Appendix to Regularized Regression Can Reintroduce Backdoor Confounding: The Case of Mass Polarization

S1 Other models

In the main text of the paper we discuss the performance of model (8). There are, of course, many other ways that we could specify models that examine the response rate and polarization relationship. In this appendix we examine the performance of the original CF1 model, OLS estimates of model (8), and three other models proposed by MP.

CF1 model the relationship between survey response rates and polarization as follows:

$$y_{it} = \beta_0 + \beta_1 rr_{it} + \beta_2 congress_t + \epsilon_{it} \tag{S1}$$

Where rr_{it} is the response rate of a survey i observed at time t and $congress_t$ is congressional polarization measured at time t. For the contact and cooperation rate versions of these models the $\beta_1 rr_{it}$ term in each model is replaced with two terms, $\beta_4 contact_{it}$ and $\beta_5 cooperation_{it}$ respectively. In their critique of CF1, MP show that, because both response rates and polarization trend over time, using model (S1) will produce spurious results for the effect of response rates on polarization, even if there is no real causal relationship between y_{it} and rr_{it}. MP suggest a number of ways to control for this possibility, and that once they do so, there is no statistical evidence of a relationship between response rates and polarization.

The first specification suggested by MP adds a $year_t$ term, which is model (8) discussed in the main paper:

$$y_{it} = \beta_0 + \beta_1 rr_{it} + \beta_2 congress_t + \beta_3 year_t + \epsilon_{it} \tag{8}$$

In the main text of the paper we report the estimates for this model using OLS and variety of regularization approaches. Here we focus on unregularized estimators.

Second, MP suggest a year fixed-effects model with a separate dummy α_t for each year in the data:

$$y_{it} = \beta_1 rr_{it} + \beta_2 congress_t + \alpha_t + \epsilon_{it} \tag{S2}$$

Third, MP suggest a year random intercepts model, which includes random intercepts, γ_t, for each year and the year mean of the response rate $\bar{rr}_t$ as a predictor in order to satisfy the requirement of conditional independence of fixed and random components of the model:[5]

$$y_{it} = \beta_1 rr_{it} + \beta_2 congress_t + \gamma_t + \beta_6 \bar{rr}_t + \epsilon_{it} \tag{S3}$$

[5] Bafumi, Joseph, and Andrew Gelman. 2007. “Fitting Multilevel Models When Predictors and Group Effects Correlate.” SSRN Electronic Journal.

Finally, MP suggest using a flexible function form for time, which we implement using a generalized additive model (GAM) with a smoothed term $S(year_t)$ for year:

$$y_{it} = \beta_0 + \beta_1 rr_{it} + \beta_2 congress_t + S(year_t) + \epsilon_{it} \tag{S4}$$

In addition to the models suggested by MP, we estimate a model suggested by Lebo and Weber (2015) for repeated cross section (RCS) data—the ARIMA-MLM. Lebo and Weber focus their discussion on the ARFIMA-MLM model, but suggest that for shorter time series (such as CF2's), the ARIMA approach is more appropriate. The ARIMA-MLM approach takes two stages. First, the RCS data is summarized at the level of each timepoint. In our case this means collapsing polls into yearly averages for each variable. Each variable in the model (independent and dependent variable) is then represented by an $ARIMA(p, d, q)$ model where p is the number of time lags for the autoregressive part of the model, d is the number of differencing steps required to achieve a stationary time series, and q is the order of the lagged errors term in the moving average model (the number of lagged parameters used to predict the current value). The values of p, d, and q are picked to minimize AIC (with small sample size correction) for each variable.

The ARIMA model for each variable is then used to predict the values for each timepoint, and these fitted values are used to render the time series stationary. The individual level observations are then filtered using the ARIMA filtered data and estimated using a multilevel model, in much the same way as model (8).

The response, contact, and cooperation rate coefficients from these models are shown in figure S1. The OLS estimates for model (S1) using CF2's data are very similar to CF1's original results. The additional observations in the CF2 data make very little difference—in all cases the CF1 estimate for $\beta_1 rr_{it}$ is well within the 95% CIs for the new estimate.

Across the models suggested by MP the CIs either overlap zero or the results are in the opposite direction to those claimed by CF2. The only results in line with the broad theory that survey response bias has inflated estimates of polarization are the contact rate coefficients in the FE (S2) and HLM (S3) models for civil rights——the type of response bias that is *not* theoretically expected to induce bias (because people who are not reached do not know the survey's content) for an issue area that CF2 say they do not expect to see affected by response bias. Even this result is inconsistent between models and is statistically insignificant in models (8), (S4), and the ARIMA-MLM. It is notable that the most advanced method of accounting for time-series issues (ARIMA-MLM) uniformly finds null results for response rate and cooperation rate.

The cumulative result of the models suggested by MP is that there is no evidence that declining response/cooperation rates have inflated our estimates of polarization.

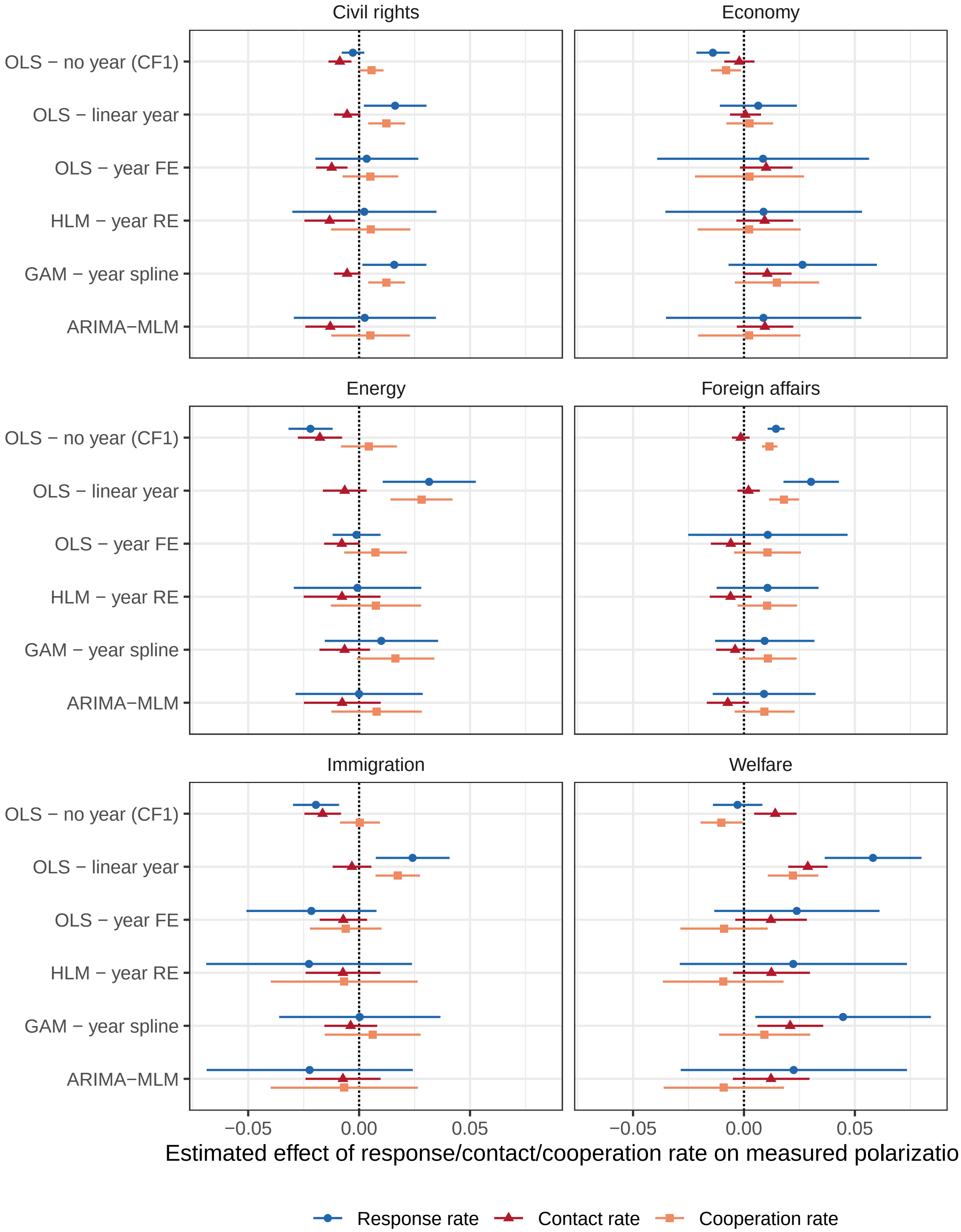


Figure S1: Key coefficients for response/cooperation/contact rates using different model specifications and estimators . See appendix R7 for full tables.

