INFERENCE FOR TWO-STEP ESTIMATORS

(M2 ETE)

Koen Jochmans

Toulouse School of Economics

Last revised on March 12, 2026

©2026 KOEN JOCHMANS

Example: Two-stage least squares

Traditional motivation for 2SLS in

$$y_i = x_i\theta + u_i, \qquad \mathbb{E}(z_i u_i) = 0$$

is that

$$x_i = z_i^\top \eta + v_i = w_i + v_i, \qquad \mathbb{E}(w_i v_i) = 0,$$

by a population least-squares fit.

Thus,

$$y_i = w_i\theta + (u_i + v_i\theta), \qquad \mathbb{E}(w_i(u_i + v_i\theta)) = 0,$$

could be estimated by ordinary least squares, if w_i were observed.

Now w_i is unobserved because η is unknown, but we can use $\hat{w}_i = z_i^\top \hat{\eta}$, the least-squares fitted values in the sample, and then estimates θ by regressing y_i on $\hat{w}_i$.

This replacement is not innocuous.

Example: Sample-selection model

Interested in θ in

$$y_i = x_i^\top \theta + u_i,$$

when we only observe outcomes y_i for units where $d_i = 1$. Classical setup has

$$d_i = \begin{cases} 1 & \text{if } z_i^\top \eta \geq v_i \\ 0 & \text{if } z_i^\top \eta < v_i \end{cases},$$

and

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \Big| \, x_i, z_i \sim \mathbf{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix} \right)$$

Standard example is wages.

Here,

$$\mathbb{E}(y_i | x_i, z_i, d_i = 1) = x_i^\top \theta + \mathbb{E}(u_i | x_i, z_i, d_i = 1) = x_i^\top \theta + \mathbb{E}(u_i | z_i, v_i \leq z_i^\top \eta)$$

and the last term is non-zero.

By joint normality, $\mathbb{E}(u_i|v_i) = \rho\sigma v_i$ and

$$\mathbb{E}(u_i|v_i \leq a) = \rho\sigma \frac{\int_{-\infty}^{a} v\,\phi(v)\,dv}{\Phi(a)} = -\rho\sigma \frac{\phi(a)}{\Phi(a)} =: -\rho\sigma\lambda(a).$$

So for the observed outcomes we can write

$$y_i = x_i^\top \theta - \rho\sigma\lambda(z_i^\top \eta) + \varepsilon_i,$$

where ε_i has conditional mean zero.

The 'Heckit' two-step estimator:

Estimates η by using the probit model for $d_i|z_i$. This gives $\hat{w}_i = \lambda(z_i^\top \hat{\eta})$.

Obtains $\hat{\theta}$ from a least-squares regression of y_i on x_i and $\hat{w}_i$

Here, $\hat{w}_i$ is an estimator of $w_i = \lambda(z_i^\top \eta)$ and the difference cannot be ignored.

Generic formulation in the regular case

θ defined by moment condition

$$\mathbb{E}(m_i(\theta, \eta)) = 0;$$

this depends on η. We will consider the just-identified setting throughout.

Suppose we have an estimator $\hat{\eta}$ for η that satisfies

$$\sqrt{n}(\hat{\eta} - \eta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_i + o_p(1)$$

for $\mathbb{E}(\psi_i) = 0$ and $\mathbb{E}(\psi_i \psi_i^\top)$ finite (called influence-function representation).

Plug-in estimator of θ satisfies

$$\sum_{i=1}^{n} m_i(\hat{\theta}, \hat{\eta}) = 0.$$

Under regularity conditions a standard mean-value expansion around θ gives

$$\sum_{i=1}^{n} m_i(\hat{\theta}, \hat{\eta}) = \sum_{i=1}^{n} m_i(\theta, \hat{\eta}) + \sum_{i=1}^{n} \frac{\partial m_i(\theta^*, \hat{\eta})}{\partial \theta^\top} (\hat{\theta} - \theta) = 0,$$

from which

$$\sqrt{n}(\hat{\theta} - \theta) = \mathbb{E}\left(\frac{\partial m_i(\theta, \eta)}{\partial \theta^\top} \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} m_i(\theta, \hat{\eta}) + o_p(1)$$

then follows.

