

OVERVIEW

In Part III, we have extended the classical linear regression model to data-generating processes that are nonnormal, nonspherical, and nonlinear. The chapters work progressively through these new situations.

1. If y_n is not normally distributed conditional on \mathbf{x}_n , then the distribution theory of the OLS estimator becomes intractable. Such nonlinear estimators as LAD may be relatively efficient, but their distributions are no more tractable. Asymptotic distribution theory provides an approximate, normal distribution for these estimators. Such approximations require fairly modest restrictions on the distribution of $\{(y_n, \mathbf{x}_n), n = 1, \dots, N\}$ and sample sizes N that are sufficiently large.
2. Given a specification of the conditional p.f. $f(y_n; \boldsymbol{\theta}_0 | \mathbf{x}_n)$, one can derive alternative, nonlinear estimators of the regression parameters for nonnormal distributions with the maximum likelihood estimator (MLE)

$$\hat{\boldsymbol{\theta}}_{\text{ML}} \equiv \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} E_N[L(\boldsymbol{\theta})]$$

where

$$L(\boldsymbol{\theta}) \equiv \log f(y_n; \boldsymbol{\theta} | \mathbf{x}_n)$$

is the conditional log-likelihood function. According to the Cramér–Rao lower bound, the variance of unbiased estimators for the parameter vector $\boldsymbol{\theta}_0$ is bounded below by $[N \cdot \mathfrak{S}(\boldsymbol{\theta}_0)]^{-1}$ where

$$\mathfrak{S}(\boldsymbol{\theta}_0) \equiv \operatorname{Var}[L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0; y_n | \mathbf{x}_n)]$$

is the information matrix and

$$L_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \equiv \frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

is the score vector.

3. In some cases, the MLE is unbiased and achieves this variance bound. More generally, the MLE is approximated by

$$\sqrt{N} \left(\hat{\boldsymbol{\theta}}_{\text{ML}} - \boldsymbol{\theta}_0 \right) \stackrel{p}{\equiv} \mathfrak{S}(\boldsymbol{\theta}_0)^{-1} \sqrt{N} E_N[L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)] \quad (23.1)$$

so that the relative efficiency of the MLE is asymptotic.

4. The method of maximum likelihood (ML) applies equally well to nonspherical distributions when one loosens the second moment assumptions of the classical model. Conditional heteroskedasticity and autoregressive serial correlation are leading examples of situations in which $\text{Var}[\mathbf{y} | \mathbf{X}] = \boldsymbol{\Omega}_0$ is not a scalar matrix. The MLE has a convenient interpretation as generalized least squares (GLS),

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} \equiv (\mathbf{X}' \boldsymbol{\Omega}_0^{-1} \mathbf{X})^{-1} \mathbf{X}' \boldsymbol{\Omega}_0^{-1} \mathbf{y}$$

5. Many exceptions to the classical first moment assumption arise as latent variable models. In these models, $E(y_n | \mathbf{x}_n)$ is no longer the simple linear function $\mathbf{x}'_n \boldsymbol{\beta}_0$ where $\boldsymbol{\beta}_0$ is the parameter vector of interest. When there is a vector of instrumental variables \mathbf{z}_n having the same dimension as \mathbf{x}_n and possessing the properties that

$$\begin{aligned} E[y_n | \mathbf{z}_n] &= E(\mathbf{x}_n | \mathbf{z}_n) \boldsymbol{\beta}_0 \\ E[\mathbf{z}_n \mathbf{x}'_n] &\text{ is nonsingular} \end{aligned}$$

then $\boldsymbol{\beta}_0$ is identified. The moment equations (or orthogonality conditions)

$$E[\mathbf{z}_n (y_n - \mathbf{x}'_n \boldsymbol{\beta}_0)] = \mathbf{0}$$

suggest the instrumental variables (IV) estimator

$$\hat{\boldsymbol{\beta}}_{\text{IV}} \equiv (\mathbf{Z}' \mathbf{X})^{-1} \mathbf{Z}' \mathbf{y}$$

Relative efficiency may be achieved among IV estimators if there are functions of \mathbf{z}_n that are MMSE predictors of \mathbf{x}_n .