S2 Random walks and correlation with time

In order to demonstrate that random walk processes are likely to be correlated with time, we simulate a random walk process starting at zero, where each step is a random draw from the standard normal distribution, such that the position of variable X at time t can be calculated as follows:

$$
\begin{aligned}
X_t &= X_0 + \sum_{t=1}^{n} Z_t \\
X_0 &= 0 \\
Z_t &\sim \mathcal{N}(0, 1)
\end{aligned}
\tag{S5}
$$

We draw 10,000 simulations of 15 periods each. To illustrate the nature of these random walks, we show a random subset of 100 of these walks in S2. As this illustration shows, although the expected value of our random walk process is zero, the walks spread out as time increases. In our case, because we have simulated the random steps to have a variance of one, and the variance sum law states that the combined variance of independent processes is their sum, $Var(X_t) = t$.

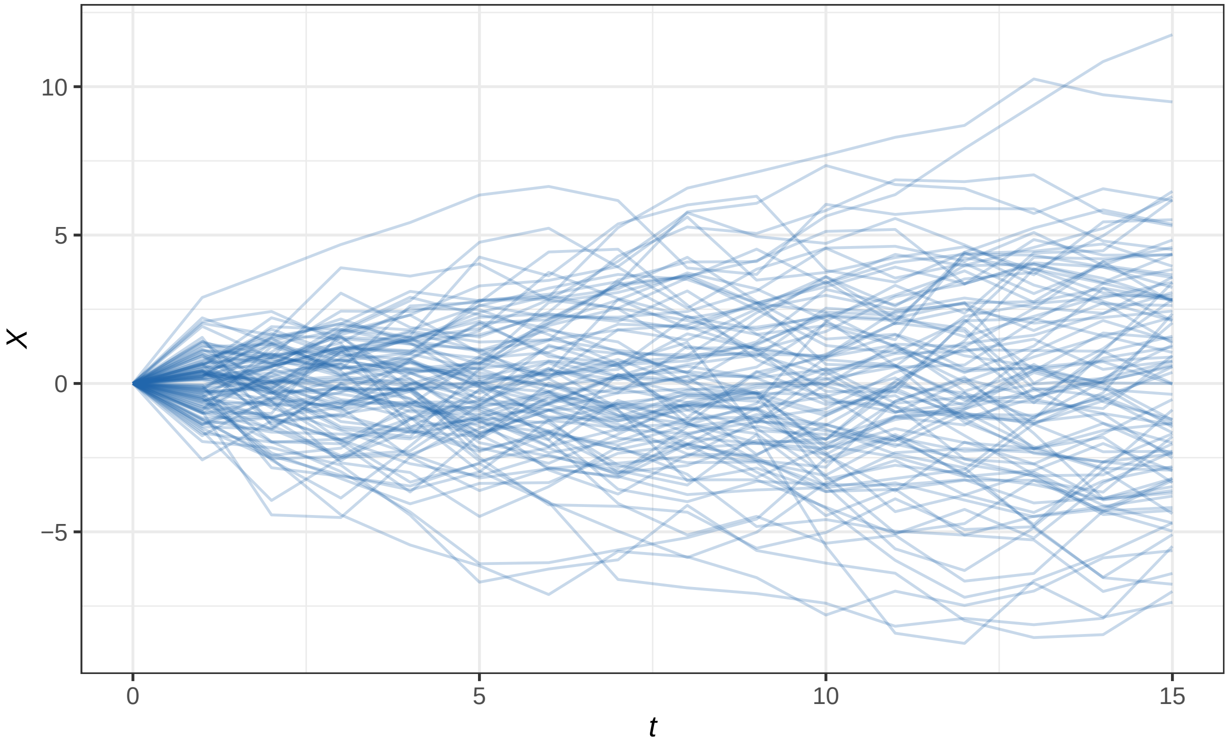


Figure S2: Random walks

Because the expected value of the random walk process is zero, the expected correlation between X and time is also zero. However, because variance increases as a function of time, many individual random walks will be highly correlated with time, as illustrated in figure S3. In aggregate these correlations will cancel out. However the average *absolute* correlation is quite high (0.6).

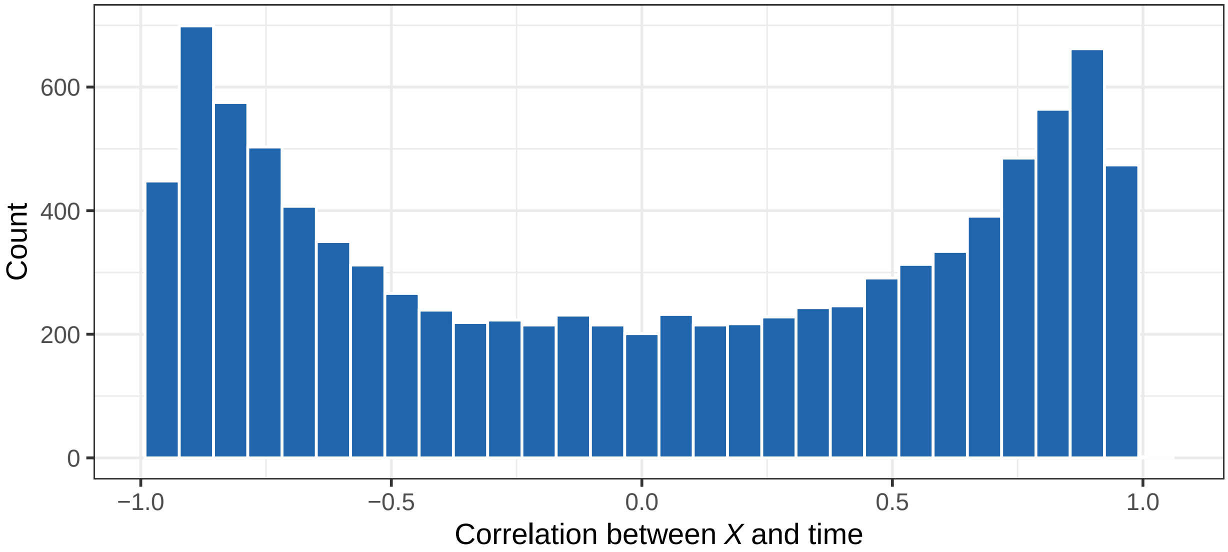


Figure S3: Random walk correlation histogram

If two variables are correlated with time, then they will also be correlated with one another, even if they are independent processes. This is illustrated in figure S4, which shows the distribution of correlations between 10,000 randomly chosen pairs of random walks in our simulations. Again the expected correlation is zero because positive and negative correlations cancel out, but the average absolute correlation is quite high (0.43).

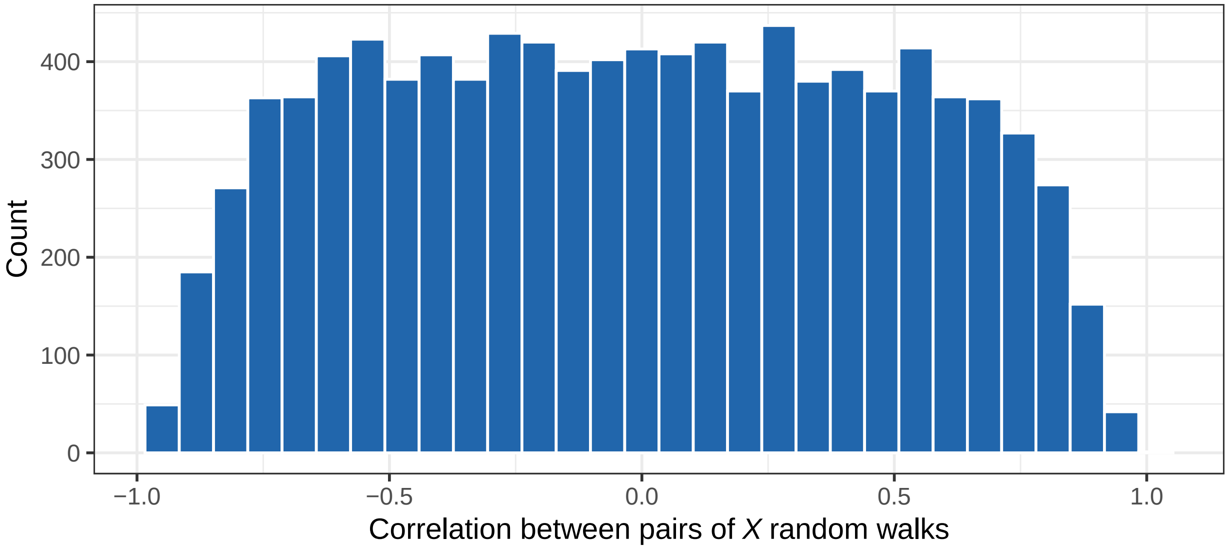


Figure S4: Random walk correlation histogram

S3 Demonstration of regularization inducing effects via the backdoor path

As we discuss in the main text, and show in our simulations, one of the consequences of using regularized regression is that it can induce effects in correlated variables that would otherwise be zero. Our own experience suggests it may not be immediately obvious to many readers why this is the case, and so to aid intuition we have constructed some simple demonstrations using simulated data.