A further expansion gives

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} m_i(\theta, \hat{\eta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} m_i(\theta, \eta) + \mathbb{E}\left(\frac{\partial m_i(\theta, \eta)}{\partial \eta^\top} \right) \sqrt{n}(\hat{\eta} - \eta) + o_p(1).$$

Combining terms and using the influence function for the first-step estimator,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{p} -\mathbb{E}\left(\frac{\partial m_i(\theta, \eta)}{\partial \theta^\top} \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(m_i(\theta, \eta) + \mathbb{E}\left(\frac{\partial m_i(\theta, \eta)}{\partial \eta^\top} \right) \psi_i \right),$$

The 'correction' term here has mean zero, so only standard errors need to be adjusted.

Derivation as joint GMM estimation

Recall

$$\mathbb{E}(m_i(\theta;\eta)) = 0.$$

Suppose that, in addition to θ, η is also characterized by the moment equation

$$\mathbb{E}(g_i(\eta)) = 0.$$

We can stack moments and solve them jointly through GMM. We then solve

$$\sum_{i=1}^{n} \begin{pmatrix} m_i(\hat{\theta},\hat{\eta}) \\ g_i(\hat{\eta}) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

The usual expansion gives

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \begin{pmatrix} m_i(\theta,\eta) \\ g_i(\eta) \end{pmatrix} + \mathbb{E} \begin{pmatrix} \frac{\partial m_i(\theta,\eta)}{\partial \theta^\top} & \frac{\partial m_i(\theta,\eta)}{\partial \eta^\top} \\ 0 & \frac{\partial g_i(\eta)}{\partial \eta^\top} \end{pmatrix} \sqrt{n} \begin{pmatrix} \hat{\theta}-\theta \\ \hat{\eta}-\eta \end{pmatrix},$$

ignoring $o_p(1)$ terms.

Inverting the system then gives the joint influence-function representation

$$\sqrt{n}\begin{pmatrix} \hat{\theta}-\theta \\ \hat{\eta}-\eta \end{pmatrix} = G\,\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix} m_i(\theta,\eta) \\ g_i(\eta) \end{pmatrix} + o_p(1)$$

for Jacobian matrix

$$G = \begin{pmatrix} -\mathbb{E}\left(\frac{\partial m_i(\theta,\eta)}{\partial \theta^\top}\right)^{-1} & \mathbb{E}\left(\frac{\partial m_i(\theta,\eta)}{\partial \theta^\top}\right)^{-1}\mathbb{E}\left(\frac{\partial m_i(\theta,\eta)}{\partial \eta^\top}\right)\mathbb{E}\left(\frac{\partial g_i(\eta)}{\partial \eta^\top}\right)^{-1} \\ 0 & -\mathbb{E}\left(\frac{\partial g_i(\eta)}{\partial \eta^\top}\right)^{-1} \end{pmatrix}.$$

In particular,

$$\sqrt{n}(\hat{\theta}-\theta) = -\mathbb{E}\left(\frac{\partial m_i(\theta,\eta)}{\partial \theta^\top}\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(m_i(\theta,\eta)+\mathbb{E}\left(\frac{\partial m_i(\theta,\eta)}{\partial \eta^\top}\right)\psi_i\right)+o_p(1)$$

for

$$\psi_i = -\mathbb{E}\left(\frac{\partial g_i(\eta)}{\partial \eta^\top}\right)^{-1} g_i(\eta),$$

which is as before.

Example: 2SLS

Here (taking scalar x_i and multiple instruments z_i),

$$m_i(\theta, \eta) = (z_i^\top \eta)(y_i - (z_i^\top \eta)\theta), \qquad g_i(\eta) = z_i(x_i - z_i^\top \eta).$$

We have

$$\mathbb{E}\left(\frac{\partial m_i(\theta, \eta)}{\partial \theta}\right) = -\mathbb{E}(w_i^2),$$

and

$$\mathbb{E}\left(\frac{\partial m_i(\theta, \eta)}{\partial \eta}\right) = \mathbb{E}\left(z_i(u_i + v_i\theta)\right) - \mathbb{E}\left(z_i w_i\right)\theta = -\mathbb{E}\left(z_i w_i\right)\theta.$$

Also, $\psi_i = \mathbb{E}(z_i z_i^\top)^{-1} z_i v_i$, and so

$$\mathbb{E}\left(\frac{\partial m_i(\theta, \eta)}{\partial \eta^\top}\right)\psi_i = -\theta\,\mathbb{E}(w_i z_i^\top)\,\mathbb{E}(z_i z_i^\top)^{-1} z_i v_i = -w_i v_i\,\theta$$

Therefore, the influence function for 2SLS is

$$\mathbb{E}(w_i^2)^{-1}\left(w_i(u_i + v_i\theta) - w_i v_i\,\theta\right) = \mathbb{E}(w_i^2)^{-1} w_i u_i.$$

The naive approach would omit the correction and use $\mathbb{E}(w_i^2)^{-1} w_i(u_i + v_i\theta)$ instead.

