Lecture 9

Regression extensions – Multivariate and nonlinear regression

Lectures in SDPE: *Econometrics I* on February 26, 2024

Markus Jäntti
Swedish Institute For Social Research
Stockholm University

Introduction

- This lecture discusses least squares based regression extensions – e.g., generalized least squares (GLS), esp. Seemingly Unrelated Regressions [SUR, chapter 11 in Hansen (2021)] and non-linear models, estimated using non-linear least squares, [NLS, chapter 23 in Hansen (2021)]
- The treatment is brief, but
  - GLS/SUR and NLS both contribute to understanding an important later topic, generalized method of moments (GMM)
  - NLS is in a way "close" to the linear projection model studied earlier

- We now consider a set of regression equations of the form

$$Y_{ji} = \boldsymbol{X}'_{ji}\beta_j + e_{ji}. \tag{1}$$

  There are $j = 1, \ldots, m$ dependent variables and $i = 1, \ldots, n$ observations with each vector of regressors $\boldsymbol{X}_{ji}$ and associated coefficient vector β_j having k_j elements; e_{ji} is a regression error and the total number of coefficients is $\overline{k} = \sum_{j=1}^{m} k_j$.
- Set up in this way, observations i are treated as independent but the variables j as correlated (and not only through $\boldsymbol{X}_{ji}$). A typical example might be the consumption of household i of different goods j.
- The dependence across variables is captured by the $m \times m$ covariance matrix $\boldsymbol{\Sigma}_i$:

$$\mathrm{E}[\boldsymbol{e}_i\boldsymbol{e}'_i] = \boldsymbol{\Sigma}_i, \tag{2}$$

  where $\boldsymbol{e}_i$ is the vector of m regression errors for observation i.

- A more compact way of writing the system is in terms of a $m \times 1$ dependent variable and regressions error, a $m \times \overline{k}$ regressor with associated $m \times 1$ coefficients:

$$\boldsymbol{y}_i = \overline{\boldsymbol{X}_i}\beta + \boldsymbol{e}_i \tag{3}$$

where the $m \times 1$ dependent variable is $\boldsymbol{y}_i = (Y_{1i}, \ldots, Y_{mi})'$ and

$$\overline{\boldsymbol{X}_i} = \begin{bmatrix} X'_{1i} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & X'_{2i} & \ldots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & X'_{mi} \end{bmatrix}, \tag{4}$$

or. . .

- . . . by stacking all n observations into $mn \times 1$ and $mn \times \overline{k}$ matrices

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_m \end{bmatrix}, \qquad \boldsymbol{e} = \begin{bmatrix} \boldsymbol{e}_1 \\ \vdots \\ \boldsymbol{e}_m \end{bmatrix}, \qquad \overline{\boldsymbol{X}} = \begin{bmatrix} \overline{\boldsymbol{X}_1} \\ \vdots \\ \overline{\boldsymbol{X}_m} \end{bmatrix}, \tag{5}$$

- . . . so we have

$$\boldsymbol{y} = \overline{\boldsymbol{X}}\beta + \boldsymbol{e} \tag{6}$$

- We might have the same regressors in all equations, so $\boldsymbol{X}_{ji} = \boldsymbol{X}_i$ and $k_j = k$, which can be written in many ways also, e.g., using a $m \times k$ parameter matrix

$$\boldsymbol{y}_i = \boldsymbol{B}\boldsymbol{X}_i + \boldsymbol{e}_i, \quad \boldsymbol{B} = (\beta_1, \ldots, \beta_m) \tag{7}$$

or as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E} \tag{8}$$

where $\boldsymbol{Y}$ and $\boldsymbol{E}$ are $n \times m$ matrices.

- With the same regressors, we can sometimes use the convenient notation involving the Kronecker product $\otimes$ that

$$\overline{\boldsymbol{X}_i} = \begin{bmatrix} \boldsymbol{X}_i' & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}_i' & \ldots & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \ldots & \boldsymbol{X}_i' \end{bmatrix} = \boldsymbol{I}_m \otimes \boldsymbol{X}_i' \tag{9}$$

- One approach to estimation is to apply least squares to each of the j equations in eq 1,

$$\widehat{\beta}_j = \left(\sum_{i=1}^{n} \boldsymbol{X}_{ji}\boldsymbol{X}'_{ji}\right)^{-1}\left(\sum_{i=1}^{n} \boldsymbol{X}_{ji}Y_{ji}\right) \tag{10}$$

and the full set of coefficients is $\widehat{\beta} = (\widehat{\beta}'_1, \ldots, \widehat{\beta}'_m)'$.