S3.1 Reopening backdoor paths with ridge regression

First, we examine the case of ridge regression, using a single simulated dataset. We simulate 130 observations using a similar DGP as our other simulations:

$$\begin{bmatrix} X \\ Z \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix} \right) \tag{S6}$$

$$Y \sim \beta_x X + \mathcal{N}(0, 0.5)$$

The only difference compared to our earlier DGP (5) is that we use a lower error variance in order to make the demonstration clearer. Key to our demonstration is that X has a positive effect on Y, and Z is negatively correlated with X but has no effect on Y. We estimate the same model as in our main simulations, using OLS and ridge regression with a λ penalty of one:

$$y_i = \beta_0 + \beta_X X_i + \beta_Z Z_i + \epsilon_i \tag{6}$$

The six panels of figure S5 illustrate the way in which ridge regression induces an apparent effect of Z on Y. Panel A shows the relationship between X and Y and the regression lines that would be fit to this data by OLS and ridge regression. While the OLS provides the best fitting line to the data, the regularized ridge line is considerably shallower (i.e. the slope is closer to zero, as you would expect from the ridge penalty). The consequences of this for the residual variance become clear if we compare panel B and C. In each case, we look at the residual variance remaining in y_i after accounting for the estimated intercept and $\hat{\beta}_X$ coefficient from model (6) (but crucially not the estimated $\hat{\beta}_z$ coefficient). We define this residual variance as $\hat{\psi}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_X X_i)$.

Panels B and C show the relationship between X and the residual variance $\hat{\psi}_i$ for the OLS and ridge regression estimates. As we would expect given the properties of OLS, $\hat{\psi}_{OLS_i}$ is uncorrelated with X. However, panel C shows that $\hat{\psi}_{regularized_i}$ is strongly (positively) correlated with X (which is what we would expect given the ridge regression line in panel A, i.e. positive values of X tend to be higher than the ridge regression line and negative values of X tend to be below it).

The consequences of the correlated ridge residual variance becomes apparent in the second row of figure S5. Panel D shows the unconditional relationship between Z and Y. Because Z and X are negatively correlated, there is a clear negative correlation between Z and Y. By construction, we know that this correlation is spurious, and that conditional on X, Z and Y should be uncorrelated. This is what we see in panel E, which shows the relationship between Z and $\hat{\psi}_{OLS_i}$—once we account for the $X \rightarrow Y$ relationship, Z and Y are essentially uncorrelated. We define the slope of this relationship $\hat{\gamma}_{OLS_Z}$ as the slope of an OLS regression of $\hat{\psi}_{OLS_i}$ on Z_i: $\hat{\psi}_{OLS_i} = \hat{\gamma}_{OLS_0} + \hat{\gamma}_{OLS_Z} Z_i$. In the case of OLS $\hat{\gamma}_{OLS_Z} = \hat{\beta}_{OLS_Z}$.

In panel F, however, we can see that Z and the $\hat{\psi}_{regularized_i}$ residual variance is negatively correlated, as we would expect given that these residuals are correlated with X, and X and Z are also correlated. The OLS line in panel F shows the estimated slope $\hat{\gamma}_{regularized_Z}$, which is the slope of an OLS regression of $\hat{\psi}_{regularized_i}$ on Z_i: $\hat{\psi}_{regularized_i} = \hat{\gamma}_{regularized_0} + \hat{\gamma}_{regularized_Z} Z_i$. In the ridge regression setting, $\hat{\gamma}_{regularized_Z} \neq \hat{\beta}_{regularized_Z}$, because $\hat{\beta}_{regularized_Z}$, shown with the dashed line, is regularized, pulling the slope closer towards zero.

In summary, because the ridge regression estimate of the relationship between X and Y leaves residuals that are correlated with X, any variable that is correlated with X (in our case Z) will also be correlated with those residuals. This process is akin to the effect of measurement error in causal inference: by inducing measurement error in accounting for the $X \rightarrow Y$ path, ridge regression partially reopens the $Z \dashleftarrow\dashrightarrow X \rightarrow Y$ backdoor path, that would otherwise be blocked by conditioning on X. Reopening this backdoor path leads to the spurious conclusion that there exists a $Z \rightarrow Y$ relationship.

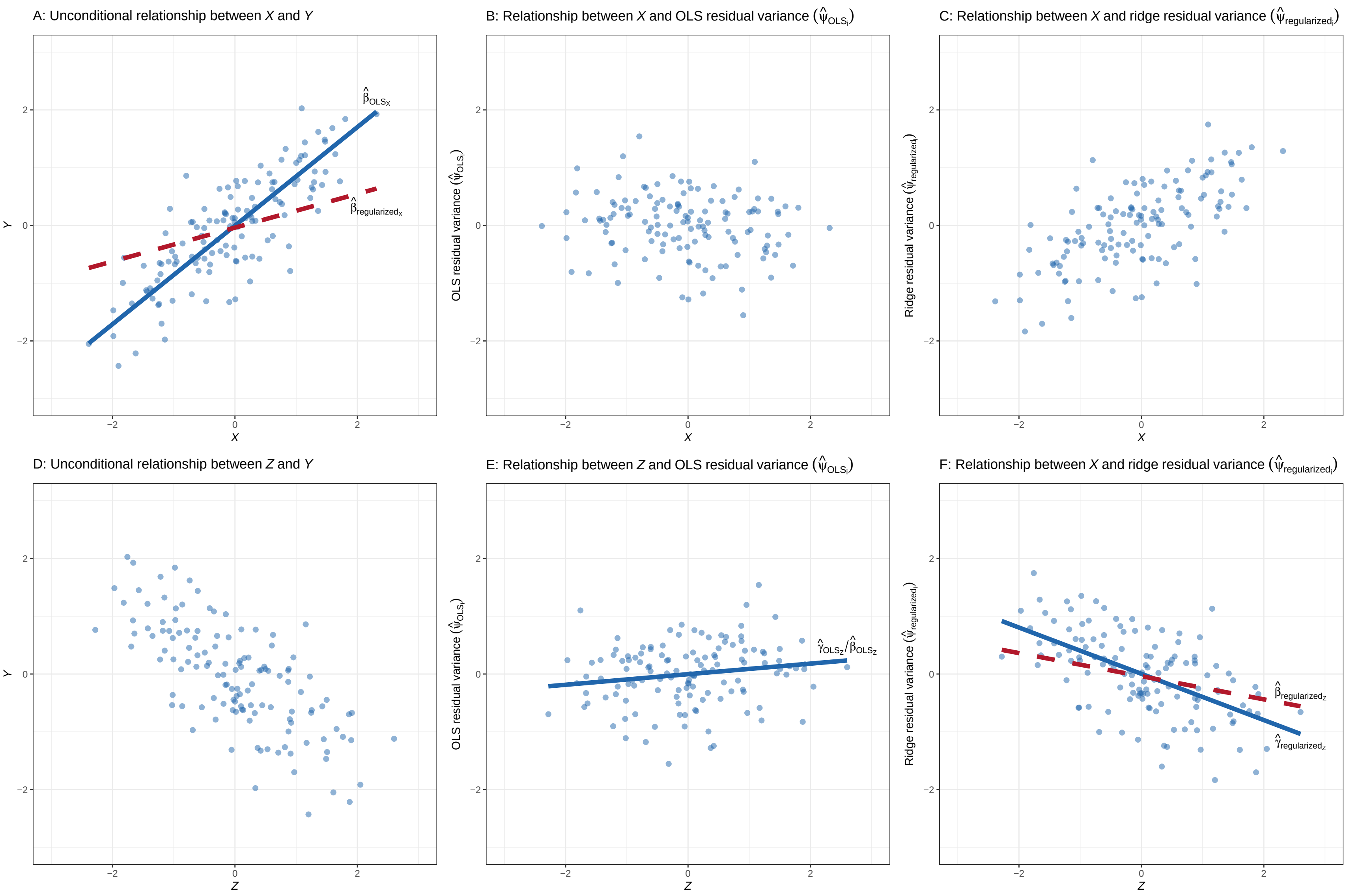


Figure S5: Process by which ridge regression spuriously estimates a significant negative effect of Z on Y.

S3.2 Differences between ridge regression and LASSO in reopening backdoor paths

Next we turn to subtle differences between ridge regression and LASSO in how they reopen backdoor paths. To do so we repeat the DGP outlined in (S6), this time varying the correlation between X and Z across a range of values from -0.9 to 0.9, simulating 500 datasets for each level of correlation, and again estimate model (6) using OLS, ridge, and LASSO with a λ of 0.05.

Figure S6 shows the mean estimate of $\hat{\beta}_{regularized_Z}$ for ridge regression and LASSO for different correlations between X and Z. Both the ridge and LASSO estimates show a consistent bias in the direction of the correlation between X and Z although the magnitude of the ridge bias is substantially larger.

