

September 2026

Orthogonal Moments in Likelihood Models

Stéphane Bonhomme, Koen Jochmans and Martin Weidner

Orthogonal Moments in Likelihood Models^{*}

Stéphane Bonhomme[†]

Department of Economics, University of Chicago

Koen Jochmans[‡]

Toulouse School of Economics, Université Toulouse Capitole

Martin Weidner[§]

Department of Economics, University College London

This version: September 25, 2026

Abstract

Many models, such as fixed-effect models for panel or network data, are hard to estimate because they feature nuisance parameters that are both numerous and estimated imprecisely. This, in general, causes an incidental-parameter problem in the estimator of the parameters of interest. The problem can be alleviated by working with an estimating equation whose expectation is insensitive to the value of the nuisance parameters. We discuss and contrast three notions of insensitivity, also called *orthogonality*, in the context of likelihood models: Neyman orthogonality, Neyman orthogonality to order q , and full orthogonality. Orthogonal moments are obtained by projecting the estimating equation on nested subspaces, which are spanned by, respectively, the scores of the nuisance parameters, the first q derivatives of the likelihood ratio with respect to the nuisance parameters, and all likelihood ratios of the model. We give explicit constructions in binary-choice, count-data, and nonlinear regression models.

JEL codes: C13, C23, C55

Keywords: bias correction, incidental parameters, network data, orthogonality, panel data

^{*}Manuscript prepared for the Journal of Applied Econometrics lecture delivered at the annual conference of the International Association for Applied Econometrics in Lisbon, Portugal, in June 2026. We thank our discussants Iván Fernández-Val and Chris Muris for insightful comments. Funded by the European Union (ERC-NETWORK-101044319) and by the French Government and the French National Research Agency under the Investissements d’Avenir program (ANR-17-EURE-0010). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

[†]sbonhomme@uchicago.edu

[‡]koen.jochmans@tse-fr.eu

[§]m.weidner@ucl.ac.uk

1 Introduction

In many estimation problems the parameters can be divided into two groups. The first group contains the target parameter; it answers the question of interest. The second group consists of nuisance parameters; they are not of primary interest but have to be estimated along the way. In program evaluation where the target is an average treatment effect the nuisance parameters may describe a propensity score. In structural settings they may be the parameters of the underlying model needed to generate policy counterfactuals. In the panel and network settings that we will focus on here they are fixed effects, one or more of them for each individual, firm, or network node.

Identification of the target parameter is typically based on a moment condition in which the nuisance parameters appear. A feasible estimating equation then requires these to be replaced by an estimator. When nuisance parameters are numerous or imprecisely estimated, as in panel and network data models with fixed effects, the estimating equation is generally biased and inference may be substantially affected.

A popular response is to replace or modify the moment function to render it less sensitive to the nuisance parameters, in a well-defined sense. In this paper we review and connect three such possibilities. The first one is *Neyman orthogonality*, introduced in [Neyman \(1959\)](#). It requires the first derivative of the expected moment with respect to the nuisance parameters to be zero at the truth. This zero expected-Jacobian condition is at the heart of the modern double/debiased machine-learning (DDML) literature as in [Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins \(2018\)](#). It introduces enough insensitivity to permit the target parameter to be estimated at the parametric rate while allowing for (sample-split) first-step estimators that converge at a rate faster than $n^{-1/4}$, where n is the sample size (see, e.g., [Newey, 1994](#)).

In this paper we focus on models that have a likelihood structure, where the density of the data depends on a fixed-dimensional parameter vector, a vector of (many) nuisance parameters, and possibly a set of exogenous covariates. In this setting a Neyman-orthogonal moment can be constructed by projecting the original moment function onto the orthogonal complement of the score of the nuisance parameters, as [Neyman \(1959\)](#) originally proposed. The DDML approach to estimation then consists of generalized method of moments (GMM) estimation based on such a

Neyman-orthogonal moment, with a sample-split estimate of the nuisance parameter plugged in.

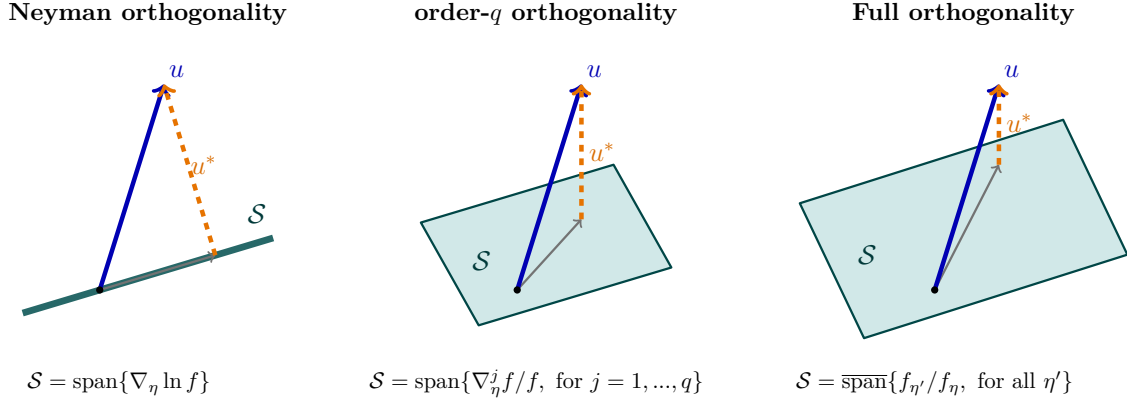
However, the faster-than- $n^{-1/4}$ convergence may be too strong a requirement in practice, when the nuisance parameters are too many or too hard to estimate. As discussed in [Bonhomme, Jochmans, and Weidner \(2024\)](#), standard panel data settings with N units, T periods, and $n = NT$ observations are a case in point. Indeed, unit fixed effects are estimated at the $T^{-1/2}$ rate, which is not asymptotically negligible relative to $n^{-1/4} = (NT)^{-1/4}$ unless N is negligible relative to T ; that is, unless the panel is so long that there is no incidental parameter problem to start with.

In panel data, it is sometimes possible to difference out the fixed effects entirely, as happens in linear panel data models (e.g., [Arellano, 2003b](#), [Chamberlain, 1992](#)) and certain classes of nonlinear models where quasi-differencing approaches are available (e.g., [Arellano and Honoré, 2001](#)). This gives rise to moment functions whose expectation is equal to zero irrespective of the value of the nuisance parameters, and to estimators that are consistent for fixed T as N tends to infinity without the need for first-step estimates of nuisance parameters or sample splitting. In panel data, informative such moments do not necessarily exist ([Chamberlain, 2010](#)). A general approach to find them is functional differencing, as introduced in [Bonhomme \(2012\)](#).