- To estimate (under homoscedasticity) the error covariance matrix, note that the residuals are
$$\widehat{\boldsymbol{e}_i} = \boldsymbol{y}_i - \overline{\boldsymbol{X}_i}\widehat{\beta}. \tag{11}$$
- The feasible estimator of the $m \times m$ variance matrix is then
$$\widehat{\boldsymbol{\Sigma}} = n^{-1} \sum_{i=1}^{n} \widehat{\boldsymbol{e}_i}\widehat{\boldsymbol{e}_i}'. \tag{12}$$

Mean and Variance of Systems Least-Squares

- In order to determine the conditional mean and variance of $\widehat{\beta}$, we make the strong assumption of conditional mean independence, $\mathrm{E}[\boldsymbol{e}_i|\boldsymbol{X}_i] = \mathbf{0}$ (here $\boldsymbol{X}_i$ is the union of all $\boldsymbol{X}_{ji}$). It follows that $\mathrm{E}[Y_{ji}|\boldsymbol{X}_i] = \boldsymbol{X}'_{ji}\beta_j$
- To obtain the mean, center the estimator:

$$\widehat{\beta} - \beta = \left(\overline{\boldsymbol{X}}'\overline{\boldsymbol{X}}\right)^{-1}\left(\overline{\boldsymbol{X}}'\boldsymbol{e}\right) = \left(\sum_{i=1}^{n}\overline{\boldsymbol{X}}'_i\overline{\boldsymbol{X}}_i\right)^{-1}\left(\sum_{i=1}^{n}\overline{\boldsymbol{X}}'_i\boldsymbol{e}_i\right). \qquad (13)$$

- Now take the conditional expectation:

$$\mathrm{E}[\widehat{\beta} - \beta|\boldsymbol{X}] = \beta - \beta = \mathbf{0}. \qquad (14)$$

Mean and Variance of Systems Least-Squares

- To get the variance of the estimator, define (cf. equation 2)

$$\mathrm{E}[\boldsymbol{e}_i \boldsymbol{e}_i' | \boldsymbol{X}_i] = \boldsymbol{\Sigma}_i. \tag{15}$$

- With independence across observations, we have

$$\begin{aligned} \mathrm{E}[\boldsymbol{e}\boldsymbol{e}' | \boldsymbol{X}] &= \mathrm{E}\left(\begin{bmatrix} \boldsymbol{e}_1 \boldsymbol{e}_1' & \boldsymbol{e}_1 \boldsymbol{e}_2' & \dots & \boldsymbol{e}_1 \boldsymbol{e}_n \\ \vdots & \ddots & \dots & \vdots \\ \boldsymbol{e}_n \boldsymbol{e}_1' & \boldsymbol{e}_2 \boldsymbol{e}_1' & \dots & \boldsymbol{e}_n \boldsymbol{e}_n \end{bmatrix} \middle| \boldsymbol{X} \right) \\ &= \begin{bmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \vdots & \ddots & \dots & \vdots \\ \boldsymbol{0} & \dots & \boldsymbol{0} & \boldsymbol{\Sigma}_n \end{bmatrix} \end{aligned} \tag{16}$$

- By independence across i we also have

$$\mathrm{Var}\left[\sum_{i=1}^{n} \overline{\boldsymbol{X}}_i' \boldsymbol{e}_i \middle| \boldsymbol{X} \right] = \sum_{i=1}^{n} \mathrm{Var}[\overline{\boldsymbol{X}}_i' \boldsymbol{e}_i | \boldsymbol{X}_i] = \sum_{i=1}^{n} \overline{\boldsymbol{X}}_i' \boldsymbol{\Sigma}_i \overline{\boldsymbol{X}}_i. \tag{17}$$

The variance of $\widehat{\beta}$ follows as:

$$\mathrm{Var}[\widehat{\beta} | \boldsymbol{X}] = \left(\overline{\boldsymbol{X}}' \overline{\boldsymbol{X}} \right)^{-1} \left(\sum_{i=1}^{n} \overline{\boldsymbol{X}}_i' \boldsymbol{\Sigma}_i \overline{\boldsymbol{X}}_i \right) \left(\overline{\boldsymbol{X}}' \overline{\boldsymbol{X}} \right)^{-1}. \tag{18}$$