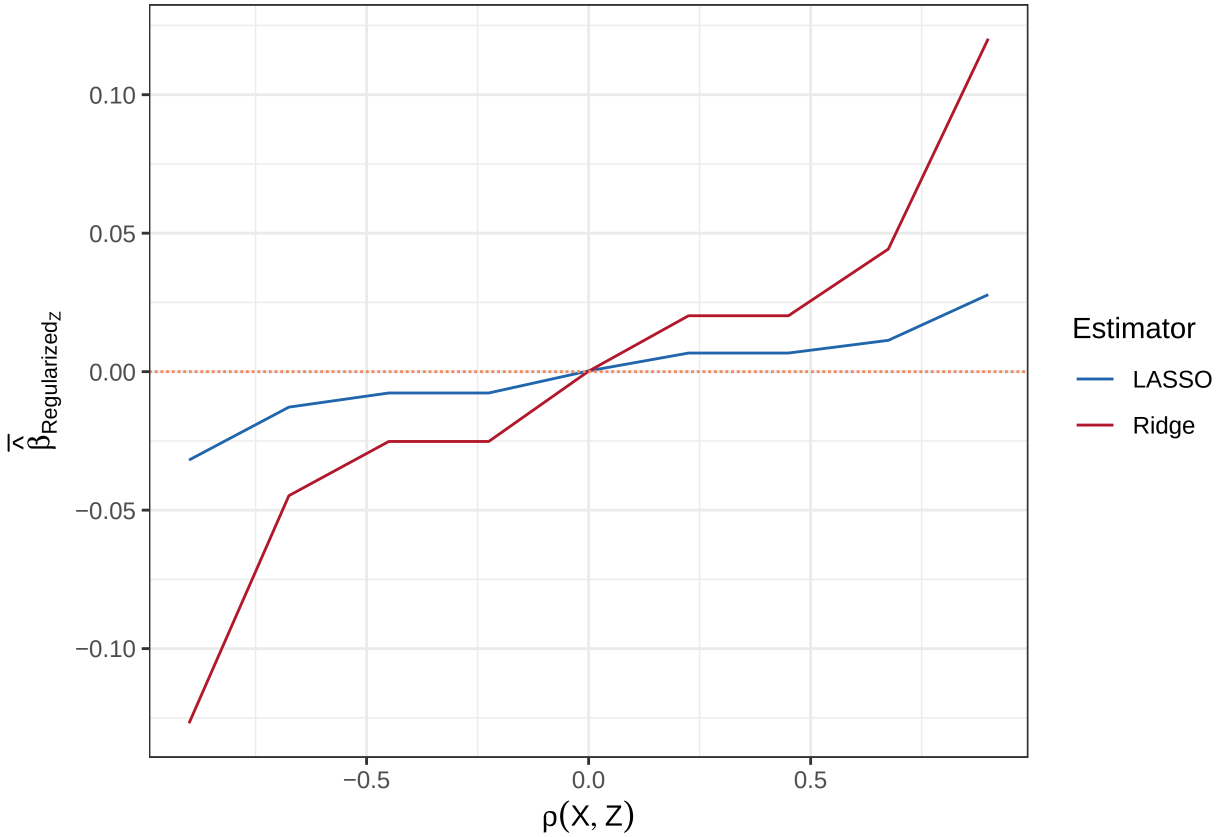


Figure S6: Average value of $\hat{\beta}_{regularized_Z}$ for different correlations between X and Z estimated with ridge regression or LASSO.

For ridge regression it is relatively straightforward to understand why the estimator will transfer some coefficient mass from the estimate of β_X to β_Z (which is always simulated as zero). Ridge regression uses a squared penalty term. This means that the estimator will "prefer" adding mass to a smaller coefficient than a larger one. Suppose that $\hat{\beta}_X$ is currently 1 $\hat{\beta}_Z$ is -0.1. That would create a ridge penalty of $1^2 + -0.1^2 = 1.01$. If we add 0.1 to $\hat{\beta}_X$, that increases the total ridge penalty to $1.1^2 + -0.1^2 = 1.22$ whereas if we added -0.1 to $\hat{\beta}_z$ it would only increase the ridge penalty to $1^2 + -0.2^2 = 1.04$. That means the ridge loss function tends to find solutions where the parameter for a correlated variable gets some mass at the expense of the larger variable's parameter.

However, it is less obvious why the LASSO estimator will spuriously add mass to $\hat{\beta}_z$ rather than increasing the magnitude of $\hat{\beta}_X$ since the LASSO penalty is proportional to the absolute value of the coefficients. Taking our previous example, adding 0.1 to $\hat{\beta}_X$, increases the total LASSO penalty to $|1.1| + |-0.1| = 1.2$ exactly the same as if we added -0.1 to $\hat{\beta}_Z$ $|1| + |-0.2| = 1.2$. In other words, the LASSO penalty should be indifferent between adding mass to $\hat{\beta}_X$ and $\hat{\beta}_Z$. Given that we simulated β_Z to be zero, it seems odd that the LASSO estimates exhibit a consistent bias.

The solution to this puzzle is that the value of $\hat{\beta}_{regularized_Z}$ reflects four things: 1) the true value of β_Z (which is zero in our simulations), 2) the variance available due to the regularization of other variables in the model, 3) the regularization of $\hat{\beta}_{regularized_Z}$, and 4) sampling variation. On any given simulation, $\hat{\beta}_{OLS_Z}$ will take a non-zero value because of sampling variation. We therefore need to look at not only the average value of $\hat{\beta}_{regularized_Z}$ but how it compares to the unregularized estimate $\hat{\beta}_{OLS_Z}$ in each simulation.

If we take the now familiar case where the correlation between X and Z is -.9, we can see how sampling variation affects the way $\hat{\beta}_{regularized_Z}$ is regularized, resulting in the bias we see above. Panel A of figure S7 shows the relationship between $\hat{\beta}_{OLS_Z}$ and $\hat{\gamma}_{regularized_Z}$ (the slope of an OLS regression of the residuals $\hat{\psi}_{regularized_i}$ on Z_i as defined above). The key point here is that as the value of $\hat{\beta}_{regularized_Z}$ becomes lower, $\hat{\gamma}_{regularized_Z}$ also becomes lower. Because of the relationship between $\hat{\beta}_{OLS_Z}$ and $\hat{\beta}_{regularized_X}$ this relationship is non-linear. In repeated samples, $\hat{\beta}_{OLS_X}$ and $\hat{\beta}_{OLS_Z}$ are positively correlated, and because $\hat{\beta}_{OLS_X}$ is larger when $\hat{\beta}_{OLS_Z}$ is larger, $\hat{\beta}_{regularized_X}$ is more heavily regularized when $\hat{\beta}_{OLS_Z}$ is larger. As a result, more residual variance is opened up when $\hat{\beta}_{OLS_Z}$ is positive, resulting in a greater difference between $\hat{\beta}_{OLS_Z}$ and $\hat{\gamma}_{regularized_Z}$ when $\hat{\beta}_{OLS_Z}$ is positive.

When the LASSO penalty is applied, the $\hat{\gamma}_{regularized_Z}$s that are closest to zero are translated into $\hat{\beta}_{regularized_Z}$s of exactly zero. The net result of this is that all the $\hat{\beta}_{OLS_Z}$s which were positive are now zero $\hat{\beta}_{regularized_Z}$s, but many of the negative $\hat{\beta}_{OLS_Z}$s are now negative $\hat{\beta}_{regularized_Z}$s.This means that even though no coefficients have flipped direction, in expectation $\hat{\beta}_{regularized_Z}$ is biased in the same direction as the correlation.