The conventional GMM derivation for 2SLS here starts from the first-order condition to the quadratic problem:

$$\left(\frac{1}{n}\sum_{i=1}^{n} x_i z_i^\top\right)\left(\frac{1}{n}\sum_{i=1}^{n} z_i z_i^\top\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i(y_i - x_i\hat{\theta}) = 0.$$

This behaves likes

$$\mathbb{E}\left(x_i z_i^\top\right)\mathbb{E}(z_i z_i^\top)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i u_i - \mathbb{E}(z_i x_i)\sqrt{n}(\hat{\theta}-\theta)\right) = o_p(1).$$

Thus, using that $\mathbb{E}(z_i z_i^\top)^{-1}\mathbb{E}(z_i x_i) = \eta$ and that $\mathbb{E}(w_i x_i) = \mathbb{E}(w_i^2)$, this equals

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_i u_i - \mathbb{E}(w_i^2)\sqrt{n}(\hat{\theta}-\theta) = o_p(1),$$

so that re-arranging gives

$$\sqrt{n}(\hat{\theta}-\theta) = \mathbb{E}(w_i^2)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_i u_i + o_p(1),$$

as before.

Orthogonality example

To motivate the sequel, consider a simple example.

Say $x_i \sim$ i.i.d. (η, θ).

Here,

$$m_i(\theta, \eta) = (x_i - \eta)^2 - \theta, \qquad g_i(\eta) = x_i - \eta.$$

This gives

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\eta})^2, \qquad \hat{\eta} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

the usual estimator of the variance.

Note that, here,

$$\mathbb{E}\left(\frac{\partial m_i(\theta, \eta)}{\partial \eta}\right) = -2\,\mathbb{E}\,(x_i - \eta) = 0,$$

so estimation of the mean does not have a (first-order) effect on the limit behavior of the variance estimator.

Orthogonality

The zero expected Jacobian condition allows to weaken the conditions on $\hat{\eta}$.

If

$$\mathbb{E}\left(\frac{\partial m_i(\theta,\eta)}{\partial \eta^\top}\right) = 0$$

we have the expansion (for univariate m_i for simplicity)

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m_i(\theta,\hat{\eta}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} m_i(\theta,\eta) + n^{1/4}(\hat{\eta}-\eta)^\top \mathbb{E}\left(\frac{\partial^2 m_i(\theta,\eta)}{\partial\eta\partial\eta^\top}\right) n^{1/4}(\hat{\eta}-\eta) + o_p(\|\hat{\eta}-\eta\|^2),$$

which yields asymptotic equivalence with the infeasible estimator as long as

$$\hat{\eta}-\eta = o_p(n^{-1/4}).$$

This therefore allows the use of much more 'complicated' first-step estimators.

Orthogonal moments

A given moment function

$$\mathbb{E}(m_i(\theta, \eta)) = 0$$

does not, in general, satisfy orthogonality.

Consider the linear adjustment

$$m_i^*(\theta, \eta) := m_i(\theta, \eta) - \alpha\, g_i(\eta).$$

Note that $\mathbb{E}(m_i^*(\theta, \eta)) = 0$ because $\mathbb{E}(m_i(\theta, \eta)) = 0$ and $\mathbb{E}(g_i(\eta)) = 0$.

