# Causal Estimation, the Cross-Lagged Regression Model, and Authoritarianism: A Short Response to Luttig (2021)

## Introduction

In an age of heightened polarization – in which the political parties have grown more ideologically extreme and voters have sorted themselves into partisan camps following their ideological leanings – individual difference variables previously linked with voter behavior are now highly collinear with partisan identification. It is precisely this reason that establishing causal relationships using observational data are difficult to estimate. In a recent book, Federico, Feldman and Weber (2021) demonstrate that authoritarianism – the tendency to prioritize social conformity over personal autonomy – is now structured by party identification.

This finding extends prior work by Hetherington and Weiler (2009) whom argue that the ascendance of racial issues, law and order, feminism, and foreign affairs mobilized authoritarians to align themselves with the Republican party. Throughtout the 1990s and early 2000s, authoritarian voters defected en masse from the Democratic Party, a process that Federico et al (2021) and Inglehart and Norris (2018) describe as a component giving rise to Donald Trump's authoritarian-inspired appeal.

In a recent piece, Luttig (2021) challenges the common wisdom that authoritarianism, a highly stable individual difference factor – that some have even shown to have a genetic component – is endogenous to party identification. Luttig (2021). Using two panel designs, Luttig demonstrates that authoritarianism is not causally prior to party preference; instead, party preference is causally prior to authoritarianism. Luttig relies on a common technique: the **Cross-Lagged Regression (CLR)** model. The CLR was pioneered in Campbell (1963) as well as Campbell and Stnaley (1963) and is an oft used, intuitive method to test **competing causal processes**. Luttig demonstrates that changes in party preference in survey wave $t-1$ affect one's authoritarianism in Wave $t$. The reverse pattern, however, does not hold: Authoritarianism at $t-1$ does not predict a positive orientation towards the Republican Party at time $t$. Partisanship is endogenous to authoritarianism, not vice-versa. Luttig (2021) demonstrates this with two representative panel-structured datasets.

One explanation for this departure of common wisdom regarding authoritarianism may be due to the nature of the authoritarianism scale, as measured by the Child Rearing Questionnaire (CRQ). The CRQ is a 4-item scale that asks respondents to choose between two competing values. While conceivable, it is difficult to imagine how a respondent's preference for "independence" or "respect for elders," or "obedience" versus "self-reliance: in children could be endogenous to their party preference. While this is plausible, the CRQ was developed precisely because of the well-argued and researched claim that measures of authoritarianism, like"Right-Wing Authoritarianism" (Altemeyer 1981) and the "Fascism (F) - Scale" (Adorno et al 1950) clearly reference political content. In the case of the CRQ, it was developed to measure authoritarianism without political content. Luttig's argument rests on a cueing process, whereby partisanship coalesces with exposure to like minded partisan elites, who may present authoritarian ideas. While this is a plausible explanation in a general sense – some elites advocating authoritarianism – it less plausible in the specific sense of the CRQ. The CRQ is not a measure of authoritarianism per se, but rather a measure of **child rearing preferences**.

In this short piece, we reconsider Luttig's findings using causal diagrams to generate various causal estimands. We argue that Luttig's findings do not adequately capture the complicated relationship between authoritarianism and its political expression. Yet, the problem does not solely reside in misplaced theory or particular data, but rather in the statistical method to identify causal processes, the Cross-Lagged Regression Model. As we will show, causal inference can be accomplished using these and similar data, but inference rests with careful theoretical reasoning about the data generating process, and the use of causal diagrams, not a data-driven appraoch commonly relied upon the the CLR.

# The Cross Lagged Regression Model

It is difficult to sort out causal mechanisms with cross-sectional observational data. It is also difficult to sort out causal mechanisms with panel data. Causal relationships require theorizing about the data generating process. Data, particularly when observational, and statistical procedures do not guarantee that researcher will identify a causal mechanism.