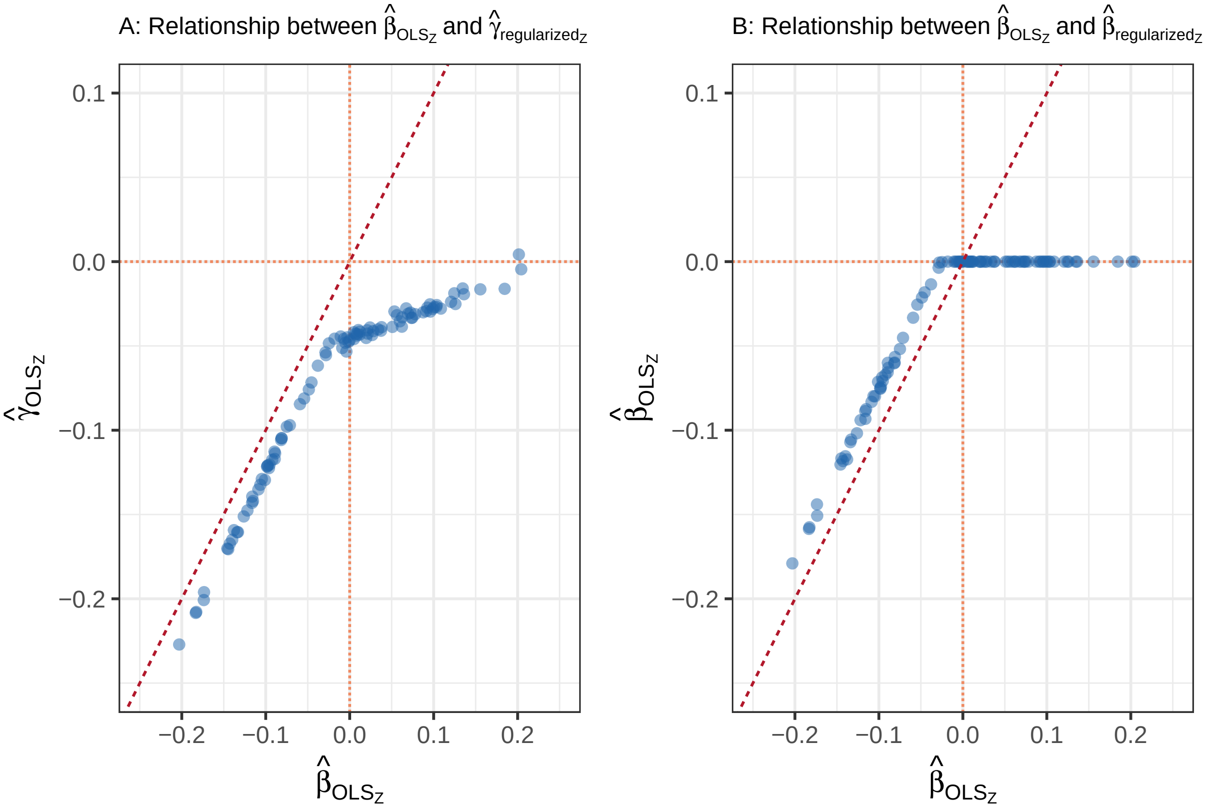


Figure S7: Relationship between estimated OLS β_{OLS_Z} coefficients and LASSO $\gamma_{regularized_Z}$ and $\beta_{regularized_Z}$ when the correlation between X and Z is -.9.

We now turn to how different correlations between X and Z affect ridge and LASSO differently. Figure S8 shows the effects of ridge regression regularization on the estimates of $\hat{\beta}_{regularized_Z}$ compared to the unregularized $\hat{\beta}_{OLS_Z}$ for different simulated correlations between X and Z. The correlation between X and Z moves the ridge estimates of β_Z in the direction of the correlation—when the correlation is negative, the ridge β_Z estimates are moved in a negative direction relative to the OLS estimates for the same data, when the correlation is positive, they are moved in a positive direction. In the case when X and Z are negatively correlated, $\hat{\beta}_{regularized_Z}$ estimated with ridge regression is consistently lower than $\hat{\beta}_{OLS_Z}$. This leads to $\hat{\beta}_{regularized_Z}$ having a larger magnitude than $\hat{\beta}_{OLS_Z}$ when $\hat{\beta}_{OLS_Z}$ is negative, and a flipped sign or smaller magnitude of the same sign when $\hat{\beta}_{OLS_Z}$ is positive. When the correlation between X and Z is positive these effects apply in the opposite direction.

Figure S9 shows the equivalent effects for $\hat{\beta}_{regularized_Z}$ estimated using LASSO. Regardless of the strength or direction of the correlation between X and Z the sign of $\hat{\beta}_{regularized_Z}$ never flips compared to $\hat{\beta}_{OLS_Z}$, and the magnitude of $\hat{\beta}_{regularized_Z}$ is always smaller than the magnitude of $\hat{\beta}_{OLS_Z}$.[6] The main impact of the correlation between X and Z is to change which estimates are regularized to zero. Changing correlations have two effects: (1) correlation skews which coefficients are regularized to zero in the opposite direction to the correlation (e.g. when the correlation is negative, positive coefficients are more likely to be regularized to zero), and (2) stronger correlations lead to a wider range of values that are regularized to zero. Although the effects of ridge and LASSO regularization look very different, the end result is very similar—in both cases the expected value of the β_Z estimate is biased in the direction of the correlation, as shown in figure S6.

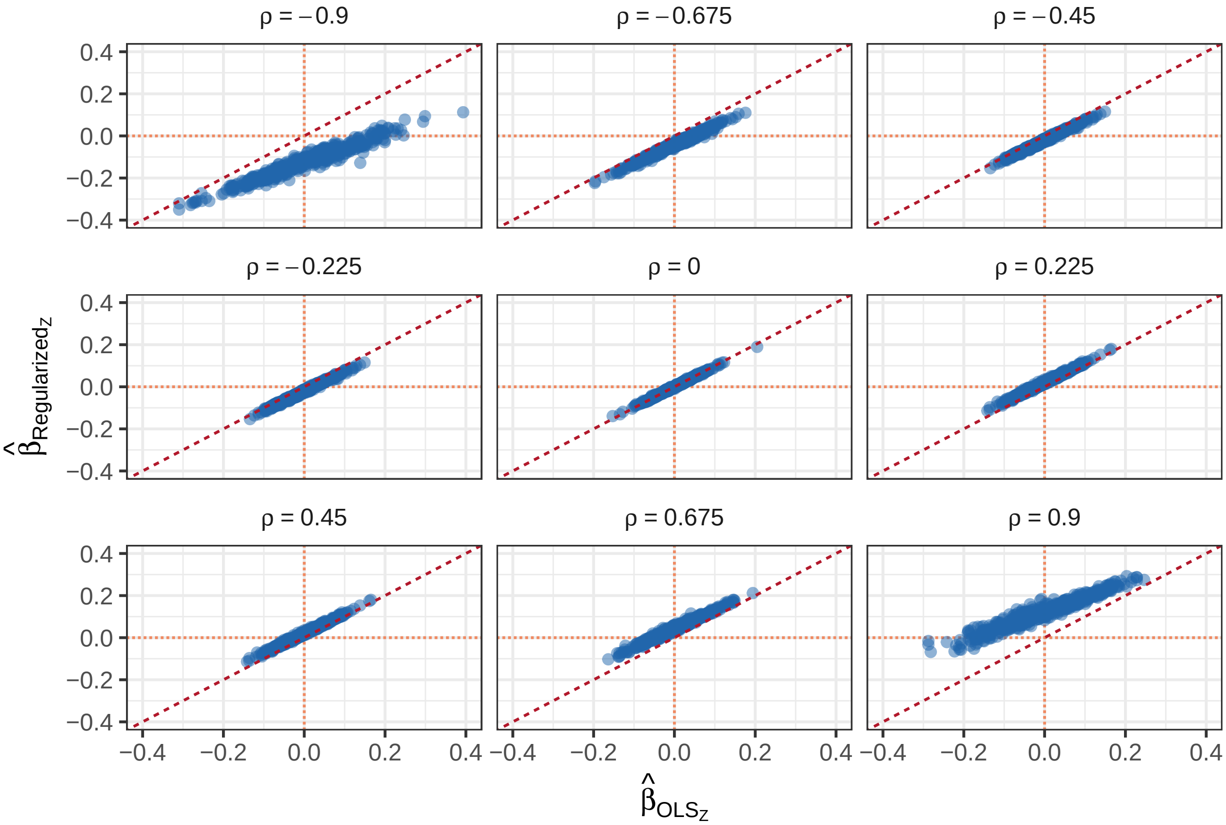


Figure S8: Relationship between estimated OLS β_Z coefficients and ridge β_Z coefficients at different correlations between X and Z.

[6] LASSO estimates can flip their sign compared to OLS very occasionally, but none of these flips occurred in the simulations we report here.