- With common regressors, matters simplify:

$$\operatorname{Var}[\widehat{\beta}|\boldsymbol{X}] = \left(\boldsymbol{I}_m \otimes (\boldsymbol{X}'\boldsymbol{X})^{-1}\right)\left(\sum_{i=1}^{n}(\boldsymbol{\Sigma}_i \otimes \boldsymbol{X}_i\boldsymbol{X}_i')\right)\left(\boldsymbol{I}_m \otimes (\boldsymbol{X}'\boldsymbol{X})^{-1}\right) \tag{19}$$

- With conditionally homoscedastic regressors ($\boldsymbol{\Sigma}_i \equiv \boldsymbol{\Sigma}$),

$$\operatorname{Var}[\widehat{\beta}|\boldsymbol{X}] = \left(\overline{\boldsymbol{X}}'\overline{\boldsymbol{X}}\right)^{-1}\left(\sum_{i=1}^{n}\overline{\boldsymbol{X}}_i'\boldsymbol{\Sigma}\overline{\boldsymbol{X}}_i\right)\left(\overline{\boldsymbol{X}}'\overline{\boldsymbol{X}}\right)^{-1}. \tag{20}$$

- And with both common regressors and homoscedastic errors,

$$\operatorname{Var}[\widehat{\beta}|\boldsymbol{X}] = \boldsymbol{\Sigma} \otimes (\boldsymbol{X}'\boldsymbol{X})^{-1} \tag{21}$$

Asymptotic Distribution

- For the asymptotic distribution, we can make do with the equation-by-equation linear projection condition, $\mathrm{E}[\boldsymbol{X}_{ji} e_{ji}] = 0$. This makes our $\widehat{\beta}_j$ *consistent* for β_j (and all of β).
- The asymptotic *marginal* distribution of each $\widehat{\beta}_j$ is normal, but we need some additional material to determine the *joint* distribution of $\widehat{\beta}$.
- By our assumptions, the vector

$$\overline{\boldsymbol{X}_i}' \boldsymbol{e}_i = \begin{bmatrix} \boldsymbol{X}_{1i} e_{1i} \\ \vdots \\ \boldsymbol{X}_{mi} e_{mi} \end{bmatrix} \tag{22}$$

is iid and has mean zero, the central limit theorem gives

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \overline{\boldsymbol{X}_i}' \boldsymbol{e}_i \xrightarrow{d} N(0, \boldsymbol{\Omega}), \tag{23}$$

where

$$\boldsymbol{\Omega} = \mathrm{E}\left[\overline{\boldsymbol{X}_i}' \boldsymbol{e}_i \boldsymbol{e}_i' \overline{\boldsymbol{X}_i}\right] = \mathrm{E}\left[\overline{\boldsymbol{X}_i}' \boldsymbol{\Sigma}_i \overline{\boldsymbol{X}_i}\right]. \tag{24}$$

- The rest follows a familiar pattern; denoting $\mathrm{E}[\overline{\boldsymbol{X}_i}'\overline{\boldsymbol{X}_i}] = \boldsymbol{Q}$, we get for the centered estimator

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{d} N(0, \boldsymbol{V}_\beta) \tag{25}$$

where $\boldsymbol{V}_\beta = \boldsymbol{Q}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}^{-1}$.

- The usefulness (and necessity) of this setup becomes apparent when we consider testing for joint hypotheses of the familiar form $\theta = \boldsymbol{r}(\beta) = \boldsymbol{r}(\beta_1, \ldots, \beta_m)$ with LS estimate $\widehat{\theta} = \boldsymbol{r}(\widehat{\beta}_1, \ldots, \widehat{\beta}_m)$, for which

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} N(0, \boldsymbol{V}_\theta) \tag{26}$$

where $\boldsymbol{V}_\theta = \boldsymbol{R}'\boldsymbol{V}_\beta\boldsymbol{R}$ and

$$\boldsymbol{R} = \frac{\partial \boldsymbol{r}(\beta)}{\partial \beta}.$$

To test for hypotheses across equations thus requires the full system error covariance.

