The ultimate result is

$$n^{1/2}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \overset{a}{\sim} N\Big(\mathbf{0}, \sigma_0^2 \operatorname*{plim}_{n\to\infty} \big(n^{-1}\boldsymbol{X}_0^\top \boldsymbol{P}_W \boldsymbol{X}_0\big)^{-1}\Big), \tag{7.34}$$

which closely resembles (7.23) for the linear case.

The **nonlinear IV estimator** based on minimizing the criterion function (7.32) was proposed by Amemiya (1974), who very misleadingly called it the **nonlinear two-stage least squares estimator**, or **NL2SLS**. In fact, it is *not* computed in two stages at all. Attempting to compute an estimator analogous to linear 2SLS would in general result in an inconsistent estimator very different from nonlinear IV.

It is illuminating to see why this is so. We must make explicit the dependence of $\boldsymbol{x}(\boldsymbol{\beta})$ on explanatory variables. Thus the model (7.31) may be rewritten as

$$\boldsymbol{y} = \boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\beta}) + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{IID}(\mathbf{0}, \sigma^2\mathbf{I}_n),$$

where $\boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\beta})$ is a vector with typical element $x_t(\boldsymbol{Z}_t, \boldsymbol{\beta})$, $\boldsymbol{Z}$ being a matrix of observations on explanatory variables, with t^{th} row $\boldsymbol{Z}_t$, some columns of which may be correlated with $\boldsymbol{u}$. The $\boldsymbol{Z}$ matrix is not necessarily $n \times k$, because there may be more or fewer parameters than explanatory variables. A 2SLS procedure would regress those columns of $\boldsymbol{Z}$ that are potentially correlated with $\boldsymbol{u}$ on the matrix of instruments $\boldsymbol{W}$ so as to obtain $\boldsymbol{P}_W\boldsymbol{Z}$. It would then minimize the objective function

$$\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{P}_W\boldsymbol{Z}, \boldsymbol{\beta})\big)^\top\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{P}_W\boldsymbol{Z}, \boldsymbol{\beta})\big). \tag{7.35}$$

This procedure would yield consistent estimates if the regression functions $x_t(\boldsymbol{Z}_t, \boldsymbol{\beta})$ were linear in all the endogenous elements of $\boldsymbol{Z}_t$. But if the regression functions were nonlinear in any of the endogenous elements of $\boldsymbol{Z}_t$, minimizing (7.35) would not yield consistent estimates, because even though $\boldsymbol{P}_W\boldsymbol{Z}$ would be asymptotically orthogonal to $\boldsymbol{u}$, $\boldsymbol{X}(\boldsymbol{Z}, \boldsymbol{\beta})\boldsymbol{P}_W$ would not be.

As a very simple example, suppose that the regression function $x_t(\boldsymbol{Z}_t, \boldsymbol{\beta})$ were βz_t^2. Thus there would be just one independent variable, which is correlated with u_t, and one parameter. The theory for linear regressions is applicable to this example, since the regression function is linear with respect to the parameter β. What is needed to obtain a consistent estimate of β is to minimize $\|\boldsymbol{P}_W(\boldsymbol{y} - \beta\boldsymbol{z}^2)\|^2$ with respect to β, where $\boldsymbol{z}^2$ means the vector with typical element z_t^2. In contrast, if one first projected $\boldsymbol{z}$ onto $\boldsymbol{W}$ in a 2SLS procedure, one would be minimizing $\|\boldsymbol{y} - \beta(\boldsymbol{P}_W\boldsymbol{z})^2\|^2$, where $(\boldsymbol{P}_W\boldsymbol{z})^2$ means the vector with typical element $(\boldsymbol{P}_W\boldsymbol{z})_t^2$. The latter minimization is evidently not restricted to the subspace $\mathcal{S}(\boldsymbol{W})$, and so it will not in general yield consistent estimates of β.

In many cases, the biggest problem with nonlinear IV procedures is how to choose $\boldsymbol{W}$. With a linear model, it is relatively easy to do so. If the equation to be estimated comes from a system of linear simultaneous equations,