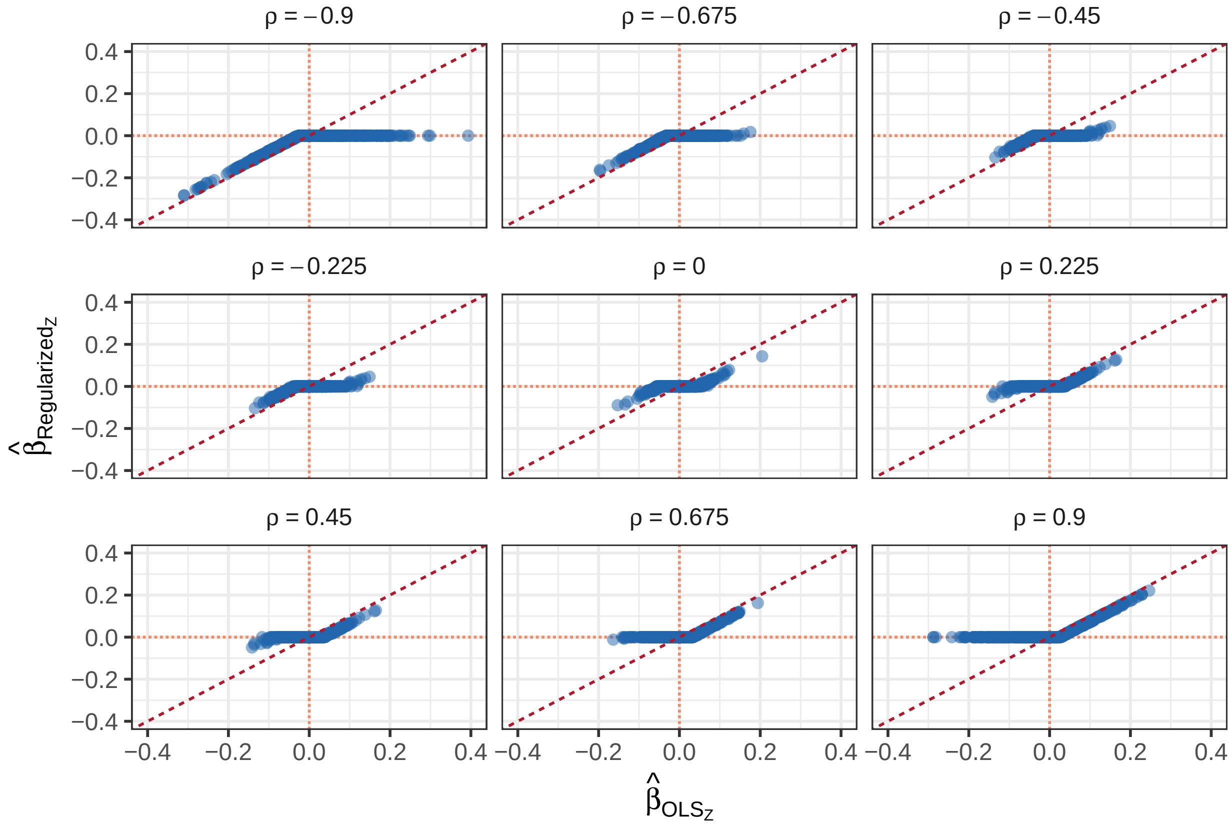


Figure S9: Relationship between estimated OLS β_Z coefficients and LASSO β_Z coefficients at different correlations between X and Z.

S4 Can we conclude that response rates do not affect polarization?

We have demonstrated that CF2's use of ridge regression is not appropriate for the problem they are addressing, that their results disappear if OLS is used, and that their use of ridge regression nearly guarantees false positives. But can we use the OLS estimator to conclude that response/contact rates do *not* affect mass polarization?

To test the statistical power available in CF2's data to detect the effect sizes they claim, we simulate further data using the same approach as outlined in appendix R3, with two differences: (1) we add an additional term $\beta_{rr_{sim}}$ to model (R2) which adds a true simulated effect for the response and cooperation rate variables (we hold the effect of contact at zero in all simulations), and (2) the variables and coefficients are scaled to produce standardized coefficients.

$$\begin{aligned} &\beta_{rr_{sim}} \sim U(0,1) \\ &y_{sim} \sim \beta_{int_{sim}} + \beta_{rr_{sim}} rr_{sim} + \beta_{con_{sim}} con_{sim} + \beta_{year_{sim}} year_{sim} + \mathcal{N}(0, RMSE_{model}) \end{aligned} \tag{S7}$$

Figure S10 shows the estimated level of statistical power (the probability of correctly returning a true positive) for each outcome according to the simulated effect of response rate and cooperation rate. The results show that for all but the largest effects, CF2's data is drastically underpowered, reaching the conventional 80% power threshold only when effects are large.[7]

[7] For effect size benchmarks, see for example, Gignac and Szodorai (2016). 'Effect size guidelines for individual differences researchers' *Personality and Individual Differences*, 102, pp. 74-78, who suggest benchmarks of 0.1 = 'small', 0.2 = 'medium', and 0.3 = 'large'.

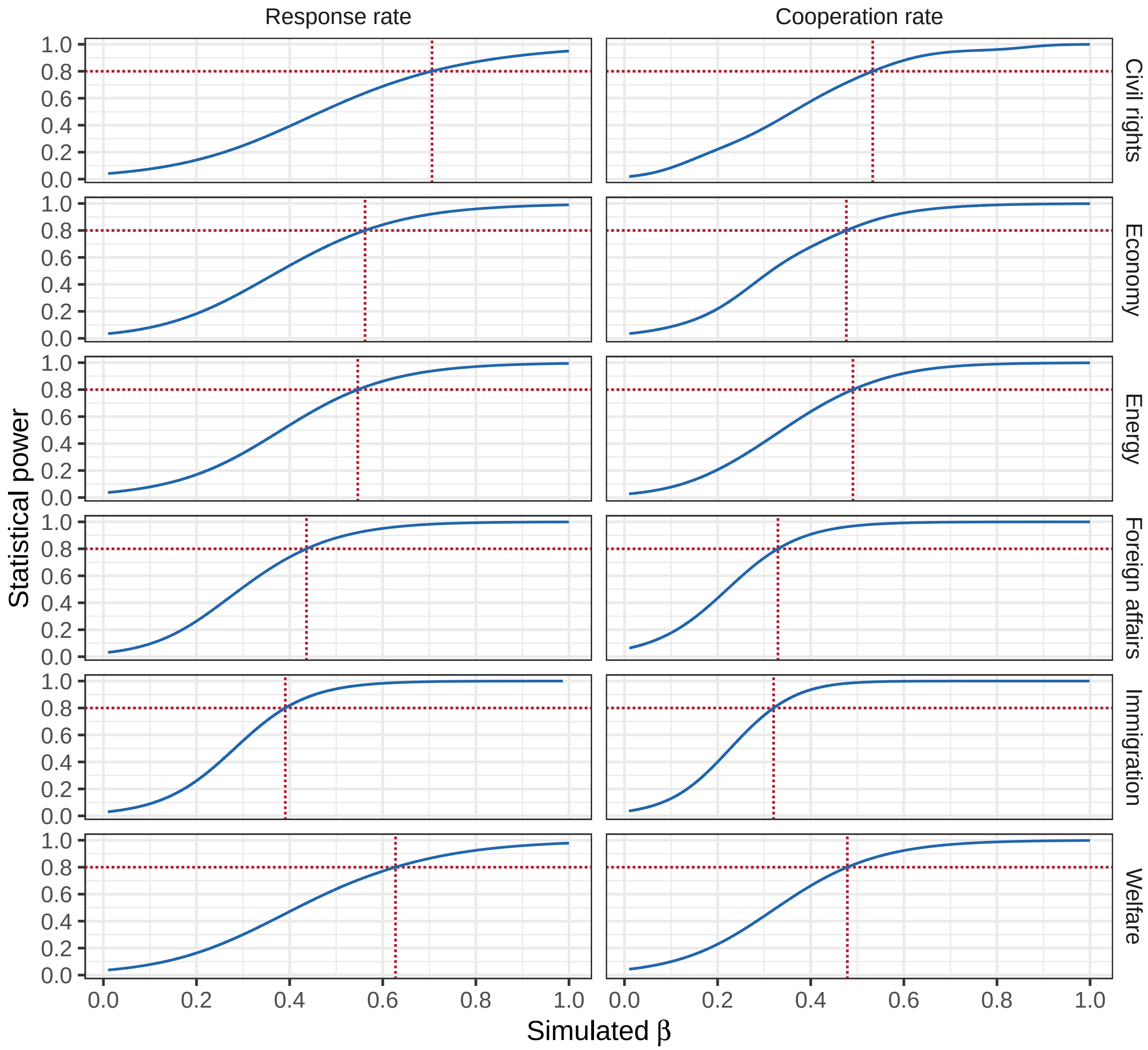


Figure S10: Estimated statistical power by effect size for an OLS estimator using simulated data based on CF2's original data.

We can also use these simulations to examine type S and type M errors.[8] A 'type S' (sign) error is an error in the sign of an estimate, that is, the probability of it being in the wrong direction if it is statistically significant. A 'type M' (magnitude) error is the extent to which statistically significant results exaggerate the true effect size. The type S error rate by effect size is shown in figure S11 and reveals that CF2's data is likely to be prone to high levels of type S errors if the underlying true effect size was reasonably small. Figure S12 shows the type M error rate and likewise indicates that CF2's data is likely to have problems

[8] Gelman and Carlin (2014) 'Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors', *Perspectives on Psychological Science*, 9(6), pp. 641-651.

with type M errors, with statistically significant results exaggerating the true effect considerably.