Now

$$\frac{\partial m_i^*(\theta, \eta)}{\partial \eta^\top} = \frac{\partial m_i(\theta, \eta)}{\partial \eta^\top} - \alpha \frac{\partial g_i(\eta)}{\partial \eta^\top}$$

and so

$$\mathbb{E}\left(\frac{\partial m_i^*(\theta, \eta)}{\partial \eta^\top}\right) = 0$$

holds for

$$\alpha = \mathbb{E}\left(\frac{\partial m_i(\theta, \eta)}{\partial \eta}\right) \mathbb{E}\left(\frac{\partial g_i(\theta, \eta)}{\partial \eta^\top}\right)^{-1}.$$

Example: 2SLS

Here,

$$m_i(\theta, \eta) = (z_i^\top \eta)(y_i - (z_i^\top \eta)\,\theta), \qquad g_i(\eta) = z_i(x_i - z_i^\top \eta),$$

and, from before,

$$\mathbb{E}\left(\frac{\partial m_i(\theta, \eta)}{\partial \eta}\right) = -\mathbb{E}(z_i w_i)\,\theta,$$

which is not zero.

Here,

$$\alpha = \mathbb{E}\left(\frac{\partial m_i(\theta, \eta)}{\partial \eta^\top}\right)\mathbb{E}\left(\frac{\partial g_i(\theta, \eta)}{\partial \eta^\top}\right)^{-1} = \theta\,\mathbb{E}(w_i z_i')\,\mathbb{E}(z_i z_i')^{-1} = \theta\,\eta'$$

and so

$$m_i^*(\theta, \eta) = (z_i^\top \eta)(y_i - (z_i^\top \eta)\,\theta) - \theta(z_i^\top \eta)(x_i - z_i^\top \eta) = (z_i^\top \eta)(y_i - x_i\theta),$$

which corresponds to the conventional GMM formulation of 2SLS based on $\mathbb{E}(z_i u_i) = 0$ with weight matrix $\mathbb{E}(z_i z_i^\top)^{-1}$, which we know yields correct inference.

Concentrated score equation

Consider the joint maximization problem

$$\max_{\theta,\eta} \mathbb{E}(\ell_i(\theta,\eta)),$$

such as a maximum-likelihood problem, for example. Take θ and η scalar.

For fixed θ, let $\eta(\theta)$ be defined through

$$\mathbb{E}\left(\frac{\partial \ell_i(\theta,\eta(\theta))}{\partial \eta}\right) = 0.$$

Then θ solves the concentrated score problem

$$\mathbb{E}\left(\frac{\partial \ell_i(\theta,\eta(\theta))}{\partial \theta}\right) = \mathbb{E}\left(\frac{\partial \ell_i(\theta,\eta)}{\partial \theta}\right) + \mathbb{E}\left(\frac{\partial \ell_i(\theta,\eta)}{\partial \eta}\right)\frac{\partial \eta(\theta)}{\partial \theta} = 0.$$

From differentiation the first-order condition for $\eta(\theta)$ we find

$$\frac{\partial \eta(\theta)}{\partial \theta} = -\mathbb{E}\left(\frac{\partial^2 \ell_i(\theta,\eta)}{\partial \eta^2}\right)^{-1}\mathbb{E}\left(\frac{\partial^2 \ell_i(\theta,\eta)}{\partial \theta \partial \eta}\right)$$

This is the efficient score,

$$m_i^*(\theta,\eta) = \frac{\partial \ell_i(\theta,\eta)}{\partial \theta} - \mathbb{E}\left(\frac{\partial^2 \ell_i(\theta,\eta)}{\partial\theta\partial\eta}\right)\mathbb{E}\left(\frac{\partial^2 \ell_i(\theta,\eta)}{\partial\eta^2}\right)^{-1}\frac{\partial \ell_i(\theta,\eta)}{\partial \eta}.$$

Note that, if ℓ_i is a log-likelihood, then

$$\mathbb{E}\left(\frac{\partial^2 \ell_i(\theta,\eta)}{\partial\eta^2}\right) = -\mathbb{E}\left(\frac{\partial \ell_i(\theta,\eta)}{\partial\eta}\,\frac{\partial \ell_i(\theta,\eta)}{\partial\eta}\right)$$

and

$$\mathbb{E}\left(\frac{\partial^2 \ell_i(\theta,\eta)}{\partial\theta\partial\eta}\right) = -\mathbb{E}\left(\frac{\partial \ell_i(\theta,\eta)}{\partial\theta}\,\frac{\partial \ell_i(\theta,\eta)}{\partial\eta}\right)$$

by the information equality.

The orthogonal score then is the (population) residual from a least-squares projection of the score for θ onto the score for η.