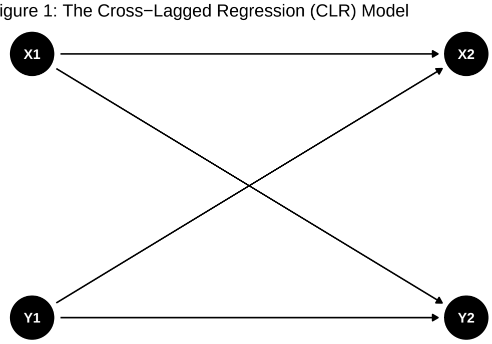
```
suppressMessages(library(dplyr))
suppressMessages(library(haven))
suppressMessages(library(dagitty))
suppressMessages(library(ggdag))
suppressMessages(library(ggplot2))
```

Consider the well-known Cross-Lagged Regression model (CLR).

```
options(repr.plot.width = 2, repr.plot.height = 2)  # Set smaller plot size
g <- dagitty("dag {
    X1 -> Y2
    Y1 -> Y2
    X1 -> X2
    Y1 -> X2

}")
coordinates(g) <- list(
    x=c(U1=0.5, Y1=1,X1=1,Y2=2,X2=2, U2=3),
    y=c(U1=1.5, Y1=1,X1=2,Y2=1,X2=2, U2=2.5) )
ggdag(g, text = TRUE) + theme_dag() + ggtitle("Figure 1: The Cross-Lagged Regression (CLR) Model")
```

Figure 1: The Cross−Lagged Regression (CLR) Model




This is a case of a Directed Acyclic Graph (DAG; Pearl 2001), as the panel structure is temporal, and the arrows are directed. The CLR is a common method to test competing causal processes. The CLR is a special case of the more general Autoregressive Distributed Lag model used in econometrics. Because there are two dependent variables, party and authoritarianism, it is common to then estimate two regression models, one with $x$ as a dependent variable, the other with $y$ as the dependent variable.

$$y_{t=2,i} = a_0 + a_1 y_{t=1,i} + a_2 x_{t=1,i} + e_{1,i} \quad (1) \tag{1}$$

$$x_{t=2,i} = b_0 + b_1 y_{t=1,i} + b_2 x_{t=1,i} + e_{2,i} \quad (2) \tag{2}$$

$$\tag{3}$$

The assumption is often that due to the temporal ordering, if one were to estimate these two equations using traditional regression techniques, a large $a_2$ coefficient represents a causal relationship from $x_1 \rightarrow y_2$; whereas a large $b_2$ coefficient translates to reversed effect, i.e., $y_1 \rightarrow x2$. If $a_2$ is large and $b_2$ is effectively zero, it is tempting to conclude that $x$ causes $y$ but $y$ does not cause $x$. In both cases, we include the lagged value of the dependent variable, $y_{t=1,i}$ and $x_{t=1,i}$, to control for the effect of the dependent variable; each unit serves as their own control, or baseline. Should $a_2$ or $b_2$ reach conventional levels of statistical significance, it is tempting to conclude that a causal effect has been identified. Economists refer to this as Granger Causality, which is a statistical test of whether the lagged value of the independent variable can predict the dependent variable after controlling for the lagged value of the dependent variable.

There are two issues that arise upon estimating this model. The first pertains to the empirical consequence of including a lagged realization of the independent variable in the regression model; the the second issue pertains to inclusion of a lagged dependent variable (Allison 1990).

Beginning with the independent variable, consider why one includes a lag and not the contemporaneous effect of $x_2$ on $y_2$ (and vice-versa). Causal inference is challenging with observational cross-sectional data, due to the researcher's inability to control for the effect of confounding covariates. This is why correct model specification is necessary. The **contemporaneous causal** effect cannot be reliably estimated because it is likely confounded by unobservable variables. It is tempting to leverage the temporal nature of the data to estimate a lagged causal effect in place of the contemporaneous causal effect.

