PLSC 503 – Spring 2024

Practical Regression

January 29, 2024

The Model

Multivariate linear regression:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

...via OLS gives:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

We know that:

- $\hat{\beta}_k = \frac{\partial \mathrm{E}(Y)}{\partial X_k}$
- Under some assumptions, OLS is BLUE / BUE

Regression Redux

```
> model<-lm(adrate~polity+subsaharan+muslperc+literacy,data=Data)
> summary(model)

Call:
lm(formula = adrate ~ polity + subsaharan + muslperc + literacy,
    data = Data)

Residuals:
     Min       1Q   Median       3Q      Max
-15.4681  -4.3947  -0.5251   3.4246  22.9358

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.39843   14.94744  -0.294   0.7702
polity      -0.01390    0.27969  -0.050   0.9606
subsaharan   3.72969    5.43093   0.687   0.4964
muslperc    -0.08689    0.06282  -1.383   0.1747
literacy     0.16575    0.09433   1.757   0.0869 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 8.264 on 38 degrees of freedom
Multiple R-squared:  0.3771,Adjusted R-squared:  0.3115
F-statistic: 5.751 on 4 and 38 DF,  p-value: 0.001013
```

Presentation: A (De)Fault-y Table

```
> M1<-lm(adrate~polity+subsaharan+muslperc+literacy,data=Data)
> M2<-lm(adrate~polity+subsaharan+muslperc,data=Data)
> M3<-lm(adrate~polity+subsaharan+literacy,data=Data)
>
> stargazer(M1,M2,M3)
```

	Dependent variable:		
	adrate		
	(1)	(2)	(3)
polity	−0.014	−0.051	−0.020
	(0.280)	(0.286)	(0.283)
subsaharan	3.730	0.530	8.268*
	(5.431)	(5.252)	(4.379)
muslperc	−0.087	−0.163***	
	(0.063)	(0.047)	
literacy	0.166*		0.256***
	(0.094)		(0.069)
Constant	−0.669	14.800**	−13.120**
	(10.410)	(5.701)	(5.298)
Observations	43	43	43
R^2	0.377	0.326	0.346
Adjusted R^2	0.312	0.275	0.295
Residual Std. Error	8.264 (df = 38)	8.483 (df = 39)	8.361 (df = 39)
F Statistic	5.751*** (df = 4; 38)	6.302*** (df = 3; 39)	6.870*** (df = 3; 39)
Note:			*p<0.1; **p<0.05; ***p<0.01

Some Guidelines (from Week One)

Tables:

- *Use column headings descriptively.*
- *Use multiple rows / columns rather than multiple tables.*
- *Learn about significant digits, and don't report more than 4-5 of them.*
- *Use a figure to replace a table when you can.*
- *Be aware of norms about *s.*

Figures:

- *Report the scale of axes, and label them.*
- *Use as much "space" as you need, but no more.*
- *Use color sparingly.*

Using Regression

Regression, Conceptually

Begin with:

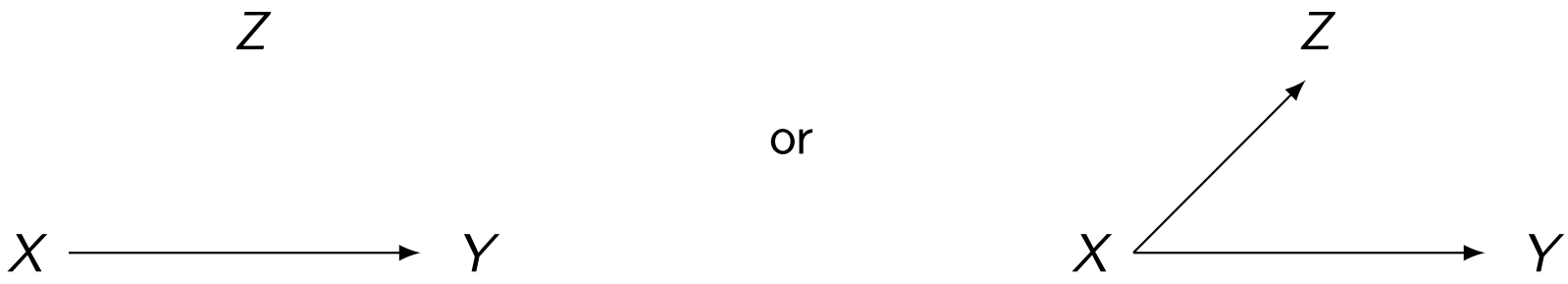
- An *outcome* Y
- A *predictor* X
- Another variable Z

We are mainly interested in $\text{Cov}(Y, X|Z)$...

- Possibly *but not necessarily* the causal relationship (cf. Berk 2010)
- Key question is *model specification*...

The Easiest Case

Things are easiest if:



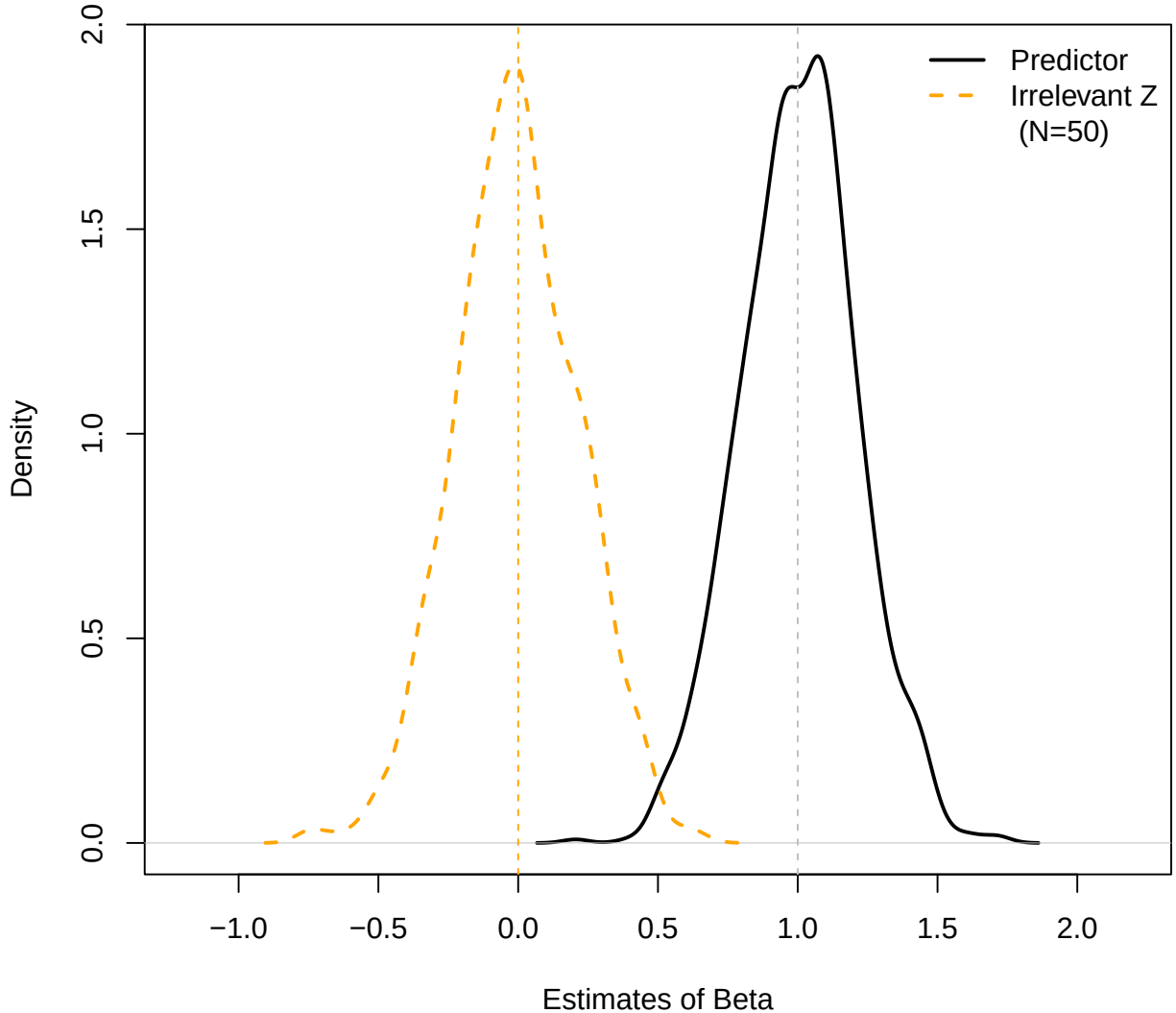
Implies that:

- Z is unimportant to understanding $\text{Cov}(X, Y)$
- $\rightarrow$ Z is *ignorable*

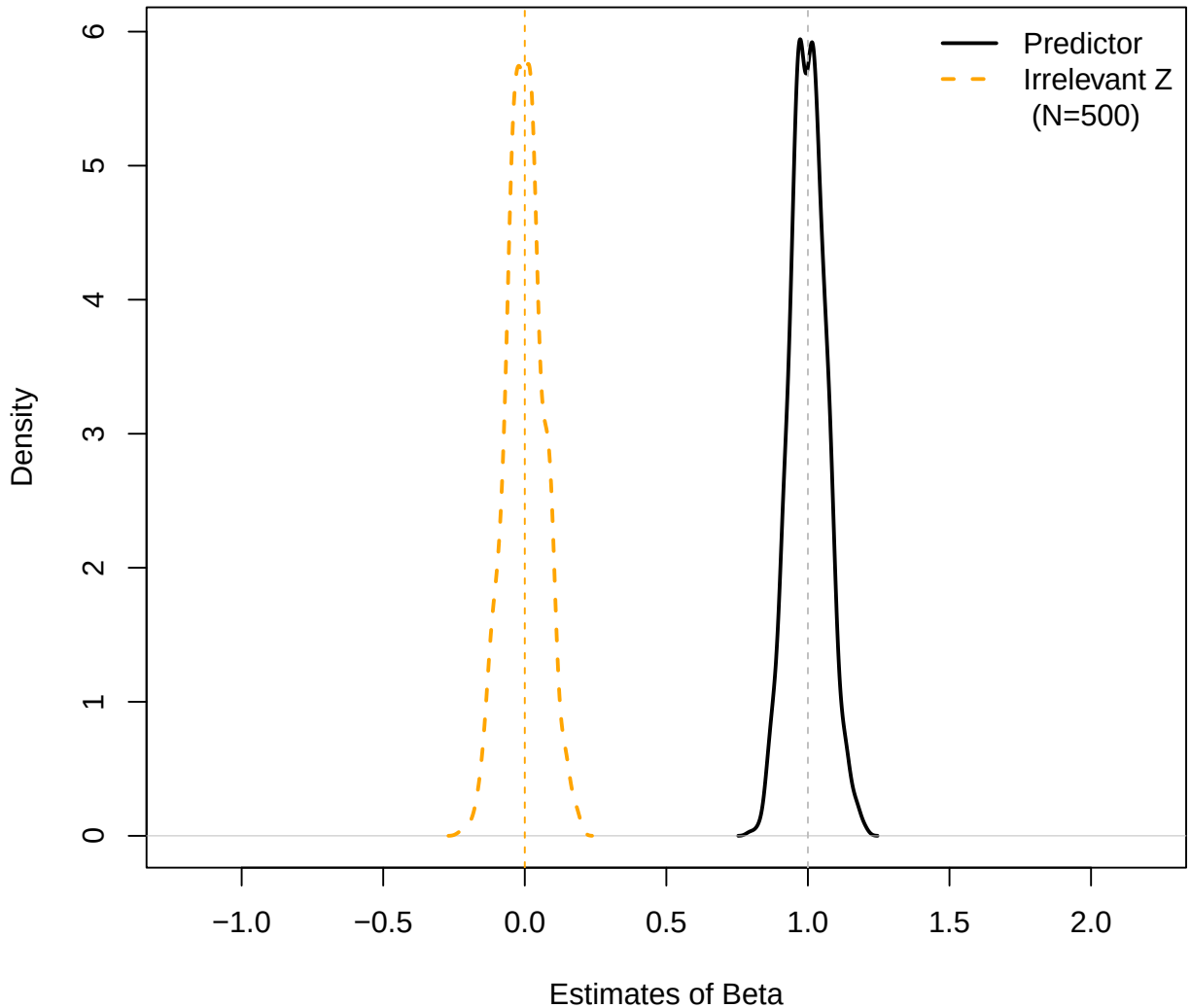
Simulations For Everyone!

```
> N<-50
> set.seed(7222009)
> X<-rnorm(N)         # Predictor
> Z<-(X+rnorm(N))/1.5 # Z "caused by" X
> Y<-X+rnorm(N)       # Outcome Y (unrelated to Z)

> print(summary(lm(Y~X)))

Coefficients:
            Estimate Std. Error t value   Pr(>|t|)
(Intercept)    0.028      0.144    0.19       0.85
X              0.978      0.162    6.05 0.00000021 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1 on 48 degrees of freedom
Multiple R-squared:  0.432,Adjusted R-squared:  0.421
F-statistic: 36.6 on 1 and 48 DF,  p-value: 0.000000212


> print(summary(lm(Y~X+Z)))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.0335     0.1460    0.23     0.82
X             0.9322     0.2161    4.31 0.000082 ***
Z             0.0659     0.2019    0.33     0.75
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1 on 47 degrees of freedom
Multiple R-squared:  0.434,Adjusted R-squared:  0.41
F-statistic:   18 on 2 and 47 DF,  p-value: 0.00000157
```

Do That 999 More Times

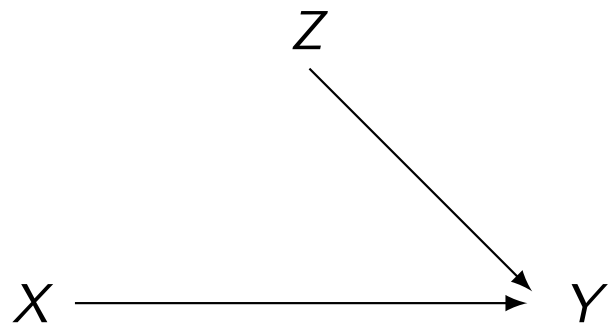


Same, But With $N = 500$



Slightly More Challenging

Suppose instead that:



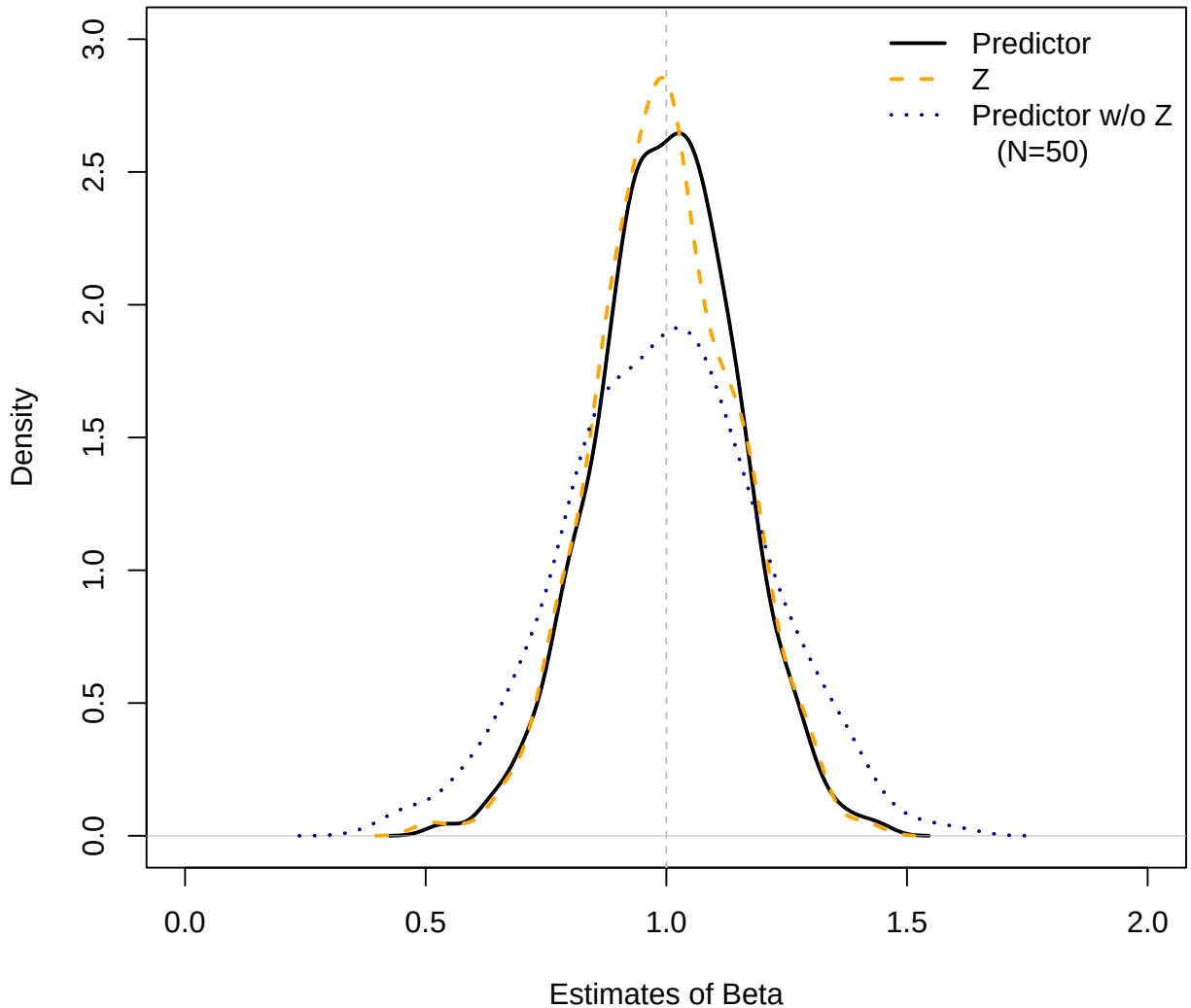
This means that:

- Z is important to / influential on understanding Y, *but*
- Z is unrelated to X...

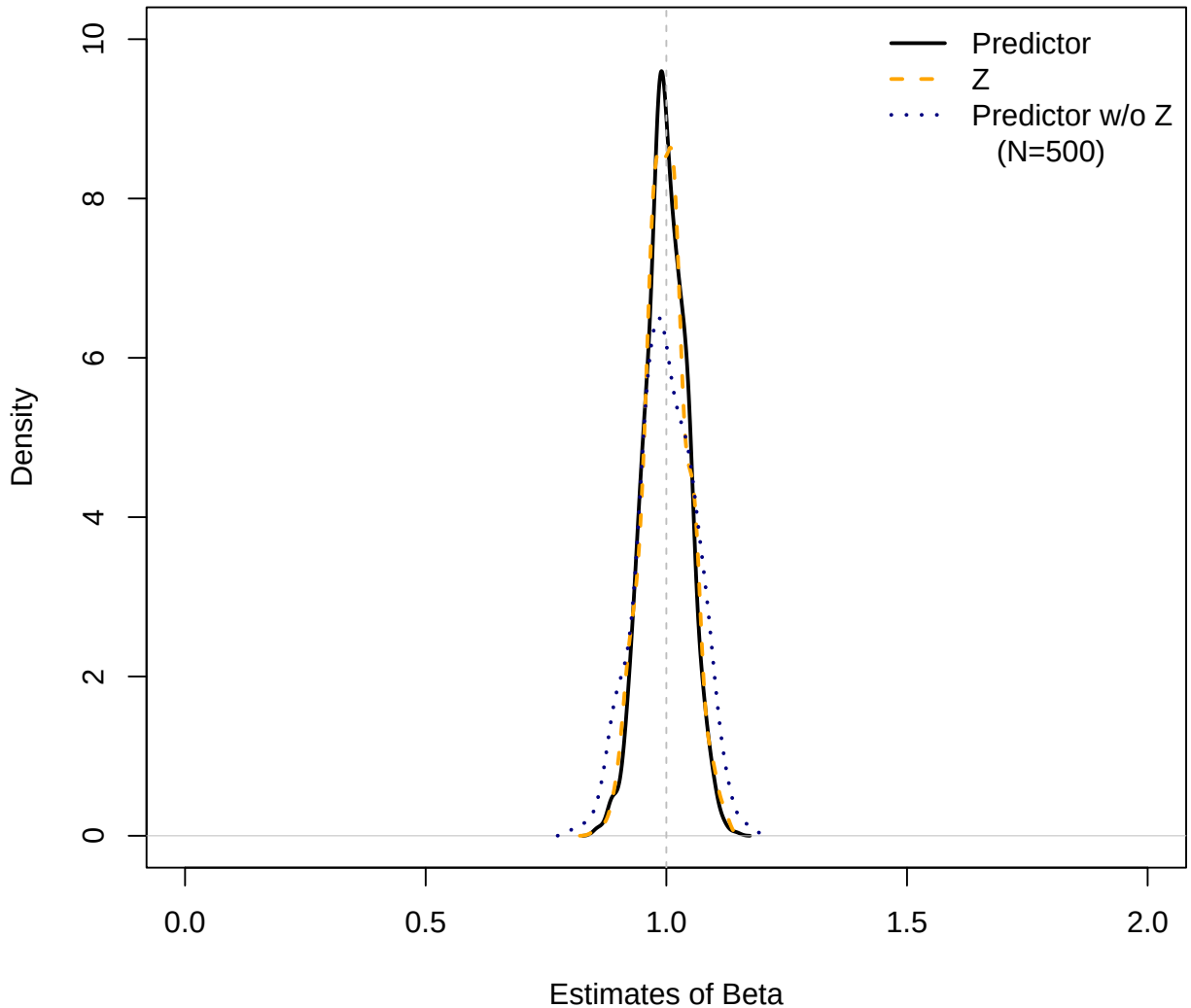
One Regression

```
> N<-50
> set.seed(7222009)
> X<-rnorm(N)          # Predictor
> Z<-rnorm(N)          # Z orthogonal to X
> Y<-X+Z+rnorm(N)      # Outcome Y


> print(summary(lm(Y~X)))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.0956     0.2167   -0.44  0.66119
X             1.0301     0.2439    4.22  0.00011 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.5 on 48 degrees of freedom
Multiple R-squared:  0.271,Adjusted R-squared:  0.256
F-statistic: 17.8 on 1 and 48 DF,  p-value: 0.000107


> print(summary(lm(Y~X+Z)))

Coefficients:
            Estimate Std. Error t value      Pr(>|t|)
(Intercept)   0.0335     0.1460    0.23          0.82
X             0.9761     0.1634    5.97 0.00000029670 ***
Z             1.0439     0.1346    7.75 0.00000000059 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1 on 47 degrees of freedom
Multiple R-squared:  0.68,Adjusted R-squared:  0.666
F-statistic:   50 on 2 and 47 DF,  p-value: 0.00000000000233
```

