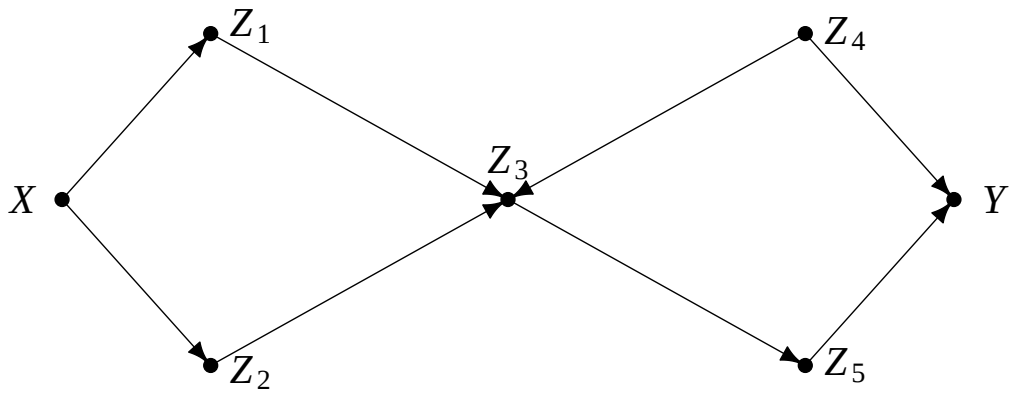
Causal Inference Using Graphs: Problem Set 1

1. Consider the Directed Acyclic Graph below and answer the following questions.



   a. What are all of the paths from X to Y?
   b. Which of these paths are blocked by conditioning on Z_3?
   c. Is there another conditioning variable that blocks the same set of paths?

2. In seminal work on Time-Series Cross-Section (TSCS) models, Beck and Katz (1995, 1996) describe some of the choices that researchers face when modeling TSCS data. As noted in Beck and Katz (1996), one of the most important choices faced by TSCS modelers is the choice of conditioning set for their regressions. In that paper, the "generic TSCS model" is written as the following:

$$y_{i,t} = \beta_0 + \beta_1 x_{i,t} + \epsilon_{i,t}; \quad i = 1, ..., n; \quad t = 1, ..., T, \tag{1}$$

   To simplify the discussion, we will assume that $x_{i,t}$ and $y_{i,t}$ are scalars and that $T = 2$ (i.e., we will assume that we have single explanatory variable within each time period, and we will assume that we only have data on two time periods).[1] In Beck and Katz (1996), one of the models discussed is the generic TSCS model with serially correlated errors.

$$\begin{aligned} y_{i,t} &= \beta_0 + \beta_1 x_{i,t} + \epsilon_{i,t}; \quad i = 1, ..., n; \quad t = 1, 2 \\ \epsilon_{i,t} &= \rho \epsilon_{i,t-1} + \nu_{i,t} \end{aligned} \tag{2}$$

   Figure 1 (a) presents the basic causal structure implied by this model. Suppose that in addition to serially correlated errors, we believe that there is serial correlation in the explanatory variable as well.

$$\begin{aligned} y_{i,t} &= \beta_0 + \beta_1 x_{i,t} + \epsilon_{i,t}; \quad i = 1, ..., n; \quad t = 1, 2 \\ \epsilon_{i,t} &= \rho \epsilon_{i,t-1} + \nu_{i,t} \\ x_{i,t} &= \delta x_{i,t-1} + \gamma_{i,t} \end{aligned} \tag{3}$$

   Figure 1 (b) presents the basic causal structure implied by this model.

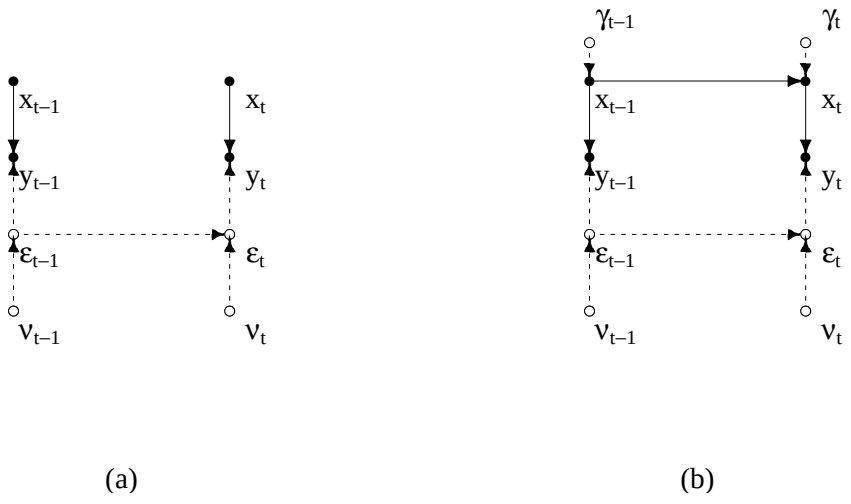
[1] The actual model (Beck and Katz 1996, pg. 3) allows $x_{i,t}$ to be a K vector of explanatory variables and allows $T > 2$.

a. Assuming that $x_{i,0} = 0$ and $\epsilon_{i,0} = 0$ for all i and assuming that $\beta_0 = 0$, $\beta_1 = 1$, $\rho = \delta = .8$, $\nu_1 \sim N(0,1)$, $\nu_2 \sim N(0,1)$, $\gamma_1 \sim N(0,1)$, $\gamma_2 \sim N(0,1)$, simulate 10,000 data sets of size n =1,000. In other words, you should simulate 10,000 data sets where each data has 1,000 observations of the variables $\{x_1, x_2, y_1, y_2\}$. For each data set run the following two regressions,

   - $y_2 \sim x_2$
   - $y_2 \sim x_2 + y_1$

   Each regression provides an estimate of β_1, so you should have two vectors each with 10,000 estimates of β_1. We know from our simulation that the true effect of x_2 on y_2 is $\beta_1 = 1$. Which of the two regressions seems to provide an unbiased estimate?

b. According to the true model (which we know is true because we simulated our data from it), the coefficient on y_1 in the model for y_2 is implicitly zero (y_1 isn't in the model for y_2). According to most econometrics textbooks, including an irrelevant variable in our regression cannot bias our estimate for β_1. Comment on what your results mean for this claim.



(a) (b)

Figure 1: *Causal graphs consistent with the model represented in Beck and Katz (1996) equation 13.*