SUPPLEMENTAL APPENDIX

1. SIMULTANEOUS CAUSALITY

UNBIASEDNESS OF THE ORDINARY LEAST SQUARES ESTIMATOR

To understand the implications of simultaneous causality for regression estimates, it is illustrative to revisit the proof of unbiasedness of the ordinary least squares (OLS) estimator. A biased estimator is an estimator whose expected value is not equal to the true value of the population parameter. For an estimator to be unbiased requires a strong assumption of exogeneity, namely mean independence: $E[U|X] = 0$. For clarity, we proceed in matrix form. We write the regression equation $Y = X\beta + U$, where our dependent variable, Y, and our error term U are $n \times 1$ vectors and X is an n x p matrix where n is the number of observations and p is the number of parameters.

We can write the OLS estimator of β in the regression equation $Y = X\beta + U$ as

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

Plugging the regression equation for Y and rearranging (note $(X'X)^{-1}X'X = I$),

$$\hat{\beta} = \beta + (X'X)^{-1}X'U.$$

Taking the expected value of both sides,

$$E[\hat{\beta}] = E[\beta] + E[(X'X)^{-1}X'U].$$

Applying the law of iterated expectations,

$$E[\hat{\beta}] = \beta + E\left[E[(X'X)^{-1}X'U|X]\right].$$

Thus, if one or more covariates in X are endogenous (in this case, $E[UX] \neq 0$), the second term will not drop out of the equation and $E[\hat{\beta}] \neq \beta$.

APPLICATION TO LYME DISEASE AND FOREST FRAGMENTATION

Above we illustrated that correlation between regressors and the error term results in biased OLS estimates. If a regressor and the dependent variable are simultaneously determined, endogeneity problems will result in unreliable estimates of the slope coefficients caused by violation of the exogeneity assumption. For instance, if the population living in the wildland urban interface (WUIpop) is positively related to Lyme disease incidence (LDI) (e.g., through fragmentation or through increased interaction with infected nymphs), then we would expect an increase in WUIpop to lead to an increase in LDI. However, what if LDI influences the number of persons living in the WUI? If causality moves in both directions, then the estimated effect of WUIpop on LDI is biased (and inconsistent) because of correlation between WUIpop and the error term.

To show this result, let the data-generating process for WUIpop in county i in a given year be defined as

$$WUIpop_i = \beta_0 + \beta_1 LDI_i + \varepsilon_i \qquad (A.1)$$

and the data-generating process for LDI be defined as

$$LDI_i = \gamma_0 + \gamma_1 WUIpop_i + \mu_i \qquad (A.2)$$

Suppose we are interested in estimating the causal effect of WUIpop on LDI (i.e., estimating the parameter γ_1 equation A.2). Substituting the right-hand side of equation (A.2) for LDI in equation A.1 yields

$$WUIpop_i = \beta_0 + \beta_1(\gamma_0 + \gamma_1 WUIpop_i + \mu_i) + \varepsilon_i.$$

Combining terms,

$$WUIpop_i = \beta_0 + \beta_1\gamma_0 + \beta_1\gamma_1 WUIpop_i + \beta_1\mu_1 + \varepsilon_i$$

$$WUIpop_i = \delta_0 + \delta_1 + \frac{\beta_1}{1 - \beta_1\gamma_1}\mu_i + \frac{1}{1 - \beta_1\gamma_1}\varepsilon_i$$

$$\text{where } \delta_0 = \frac{\beta_0}{1 - \beta_1\gamma_1} \text{ and } \delta_1 = \frac{\beta_1\gamma_0}{1 - \beta_1\gamma_1}.$$

Multiplying both sides by μ_i and taking the expected value yields

$$E[WUIpop_i \cdot \mu_i] = \delta_0 E[\mu_i] + \delta_1 E[\mu_i] + \frac{\beta_1}{1 - \beta_1\gamma_1}E[\mu_i \cdot \mu_i] + \frac{1}{1 - \beta_1\gamma_1}E[\varepsilon_i \cdot \mu_i].$$

