -1, 1}, PlotStyle → {Dashing[Large], Dashing[None], Thin},
  PlotLegends → Placed[{"OLS", "Control", "Mayer"}, Bottom], PlotRange → {0, 2}],
 {a, -.5, .5}, {b, -.5, .5}, {c, -.5, .5}, {e, -.5, .5}]
```



Graph absolute bias *factors*, OLS (ref), Control (blue), Difference (orange)

```
Manipulate[Plot[
 {Abs[((b + a c - (e + d c + a d b) (b e + a c e + a d))/(1 - (b e + a c e + a d)^2) - b)/(a c)], Abs[((b + a c - (b a d + e + c d))/(1 - (b e + a c e + a d)) - b)/(a c)], 1},
 {d, -1, 1}, PlotRange → {0, 4}, PlotStyle → {Dashing[Large], Dashing[None], Thin},
 AxesLabel → {d, Abs[S]}, PlotLegends → Placed[{"Control", "Mayer"}, Bottom]],
{a, -.5, .5}, {b, -.5, .5}, {c, -.5, .5}, {e, -.5, .5}]
```