Yet, Bellamare et al (2017) notes that a lagged realization of the independent variable does not necessarily improve one's ability to identify a causal effect. Instead, it moves the problem to identifying a lagged causal effect, effectively ignoring causal mechanisms that operate contemporaneously. Likewise, the inclusion of a lagged dependent variable is also problematic, for reasons well-documented in the literature (see Allison 1990). Introductory statistics textbooks present it in the context of autocorrelation, whereby including a lagged realization of the dependent variable will correlate with the error terms in equations (1) and (2), subsequently biasing parameter estimates and standard errors.

These problems can be particularly pernicious and difficult to resolve with two-wave panel designs (Hamaker, Kuipers and Grasman (2015). In particular, when variables in the model are stable, which is most certainly the case with authoritarianism and party/candidate evaluation, it becomes increasingly difficult to identify causal effects. For this reason, scholars have increasingly moved to adopting the Random-Intercept Cross-Lagged panel regression model (RI-CLPM), in which the paramters in equation (1) and (2) are estimated, but the intercept terms randomly vary across units, accounting for with-person variation. Unfortunately, this technique is only available when more than two-waves of panel data are available, a circumstance that does not apply to Luttig (2021).

We advocate a solution that is more qualitative in nature, by leveraging causal diagrams informed by *a priori* assumptions to generate causal estimands of interest. As we will show, the two-wave CLR model afford us limited ability to identity causal effects, and through careful consideration of the causal diagrams to identity the effects of authoritarianism on partisanship (and vice-versa), we find little evidence to support Luttig's claim that authoritarianism is endogenous to party. Before outlining our approach, we first discuss the logic of causal diagrams to identify causal effects.

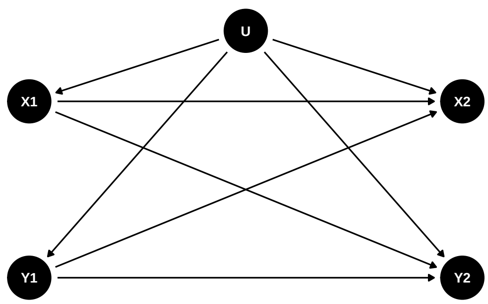
# Causal Inference using Directed Acyclic Graphs

Identitying causal effects with observational data is difficult due to the presence of confounding variables. The problem is not limited to cross-sectional data, and also apply to observational panel data (though occasionally for different reasons). Consider the causal diagram in Figure 2, a slight modification on the CLR model presented in Figure 1. In this case, we have added a confounder set $U$ that is correlated with both $x_1$ and $y_1$, as well as $x_2$ and $y_2$.

```r
g <- dagitty("dag {
    X1 -> Y2
    Y1 -> Y2
    X1 -> X2
    Y1 -> X2
    U -> X1
    U -> Y1
    U -> X2
```

```
    U -> Y2
    U <-> U
}")
coordinates(g) <- list(
    x=c( Y1=1,X1=1,Y2=2,X2=2, U=1.5),
    y=c( Y1=1,X1=2,Y2=1,X2=2, U=2.4) )
ggdag(g, text = TRUE) + theme_dag() + ggtitle("Figure 2: The Cross-Lagged Regression (CLR) Model.\nConf
```

Figure 2: The Cross−Lagged Regression (CLR) Model.
Confounding by U



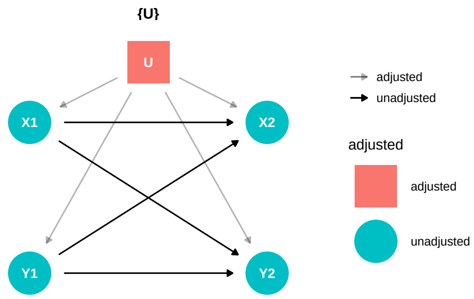

In this case, the confounder set creates a "backdoor path" from U to $x_2$ and $y_2$ through $x_1$ and $y_1$. Ignoring the confounder set $U$ will bias the estimates of the coefficients estimated in the model; this is the notion of "spurious regression" whereby we may even estimate a relationship when none exist.