Example: Classical linear regression

$$y_i = x_i\theta + w_i^\top \eta + \varepsilon_i, \qquad \varepsilon_i | x_i, w_i \sim \mathbf{N}(0, \sigma^2).$$

The scores at the truth are

$$\frac{x_i \varepsilon_i}{\sigma^2}, \qquad \frac{z_i \varepsilon_i}{\sigma^2},$$

and so

$$\alpha = \mathbb{E}(\varepsilon_i^2\, x_i w_i^\top)\, \mathbb{E}(\varepsilon_i^2\, w_i w_i^\top)^{-1} = \mathbb{E}(x_i w_i^\top)\, \mathbb{E}(w_i w_i^\top)^{-1}.$$

Then

$$m_i(\theta, \eta) = x_i(y_i - x_i\theta - w_i^\top \eta)$$

but

$$m_i^*(\theta, \eta) = (x_i - z_i^\top \alpha)(y_i - x_i\theta - w_i^\top \eta).$$

This is just the usual partialling-out to multiple regression.

High-dimensional regression

Write $v_i = x_i - w_i^\top \alpha$.

The orthogonalized estimating equation for θ can be expanded as

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(x_i - w_i^\top \hat{\alpha})(y_i - x_i\theta - w_i^\top \hat{\eta}) = \frac{1}{n}\sum_{i=1}^{n} v_i \varepsilon_i \\ - \frac{1}{n}\sum_{i=1}^{n} v_i\, w_i^\top (\hat{\eta} - \eta) \\ - \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i\, w_i^\top (\hat{\alpha} - \alpha) \\ + (\hat{\alpha} - \alpha)^\top \frac{1}{n}\sum_{i=1}^{n} w_i w_i^\top (\hat{\eta} - \eta)\end{aligned}$$

by virtue of orthogonality with respect to both η and α.

If $\dim w_i$ grows with n the second and third right-hand side term can be delicate to deal with.

To see this suppose that

$$\hat{\eta} = \eta + \frac{1}{n}\sum_{i=1}^{n} \psi_i$$

for some mean-zero ψ_i.

Then

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} v_i\, w_i^\top (\hat{\eta} - \eta)\right) = \mathbb{E}\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} v_i\, w_i^\top \psi_j\right) = \frac{\mathbb{E}(v_i w_i^\top \psi_i)}{n},$$

which is generally

$$O\left(\frac{\dim w_i}{n}\right).$$

So, the bias will only be negligible asymptotically if $\dim w_i/\sqrt{n} \to 0$.

An example is the panel data problem, where $\dim w_i \propto N$ and $n = N \times T$, so that the bias becomes $O(T^{-1})$ and we require that $N = o(T)$ for inference to be correct.

Sampling splitting

Split the data into (say) $2 \times n/2$ observations (with n taken to be even here).

Use the second half of data to compute $\hat{\eta}$ and $\hat{\alpha}$ and the first half to evaluate the moment equation.

The subsamples are independent.

Then

$$\frac{1}{n/2} \sum_{i=1}^{n/2} v_i w_i^\top \left(\hat{\eta}_2 - \eta\right)$$

has mean zero and conditional variance

$$\left(\hat{\eta} - \eta\right)^\top \frac{2}{n} \mathbb{E}(w_i v_i v_i w_i^\top) \left(\hat{\eta}_2 - \eta\right) = O_p\left(\frac{\|\hat{\eta}_2 - \eta\|^2}{n}\right)$$

and so is asymptotically negligible as long as $\|\hat{\eta}_2 - \eta\| = o_p(1)$.

Cross-fitting

Sample splitting is inefficient because we evaluate the moment condition in only $n/2$ observations.

So variance is twice as large.

Flip the roles of the splits, and average the resulting moment equation:

$$\frac{1}{n}\sum_{i=1}^{n/2}(x_i - w_i^\top \hat{\alpha}_2)(y_i - x_i\theta - w_i^\top \hat{\eta}_2) + \frac{1}{n}\sum_{i=n/2+1}^{n}(x_i - w_i^\top \hat{\alpha}_1)(y_i - x_i\theta - w_i^\top \hat{\eta}_1)$$

and solve this for θ to get a 'cross-fitted' estimator.

Nothing specific about using half splits.

Seemingly efficient would be to use a leave-one-out strategy:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - w_i^\top \hat{\alpha}_{-i})(y_i - x_i\theta - w_i^\top \hat{\eta}_{-i}),$$

where $\hat{\alpha}_{-i}$ and $\hat{\eta}_{-i}$ are constructed using all but observation i.