Many Regressions

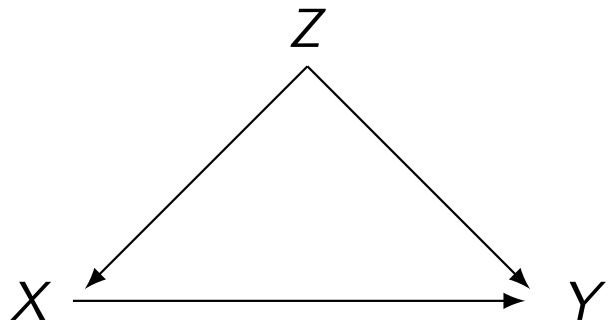


Same, But With $N = 500$



Confounding

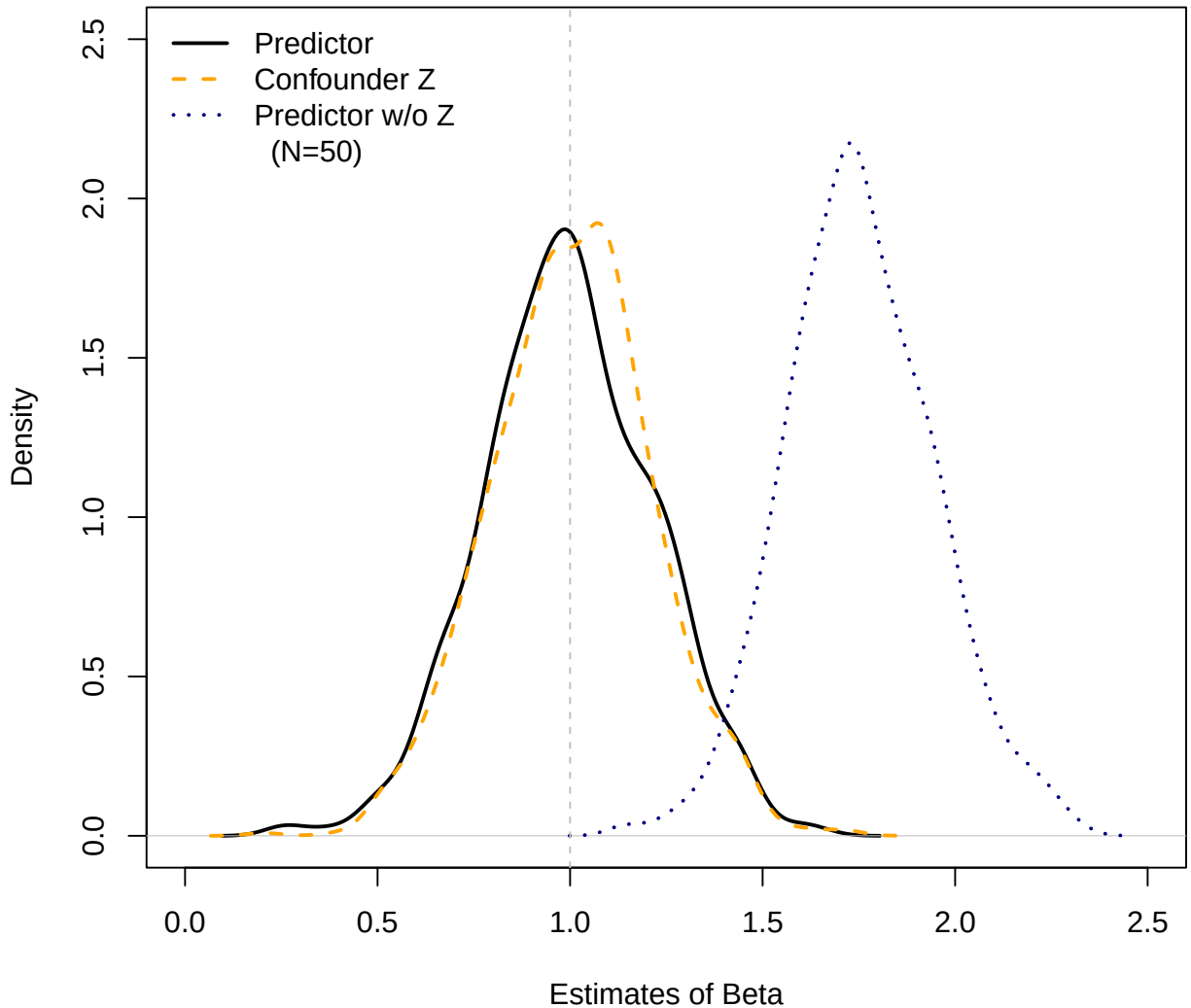
The classic example is:



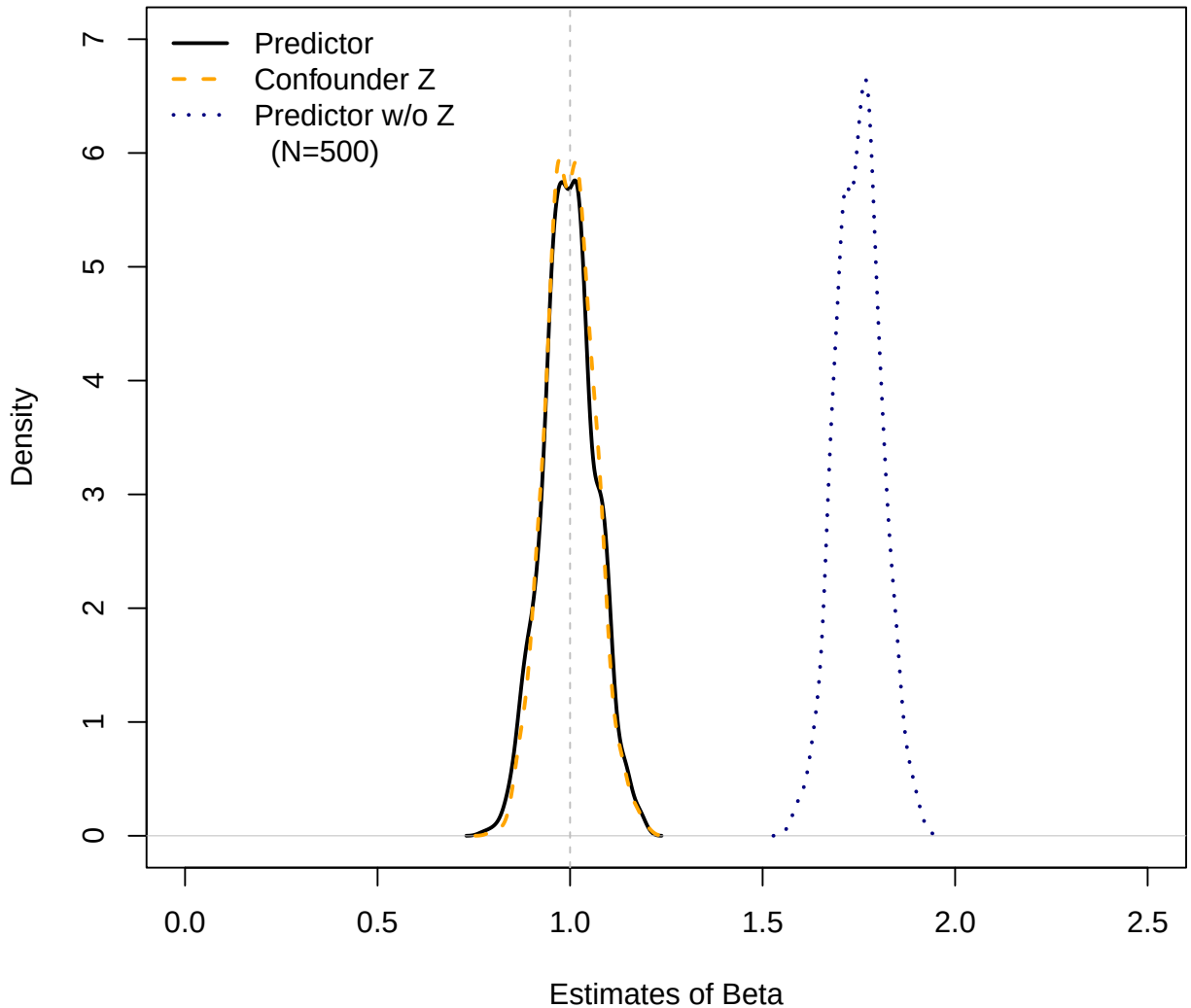
This means that:

- Z is important to / influential on both X and Y
- The marginal association $\text{Cov}(X, Y|Z)$ (obviously) depends on Z...
- More specifically, $\text{Cov}(X, Y|Z) \neq \text{Cov}(X, Y)$

So Much Confounding

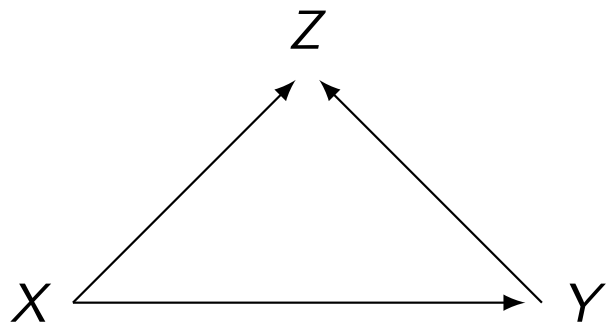


Same, But With $N = 500$



"Collider Bias"

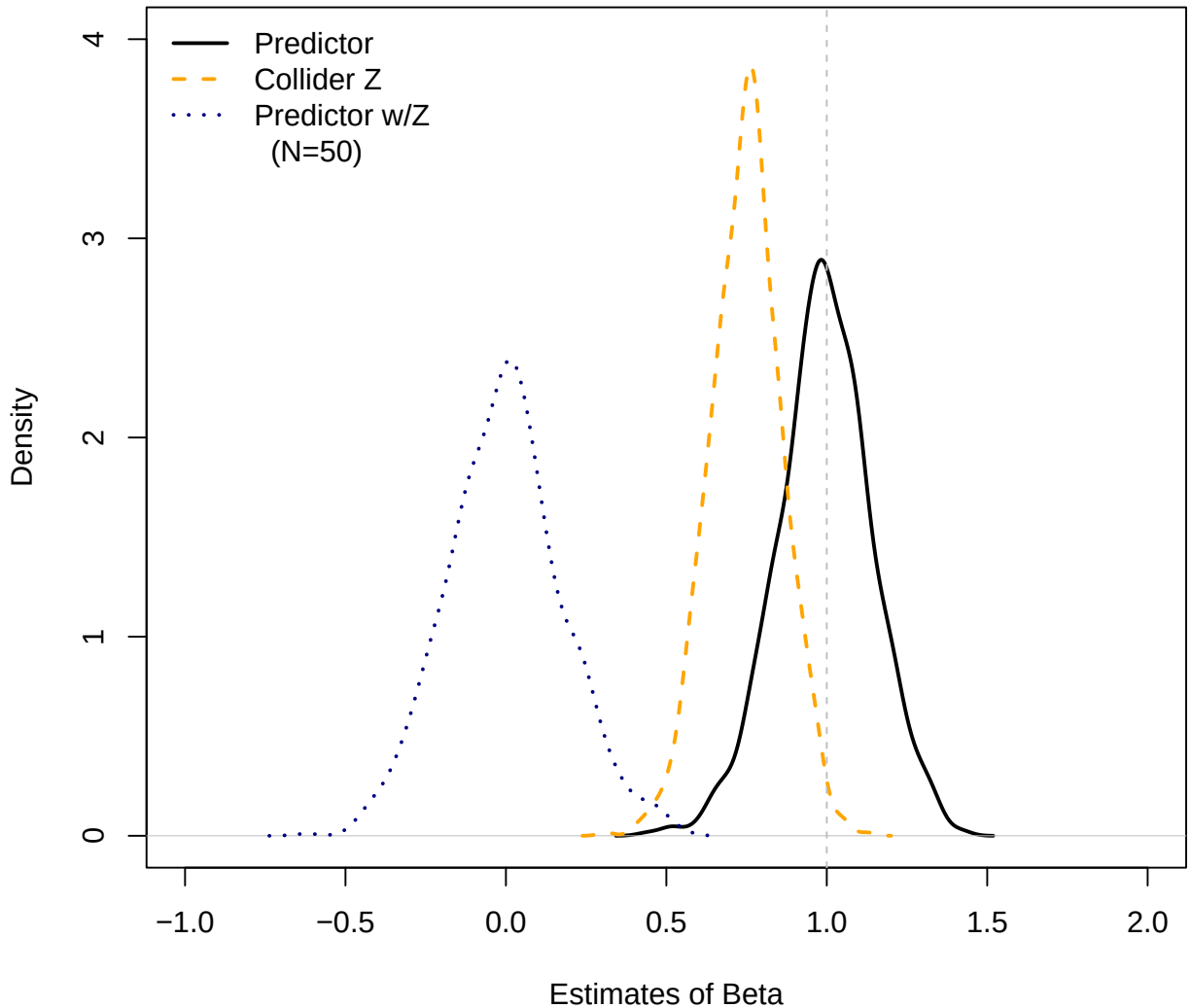
Z is a "collider":