The solution advocated in most textbooks is to include the confounder set $U$ in one's regression model. This is also known as "back-door adjustment" as we condition on this confounder set blocking any paths that run through $U$. This is reason researchers often include a set of control variables, occasionally an unwieldy set, a "kitchen sink" of numerous variables that may or may not be controls (King, 20??). Luttig (2021), for instance, controls for sex, education, income, race/ethnicity. These variables are likely to be correlated with both authoritarianism and party/candidate evaluation, and thus should be included in the regression model.

The figure below demonstrates that including the confounder set $U$ blocks the backdoor path from $x_2$ to $y_2$ through $x_1$ and $y_1$.

```
ggdag_adjustment_set(g, exposure = "Y1", outcome = "Y2", shadow = "TRUE") + theme_dag() + ggtitle("Figu
```

Figure 3: We need to condition on:



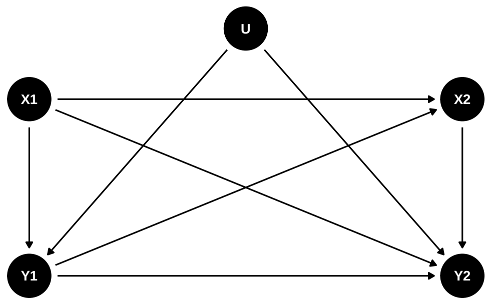

But what if prior research indicates that $U$ isn't a common cause of both $x$ and $y$, but rather a cause of $y$. Also, assume a causal arrow between $x_1$ and $y_1$ and $x_2$ and $y_2$? This is contemporaneous causal correlation, which is not absent from data because a temporal component is specified. So, we also draw an arrow from $x_1 \rightarrow y_1$ and $x_2 \rightarrow y_2$.

Now the causal diagram becomes:

```
g <- dagitty("dag {
    X1 -> Y2
    X2 -> Y2
    Y1 -> Y2
```

```
    X1 -> X2
    Y1 -> X2
    U -> Y1
    U -> Y2
    X1 -> Y1
}")
coordinates(g) <- list(
    x=c( Y1=1,X1=1,Y2=2,X2=2, U=1.5),
    y=c( Y1=1,X1=2,Y2=1,X2=2, U=2.4) )
ggdag(g, text = TRUE) + theme_dag() + ggtitle("Figure 4: The Cross-Lagged Regression (CLR) Model.\nConte
```

Figure 4: The Cross−Lagged Regression (CLR) Model.
Contemporaneous Causal Effect




In this case, the problem is somewhat different. $y_1$ and $y_2$ are "colliders" in the model, and conditioning on $y_1$ – the lag – will open a backdoor path from $U$ to $y_2$ and $x_2$, a problem known as "collider bias" (Pearl 2001). In this case, we are better served by excluding $y_1$ from the estimated equation, adopting a simpler model, $y_2 = \beta_1 x_1 + \epsilon$.

### Emprical Problems

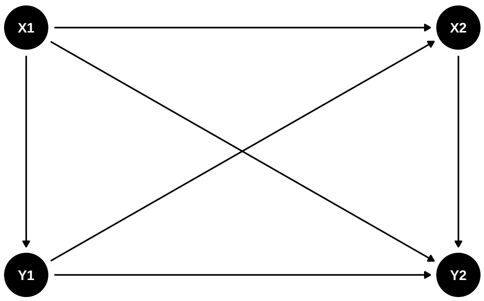
Analysts who use the CLR attempt to sidestep contemporaneous causality – a difficult to defend concept when using observational data – by using lagged effects. The idea is relatively simple and elegant: if we measure $x$ at time $t$ and $y$ at time $t+1$, then we can be confident that $x$ is not caused by $y$. Any effect of of $x$ that precedes $y$ must be causal. While this may appear reasonable at face-value, it may be fallacious for several reasons.