Motivated by these differencing approaches in panel data settings, the second notion of insensitivity that we consider here is *full orthogonality*, in the sense that the expected moment is zero at every value of the nuisance parameters. We show that the projection approach of functional differencing can be applied to general nuisance parameters, not only fixed effects in panel data. This broadens the observation in [Bonhomme and Dano \(2024\)](#), who consider functional differencing in network settings. In contrast to Neyman orthogonality, full orthogonality is a global property. We show that fully orthogonal moments can be obtained by projecting the original moment function, similarly to Neyman’s original proposal, with the difference that the projection is orthogonal to the closed span of likelihood ratios under the model, hence infinite-dimensional in general.

The third notion, *Neyman orthogonality to order q* ([Mackey, Syrgkanis, and Zadik, 2018](#)), lies between the other two. Here, all derivatives to order q of the moment vanish at true parameter values. This notion is a local one, and it coincides with usual Neyman orthogonality when $q = 1$. While estimation using order- q orthogonal moments requires first-step estimates of nuisance parameters and sample splitting,

Figure 1: Achieving orthogonality by projection



Notes: u denotes the original moment, u^* the orthogonal moment, $\mathcal{S}$ is the space that defines the projection: scores with respect to the nuisance parameter η for Neyman orthogonality, derivatives of likelihood ratios (or Bhattacharyya basis) for order- q orthogonality, and likelihood ratios for full orthogonality. All these concepts will be defined in detail in the text.

the rate requirement is less restrictive compared to Neyman orthogonality: a rate faster than $n^{-1/2(q+1)}$ is sufficient to ensure asymptotic unbiasedness. For example, in panel data settings where unit fixed effects are estimated at the $T^{-1/2}$ rate, the condition requires N to be negligible relative to T^q . This is less restrictive the larger q is, and it corrects for first-order incidental parameter bias as soon as $q \geq 2$. This property is similar to large- N, T bias reduction approaches in the panel literature (see, e.g., [Hahn and Newey, 2004](#) and [Dhaene and Jochmans, 2015a,b](#)).

An explicit construction of such order- q orthogonal moment functions is given in [Bonhomme, Jochmans, and Weidner \(2024\)](#). It consists in projecting the original moment function orthogonally to the span of the first q elements of the [Bhattacharyya \(1946\)](#) basis (which are generalized score functions). This construction generalizes the projected-score approach ([Small and McLeish, 1989](#), [Waterman and Lindsay, 1996](#)) to general moments and target parameters, and it can be interpreted as constructing a higher-order influence function ([Robins, Li, Tchetgen Tchetgen, and van der Vaart, 2008](#)) in models with a likelihood structure.

In our likelihood setting, each of the three notions of orthogonality arises from the same operation. In each case, we project the original moment function u on a

subspace $\mathcal{S}$ spanned by functions of the data and keep the residual u^* , which is the orthogonal moment. We provide a graphical illustration in Figure 1. For Neyman orthogonality, the subspace $\mathcal{S}$ is the span of the nuisance score. For orthogonality to order q it is the span of the first q generalized score functions. For full orthogonality it is the closed span of all likelihood ratios of the model. The three subspaces are nested, and so the trade-off between each of the three forms of orthogonality is easy to see: the larger the subspace, the more of the original moment is removed. This increases robustness to estimation noise in the nuisance parameters but, at the same time, decreases the information about the target parameter that remains present. This geometric principle allows us to connect seemingly-different procedures that have appeared in the literature.

While our aim is largely expository, several of the connections we draw appear to be new. We characterize full orthogonality as orthogonality to the closed span of the model’s likelihood ratios, which yields a projection-based construction of fully-orthogonal moments in general likelihood models that generalizes the scope of the functional differencing approach. We show that this construction coincides with Neyman’s projection device, applied to a semiparametric mixture model in which the distribution of the nuisance parameter is itself the nuisance parameter. We also show that, under an analyticity condition, a moment that is orthogonal to every order q is fully orthogonal, so that the three notions can be viewed as a single increasing sequence of projections. Finally, we show that the orthogonality notions and projections can be naturally adapted to handle more general target parameters such as average effects, for example.

Throughout the presentation we use several examples: a binary-choice logit model, a panel data Poisson model, and a nonlinear regression model with Gaussian errors. The logit case is sufficiently general to cover the classic fixed-effect panel-data model going back to [Rasch \(1961\)](#), the β -model of network formation ([Chatterjee, Diaconis, and Sly, 2011](#), [Graham, 2017](#)), as well as generic high-dimensional logistic regression ([Sur and Candès, 2019](#)). An interesting feature of this model is that all projections can be written down in closed form. For example, Neyman orthogonalization amounts to variance-weighted demeaning, whereas higher-order orthogonalization consists in projecting on powers of the sufficient statistic. For the panel logit in particular, the construction naturally comes to an end when q is equal to the length of the panel, in which case it delivers the conditional-logit score of [Rasch \(1961\)](#), [Andersen](#)

(1970), and Chamberlain (1980). The panel Poisson example shows a case where Neyman orthogonality automatically achieves full orthogonality. Lastly, the nonlinear regression example differs from the other two in that it belongs to a *curved* exponential family (Efron, 1975), and does not possess a low-dimensional sufficient statistic. Yet, higher-order projections remain tractable in this case.

2 Three notions of orthogonality

2.1 Moment conditions and plug-in bias

We will work in a setting where we are interested in a target parameter $\theta_0 \in \Theta$, known to be identified through a vector of d_u moment conditions

$$\mathbb{E}[u(y, x, \theta_0, \eta_0)] = 0. \quad (2.1)$$

Here the vectors y and x are data and represent outcomes and regressors, respectively, $\eta_0 \in \mathcal{H} \subseteq \mathbb{R}^{d_\eta}$ is a nuisance parameter, and the function u is known. Estimation problems of this kind are common. For example, a propensity score or a conditional mean of potential outcomes plays the role of η_0 in program evaluation, where an average treatment effect may be the target θ_0 . Difference-in-differences settings, both linear and nonlinear, collect unit and time effects into η_0 and a policy parameter into θ_0 . In structural analysis η_0 would usually contain the parameters of the underlying structural model, whereas θ_0 would then be a policy counterfactual of interest. In applications, estimands that depend on η_0 , such as average treatment effects or welfare quantities, are also of interest, and we will study them in Section 6.