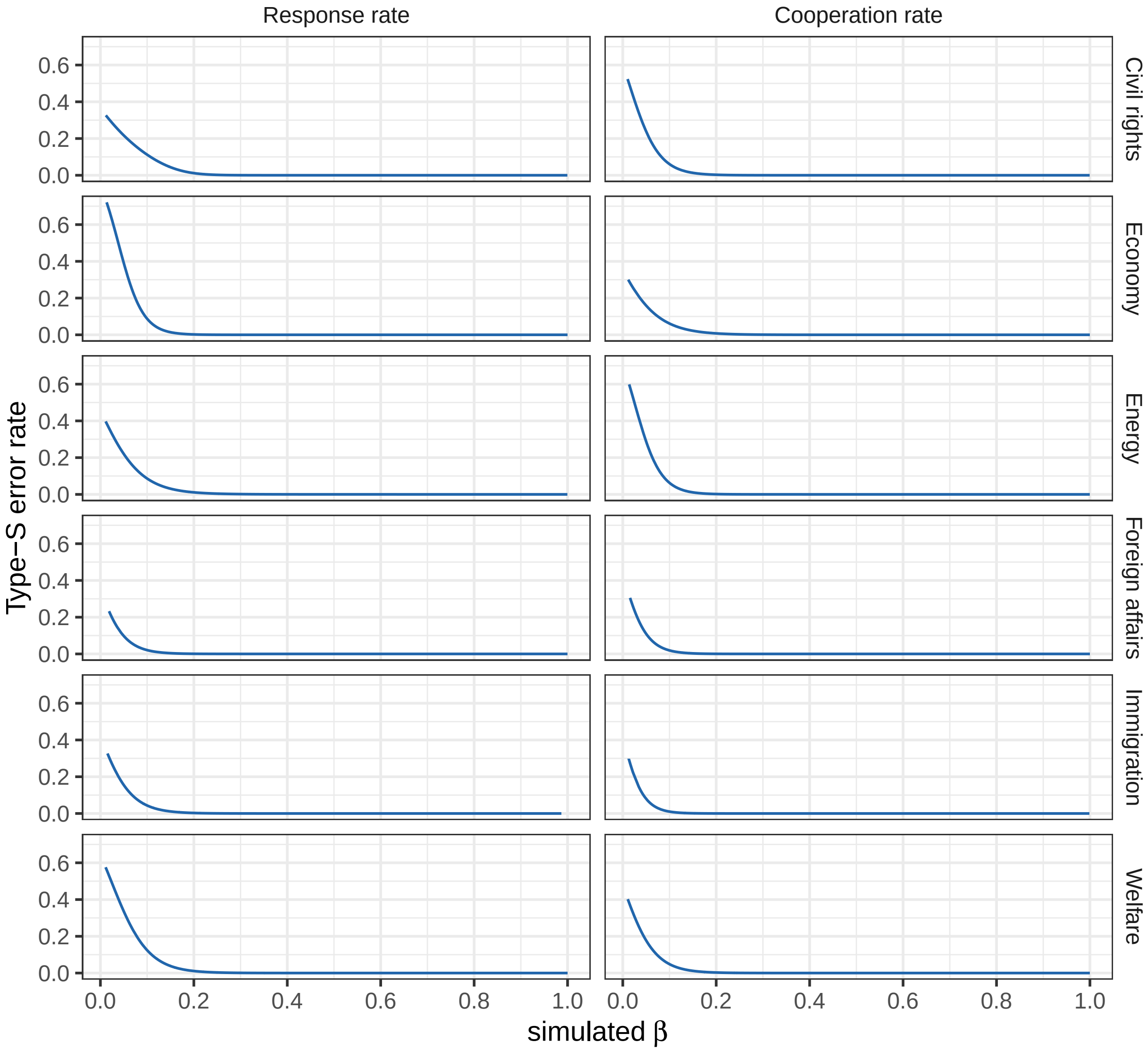


Figure S11: Estimated type S error rate by effect size for an OLS estimator using simulated data based on CF2's original data.

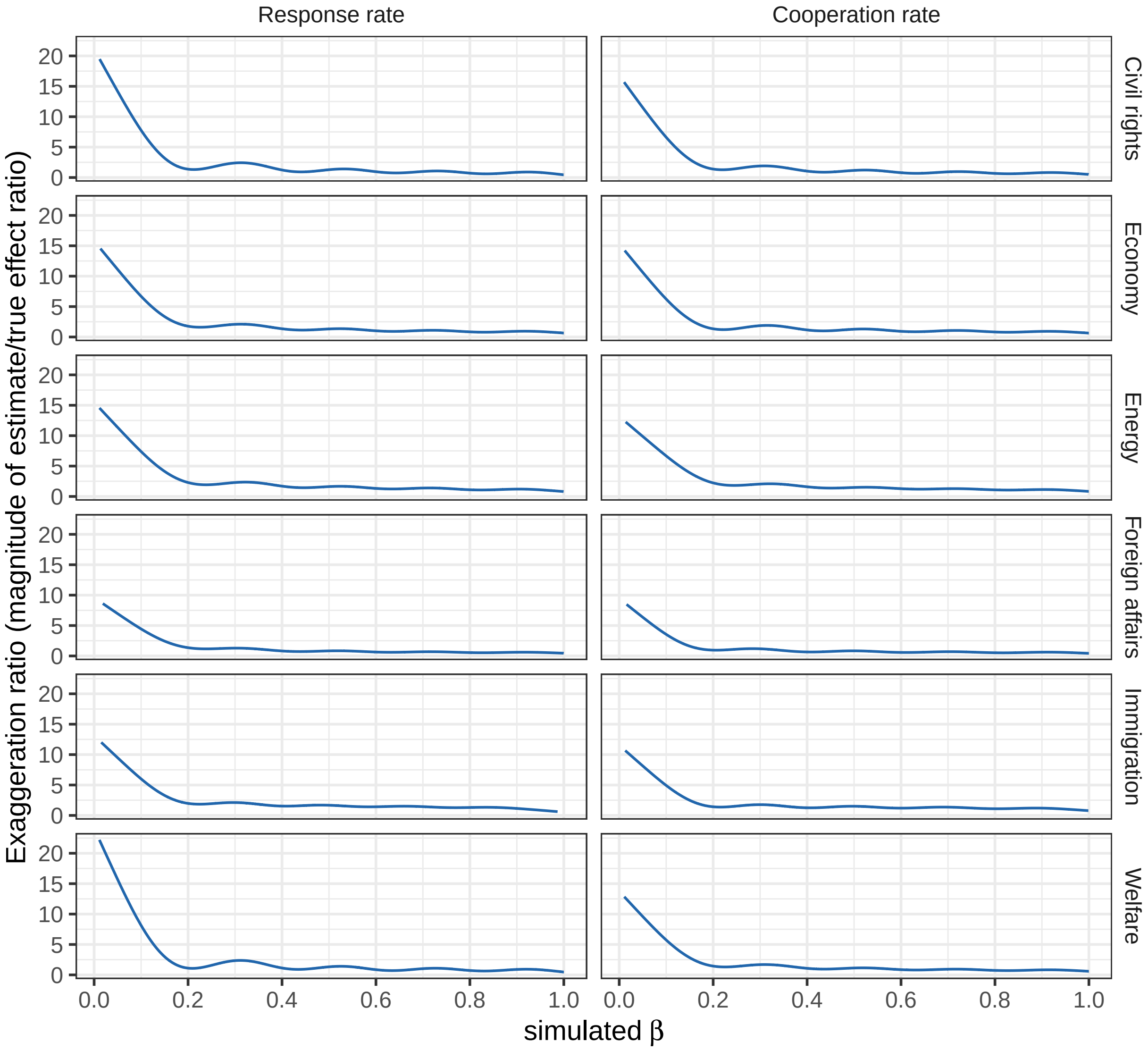


Figure S12: Estimated type M error rate by effect size for an OLS estimator using simulated data based on CF2's original data.

CF2's data is underpowered for two reason, first, the sample size is small, and second, as CF2 acknowledge, they are estimating models with collinear variables, which inflates the standard errors. The results of these simulations are clear: CF2's data is insufficient to test for all but the largest effects. We do not believe there are good reasons to think that the effects of response and cooperation rates on measured polarization are large enough to be confident that the data is not returning inflated estimates (at best) or even estimates with the wrong sign. In fact, these power estimates are likely to be overly conservative as they assume that the correct DGP is a simple linear time-trend rather than a more complex structure such as ARIMA-MLM that has even higher data requirements. In short, we cannot draw conclusions about the effect of survey response rates on mass polarization either way.

S4.1 Other causal structures and omitted variable bias

In addition to power concerns, there may also be problems with omitted variable bias. In the main text, we discuss polarization as the key dependent variable. However, it may be more appropriate to split polarization into its actual level and observed level. When we do this, it may also be appropriate to allow that polarization could affect cooperation rates (if we posit that polarized individuals are more likely to answer surveys, then higher levels of polarization should increase cooperation rates). This DAG is shown in equation (S8).

$$\begin{tikzcd} time \arrow[r] \arrow[dr] & polarization_{actual} \arrow[r] \arrow[d] & polarization_{observed} \\ & cooperation \arrow[ur] & \end{tikzcd} \tag{S8}$$

This DAG creates an additional issue for CF2's model because the effect of *cooperation* on measured $polarization_{observed}$ is only causally identified if we condition on *time* and $polarization_{actual}$. We have discussed numerous methods for appropriately conditioning on *time*, but this still leaves the question of adjusting for $polarization_{actual}$. CF2 (and our reanalysis) includes a control for congressional polarization (which should not be affected by response rate mechanisms). If this control is sufficient to adjust for $polarization_{actual}$, then the estimate would still be causally identified. However, to the extent that this control is inadequate, the estimates will be subject to omitted variable bias. The size of this bias depends on the strength of the relationship between $polarization_{actual}$ and $polarization_{observed}$ (presumably strong) and the relationship between $polarization_{actual}$ and *cooperation* (likely relatively weak).