- To estimate the covariance matrix of the estimator, we rely in the general case of conditional heteroscedasticity and non-identical covariates as

$$\widehat{\boldsymbol{V}}_{\widehat{\beta}} = \left(\overline{\boldsymbol{X}}'\overline{\boldsymbol{X}}\right)^{-1}\left(\sum_{i=1}^{n}\overline{\boldsymbol{X}}_i'\widehat{\boldsymbol{e}}_i\widehat{\boldsymbol{e}}_i'\overline{\boldsymbol{X}}_i\right)\left(\overline{\boldsymbol{X}}'\overline{\boldsymbol{X}}\right)^{-1}. \tag{27}$$

- Under the standard assumptions made for the single-equation case,

$$n\widehat{\boldsymbol{V}}_{\widehat{\beta}} \xrightarrow{p} \boldsymbol{V}_{\beta}. \tag{28}$$

Seemingly Unrelated Regression

- A special case of multivariate regression is Seemingly Unrelated Regression (SUR) where "seemingly" unrelated observations are nonetheless related through common shocks.
- Consider the conditionally homoscedastic regression assuming conditional mean independence,

$$\boldsymbol{y}_i = \overline{\boldsymbol{X}}_i\beta + \boldsymbol{e}_i, \quad \mathrm{E}[\boldsymbol{e}_i|\boldsymbol{X}_i] = 0, \quad \mathrm{E}[\boldsymbol{e}_i\boldsymbol{e}_i'|\boldsymbol{X}_i] = \boldsymbol{\Sigma} \tag{29}$$

- The generalized least squares estimator of β is

$$\begin{aligned}\widetilde{\boldsymbol{\beta}} &= \left(\sum_{i=1}^{n}\overline{\boldsymbol{X}}_i'\boldsymbol{\Sigma}^{-1}\overline{\boldsymbol{X}}_i\right)^{-1}\left(\sum_{i=1}^{n}\overline{\boldsymbol{X}}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_i\right) \\ &= \left(\overline{\boldsymbol{X}}'(\boldsymbol{I}_n \otimes \boldsymbol{\Sigma}^{-1})\overline{\boldsymbol{X}}\right)^{-1}\left(\overline{\boldsymbol{X}}'(\boldsymbol{I}_n \otimes \boldsymbol{\Sigma}^{-1})\boldsymbol{y}\right).\end{aligned} \tag{30}$$

- Once you replace the unknown $\boldsymbol{\Sigma}$ by its estimator, this is a feasible GLS estimator better known as SUR:

$$\widehat{\beta}_{sur} = \left(\overline{\boldsymbol{X}}'(\boldsymbol{I}_n \otimes \widehat{\boldsymbol{\Sigma}}^{-1})\overline{\boldsymbol{X}}\right)^{-1}\left(\overline{\boldsymbol{X}}'(\boldsymbol{I}_n \otimes \widehat{\boldsymbol{\Sigma}}^{-1})\boldsymbol{y}\right) \tag{31}$$

An embarrassing example

- In a paper coauthored by yours truly (Jäntti, Pirttilä, and Selin 2015), we estimate among other things labour supply by regressing hours h worked on log net wages $\ln w(1-\tau)$ and some controls,
$$h = \beta_1 \ln w(1-\tau) + \cdots + \beta_k + e. \tag{32}$$
- Our interest is not in β_1 but in the elasticity of hours wrt net wages, which in this semi log specification is
$$\eta = \frac{\beta_1}{\mathrm{E}[h]}.$$
- In the paper, we treat $\widehat{\mathrm{E}[h]}$ as constant, which it is not, since it is an estimator and, moreover, it is an estimator that is correlated with $\widehat{\beta}_1$.
- We should have setup a system as
$$\begin{aligned} h &= \ln w(1-\tau)\beta_{11} + \cdots + \beta_{1k} + e_1 \\ h &= \beta_{21} + e_2 \end{aligned} \tag{33}$$
and estimated the elasticity by
$$\widehat{\eta} = \frac{\widehat{\beta}_{11}}{\widehat{\beta}_{21}}$$
and worked out the standard error of this using the delta method (but did not!).