### Contemporaneous Causal Effects

```
g <- dagitty("dag {
    X1 -> Y2
    X1 -> Y1[style=dashed]
    Y1 -> Y2
    X1 -> X2
    Y1 -> X2
    X2 -> Y2

}")
coordinates(g) <- list(
    x=c( Y1=1,X1=1,Y2=2,X2=2, U=1.5),
    y=c( Y1=1,X1=2,Y2=1,X2=2, U=2.4) )
ggdag(g, text = TRUE) + theme_dag() + ggtitle("Figure XX: The Cross-Lagged Regression (CLR) Model.\nCont
```

Figure XX: The Cross−Lagged Regression (CLR) Model.
Contemporaneous Causal Effect

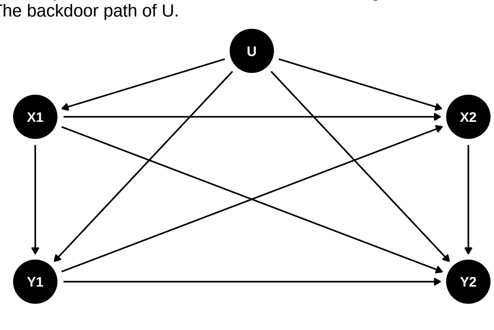



There is nothing inherently problematic about this causal effect. But the CLR in this circumstance is capturing an effect that does not necessarily comport to what Luttig (2021) describes. By conditioning on both $X1$ and $Y1$ though, we are not obtaining the total effect of $X1$ on $Y2$. Instead, we are capturing the **direct effect** of $X1$ on $Y2$ after accounting for $Y2$, and where $Y2$ mediates the relationship. $Y1$ is an intervening, mediating variable; it structures the effect of $X1$. Put another way, $X1$ may in fact have a rather large and substantial effect on $X2$. If we think that authoritarianism does not directly influence party evaluations at time $t+1$ because its effects are mediated by party evaluations at time $t$, then we are not capturing the total effect of authoritarianism on party evaluations in the CLR. In fact, this story is largely consistent with Federico, Feldman and Weber (2023) who show that authoritarianism does not have direct effects in many circumstances, precisely because authoritarians have sorted themselves into partisan camps and have adopted a partisan lens in which the world is viewed.

The problem, however, is when unobserved variables also predict $Y1$. Such is the case below,

```
g <- dagitty("dag {
    X1 -> Y2
    X2 -> Y2
    Y1 -> Y2
    X1 -> X2
    Y1 -> X2
    U -> Y1
    U -> Y2
    U -> X1
    U -> X2
    X1 -> Y1
}")
coordinates(g) <- list(
    x=c( Y1=1,X1=1,Y2=2,X2=2, U=1.5),
    y=c( Y1=1,X1=2,Y2=1,X2=2, U=2.4) )
ggdag(g, text = TRUE) + theme_dag() + ggtitle("Figure 4: The Cross-Lagged Regression (CLR) Model.\nConte
```

Figure 4: The Cross−Lagged Regression (CLR) Model.
Contemporaneous Causal Effect and Confounding.
The backdoor path of U.




The problem is now not just in interpretation, but also estimation. Adding a causal arrow from U to Y1 and Y2, we have a confounding variable. And this confounding variable creates a bias in the CLR model.

By conditioning on $Y1$ – a collider – we open a backdoor path from $U$ to $Y2$ and $X2$. The effect of these variables, along with authoritarianism are captured in the estimated coefficient. This, however, is clearly not what we want. We would like to know the effect of authoritarianism on party evaluations, not the effect of authoritarianism, $Y1$ and $U$ on party evaluations. This is case of "m-bias", because we are conditioning on an effect that is a consequence of $X1$ but also unbobserved variables $U$. The CLR model will estimate the effect of $X1$ on $Y2$, but it will be an incorrect estimate. The bias will bend in the direction of the confounding variable in relation to the observed variables. If $U$ is positively correlated with $X1$ and $Y2$, then the CLR will overestimate the effect of $X1$ on $Y2$. If $U$ is negatively correlated with $X1$ and $Y2$, then the CLR will underestimate the effect of $X1$ on $Y2$. In either case, the CLR will be biased.