6. Alternatively, the conditional first moment restriction may be explicitly nonlinear in $\boldsymbol{\beta}_0$, as in

$$E[y_n - \mu(\boldsymbol{\beta}_0; \mathbf{x}_n) | \mathbf{z}_n] = 0$$

or yet more generally,

$$E[\mathbf{g}(\boldsymbol{\beta}_0; y_n, \mathbf{x}_n) | \mathbf{z}_n] = 0$$

One can estimate $\boldsymbol{\beta}_0$ with the generalized method of moments (GMM), a method that contains elements of nonlinear least squares (NLS), GLS, and IV. Specifically,

$$\hat{\boldsymbol{\beta}}_{\text{GMM}} \equiv \underset{\boldsymbol{\beta}}{\text{argmin}} E_N[\mathbf{g}(\boldsymbol{\beta})]' \mathbf{C}_N E_N[\mathbf{g}(\boldsymbol{\beta})]$$

where \mathbf{C}_N is usually a consistent estimator of $\text{Var}[\mathbf{g}(\boldsymbol{\beta}_0)]^{-1}$.

Running through this epic of generalizations of the classical linear model are several themes.

1. In every case, the estimators are *asymptotically linear*. That is,

$$\sqrt{N} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \right) \stackrel{p}{\equiv} \sqrt{N} E_N[\boldsymbol{\psi}(U_n)]$$

where $E[\boldsymbol{\psi}(U_n)] = \mathbf{0}$ and $\text{Var}[\boldsymbol{\psi}(U_n)]$ exists. Because of this property and a central limit theorem, the estimators are asymptotically normally distributed with an asymptotic variance equal to $\text{Var}[\boldsymbol{\psi}(U_n)]$.

2. The asymptotic linearity of the estimators coincides with interpreting all of the estimation procedures as minimization of generalized distance. In this way, the estimators generalize OLS and their statistical theory is analogous.

(a) GLS is the most direct generalization:

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = \underset{\boldsymbol{\beta}}{\text{argmin}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \boldsymbol{\Omega}_0^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

This is equivalent to the minimum distance problem

$$\hat{\boldsymbol{\mu}}_{\text{GLS}} = \underset{\boldsymbol{\mu} \in \text{Col}(\mathbf{X})}{\text{argmin}} (\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Omega}_0^{-1} (\mathbf{y} - \boldsymbol{\mu})$$

Alternatively, GLS is OLS after a linear transformation of the data by $\boldsymbol{\Omega}_0^{-1/2}$.

- (b) According to (23.1), the log-likelihood function is approximated by

$$N E_N[L(\boldsymbol{\theta}_N) - L(\boldsymbol{\theta}_0)] \stackrel{p}{\approx} -\frac{1}{2} (\boldsymbol{\psi}_N - \boldsymbol{\beta}_N)' \mathfrak{S}(\boldsymbol{\theta}_0) (\boldsymbol{\psi}_N - \boldsymbol{\beta}_N) + \frac{1}{2} \boldsymbol{\psi}_N' \mathfrak{S}(\boldsymbol{\theta}_0) \boldsymbol{\psi}_N \quad (23.2)$$

where

$$\begin{aligned} \boldsymbol{\beta}_N &\equiv \sqrt{N} (\boldsymbol{\theta}_N - \boldsymbol{\theta}_0) \\ \boldsymbol{\psi}_N &\equiv \mathfrak{S}(\boldsymbol{\theta}_0)^{-1} \sqrt{N} E_N L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0) \end{aligned}$$

and $\boldsymbol{\beta}_N$ is bounded. Maximizing $L(\boldsymbol{\theta}_N)$ over $\boldsymbol{\theta}_N$ is asymptotically equivalent to minimizing the leading generalized distance in $\boldsymbol{\beta}_N$.