Consider a sample of independent observations (y_i, x_i) , and let $\hat{\eta}$ be an estimator of the nuisance parameter. It is then natural to estimate θ_0 by the *plug-in* generalized method of moments (GMM) estimator (Hansen, 1982)

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \left\| \sum_i u(y_i, x_i, \theta, \hat{\eta}) \right\|_{\Omega}^2, \quad (2.2)$$

for a chosen positive-definite weight matrix Ω , where $\|a\|_{\Omega}^2 = a^\top \Omega a$. We denote the total number of observations as n .¹

¹In $N \times T$ panel-data settings an observation is the set of T time-series observations on a given

However, estimation noise in $\hat{\eta}$ introduces bias in the estimating equation for θ_0 , in that

$$\mathbb{E}[u(y, x, \theta_0, \hat{\eta})] \neq 0.$$

This bias can be substantial when $\hat{\eta}$ is very imprecise; for example, this occurs when η_0 is high-dimensional as in models with fixed effects. Bias in the moment equation generally implies that the plug-in estimator $\hat{\theta}$ is biased as well. This, then, invalidates conventional inference procedures based on it.

One way to proceed is to work with a modified version of the moment function u , say u^* , which still satisfies

$$\mathbb{E}[u^*(y, x, \theta_0, \eta_0)] = 0, \quad (2.3)$$

but which at the same time is less sensitive to estimation noise in the nuisance parameter. Then, estimation is simply based on (2.2) using u^* instead of u . We will consider three notions of insensitivity, or *orthogonality*.

2.2 Definitions

Each of the three notions of orthogonality that we consider delivers a modified estimating equation. We will denote these as u_1^* for Neyman orthogonality, u_q^* for orthogonality to order q , and u_∞^* for full orthogonality. In addition to (2.3), these functions are designed to satisfy, respectively,

$$\mathbb{E}[\nabla_\eta u_1^*(y, x, \theta_0, \eta_0)] = 0, \quad (2.4)$$

where $\nabla_\eta u$ is the $d_\eta \times d_u$ matrix of partial derivatives in η ,

$$\mathbb{E}[\nabla_\eta^j u_q^*(y, x, \theta_0, \eta_0)] = 0 \quad j = 1, \dots, q, \quad (2.5)$$

where $\nabla_\eta^j u$ denotes the matrix of partial derivatives at order j , again with d_u columns, and, finally,

$$\mathbb{E}[u_\infty^*(y, x, \theta_0, \eta)] = 0 \quad \text{for all } \eta \in \mathcal{H}. \quad (2.6)$$

unit, out of a total of N units. Thus, in that case (2.2) concerns N summands, whereas the total sample size is $n = NT$. In a single network involving N nodes the sum collapses to a single term, whereas the number of observations is $n = N(N-1)/2$ (in the undirected case). All rate conditions below are written in terms of n .

Neyman orthogonality, in (2.4), is the weakest of the three requirements, and general recipes exist to transform a given u into u_1^* . Full orthogonality, in (2.6), in contrast is the strongest. However, informative fully-orthogonal moments do not always exist. Orthogonality to a certain order q , in (2.5), then interpolates between these two cases. We will study the three notions of orthogonality in Sections 3, 4, and 5, respectively, within the scope of the likelihood framework, as in Neyman’s (1959) original contribution. We describe this framework next.

2.3 Model and examples

In the conditional likelihood framework, the distribution of y given x belongs to the family

$$\{f(y|x; \theta, \eta) : \theta \in \Theta, \eta \in \mathcal{H}\}$$

and so is known up to values of the parameters θ and η . We will write $\mathbb{E}_{x;\theta,\eta}$ to denote expectations under $f(\cdot|x; \theta, \eta)$ and, as before, indicate true parameter values by θ_0 and η_0 . Here, the parameter of interest is θ_0 and one natural option for the function u would be the score function $\nabla_\theta \ln f$, where we use ∇_θ to denote the partial-derivative operator with respect to θ . Other choices for u are possible, for example when the score is difficult to calculate.

We now introduce the three examples we use as illustrations: a logistic regression model, a Poisson model with fixed effects, and a nonlinear regression model.

Example 1 (Logit). Consider a generic logistic regression

$$y = \mathbf{1} \{x_1 \theta_0 + x_2 \eta_0 + \varepsilon \geq 0\}, \quad \varepsilon \mid x_1, x_2 \sim \text{i.i.d. Logistic}, \quad (2.7)$$

where $y \in \{0, 1\}^n$, $x_1 \in \mathbb{R}^{n \times d_\theta}$, $x_2 \in \mathbb{R}^{n \times d_\eta}$, and the indicator acts coordinate-wise. We focus on situations where d_θ is moderate and d_η is large, and here our target parameter is θ_0 . A useful property of the model is that $s = x_2^\top y$ is a sufficient statistic for η_0 at a given θ_0 . Model (2.7) nests several popular specifications in the literature, as we now illustrate.

A first special case of model (2.7) is the panel binary-choice model. Given a random sample on individuals $i = 1, \dots, N$, followed over time periods $t = 1, \dots, T$,

the panel logit model with a scalar outcome and individual-specific fixed effects is

$$y_{it} = \mathbf{1} \left\{ x_{it}^\top \theta_0 + \eta_{i0} + \varepsilon_{it} \geq 0 \right\}, \quad \varepsilon_{it} \mid x_{i1}, \dots, x_{iT} \sim \text{i.i.d. Logistic.} \quad (2.8)$$

In this model, the maximum likelihood estimator is inconsistent for T fixed as N tends to infinity (Chamberlain, 1980). Likewise, the plug-in estimator $\hat{\theta}$ based on first-step estimates $\hat{\eta}_i$ is inconsistent too.

