Regression methods for systems of equations

Introduction

Due to the prevalence of feedback effects within forest systems, forest modelers must often deal with interdependence among the variables they are trying to predict. We can choose to ignore these interdependencies and fit each equation separately using ordinary least squares (OLS). A more appropriate method, however, would be to fit the series of equations simultaneously. Fitting a system of equations simultaneously can be achieved by using one of four main estimation methods, depending on the characteristics of the equations.

Estimation Methods

1) **Two stage least square regression (2SLS)**: This method of estimation should be used when we have a system of equations where at least one y-variable occurs on the left-hand-side (LHS) of an equation and on the right-hand-side (RHS) of at least one other equation in the system. For example:

$$Y_{1t} = \alpha_0 + \alpha_1 Y_{2t} + \alpha_2 X_{1t} + \alpha_3 X_{2t} + \epsilon_{1t}$$

$$Y_{2t} = \beta_0 + \beta_1 Y_{1t} + \beta_2 X_{3t} + \beta_3 X_{2t} + \epsilon_{2t}$$

Since, both Y-variables in the system occur on the LHS and RHS, they are known as endogenous variables, while the X-variables are exogenous variables. Because the Y-variables are jointly determined there is a violation of independence of between the endogenous variable on the RHS and the error term. Fitting the equations using OLS results in simultaneity bias. If the error terms are not correlated across equations and are i.i.d. (independent and identically distributed), then 2SLS may be used whereby:

   a. a first-stage estimate of the Y-variables is obtained through a linear model using only the X-variables, plus any other variables that are good predictors of Y, but are not correlated with the error term (called **instrumental variables**). This filters out the correlation of the error with the endogenous variable.

   b. a second stage estimate of the Y-variables is obtained using the original equations, but replacing the endogenous variables on the RHS with the first stage estimates of the Y-variables.

2) **Seemingly unrelated regression (SUR):** If the Y-variables in the system of equations only occur on the LHS, but are inherently interdependent and estimated using a similar set of X-variables, then the error terms are likely to be correlated across equations. If this is the case, and the errors are iid, then SUR may be used, whereby:

   a. An estimate of the cross-equation error-covariance matrix is obtained by fitting the equations using OLS.

   b. The equations are then refit using the estimated error-covariance matrix in an Estimated Generalized Least Squares (EGLS) fit.

3) **Three stage least squares regression (3SLS):** If conditions described in 1) and 2) hold, and the errors are iid, then 3SLS may be used to remove simultaneity bias and improve the efficiency of the estimators by taking into account the cross-equation error-covariance matrix. The steps of a 3SLS are:

    a. a first-stage estimate of the Y-variables is obtained through a linear model as described in 1a.

    b. a second stage estimate of the Y-variables is obtained using the original equations, but replacing the endogenous variables on the RHS with the first stage estimates of the Y-variables.

    c. An estimate of the cross-equation error-covariance matrix is obtained following the second stage fit.

    d. In the third stage, the equations from the second stage are refit using the estimated error-covariance matrix in an Estimated Generalized Least Squares (EGLS) fit.

4) **Multi-stage least squares regression (MSLS):** If conditions described in 1) and 2) hold, and the errors are **NOT** iid, then multi-stage least squares regression may be used. For an in depth description of MSLS, see LeMay (1990). This method of fitting systems of equations is not explicitly included in the R 'systemfit' package.

Note: In the case of SUR and 3SLS, an iterative approach to estimating the cross equation error-covariance matrix may be used. Under the iterative approach, the equations are refit several times using an EGLS fit, each time using a new (and hopefully improved) estimate of the cross equation error-covariance matrix. Lastly, as an alternative to the instrumental variables method used in 2SLS and 3SLS, limited information maximum likelihood (LIML) for 2SLS and full information maximum likelihood (FIML) for 3SLS may be used. This, however, assumes that the errors are normally distributed. The iterative method is available in the R 'systemfit' package, while the LIML and FIML are not.

Nonlinear systems of equations

If the equations within the system are nonlinear in their parameters, then nonlinear systems regression may be used. These include nonlinear two stage regression (N2SLS) and nonlinear three stage regression (N3SLS). Conceptually, the approach to fitting the nonlinear systems of equations is the similar to 2SLS and 3SLS. Computationally, however, there are some differences with regard to how the instrumental variables equation is used to purge the correlation of the error with the endogenous variable (Jorgenson and Laffont 1974; Judge et al. 1985). Nonlinear systems regression may be performed using the 'systemfit' package in R, however, this is a relatively new function and equation fits are extremely sensitive the starting values of parameters (Henningsen and Hamann 2007).

Examples of a linear system of equations using R:

During the R showcase, I will present two examples for fitting systems of equations:

Example 1) In this example, we wish to obtain estimates of crown radius (crad) at the base of the tree crown and crown height (ch) by fitting models for crad and ch as a system of equations. The equations

are fit individually using OLS, and then fit as a system using 2SLS and 3SLS. Separate instrumental variables equations are used to filter out the correlation of the endogenous variables with the error terms.

Example 2) For this example, a seemingly unrelated regression is used to fit component tree biomass equations for stem wood biomass, foliage biomass, and branch biomass. The equations are first fit using OLS, and then fit using SUR.

References

Henningsen, A., and Hamann, J.D. 2007. systemfit: A package for estimating systems of simultaneous equations in R. J. of Stat. Soft., 23(4): 1-40.

Jorgenson, D.W. , and Laffont. J.-J. 1974. Efficient estimation of nonlinear simultaneous equations with additive disturbances. Ann. Soc. Econ. Meas. 3: 615-640.

Judge, G.G., Hill, R.C., Griffiths, W.E., Lutkepohl, H., and Lee, T.C. 1985. The theory and practice of econometrics. 2nd ed. John Wiley and Sons, New York.

LeMay, V.M. 1990. MSLS: a linear least squares technique for fitting a simultaneous system of equations with a generalized error structure. Can. J. For. Res. 20: 1830-1839.