- (c) GMM has a similar underlying approximation. Given

$$\sqrt{N} E_N[\mathbf{g}(\boldsymbol{\theta}_N)] \stackrel{p}{\approx} \sqrt{N} E_N[\mathbf{g}(\boldsymbol{\theta}_0)] + E[\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)] \sqrt{N} (\boldsymbol{\theta}_N - \boldsymbol{\theta}_0)$$

and a weighting matrix \mathbf{C}_N we obtain

$$N \cdot E_N[\mathbf{g}(\boldsymbol{\theta}_N)]' \mathbf{C}_N E_N[\mathbf{g}(\boldsymbol{\theta}_N)] \stackrel{p}{\approx} (\mathbf{y}_N - \mathbf{X}\boldsymbol{\beta}_N)' \mathbf{C}_N (\mathbf{y}_N - \mathbf{X}\boldsymbol{\beta}_N) \quad (23.3)$$

where

$$\begin{aligned} \mathbf{y}_N &\equiv \sqrt{N} E_N[\mathbf{g}(\boldsymbol{\theta}_0)] \\ \mathbf{X} &\equiv -E[\mathbf{g}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)] \end{aligned}$$

Asymptotic approximations make this general simplification possible. They also endow the vector to be fitted with a multivariate normal distribution, making the approximate, asymptotic distribution theory analogous to the exact theory for OLS and a conditionally normally distributed dependent variable.

3. Moment conditions underlie the distance measures and projection characterizes their minimization.

(a) The GLS fitted vector

$$\mathbf{y} - \hat{\boldsymbol{\mu}}_{\text{GLS}} \perp \text{Col}(\boldsymbol{\Omega}_0^{-1} \mathbf{X}) \quad \Leftrightarrow \quad \hat{\boldsymbol{\mu}}_{\text{GLS}} = \mathbf{P}_{\mathbf{X} \perp \boldsymbol{\Omega}_0^{-1} \mathbf{X}} \mathbf{y}$$

is a nonorthogonal projection onto $\text{Col}(\mathbf{X})$ that takes into account differences in variances and nonzero covariances in an optimal way for minimum variance estimation.

(b) Like GLS, the IV fitted vector

$$\mathbf{y} - \hat{\boldsymbol{\mu}}_{\text{IV}} \perp \text{Col}(\mathbf{Z}) \quad \Leftrightarrow \quad \hat{\boldsymbol{\mu}}_{\text{IV}} = \mathbf{P}_{\mathbf{X} \perp \mathbf{Z}} \mathbf{y}$$

is a nonorthogonal projection onto $\text{Col}(\mathbf{X})$. Unlike GLS, the direction of the projection may be critical to the consistency of the resultant estimator.

(c) Given identification and consistency, ML distribution theory rests on the score identity

$$E[L_{\theta}(\boldsymbol{\theta}_0)] = \mathbf{0}$$

and the variance of the score, the information matrix. Projection is trivial in the unrestricted case because the optimal $\boldsymbol{\beta}_N$ in (23.2) is actually equal to $\boldsymbol{\psi}_N$, giving (23.1). If restrictions apply to $\boldsymbol{\beta}_N$, then nonorthogonal projection is optimal as in restricted least squares (RLS).

(d) GMM is analogous to GLS, as (23.3) shows.

4. Hypothesis tests also rest on generalized distance, measuring the distance between different estimators of the parameters.
5. The approximate quadratic structure of these econometric problems is exploited in many numerical optimization methods as well.

Not all of the estimation theory is captured by a method of moments, however. There are important differences among the estimation methods that arise primarily with respect to parameter identification and estimator consistency. Identification of parameters and consistency of the MLE rests on properties of the likelihood function, not primarily the score function. In contrast, GMM identification and consistency fall upon properties of the moment functions.

At the end, we have highlighted the role of latent models in econometrics. Our models of nonnormal and nonspherical distributions are largely specifications for capturing observable phenomena, whereas the models motivating IV involve unobservable variables. There are, of course, latent models for nonnormal and nonspherical behavior as well. Such models are an essential tool in economics and econometrics. In Part IV we will describe several important examples.