A second special case is a logistic model of network formation. Consider a process of conditionally-independent undirected link formation where each pair of agents $i < j$ decide whether to form a link according to the binary-choice model

$$y_{ij} = \mathbf{1} \left\{ x_{ij}^\top \theta_0 + \eta_{i0} + \eta_{j0} + \varepsilon_{ij} \geq 0 \right\}, \quad \varepsilon_{ij} \mid x \sim \text{i.i.d. Logistic.} \quad (2.9)$$

Homophily is captured by the pair-level covariates gathered in $x = (x_{ij})_{i < j}$, and degree heterogeneity by the agent effects gathered in $\eta_0 = (\eta_{10}, \dots, \eta_{N0})^\top$, where there are N agents and $n = N(N-1)/2$ pairs. The maximum likelihood estimator $\hat{\theta}$ is consistent as N tends to infinity but the asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$ is not centered at zero (Graham, 2017, Jochmans, 2018).

More generally, model (2.7) also nests logistic regression models with a high-dimensional matrix x_2 of controls. In such settings, maximum likelihood is unreliable, or fails to exist altogether, when d_η is not small relative to n , as shown by Sur and Candès (2019). Models of binary choice on bipartite networks also share this structure (Bonhomme and Dano, 2024).

Example 2 (Poisson with fixed effects). For a panel of counts, a common specification is the Poisson model (see, e.g., Hausman, Hall, and Griliches 1984; Blundell, Griffith, and Windmeijer 2002). For units $i = 1, \dots, N$ and time periods $t = 1, \dots, T$ we specify

$$y_{it} \mid x_{i1}, \dots, x_{iT} \sim \text{independent Poisson}(\mu_{it}), \quad (2.10)$$

where $\mu_{it} = \exp(x_{it}^\top \theta_0 + \eta_{i0})$ is the conditional mean. This model has the special property that the Neyman-orthogonal score is in fact fully orthogonal, as we will see in Section 3.

Example 3 (Nonlinear regression). Our third example is a nonlinear model for a

continuous outcome $y \in \mathbb{R}^n$,

$$y = m(x, \beta_0, \eta_0) + \sigma_0 \varepsilon, \quad \varepsilon \mid x \sim \text{i.i.d. standard Gaussian}, \quad (2.11)$$

for a known function m , a matrix of covariates x , η_0 a high-dimensional vector of parameters, and $\theta_0 = (\beta_0^\top, \sigma_0^2)^\top$ a low-dimensional parameter. Model (2.11) contains as special cases the model with two-way (e.g., worker and firm) heterogeneity of [Abowd, Kramarz, and Margolis \(1999\)](#) and its nonlinear generalization introduced in [Crippa \(2025\)](#), under the assumption that errors are Gaussian. The model also nests models of team production (as in [Ahmadpoor and Jones, 2019](#), for example). Unlike the first two examples, the nonlinear regression model is not part of the standard exponential family in general, and it does not admit a low-dimensional sufficient statistic for η_0 .

3 Neyman orthogonality

3.1 Motivation

Recall that a function u_1^* is Neyman orthogonal if two conditions hold: it has mean zero at the truth, as in (2.3), and its derivative in η has mean zero at the truth, as in (2.4). To motivate this second condition, suppose for the presentation that $d_\eta = d_u = 1$, so both the nuisance parameter and moment condition are scalar. Expand u_1^* around the nuisance parameter, take expectations, and exploit (2.3) to arrive at

$$\mathbb{E}[u_1^*(y, x, \theta_0, \hat{\eta})] = \mathbb{E}[\nabla_\eta u_1^*(y, x, \theta_0, \eta_0)(\hat{\eta} - \eta_0)] + (\text{higher-order terms}), \quad (3.1)$$

where here expectations are taken at θ_0, η_0 , and the population distribution of x .

We will assume that the estimator $\hat{\eta}$ is independent of the estimation sample. Under random sampling, a way to enforce independence is to estimate $\hat{\eta}$ on one part of the sample, and to estimate the target parameter $\hat{\theta}$, given the first-step estimate $\hat{\eta}$, based on the left-out observations. This *sample-splitting* strategy is popular in the literature, and variability can be reduced by using multiple splits and averaging the resulting estimates, through cross-fitting ([Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins, 2018](#)).

The dominant term on the right-hand side of (3.1) can then be written as

$$\underbrace{\text{cov} [\nabla_{\eta} u_1^*(y, x, \theta_0, \eta_0), (\hat{\eta} - \eta_0)]}_{=0 \text{ by independence (sample-splitting)}} + \underbrace{\mathbb{E} [\nabla_{\eta} u_1^*(y, x, \theta_0, \eta_0)] \mathbb{E} [(\hat{\eta} - \eta_0)]}_{=0 \text{ under (2.4)}},$$

where the first term is zero by independence and the second term is also zero since u_1^* is Neyman orthogonal, i.e., (2.4) holds.

As such, Neyman orthogonality removes the leading term from the bias expansion (3.1). The remaining bias is of the order of the mean-squared error $\mathbb{E}(\|\hat{\eta} - \eta_0\|^2)$ in general. It will be asymptotically negligible provided that it goes to zero faster than $n^{-1/2}$. This requires the first-step estimator $\hat{\eta}$ to converge at a rate faster than $n^{-1/4}$. Hence, with Neyman orthogonality and sample splitting, bias is asymptotically negligible even though nuisance parameters are estimated at a slower-than-parametric rate.

3.2 Neyman's (1959) orthogonality construction

In our likelihood setting, the requirement of Neyman orthogonality can be expressed in terms of the correlation between u_1^* and the score for the nuisance parameter. To see this, we first adapt the Neyman orthogonality property to the current setting, by requiring that it holds conditionally on x and at all values of the parameters. That is, we call u_1^* Neyman orthogonal if

$$\mathbb{E}_{x;\theta,\eta} [\nabla_{\eta} u_1^*(y, x, \theta, \eta)] = 0 \quad \text{for all } \theta \in \Theta, \eta \in \mathcal{H}, \text{ almost every } x. \quad (3.2)$$

We will also require u_1^* to be unbiased conditionally on x and at all values of the parameters; that is,

$$\mathbb{E}_{x;\theta,\eta} [u_1^*(y, x, \theta, \eta)] = 0 \quad \text{for all } \theta \in \Theta, \eta \in \mathcal{H}, \text{ almost every } x. \quad (3.3)$$

The construction below will guarantee (3.3) and (3.2), which are slightly stronger than (2.3) and (2.4), respectively.