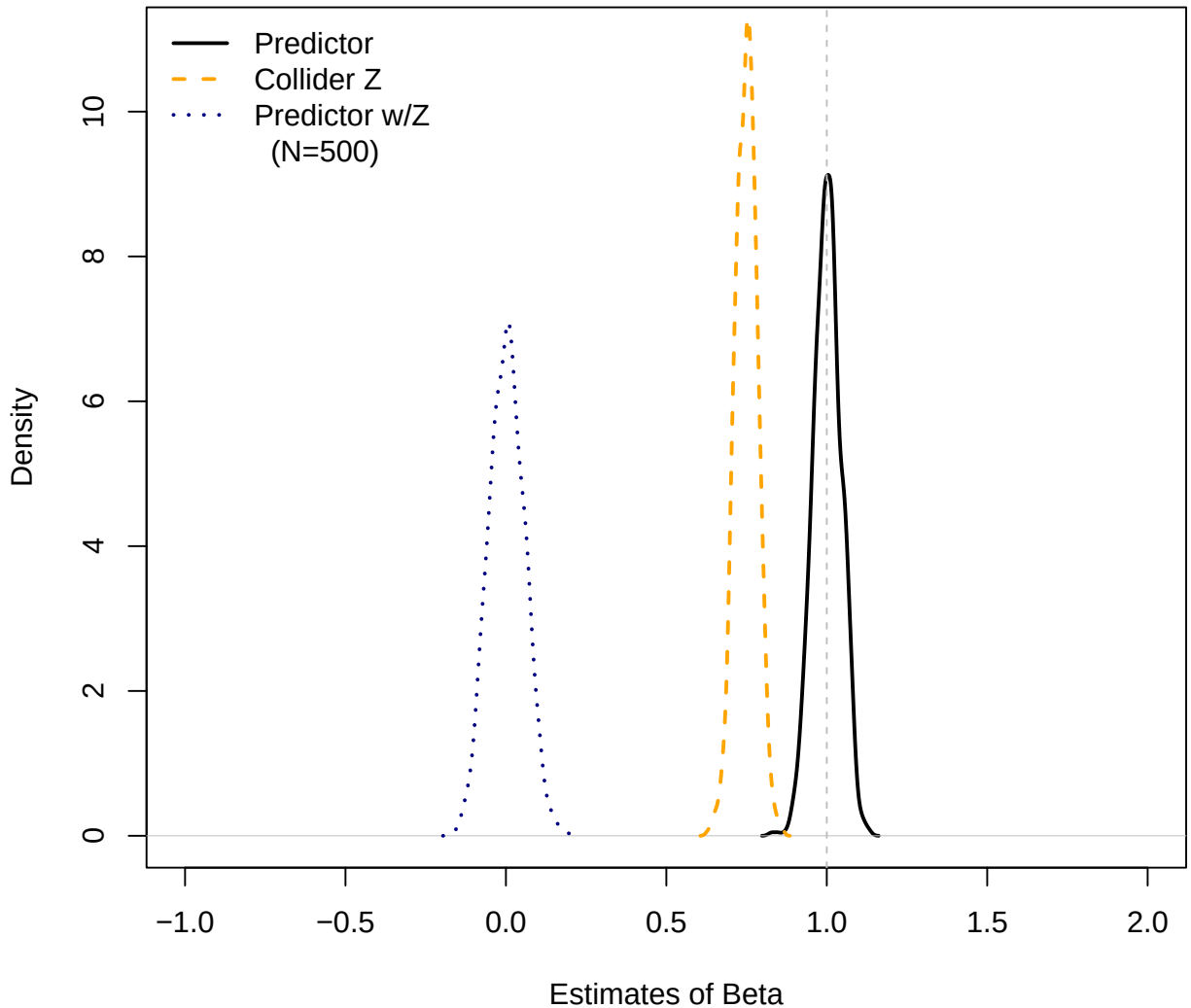
This means that:

- Z is influenced *by* both X and Y
- Once again, $\text{Cov}(X, Y|Z) \neq \text{Cov}(X, Y)$
- Sometimes referred to as Berkson's paradox

So Much Colliding

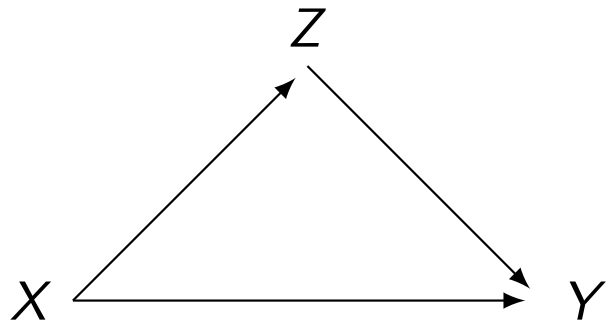


Same, But With $N = 500$



Mediators

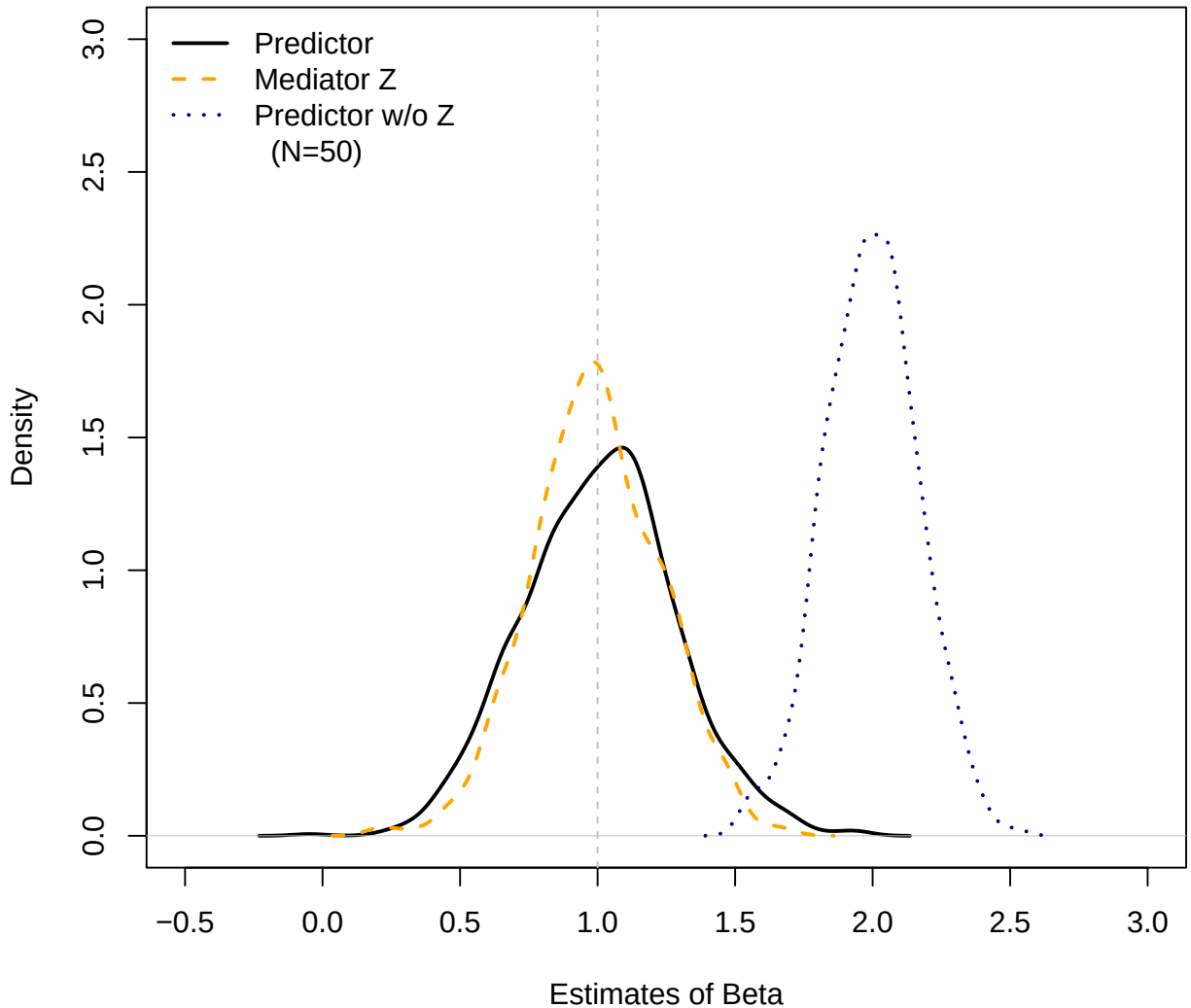
Z is a "mediator":