If the common assumptions of the linear regression model hold, in particular that μ_I has a mean of zero, ε_i is uncorrelated with μ_i, and μ_i has non-zero variance σ^2_μ,

$$E[WUIpop_i \cdot \mu_i] = \frac{\beta_1}{1 - \beta_1\gamma_1}E[\mu_i \cdot \mu_i] = \frac{\beta_1}{1 - \beta_1\gamma_1}\sigma^2_\mu \neq 0$$

As a result of the non-zero correlation between WUIpop and the error term μ_i, estimating equation (A.2) would lead to a biased estimate of γ_1. Similarly, we can show that there is non-zero correlation between LDI and ε_i by substituting the right-hand side of equation (A.1) for WUIpop in equation (A.2). This would result in a biased estimate of β_1. Thus, separate OLS estimation of either equation (A.1) or equation (A.2) produces biased estimates when WUIpop and LDI are simultaneously determined.

2. FIXED EFFECTS VERSUS RANDOM EFFECTS MODELS

STATISTICAL MODELS

A cross-sectional model uses data from one time period to quantify the effect of independent variables such as forest fragmentation on a dependent variable such as Lyme disease incidence. The advantage of a cross-sectional regression is that it allows for a larger range of adaptations at high or low levels of Lyme disease. However, the cross-sectional model is only valid if the estimated effect of forest fragmentation on LDI is unbiased or consistent.[1]

There are a number of factors that potentially influence LDI and the WUI population, including the number and locations of roads and recreational areas. Ignoring or imprecisely measuring these factors may result in biased estimation

of the effect of WUIpop (and all other regressors) on LDI. The magnitude and sign of the omitted variable bias is difficult to decipher in a multiple regression model.[1]

With data for multiple time periods, it is possible to control for unobserved factors that may bias the cross-sectional regression. In a fixed-effects model, a separate intercept term is estimated for each county, thereby controlling for any time-invariant factors that influence the dependent variable. Or, the fixed effect removes the time-invariant factors from the error term, limiting the potential for correlation with regressors in the model.

A fixed-effects version of LDI model in equation (A.2) is given by

$$LDI_{it} = \gamma_0 + \gamma_1 WUIpop_{it} + c_i + \mu_{it} \quad (A.3)$$

The difference with equation (A.2) is that LDI, WUIpop, and μ are now indexed by county i and year t and the term c_i is included to capture additional differences among counties. In particular, the c_i term measures the combined influence of all time-invariant factors on LDI. The inclusion of the c_i term is equivalent to de-meaning the data (i.e., $n^1 \sum_t X_{it}$ is subtracting from each variable X_{it}), which implies that the model parameters are identified from the deviations of variables around their mean.

Assumptions regarding the relationship between the county effect, c_i, and the observable variables such as WUIpop define the difference between fixed and random effects models. The key assumption of the random effects model is that the unobserved effects (the c_i) are uncorrelated with the explanatory variable WUIpop. The random effects model can be written as

$$LDI_{it} = \gamma_0 + \gamma_1 WUIpop_{it} + v_{it} \quad (A.4)$$

where $v_{it} = c_i + \varepsilon_{it}$. Whereas consistent estimation of the fixed effects model in equation (A.3) requires the errors (μ) to be uncorrelated with WUIpop, the random effects model imposes the additional assumption that the c terms are uncorrelated with WUIpop (and with μ).

In many applications, the assumption that unobserved time-invariant effects are uncorrelated with the explanatory variables is unreasonable. As discussed above, roads or state regulations may be unobserved, but it is likely that access and development laws are correlated with the proportion of a county's population residing in the WUI. The benefit of the fixed effects model is that it does not place restrictions on the relationship between the time-invariant unobserved effects and the observable covariates. In contrast, if c_i is correlated with the covariates and included as part of the error term (i.e., as a random effect), coefficient estimates will be biased and inconsistent. Even when the random effects assumptions hold, the fixed effects estimator will be consistent and unbiased, though less efficient than the random effects estimator.

SUPPLEMENTAL APPENDIX REFERENCE

1. Wooldridge JM, 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press, 247–282.