To proceed, note that (3.3) is equivalent to

$$\int u_1^*(y, x, \theta, \eta) f(y \mid x; \theta, \eta) dy = 0 \quad \text{for all } \theta \in \Theta, \eta \in \mathcal{H}, \text{ almost every } x. \quad (3.4)$$

Differentiation with respect to η produces

$$\mathbb{E}_{x;\theta,\eta} [\nabla_{\eta} u_1^*(y, x, \theta, \eta)] + \mathbb{E}_{x;\theta,\eta} [\nabla_{\eta} \ln f(y | x; \theta, \eta) u_1^*(y, x, \theta, \eta)^{\top}] = 0.$$

Hence, letting

$$w_1(y, x, \theta, \eta) = \nabla_{\eta} \ln f(y | x; \theta, \eta),$$

it follows that the Neyman orthogonality property (3.2) is equivalent, under (3.3), to

$$\mathbb{E}_{x;\theta,\eta} [u_1^*(y, x, \theta, \eta) w_1(y, x, \theta, \eta)^{\top}] = 0 \quad \text{for all } \theta \in \Theta, \eta \in \mathcal{H}, \text{ almost every } x. \quad (3.5)$$

The characterization in (3.5) readily points to the practical construction of Neyman-orthogonal moments, as initially proposed in [Neyman \(1959\)](#). Denote the orthogonal projection of a function v onto a space $\mathcal{S}$ by

$$\text{Proj}_{x;\theta,\eta} [v \parallel \mathcal{S}] = \underset{h \in \mathcal{S}}{\text{argmin}} \mathbb{E}_{x;\theta,\eta} [\|v(y) - h(y)\|^2]. \quad (3.6)$$

The orthogonal moment u_1^* is simply constructed by projecting the original u onto the span of w_1 and keeping the residual,

$$u_1^*(\cdot, x, \theta, \eta) = u(\cdot, x, \theta, \eta) - \text{Proj}_{x;\theta,\eta} [u(\cdot, x, \theta, \eta) \parallel \text{span} \{w_1(\cdot, x, \theta, \eta)\}]. \quad (3.7)$$

The projection guarantees that (3.5) holds. Moreover, since $\mathbb{E}_{x;\theta,\eta} [w_1(y, x, \theta, \eta)] = 0$, (3.3) holds as well whenever the original moment function u is unbiased at all parameter values.

Finally, since η is finite-dimensional, the projection is finite-dimensional as well and the orthogonal moment u_1^* in (3.7) takes the explicit form

$$u_1^*(y, x, \theta, \eta) = u(y, x, \theta, \eta) - \Lambda_1(x; \theta, \eta) w_1(y, x, \theta, \eta), \quad (3.8)$$

with least-squares projection coefficient

$$\Lambda_1(x; \theta, \eta) = \mathbb{E}_{x;\theta,\eta} [u(y, x, \theta, \eta) w_1(y, x, \theta, \eta)^{\top}] \mathbb{E}_{x;\theta,\eta} [w_1(y, x, \theta, \eta) w_1(y, x, \theta, \eta)^{\top}]^{-1},$$

provided the Gram matrix $\mathbb{E}_{x;\theta,\eta} [w_1(y, x, \theta, \eta) w_1(y, x, \theta, \eta)^{\top}]$ is non-singular.

3.3 Examples

Example 2 (continued). Take a single unit from the panel Poisson model (2.10) and let $\mu_t = e^{x_t^\top \theta + \eta} = e^{x_t^\top \theta} e^\eta$. Here, taking the original moment function u to be the score with respect to θ , we have

$$w_1 = \nabla_\eta \ln f = \sum_{t=1}^T (y_t - \mu_t), \quad u = \nabla_\theta \ln f = \sum_{t=1}^T x_t (y_t - \mu_t).$$

By the mean-variance equality of the Poisson distribution, the least-squares projection coefficient is

$$\Lambda_1 = \frac{\sum_t \mu_t x_t}{\sum_t \mu_t} = \sum_t p_t x_t = \bar{x}, \quad p_t = \frac{e^{x_t^\top \theta}}{\sum_\tau e^{x_\tau^\top \theta}},$$

which is a weighted mean of the regressors. Hence,

$$u_1^* = \sum_{t=1}^T (x_t - \bar{x}) (y_t - \mu_t), \quad (3.9)$$

in which the weighted covariate mean $\bar{x}$ is free of η .

In fact, in this model u_1^* itself is free of η . To see this, note that

$$\begin{aligned} u_1^* &= \sum_{t=1}^T (x_t - \bar{x}) (y_t - \mu_t) \\ &= \sum_{t=1}^T (x_t - \bar{x}) y_t - e^\eta \sum_{t=1}^T (x_t - \bar{x}) e^{x_t^\top \theta} \\ &= \sum_{t=1}^T (x_t - \bar{x}) y_t - e^\eta \left(\sum_{\tau=1}^T e^{x_\tau^\top \theta} \right) \sum_{t=1}^T p_t (x_t - \bar{x}) \\ &= \sum_{t=1}^T (x_t - \bar{x}) y_t, \end{aligned} \quad (3.10)$$

where the last equality uses that $\sum_t p_t (x_t - \bar{x}) = 0$. Consequently, for every η' ,

$$\mathbb{E}_{x;\theta,\eta} [u_1^*(y, x, \theta, \eta')] = \sum_{t=1}^T (x_t - \bar{x}) \mu_t = e^\eta \left(\sum_{\tau=1}^T e^{x_\tau^\top \theta} \right) \sum_{t=1}^T p_t (x_t - \bar{x}) = 0, \quad (3.11)$$

and the Poisson model's Neyman-orthogonal score is fully orthogonal in the sense of (2.6).

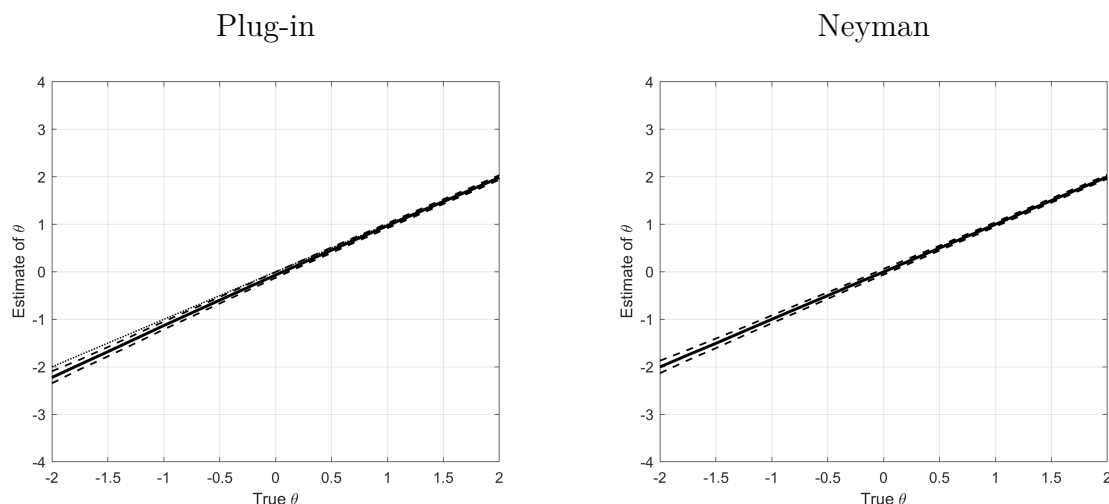
This robustness property arises from multiplicative separability, and relates to