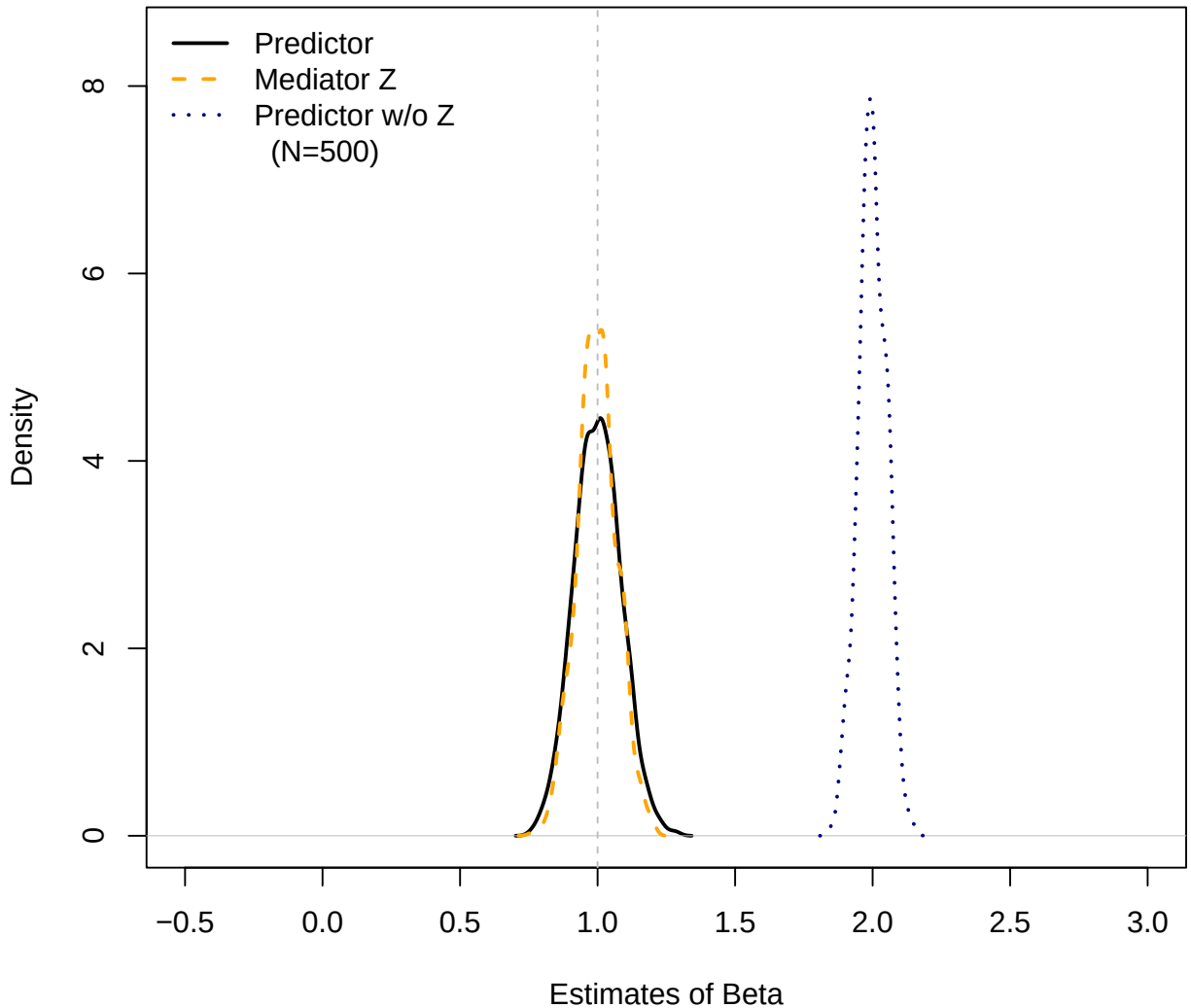
This means that:

- Z is influenced by X; Y by X and Z
- Once again, $\text{Cov}(X, Y|Z) \neq \text{Cov}(X, Y)$
- Think of $\text{Cov}(X, Y)$ = "total effect" and $\text{Cov}(X, Y|Z)$ = "direct effect"

Mediation Illustrated



Same, But With $N = 500$



Lessons...

Some takeaways... *In a linear model*:

- Variables that are irrelevant (to Y), are irrelevant...
- Variables that are relevant to Y but unrelated to X need not be modeled
- Confounders require that we condition on them, or else there's bias
- Colliders require that we *do not* condition on them, or else there's bias
- Mediators may or may not be good to condition on...

Mostly: **Model specification is hard.**

Readings...

Things to ponder:

- Berk (2010): "Types" of regressions...
- Keele et al. (2020): Interpreting "control" variables
- Rainey (2014): Negligible effects...
- Spirling & Stewart (2022): "Inference to the Best Explanation"
- Westreich & Greenland (2013): The "Table 2 Fallacy"

For Discussion: Reuveny and Li (2003)

Comparative Political Studies

Impact Factor: **3.955**
5-Year Impact Factor: **5.171**

Available access | Research article | First published online July 1, 2016

Economic Openness, Democracy, and Income Inequality: An Empirical Analysis

Rafael Reuveny and Quan Li View all authors and affiliations

Volume 36, Issue 5 | https://doi-org.ezaccess.libraries.psu.edu/10.1177/0010414003036005004

Contents | PDF / ePub | More

Abstract

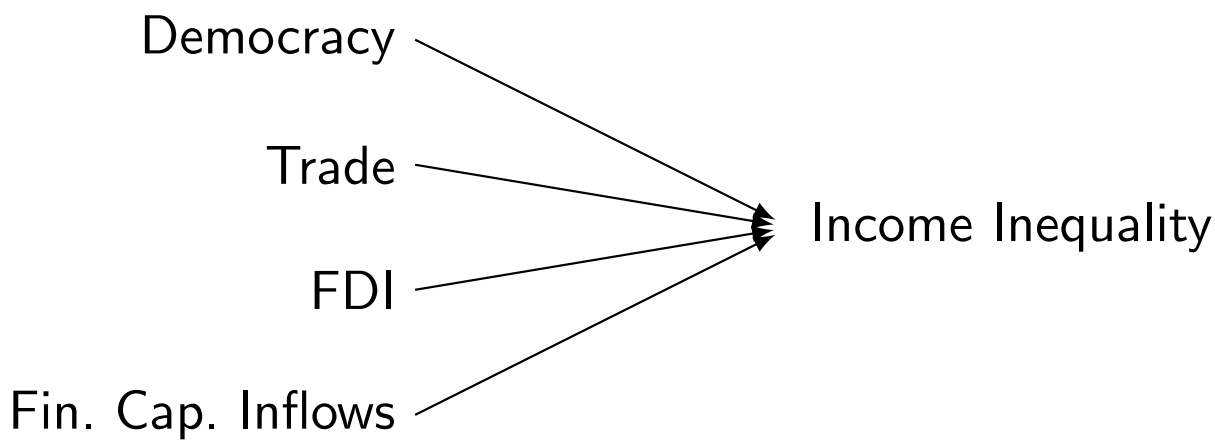
Scholars have studied effects of economic openness and democracy on national income inequality in two literatures. In democracy studies, scholars agree democracy reduces inequality but empirical evidence is ambiguous. In globalization studies, effects of economic openness on inequality are debated but have not been rigorously examined. This article is the first systematic statistical study of the effects of both economic openness and democracy on income inequality. These effects need to be studied together. The authors measure national income inequality from a comprehensive Gini coefficient data set. Economic openness is measured from trade flows, foreign direct investment inflows, and financial capital inflows. The period studied is 1960 to 1996, the unit of analysis is a country decade, and the sample includes 69 countries. The authors find that democracy and trade reduce income inequality, foreign direct investments increase income inequality, and financial capital does not affect income inequality. Policy implications are discussed.

Reuveny and Li: Hypotheses

Hypotheses:

Hypothesis 1: Democracy reduces income inequality.
Hypothesis 2: Trade increases income inequality in DCs and reduces it in LDCs.
Hypothesis 3: FDIs reduce income inequality.
Hypothesis 4: Financial capital inflows reduce income inequality.