Higher-order orthogonality

Return to original problem.

Take η to be scalar for notational simplicity.

Consider a higher-order expansion of the moment condition

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m_i(\theta,\hat{\eta}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} m_i(\theta,\eta) \\ + \quad \mathbb{E}\left(\frac{\partial\ m_i(\theta,\eta)}{\partial\eta}\right)\sqrt{n}(\hat{\eta}-\eta) \\ + \frac{1}{2}\mathbb{E}\left(\frac{\partial^2 m_i(\theta,\eta)}{\partial\eta^2}\right)\sqrt{n}(\hat{\eta}-\eta)^2 \\ \cdots \\ + \frac{1}{q!}\mathbb{E}\left(\frac{\partial^q m_i(\theta,\eta)}{\partial\eta^q}\right)\sqrt{n}(\hat{\eta}-\eta)^q + O_p(\sqrt{n}|\hat{\eta}-\eta|^{q+1})\end{aligned}$$

where we can rely on sample splitting to handle the deviations of sample means from population means.

Orthogonality to order q means that

$$\mathbb{E}\left(\frac{\partial^p m_i(\theta,\eta)}{\partial \eta^p}\right) = 0, \qquad 1 \leq p \leq q.$$

Then it suffices that $\sqrt{n}|\hat{\eta} - \eta|^{q+1} = o_p(1)$, which translates to

$$\hat{\eta} - \eta = o_p(n^{-1/2(q+1)}).$$

We thus obtain increased robustness to the presence of estimation noise in the nuisance parameter.

Parametric problems

Consider setting where distribution of data is known up to (θ, η).

From before we know that first-order orthogonality is achieved by the efficient score

$$\frac{\partial \ell_i(\theta, \eta)}{\partial \theta} - \alpha \frac{\partial \ell_i(\theta, \eta)}{\partial \eta}.$$

for

$$\alpha = \mathbb{E}\left(\frac{\partial \ell_i(\theta, \eta)}{\partial \theta} \frac{\partial \ell_i(\theta, \eta)}{\partial \eta^\top}\right) \mathbb{E}\left(\frac{\partial \ell_i(\theta, \eta)}{\partial \eta} \frac{\partial \ell_i(\theta, \eta)}{\partial \eta^\top}\right)^{-1}$$

We can exploit the parametric context to write

$$\alpha(\theta, \eta) = \mathbb{E}_{\theta,\eta}\left(\frac{\partial \ell_i(\theta, \eta)}{\partial \theta} \frac{\partial \ell_i(\theta, \eta)}{\partial \eta^\top}\right) \mathbb{E}_{\theta,\eta}\left(\frac{\partial \ell_i(\theta, \eta)}{\partial \eta} \frac{\partial \ell_i(\theta, \eta)}{\partial \eta^\top}\right)^{-1}$$

which is known up to the model parameters, to use

$$\frac{\partial \ell_i(\theta, \eta)}{\partial \theta} - \alpha(\theta, \eta) \frac{\partial \ell_i(\theta, \eta)}{\partial \eta}.$$

This construction does not require the introduction of additional nuisance parameters.

Second-order orthogonality

We projected on the score for the nuisance, $\partial \ell_i(\theta, \eta)/\partial \eta$, because we know that

$$\mathbb{E}_{\theta,\eta}\left(\sum_{i=1}^{n} \frac{\partial \ell_i(\theta,\eta)}{\partial \eta}\right) = 0.$$

We also know that, by taking derivatives here, we get

$$\int \sum_{i=1}^{n} \frac{\partial^2 \ell_i(\theta,\eta)}{\partial \eta^2} \prod_{i=1}^{n} e^{\ell_i(\theta,\eta)} + \sum_{i=1}^{n} \frac{\partial \ell_i(\theta,\eta)}{\partial \eta} \frac{\partial \prod_{i=1}^{n} e^{\ell_i(\theta,\eta)}}{\partial \eta} \, d\,\text{data} = 0$$

which we can re-arrange to get

$$\mathbb{E}_{\theta,\eta}\left(\sum_{i=1}^{n} \frac{\partial^2 \ell_i(\theta,\eta)}{\partial \eta^2} + \left(\sum_{i=1}^{n} \frac{\partial \ell_i(\theta,\eta)}{\partial \eta}\right)^2\right) = 0.$$

This is the usual information inequality.