the property that the Poisson fixed-effects maximum likelihood estimator is immune to the incidental parameter problem (Hausman, Hall, and Griliches, 1984, Blundell, Griffith, and Windmeijer, 2002, Lancaster, 2002, Dhaene and Jochmans, 2015a).

Figure 2 illustrates this observation, based on a Monte Carlo simulation exercise with $N = 1000$ and $T = 5$. There is a single binary covariate, and the parameter θ is shown on the x-axis. On the y-axis we report the mean (in solid), 5th and 95th percentiles (in dashed) based on 1000 simulations. Details on the design and estimation, including sample splitting, can be found in Appendix C.

The left graph shows the plug-in estimator. The plug-in moment is not orthogonal, and whether the resulting estimator is biased therefore depends on the first-step estimator. In our design, the bias arises because units with no positive split-sample count are dropped, which makes $\hat{e}^{\hat{\eta}_i}$ an upward-biased estimate of e^{η_i} among retained units. While the bias is moderate, it is somewhat larger at lower values of θ_0 . In contrast, the Neyman-orthogonal estimator shown in the right graph does not depend on $\hat{\eta}$ at all, and is unaffected.

Figure 2: Plug-in and Neyman-orthogonal estimators in a Poisson model



Notes: Monte Carlo mean (solid), 5th and 95th percentiles (dashed), and the 45-degree line (dotted), as functions of θ_0 . $N = 1000$, $T = 5$, $T_{\text{split}} = 2$ periods for sample splitting (without cross-fitting), and 1000 replications. Units with no positive split-sample count are dropped. The Neyman-orthogonal estimator does not require first-step nuisance estimates, but here we still drop units with no positive split-sample count.

Example 1 (continued). In the panel logit model (2.8), the scores for the slope coefficients and fixed effect have the same structure as in the Poisson model,

$$w_1 = \nabla_\eta \ln f = \sum_{t=1}^T (y_t - \pi_t), \quad u = \nabla_\theta \ln f = \sum_{t=1}^T x_t (y_t - \pi_t),$$

where $\pi_t = \exp(x_t^\top \theta + \eta) / (1 + \exp(x_t^\top \theta + \eta))$. Let $v_t = \pi_t(1 - \pi_t)$ be the conditional variance of y_t . Then projecting u orthogonally to w_1 delivers

$$u_1^* = \sum_{t=1}^T (x_t - \bar{x}) (y_t - \pi_t), \quad \bar{x} = \frac{\sum_{t=1}^T v_t x_t}{\sum_{t=1}^T v_t}. \quad (3.12)$$

The key difference with the previous example is that, now, the weights v_t in $\bar{x}$ still depend on η , while the weights p_t in the Poisson model did not.

For fixed θ , η , and x , let

$$G_{x;\theta,\eta}(\eta') = \mathbb{E}_{x;\theta,\eta} [u_1^*(y, x, \theta, \eta')] = \sum_{t=1}^T (x_t - \bar{x}(\eta')) (\pi_t(\eta) - \pi_t(\eta')), \quad (3.13)$$

where we have made the dependence of π_t on η explicit. We have $G_{x;\theta,\eta}(\eta) = 0$. Furthermore, by Neyman orthogonality, it equally holds that $\partial_{\eta'} G_{x;\theta,\eta}(\eta')|_{\eta'=\eta} = 0$. However, we have (see Appendix A.1 for a proof)

$$\partial_{\eta'}^2 G_{x;\theta,\eta}(\eta')|_{\eta'=\eta} = \sum_{t=1}^T v_t (1 - 2\pi_t) (x_t - \bar{x}), \quad (3.14)$$

where all quantities are evaluated at η . This quantity differs from zero in general. For example, for $T = 2$ and a scalar covariate we have

$$\partial_{\eta'}^2 G_{x;\theta,\eta}(\eta')|_{\eta'=\eta} = \frac{2(x_1 - x_2) v_1 v_2 (\pi_2 - \pi_1)}{v_1 + v_2}, \quad (3.15)$$

which is non-zero unless the covariate is constant over time or $\theta = 0$. The mean of u_1^* is thus still subject to quadratic noise in $\eta' - \eta$. In this example the nuisance parameter estimate converges at the $T^{-1/2}$ rate, which then translates into a bias of order T^{-1} . This is the same order as the bias of the standard maximum likelihood estimator (Arellano, 2003a).

The panel logit example reveals that we need $N = o(T)$ for Neyman-orthogonal

moments to alleviate the incidental-parameter bias in panel data. This rate corresponds to panels that are substantially longer than they are wide. Furthermore, it is also the same rate required for maximum likelihood to be asymptotically unbiased (see, e.g., [Hahn and Newey 2004](#)). As such, first-order orthogonalization of the likelihood score does not improve the order of the bias over estimation based on the score itself.

It is worth noting that the projection underlying [\(3.12\)](#) can be described in terms of the sufficient statistic. Indeed,

$$w_1 = \sum_{t=1}^T (y_t - \pi_t) = s - \mathbb{E}_{x;\theta,\eta}[s], \quad s = \sum_{t=1}^T y_t,$$

where s is sufficient for η at a given θ . The nuisance score and the centered sufficient statistic therefore span the same space, and u_1^* is the residual from projecting u orthogonally to s . The same holds in the general logit model [\(2.7\)](#), where $s = x_2^\top y$ and

$$w_1 = \nabla_\eta \ln f = x_2^\top (y - \pi) = s - \mathbb{E}_{x;\theta,\eta}[s].$$

The same property also holds in the Poisson model of [Example 2](#), where $w_1 = \sum_{t=1}^T (y_t - \mu_t) = s - \mathbb{E}_{x;\theta,\eta}[s]$ for the same sufficient statistic $s = \sum_{t=1}^T y_t$, with the difference that, as [\(3.11\)](#) shows, in that model Neyman's projection yields full orthogonality.