This problem is well-established in the causal effects literature, as it represents the potentially problematic consequences of conditioning on a post-treatment variable, in this case, $Y1$, since $X1 \rightarrow Y1$. If other variables are left unspecified, then this "conditioning" exercise gives misleading results.

Together, not only must we consider the types of effects in our model – a matter of interpretation – we must also reason about the underlying causal structure in the data. Failing to consider the unobervable data may undermine our ability to accurately estimate both indirect and direct effects.

### Lagged Effects

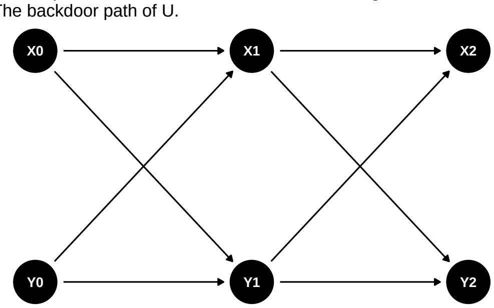
Let us assume there are no contemporaneous causal effects, and the CLR is generally correctly specified, wherein the only causal parameters are lagged effects. Figure 1, in other words, accurately captures the causal structure of the data. The assumption of the model is that $X1$ and $y1$ are exogeneous: They both predict $Y2$ but do not predict each other. This is generally implausible. Imagine a circumstance, for instance, where we were to observe an additional wave of panel data, where,

```
g <- dagitty("dag {
    X1 -> Y2
    Y1 -> Y2
    X1 -> X2
    Y1 -> X2
    X0 -> X1 [color=gray]
    Y0 -> Y1 [color=gray]
    X0 -> Y1 [color=gray]
    Y0 -> X1 [color=gray]
}")

coordinates(g) <- list(
    x=c( Y0 = 0, X0 = 0, Y1=1,X1=1,Y2=2,X2=2, U=1.5),
    y=c(Y0 = 1, X0 = 2, Y1=1,X1=2,Y2=1,X2=2, U=2.4) )

ggdag(g, text = TRUE) + theme_dag() + ggtitle("Figure 4: The Cross-Lagged Regression (CLR) Model.\nConte
```

Figure 4: The Cross−Lagged Regression (CLR) Model.
Contemporaneous Causal Effect and Confounding.
The backdoor path of U.




In this case, the CLR model is not identified. The reason is that $X1$ and $Y1$ are not exogeneous. They are both predicted by $X0$ and $Y0$. Conditioning on $Y1$ and $X1$ in our model opens a backdoor path from $X0$

and $Y0$ to $Y2$ and $X2$. Thus, we should acknowledge that – again, because the data are observational in nature – the parameters in the CLR model may not coincide to the parameters or effects as interpreted in a non-causal regression model The non-exogeneity of the lagged variables is a problem for the CLR model. The problem becomes intractable when some common cause of these variables is ignored. A confounder variable

The idea is that the lagged variable is measured prior to the outcome variable, and thus the lagged variable cannot be caused by the outcome variable. This is the logic of the CLR model, and the reason why the CLR model is often used in panel data settings.

## Data, Models and Estimation

As Luttig notes:

**But previous analyses of the authoritarian divide separating Republicans from Democrats, and Trump supporters from opponents, rest primarily on correlational analyses that are unable to determine whether authoritarianism shapes partisan preferences, or whether political attitudes may instead shape how respondents answer the authoritarianism questionnaire. This is problematic, as recent research raises questions about the validity, over-time stability, and exogeneity of the child rearing measure of authoritarianism** (Luttig, 2021, 3).