- Consider the CEF (with a single, i.e., scalar, X):

$$\mathrm{E}[Y|X] = m(X) = \theta_1 + \theta_2 \exp^{\theta_3 X} \tag{34}$$

- This is not linear in the parameters and the coefficients can not be estimated by OLS.
- We can formulate this as a non-linear regression:

$$Y = \theta_1 + \theta_2 \exp^{\theta_3 X} + e \tag{35}$$

- The sum of squared deviations for sample data is

$$S_n(\theta) = \frac{1}{2}\sum_{i=1}^{n} e_i^2 = \frac{1}{2}\sum_{i=1}^{n} [Y_i - m(X_i, \theta)]^2. \tag{36}$$

- The NLS estimator is the solution to the first-order condition:

$$\frac{\partial S(\theta)}{\partial \theta} = -\sum_{i=1}^{n} [Y_i - m(X_i, \theta)] \frac{\partial m(X_i, \theta)}{\partial \theta} = \mathbf{0}. \tag{37}$$

- For our example equation 34 these are:

$$\begin{aligned}
\frac{\partial S(\theta)}{\partial \theta_1} &= -\sum_{i=1}^{n} [Y_i - \theta_1 - \theta_2 e^{\theta_3 X_i}] = 0 \\
\frac{\partial S(\theta)}{\partial \theta_2} &= -\sum_{i=1}^{n} [Y_i - \theta_1 - \theta_2 e^{\theta_3 X_i}] e^{\theta_3 X_i} = 0 \\
\frac{\partial S(\theta)}{\partial \theta_3} &= -\sum_{i=1}^{n} [Y_i - \theta_1 - \theta_2 e^{\theta_3 X_i}] \theta_2 X_i e^{\theta_3 X_i} = 0
\end{aligned} \tag{38}$$

These do not have a closed form solution, so θ needs to be estimated using iterative numerical methods.

Non-linear Least Squares (NLS)

- The NLS estimator is an example of an “m-estimator” (see Hansen 2021, ch 22)
- This class includes the LS estimators we have considered hitherto
- For the asymptotic distribution of the NLS estimator (and more generally, m-estimators) requires some additional assumptions (see Hansen 2021, p 777):
  - the parameter set is compact
  - the moment function is finite and continuous in the parameters and bound
  - the criterion function converges uniformly in the parameter set

- For differentiable $m()$, with the vector of partial derivatives

$$\boldsymbol{m}_\theta(\boldsymbol{X}, \theta) = \frac{\partial m(\boldsymbol{X}, \theta)}{\partial \theta} \tag{39}$$

the non-linear least squares (NLS) estimate is asymptotically normal with

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{V}_\theta) \tag{40}$$

with $\boldsymbol{V}_\theta = \mathrm{E}[\boldsymbol{m}_\theta \boldsymbol{m}'_\theta]^{-1} \mathrm{E}[\boldsymbol{m}_\theta \boldsymbol{m}'_\theta e^2] \mathrm{E}[\boldsymbol{m}_\theta \boldsymbol{m}'_\theta]^{-1}$ which can be estimated using the NLS residuals and the "plug-in" estimate of $\mathrm{E}[\boldsymbol{m}_\theta \boldsymbol{m}'_\theta]$ at sample data points and the NLS estimate of θ.

Testing for Omitted NonLinearity

- If the worry is that a given linear regression omits non-linear regressors, this can easily be tested using conventional methods.
- Suppose we first estimate the regression

$$Y = \boldsymbol{X}'\beta + e. \tag{41}$$

$\boldsymbol{Z} = \boldsymbol{h}(\boldsymbol{X})$ is a set of non-linear functions of $\boldsymbol{X}$. Omitted non-linearity can be tested by estimating

$$Y = \boldsymbol{X}'\widetilde{\beta} + \boldsymbol{Z}'\widetilde{\gamma} + \tilde{e} \tag{42}$$

and testing $\gamma = \mathbf{0}$ using a Wald test.

- A variant is the RESET test, which uses fitted values from the "short" regression $\hat{Y}_i = \boldsymbol{X}_i'\widehat{\beta}$ to form $\boldsymbol{Z}_i' = (\hat{Y}_i^2, \ldots, \hat{Y}_i^m)$,

$$Y_i = \boldsymbol{X}_i'\widetilde{\beta} + \boldsymbol{Z}_i'\widetilde{\gamma} + \tilde{e}_i. \tag{43}$$

Again, the Wald statistic of the hypothesis that $\gamma = \mathbf{0}$ (with a χ^2_{m-1} distribution) is a test for omitted non-linearity.

Hansen, Bruce E (2021). *Econometrics*. Madison, WI: University of Wisconsin.

Jäntti, Markus, Pirttilä, Jukka, and Selin, Håkan (2015). "Estimating labour supply elasticities based on cross-country micro data: A bridge between micro and macro estimates?" *Journal of Public Economics* 127, pp. 87–99. DOI: doi:10.1016/j.jpubeco.2014.12.006. URL: http://www.sciencedirect.com/science/article/pii/S0047272714002527.