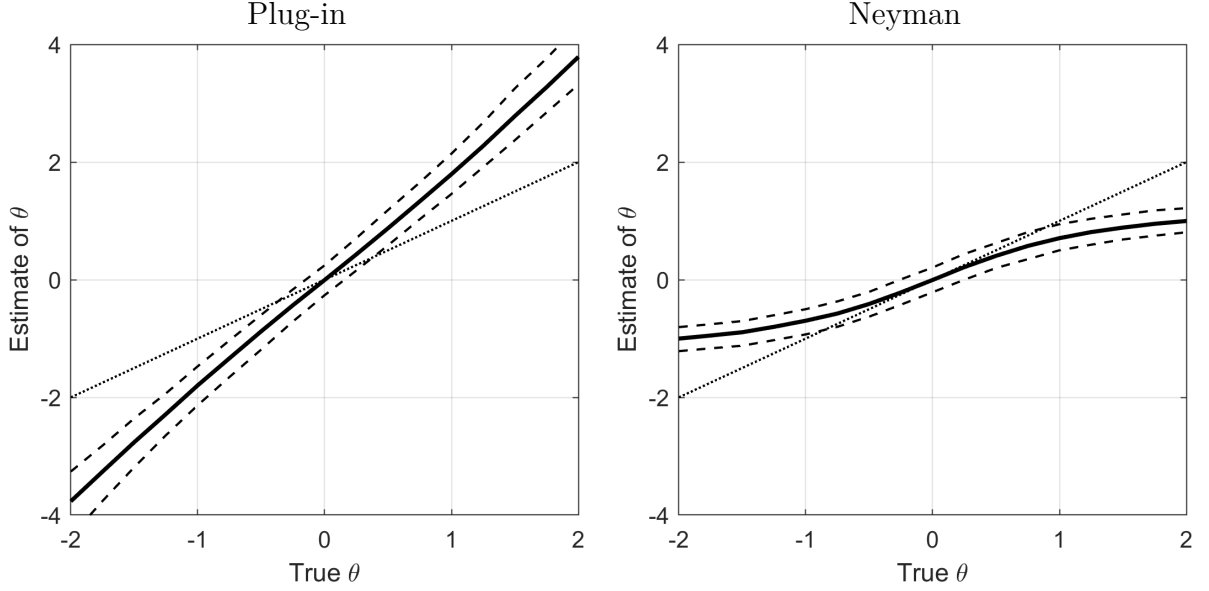
[Figure 3](#) illustrates this in a logit model. Details on the design and estimation, including sample splitting, can be again found in [Appendix C](#). The left graph shows that the plug-in estimator is biased, and the bias increases with $|\theta_0|$. Moreover, the right graph shows that the Neyman-orthogonal estimator remains biased, although the bias is in the other direction (in this case, towards zero) compared to plug-in. This confirms that the Neyman-orthogonal estimator is not fully orthogonal in the logit model, and shows that the bias can be large in practice.

4 Fully-orthogonal moments

4.1 Definition and motivation

In the likelihood setting, we require full orthogonality to hold conditionally on x and at all values of the parameters. We thus define a moment function u_∞^* to be fully

Figure 3: Plug-in and Neyman-orthogonal estimators in a logit model



Notes: see notes to Figure 2. Units whose split-sample outcomes are all zero or all one are dropped.

orthogonal if

$$\mathbb{E}_{x;\theta,\eta}[u_{\infty}^*(y, x, \theta, \eta')] = 0 \quad \text{for all } \theta \in \Theta, \eta \in \mathcal{H}, \eta' \in \mathcal{H}, \text{ almost every } x. \quad (4.1)$$

When (4.1) holds, the mean of u_{∞}^* is completely insensitive to the value of the nuisance parameter at which it is evaluated. Note that (4.1) implies $\mathbb{E}_{x;\theta,\eta}[u_{\infty}^*(y, x, \theta, \eta)] = 0$, so a separate unbiasedness requirement is not needed.

Because (4.1) holds at every η' , one may fix η' at any convenient value. We then write $u_{\infty}^*(y, x, \theta)$ for the resulting function, so that full orthogonality reads

$$\mathbb{E}_{x;\theta,\eta}[u_{\infty}^*(y, x, \theta)] = 0 \quad \text{for all } \theta \in \Theta, \eta \in \mathcal{H}, \text{ almost every } x. \quad (4.2)$$

In some models a fully-orthogonal moment can be chosen so as not to depend on the value at which η' was fixed, as with the panel data moments in differences discussed in the introduction. In general, the projection construction of the next subsection requires setting η' to a particular value.

The attractiveness of a fully-orthogonal estimating equation is immediate. Unlike in the Neyman-orthogonal case, GMM estimation based on a fully-orthogonal moment

does not require a first-step estimator of η_0 converging at a specific rate: by (4.1) the moment condition holds at every value of the nuisance parameter. Standard GMM theory can then be applied to establish the asymptotic properties of the resulting estimator. In panel data, this leads to estimators that are consistent for fixed T as N tends to infinity, in contrast with estimators based on Neyman orthogonality alone.

4.2 Orthogonality construction

Condition (4.2) is

$$\int u_{\infty}^*(y, x, \theta) f(y | x; \theta, \eta) dy = 0 \quad \text{for all } \theta \in \Theta, \eta \in \mathcal{H}, \text{ almost every } x. \quad (4.3)$$

We will now show that (4.3) is equivalent to u_{∞}^* being uncorrelated with the closed span of likelihood ratios of the model. This characterization, which is reminiscent of (3.5) in the Neyman case, will suggest a construction of fully-orthogonal moments based on a projection argument.