We will show that more careful reasoning about the data and CLR necessitates a reconsideration of Luttig's findings.

There are three datasets, all located in /data

- /data/panel_data.2016.rda. This is the 2016-2020 ANES
- /data/dataverse_files/merge201214nes.dta This is the 2012-13 ANES reported in Luttig (2021)
- /data/dataverse_files/2016 panel study.dta This is the SSI data reported in Luttig (2021)

/data/dataverse_files/ also includes Luttig's (2021) replication code.

## Replication of Luttig (2021)

Models and simulations are included in /models

- clr.R is the CLR model reported by Luttig (2021), but also includes the 2016-2020 ANES.
- gEst.R includes a a simple g-estimator to estimate the causal effect of authoritarianism on candidate evaluation. The logic here is relatively straightforward. This is explained well in Hernan and Robins (2023). The estimation steps are:
  1) Use background variables to generate a propensity score, create inverse probability weights of treatment.
  2) Estimate the treatment effect using the inverse probability weights.
  3) Calculate the average treatment effect of authoritarianism on candidate evaluation, marginalizing over the data.
- Much of the problem with the CLR in Luttig (2021) is that.
  1) Match ANES1213 respondents with ANES2016 respondents. Use the match to impute 2012 scores for the 2016 data, as-if such data existed.
  2) Estimate a RI-CLR model using the imputed data, to account for within-unit variation.
- Simulate data and conduct sensitivity analysis

## Simulate Data

Assume this model, where Z represent a set of exogenous covariates, and U represents a set of unobserved confounders.

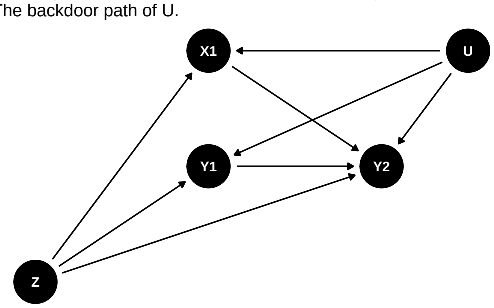
```r
g <- dagitty("dag {
    X1 -> Y2
    Y1 -> Y2
    Z  -> Y2 [color=gray]
    Z  -> Y1 [color=gray]
    Z  -> X1 [color=gray]
    U  -> Y2 [color=gray]
    U  -> Y1 [color=gray]
    U  -> X1 [color=gray]

    }")

coordinates(g) <- list(
    x=c(Z = 0,  Y1=1,X1=1,Y2=2, U=2.5),
    y=c(Z = 0,  Y1=1,X1=2,Y2=1, U=2) )

ggdag(g, text = TRUE) + theme_dag() + ggtitle("Figure 4: The Cross-Lagged Regression (CLR) Model.\nConte
```

Figure 4: The Cross−Lagged Regression (CLR) Model.
Contemporaneous Causal Effect and Confounding.
The backdoor path of U.




Conditioning on Y1 or X1 opens a backdoor path from U to Y2. The structural model:

$$y_{1,i} = a_1 z_i + a_2 u_i + e_{1,i} \quad (1) \quad (4)$$

$$x_{1,i} = b_1 z_i + b_2 u_i + e_{2,i} \quad (2) \quad (5)$$

$$y_{2,i} = c_1 y_{1,i} + c_2 x_{1,i} + c_3 z_i + c_4 u_i + e_{3,i} \quad (3) \quad (6)$$

$$(7)$$

If both U and Z are observed, then the causal estimand is

$$E(Y^{X1=1} - Y^{X1=0}) = E_{Z,U}[E(Y|X1 = 0, Z, U) - [E(Y|X1 = 0, Z, U)] = c_2$$

$$E_Z[E(Y|X1 = 0, Z) - [E(Y|X1 = 0, Z)] = c_2$$