Use shorthand

$$s(\theta, \eta) = \sum_{i=1}^{n} \frac{\ell_i(\theta, \eta)}{\partial \theta}.$$

Can then write

$$f_1(\theta, \eta) = \sum_{i=1}^{n} \frac{\partial \ell_i(\theta, \eta)}{\partial \eta} = \frac{\partial \prod_{i=1}^{n} e^{\ell_i(\theta,\eta)} / \partial \eta}{\prod_{i=1}^{n} e^{\ell_i(\theta,\eta)}}$$

$$f_2(\theta, \eta) = \sum_{i=1}^{n} \frac{\partial^2 \ell_i(\theta, \eta)}{\partial \eta^2} + \left(\sum_{i=1}^{n} \frac{\partial \ell_i(\theta, \eta)}{\partial \eta} \right)^2 = \frac{\partial^2 \prod_{i=1}^{n} e^{\ell_i(\theta,\eta)} / \partial \eta^2}{\prod_{i=1}^{n} e^{\ell_i(\theta,\eta)}}$$

and consider whether we can find coefficients $\alpha_1(\theta, \eta)$ and $\alpha_2(\theta, \eta)$ so that

$$s(\theta, \eta) - \alpha_1(\theta, \eta)\, f_1(\theta, \eta) - \alpha_2(\theta, \eta)\, f_2(\theta, \eta)$$

is orthogonal to order two.

First-order orthogonality requires that

$$\begin{aligned}\frac{\partial s(\theta,\eta)}{\partial \eta} &- \alpha_1(\theta,\eta)\frac{\partial f_1(\theta,\eta)}{\partial \eta} - \frac{\partial \alpha_1(\theta,\eta)}{\partial \eta} f_1(\theta,\eta) \\ &- \alpha_2(\theta,\eta)\frac{\partial f_2(\theta,\eta)}{\partial \eta} - \frac{\partial \alpha_2(\theta,\eta)}{\partial \eta} f_2(\theta,\eta)\end{aligned}$$

has mean zero at true values.

Take expectations and note how derivatives of α_1 and α_2 vanish to arrive that

$$\mathbb{E}_{\theta,\eta}\left(\frac{\partial s(\theta,\eta)}{\partial \eta}\right) - \alpha_1(\theta,\eta)\,\mathbb{E}_{\theta,\eta}\left(\frac{\partial f_1(\theta,\eta)}{\partial \eta}\right) - \alpha_2(\theta,\eta)\,\mathbb{E}_{\theta,\eta}\left(\frac{\partial f_2(\theta,\eta)}{\partial \eta}\right) = 0.$$

Solving for α_1 gives

$$\alpha_1(\theta,\eta) = \left\{\mathbb{E}_{\theta,\eta}\left(\frac{\partial s(\theta,\eta)}{\partial \eta}\right) - \alpha_2(\theta,\eta)\,\mathbb{E}_{\theta,\eta}\left(\frac{\partial f_2(\theta,\eta)}{\partial \eta}\right)\right\}\mathbb{E}_{\theta,\eta}\left(\frac{\partial f_1(\theta,\eta)}{\partial \eta}\right)^{-1}.$$

Note that we can re-write this expression as

$$\begin{aligned}\alpha_1(\theta,\eta) = & \quad \mathbb{E}_{\theta,\eta}\left(\frac{\partial s\ (\theta,\eta)}{\partial \eta}\right)\mathbb{E}_{\theta,\eta}\left(\frac{\partial f_1(\theta,\eta)}{\partial \eta}\right)^{-1} \\ & - \alpha_2(\theta,\eta)\,\mathbb{E}_{\theta,\eta}\left(\frac{\partial f_2(\theta,\eta)}{\partial \eta}\right)\mathbb{E}_{\theta,\eta}\left(\frac{\partial f_1(\theta,\eta)}{\partial \eta}\right)^{-1}\end{aligned}$$

The ratios on the right-hand side are projection coefficients of s on f_1 and of f_2 on f_1, respectively.