So:



Controls (part I):

The model includes control variables frequently used in previous studies. GDP per capita (GDPPC) is expressed in purchasing power parity – adjusted international dollars. Kuznets (1955) hypothesized that below some level of GDPPC, income inequality rises with GDPPC; above this level, income inequality declines with GDPPC. This pattern is known as the Kuznets curve. Previous empirical results on the Kuznets curve are mixed. Ahluwalia, Carter, and Chenery (1979) and Higgins and Williamson (1999), for example, find evidence supporting the Kuznets hypothesis. Deininger and Squire (1998) found no supporting evidence. If Kuznets was right, the coefficients of GDPPC and $GDPPC^2$ should be positive and negative, respectively. The level of education and the share of agriculture in GDP, which can also affect income inequality, are indirectly included in the model. Both variables tend to be highly correlated with GDPPC.

Controls (part II):

Our third control variable is the one-decade-lagged Gini coefficient—the level of past income inequality. The inclusion of this variable is consistent with the observed tendency of inequality to persist over time. Several theoretical reasons account for this tendency. First, wealth concentration typically correlates positively with political influence, generating arrangements favoring wealth owners.[19] Second, people tend to marry those from the same socioeconomic group. Consequently, the children of the rich (or poor) group remain in the original group, perpetuating income differences across groups. Third, in cases where the poor and the rich belong to different ethnic groups,

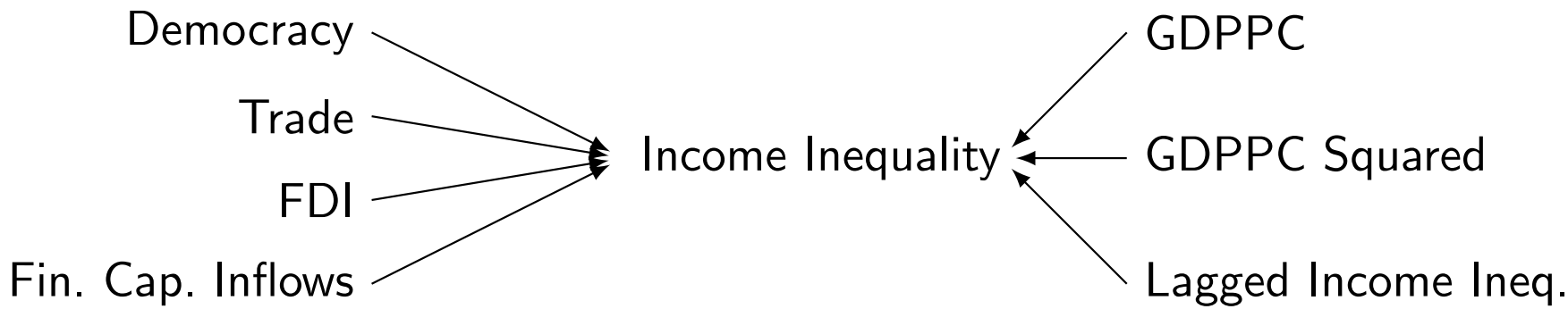
Reuveny and Li: Controls (continued)

Controls (part III):

racial discrimination can institutionalize the current income distribution (Lewis, 1994). Fourth, education can promote upward social mobility, but acquiring education is costly. Poor people tend to have more children than do rich people (Heerink, 1994). Thus, education spending per child tends to be smaller for the poor, ensuring a vicious circle. The poor remain less educated and earn less, and income inequality persists (Dasgupta, 1993).

The inclusion of past income inequality in the model also helps to control for the effect of potentially relevant but omitted structural variables, such as the ethnic and demographic structures of society. Many studies adopt this modeling strategy (e.g., Bollen, 1979; E. N. Muller, 1995; E. N. Muller & Seligson, 1994). As Burkhart and Lewis-Beck (1994) put it, "With such a pervasive control in place, it is more difficult for spurious effects to be reported" (p. 905).

Suggests (?):



Reuveny and Li: Table 1

Table 1
Income Inequality, Democracy and Economic Openness

	All	Less Developed Countries	Organization for Economic Cooperation and Development
Democracy level	–0.0125***	–0.0112***	–0.0125**
	(0.0033)	(0.0037)	(0.0056)
Trade openness	–0.0013**	–0.0013**	–0.0026**
	(0.0006)	(0.0008)	(0.0013)
Portfolio inflow	0.0074	0.0340	0.0059
	(0.0166)	(0.0367)	(0.0157)
FDI inflow	0.0632***	0.0518**	0.0590***
	(0.0229)	(0.0290)	(0.0218)
GDPPC	–1.91e – 06	3.92e – 05**	8.51e – 07
	(1.08e – 05)	(2.20e – 05)	(1.54e – 05)
$GDPPC^2$	1.90e – 10	–2.98e — 09**	2.07e – 10
	(4.40e – 10)	(1.28e – 09)	(4.85e – 10)
Past income inequality	0.7181***	0.7163***	0.4901***
	(0.0590)	(0.0693)	(0.0755)
Constant	–0.1307***	–0.1898***	–0.2718***
	(0.0402)	(0.0510)	(0.0682)
Observations	142	99	43
Adjusted R^2	0.69	0.62	0.52

Note: Huber-White robust standard errors are in parentheses and are adjusted for country. *N* is the number of observations in each sample. This number differs from the number of decades multiplied by the number of countries in a sample due to missing data and the inclusion of the lagged dependent variable. FDI = foreign direct investments; GDPPC = GDP per capita.
$^{**}p = .05$. $^{***}p = .01$.

Discussion, I:

The effect of democracy on income inequality is statistically significant at the 5% level for the DC sample and at the 1% level for the LDC and all-country samples. The effect of democracy is always negative, indicating that democracy reduces the level of income inequality. When using better data of income inequality (relative to previous studies of the effect of democracy on inequality), including economic openness in the model, and controlling for the Kuznets (1955) curve and past income inequality, our results support Hypothesis 1: Democracy reduces income inequality within countries.

Discussion, II:

The effect of trade openness on income inequality is negative and statistically significant at the 5% level for all the samples, indicating that trade openness reduces income inequality. This result supports Hypothesis 2 with regard to LDCs but not DCs. As discussed in the Effects of Democracy section, trade generates both inequality-increasing and inequality-decreasing effects. Hence, our findings can be interpreted as representing the net effect of trade on income inequality within countries, which in turn reduces income inequality.

The effect of FDI inflows on income inequality is positive and statistically significant at the 5% level for the DC and the LDC samples and at the 1% level for all countries. This result shows that FDI inflows increase income inequality. Again, FDI can generate both inequality-increasing and inequality-decreasing effects. According to our results, the net effect of FDI is to raise income inequality. We need to reject Hypothesis 3.

Reuveny and Li: Discussion (continued)

Discussion, III:

The effect of portfolio inflow on income inequality is positive in all samples, but it is never statistically significant. These results are consistent with the observation that the rise in portfolio investment inflows is a relatively recent phenomenon. If financial market integration continues to deepen, portfolio investments may significantly affect income inequality in the future. In any case, our results do not support Hypothesis 4.

Discussion, IV:

The effects of GDPPC and $GDPPC^2$ on income inequality are statistically significant in the LDC sample at the levels of 5% and 1%, respectively. GDPPC has a positive effect on income inequality, whereas $GDPPC^2$ has a negative effect. These results support the existence of a Kuznets (1955) curve for the LDCs. In the DCs and the all-country samples, the Kuznets curve is not statistically significant. The insignificance of the Kuznets curve for the DCs is to be expected. The Western European countries experienced a Kuznets curve transformation in the late 19th and early 20th centuries (Kuznets, 1955). Because this period is not included in our sample, we do not detect a Kuznets curve effect for the DCs.