Since $polarization_{actual}$ is theorized to be positively related to *cooperation* and $polarization_{observed}$, the effect of omitting $polarization_{actual}$ would be that we estimate a more positive relationship between *cooperation* and $polarization_{observed}$ than is actually the case. This is a further reason why we refrain from making strong substantive claims based on the null or positive results from our revised models/estimators.

S5 Accounting for scaling when converting between ridge regression λ and Bayesian priors

The conversion between Bayesian priors and equivalent ridge regression λ is complicated by the idiosyncratic scaling used in many implementations including the *lmridge* package. Specifically, the *lmridge* documentation states that the predictors are scaled "to correlation form, such that the correlation matrix has unit diagonal elements". However, the advice regarding priors typically assumes that all predictor and outcome variables are scaled to have mean 0 and standard deviation 1 (often referred to as unit-scaled).

To demonstrate how different those scalings are in practice, we simulate a dataset of three predictors with varying standard deviations: $x_1 \sim \mathcal{N}(0,3)$, $x_2 \sim \mathcal{N}(0,5)$, $x_3 \sim \mathcal{N}(0,0.5)$. We then create our dependent variable $y = 0.25x_1 + 0.5x_2 + 0.75x_3 + \sigma^2$, where σ^2 $\mathcal{N}(0,1)$.

Table S1 shows the original standard deviations, *lmridge* scaled standard deviations, and unit-scaled standard deviations for the four variables in our simulation.

Table S1: Standard deviations of variables under different scalings

Scenario	y	x1	x2	x3
Raw	3.122	3.293	5.444	0.462
Standard lmridge scaling	3.122	0.082	0.082	0.082
Unit scaling	1.000	1.000	1.000	1.000

For illustration, we estimate the linear model using *lmridge* with $\lambda = 3$ and the standard *lmridge* scaling. But how does $\lambda = 3$ on the *lmridge* scaling translate into the standard deviation (σ) of a normal distribution on unit-scaling? We first use the following formula to convert from λ to the standard deviation of a Bayesian prior:

$$\sigma = \sqrt{\frac{Var(\hat{\epsilon})}{\lambda}}$$

where $Var(\hat{\epsilon})$ is the variance of the residuals from the *lmridge* estimate.

Substituting in the values from the *lmridge* estimate

$$\sigma = \sqrt{\frac{5.67}{3}} = 1.37$$

We can confirm that a Bayesian estimate produces nearly identical results to the ridge regression estimates in table S2.

Table S2: Ridge regression estimates of linear model on simulated data with Bayesian equivalent estimates.

	Ridge	Bayes
x1	2.65	2.62
x2	8.55	8.52
x3	2.29	2.26

However, both estimates here use the ridge scaling we saw in table S1. We must convert σ to its unit-scaled equivalent. To do this, we multiply σ by the standard deviation of the predictor variables and divide by the dependent variable:

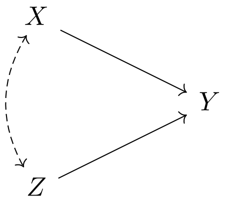
$$\sigma_{unit} = \frac{\sigma * sd(x_1)}{sd(y)}$$

Substituting in the values from our simulation, we find that the original ridge regression with $\lambda = 3$ and *lmridge* default scaling is equivalent to a unit-scaled normal prior with mean 0 and standard deviation of 0.0361: a very informative prior.

S6 Using LASSO for variable selection

Rather than using LASSO to estimate model parameters, an alternative is to use it for variable selection.[9] However, this approach is also prone to reintroducing backdoor confounding, albeit in a slightly different way to that which we discuss in the main body of the paper. In brief, using LASSO for variable selection is a two stage process. First we estimate the model using LASSO, and then make an unregularized estimate of a separate model of the sub-set of variables that are not reduced to zero in the first stage LASSO.

The potential problem with this approach is easy to show. Imagine we have a DGP similar to the one shown earlier in DAG (4), but this time there is also a $Z \rightarrow Y$ causal path, as shown in (S9).



(S9)

If we fail to condition on one of X or Z, our estimate of the effect of the other variable will be confounded by the $Z \leftarrow\dashrightarrow X \rightarrow Y$ backdoor path. If we allow LASSO to choose which variables to include in the model, then in expectation our estimate of the effect of X will be biased proportional to the probability of Z being included in the model (and vice versa) and the correlation between X and Z. We can show that this is the case with another simulation, which follows a similar DGP to our earlier simulations, but where $\beta_X = 0.8$, $\beta_Z = 0.2$ (the correlation between X and Z is again set at -0.9). We generate 10,000 datasets and randomly select a λ penalty between 0 and 1.5 (by which point LASSO always sets β_Z to 0).

The results of these simulations are shown in figure S13. As the λ penalty increases, the probability Z is selected by the LASSO decreases steadily until it approaches zero. At the same time, as λ increases, our estimates of β_X are steadily inflated (because the models increasingly do not control for Z), before flattening out at the point Z is almost always omitted from the model.

[9] See for example, Zhao, Witten, and Shojaie (2021), In Defense of the Indefensible: A Very Naïve Approach to High-Dimensional Inference, *Statistical Science* 36(4): 562-577.

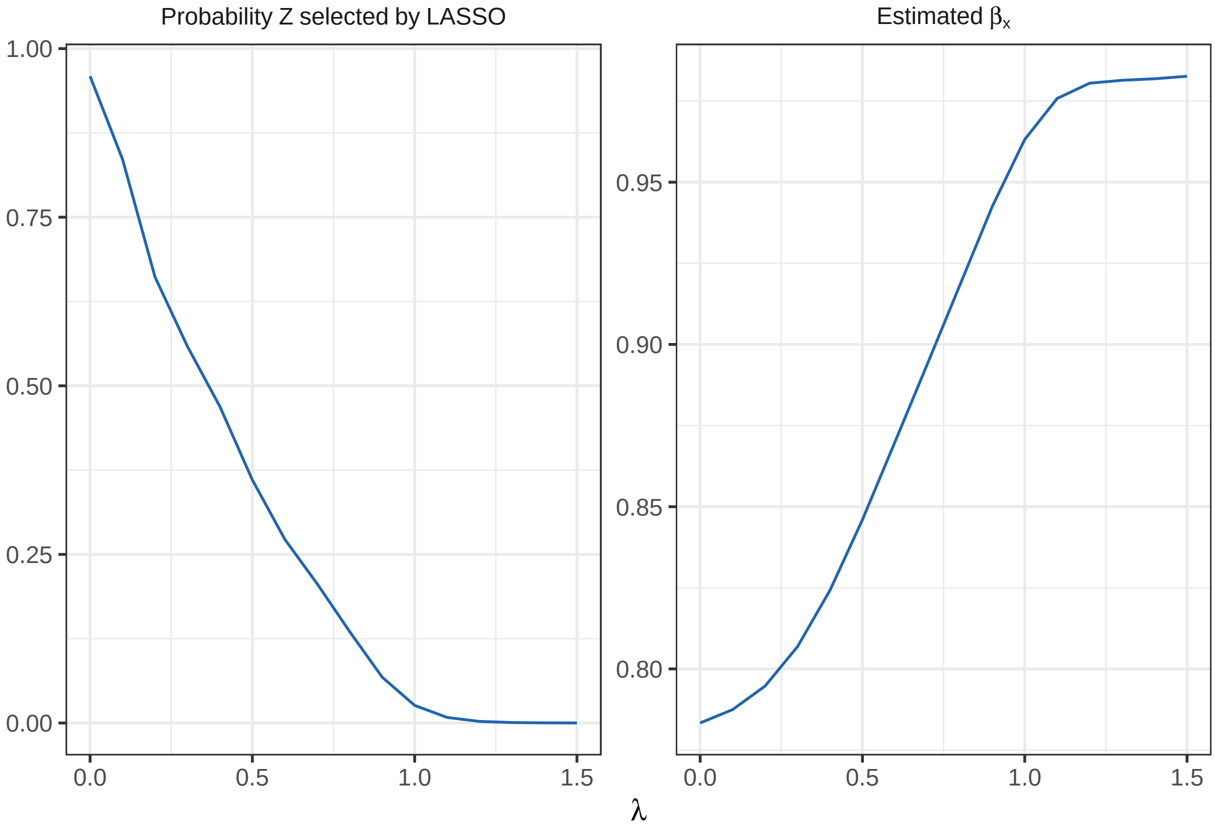


Figure S13: Probability Z is selected and estimated β_X by LASSO λ penalty.