Plugging back into our original problem gives

$$s^*(\theta,\eta) - \alpha_2(\theta,\eta)\, f_2^*(\theta,\eta),$$

where

$$s^*(\theta,\eta) = s(\theta,\eta) - \mathbb{E}_{\theta,\eta}\left(\frac{\partial s(\theta,\eta)}{\partial \eta}\right)\mathbb{E}_{\theta,\eta}\left(\frac{\partial f_1(\theta,\eta)}{\partial \eta}\right)^{-1} f_1(\theta,\eta)$$

and

$$f_2^*(\theta,\eta) = f_2(\theta,\eta) - \mathbb{E}_{\theta,\eta}\left(\frac{\partial f_2(\theta,\eta)}{\partial \eta}\right)\mathbb{E}_{\theta,\eta}\left(\frac{\partial f_1(\theta,\eta)}{\partial \eta}\right)^{-1} f_1(\theta,\eta)$$

Taking first derivatives gives

$$\frac{\partial s^*(\theta,\eta)}{\partial \eta} - \alpha_2(\theta,\eta)\,\frac{\partial f_2^*(\theta,\eta)}{\partial \eta} - \frac{\partial \alpha_2(\theta,\eta)}{\partial \eta}\,f_2^*(\theta,\eta)$$

which is mean zero because both s^* and f_2^* are orthogonal to order one.

Taking second derivatives gives

$$\frac{\partial^2 s^*(\theta,\eta)}{\partial \eta^2} - \alpha_2(\theta,\eta)\,\frac{\partial^2 f_2^*(\theta,\eta)}{\partial \eta^2} - 2\frac{\partial \alpha_2(\theta,\eta)}{\partial \eta}\,\frac{\partial f_2^*(\theta,\eta)}{\partial \eta} - \frac{\partial^2 \alpha_2(\theta,\eta)}{\partial \eta^2}\,f_2^*(\theta,\eta)$$

Taking expectations and equalizing to zero gives

$$\mathbb{E}_{\theta,\eta}\left(\frac{\partial^2 s^*(\theta,\eta)}{\partial \eta^2}\right) - \alpha_2(\theta,\eta)\,\mathbb{E}_{\theta,\eta}\left(\frac{\partial^2 f_2^*(\theta,\eta)}{\partial \eta^2}\right) = 0$$

from which we find

$$\alpha_2(\theta,\eta) = \mathbb{E}_{\theta,\eta}\left(\frac{\partial^2 s^*(\theta,\eta)}{\partial \eta^2}\right)\mathbb{E}_{\theta,\eta}\left(\frac{\partial^2 f_2^*(\theta,\eta)}{\partial \eta^2}\right)^{-1}.$$

We can again use Bartlett identifies to clean up the expression.

This gives

$$s(\theta,\eta) - \begin{pmatrix} \alpha_1(\theta,\eta) & \alpha_2(\theta,\eta) \end{pmatrix} \begin{pmatrix} f_1(\theta,\eta) \\ f_2(\theta,\eta) \end{pmatrix},$$

for

$$\begin{pmatrix} \alpha_1(\theta,\eta) \\ \alpha_2(\theta,\eta) \end{pmatrix} = \mathbb{E} \begin{pmatrix} f_1(\theta,\eta)f_1(\theta,\eta) & f_1(\theta,\eta)f_2(\theta,\eta) \\ f_2(\theta,\eta)f_1(\theta,\eta) & f_2(\theta,\eta)f_2(\theta,\eta) \end{pmatrix}^{-1} \mathbb{E} \begin{pmatrix} f_1(\theta,\eta)s(\theta,\eta) \\ f_2(\theta,\eta)s(\theta,\eta) \end{pmatrix},$$

which is again a regression residual.

Orthogonality to order q

Collect the leading q scaled likelihood derivatives $f_1, f_2, \ldots, f_q$ (for given values θ, η) in f.

Use the regression residual

$$s^* = s - \alpha f = 0$$

as empirical moment equation.

Note that

$$\frac{\partial \mathbb{E}_{\theta,\eta}(s^*(\theta,\eta))}{\partial \theta} = \mathbb{E}_{\theta,\eta}\left(\frac{\partial s^*(\theta,\eta)}{\partial \theta}\right) - \mathbb{E}_{\theta,\eta}(s^*(\theta,\eta)s(\theta,\eta)) = 0.$$

Then, using that $\mathbb{E}_{\theta,\eta}(s^*(\theta,\eta)s(\theta,\eta)) = \mathbb{E}_{\theta,\eta}(s^*(\theta,\eta)s^*(\theta,\eta))$ we obtain a generalized information equality.