To proceed, fix some covariate value x and parameter values $\theta \in \Theta$ and $\eta \in \mathcal{H}$. We denote the likelihood ratios of the model as

$$r_{\eta'}(y, x, \theta, \eta) = \frac{f(y | x; \theta, \eta')}{f(y | x; \theta, \eta)}, \quad \eta' \in \mathcal{H}.$$

Throughout this section we assume that, for every x and $\theta \in \Theta$, the distributions $f(\cdot | x; \theta, \eta)$, $\eta \in \mathcal{H}$, are mutually absolutely continuous, so that the likelihood ratios $r_{\eta'}$ are well defined and the support of y does not vary with η . For the projection arguments below we further assume that $r_{\eta'}(\cdot, x, \theta, \eta) \in L^2(f(\cdot | x; \theta, \eta))$ for all $\eta, \eta' \in \mathcal{H}$.

With this definition, we have

$$\begin{aligned} (4.3) &\Leftrightarrow \int u_{\infty}^*(y, x, \theta) f(y | x; \theta, \eta') dy = 0 \quad \text{for all } \eta' \in \mathcal{H} \\ &\Leftrightarrow \int u_{\infty}^*(y, x, \theta) \frac{f(y | x; \theta, \eta')}{f(y | x; \theta, \eta)} f(y | x; \theta, \eta) dy = 0 \quad \text{for all } \eta' \in \mathcal{H} \\ &\Leftrightarrow \int u_{\infty}^*(y, x, \theta) r_{\eta'}(y, x, \theta, \eta) f(y | x; \theta, \eta) dy = 0 \quad \text{for all } \eta' \in \mathcal{H}. \end{aligned}$$

This shows that u_∞^* is fully orthogonal in the sense of (4.2) if and only if

$$\mathbb{E}_{x;\theta,\eta} [u_\infty^*(y, x, \theta) r_{\eta'}(y, x, \theta, \eta)] = 0 \quad \text{for all } \theta \in \Theta, \eta \in \mathcal{H}, \eta' \in \mathcal{H}, \text{ almost every } x. \quad (4.4)$$

Condition (4.4) is the full-orthogonality counterpart to the Neyman characterization (3.5). It shows that u_∞^* is fully orthogonal if and only if it is uncorrelated with all likelihood ratios $r_{\eta'}$ for $\eta' \in \mathcal{H}$. Hence, collecting the likelihood ratios into the closed linear span (for given x, θ, η),

$$\mathcal{S}_{x;\theta,\eta} = \overline{\text{span}} \{ y \mapsto r_{\eta'}(y, x, \theta, \eta) : \eta' \in \mathcal{H} \} \subseteq L^2(f(\cdot | x; \theta, \eta)), \quad (4.5)$$

it follows from (4.4) and continuity of the inner product that u_∞^* is fully orthogonal if and only if it is orthogonal in L^2 to the linear space $\mathcal{S}_{x;\theta,\eta}$.

This characterization suggests the following construction, based on a projection argument. Given any moment function u , the projection residual, which now carries the reference value η as an argument,

$$u_\infty^*(\cdot, x, \theta, \eta) = u(\cdot, x, \theta, \eta) - \text{Proj}_{x;\theta,\eta}(u(\cdot, x, \theta, \eta) \| \mathcal{S}_{x;\theta,\eta}), \quad (4.6)$$

where $\text{Proj}_{x;\theta,\eta}[v \| \mathcal{S}]$ is given by (3.6), is fully orthogonal and, among all fully-orthogonal functions, minimizes the $L^2(f(\cdot | x; \theta, \eta))$ distance to u .

Note that (4.6) depends on the value η , through the inner product used in the projection and, when u depends on it, through the moment function itself. This dependence is of a different nature from the one in the Neyman construction. There, η is the unknown nuisance parameter and must be replaced by an estimate. Here, η is merely a reference value that we are free to choose: by (4.3), full orthogonality of a given function involves no reference value, so every choice of η produces a valid element of the same set. The choice does affect which element is selected, hence the precision of the resulting estimator, but is immaterial for the full orthogonality property.

Unlike the Neyman case where the projection is finite-dimensional (since we focus on a finite-dimensional η), the projection in (4.6) is infinite-dimensional in general. The moment function u_∞^* can be expressed as the solution to an operator equation. We provide details in Appendix B.

4.3 Examples

Example 2 (continued). In the panel Poisson model (2.10), a fully-orthogonal moment based on $u = \nabla_\theta \ln f$ is $u_\infty^* = u - \mathbb{E}_{x;\theta,\eta}[u | s]$, where $s = \sum_{t=1}^T y_t$. A key property of this model is that the conditional expectation $\mathbb{E}_{x;\theta,\eta}[u | s]$ is linear. Indeed, conditionally on s the vector y is multinomial with probabilities p_t , so that $\mathbb{E}_{x;\theta,\eta}[u | s] = \bar{x}(s - \sum_t \mu_t)$, which lies in the span of w_1 . Hence, u_∞^* coincides with the Neyman-orthogonal score u_1^* in (3.9).

Example 1 (continued). Consider the general logistic regression model (2.7), for a vector of outcomes $y \in \{0, 1\}^n$. The likelihood is

$$f(y | x; \theta, \eta) = \prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i}, \quad \pi_i = \frac{\exp(x_{1i}^\top \theta + x_{2i}^\top \eta)}{1 + \exp(x_{1i}^\top \theta + x_{2i}^\top \eta)},$$

the rows of the matrices x_1 and x_2 being $x_{1i}^\top$ and $x_{2i}^\top$ for $i = 1, \dots, n$. In this exponential family in η at fixed θ , with sufficient statistic $s = x_2^\top y \in \mathbb{R}^{d_\eta}$, the likelihood ratios are given by

$$r_{\eta'}(y, x, \theta, \eta) = \frac{f(y | x; \theta, \eta')}{f(y | x; \theta, \eta)} = c(x, \theta, \eta', \eta) \exp\left(s^\top (\eta' - \eta)\right), \quad (4.7)$$

where

$$c(x, \theta, \eta', \eta) = \prod_{i=1}^n \frac{1 + \exp(x_{1i}^\top \theta + x_{2i}^\top \eta)}{1 + \exp(x_{1i}^\top \theta + x_{2i}^\top \eta')}$$

does not depend on y . The likelihood ratios are thus, up to a normalization, exponentials in the sufficient statistic s , and

$$\mathcal{S}_{x;\theta,\eta} = \overline{\text{span}} \left\{ y \mapsto \exp\left(s^\top \tau\right) : \tau \in \mathbb{R}^{d_\eta} \right\}. \quad (4.8)$$

Since s has finite support this span is the set of all functions of s , so that

$$u_\infty^*(y, x, \theta, \eta) = u(y, x, \theta, \eta) - \mathbb{E}_{x;\theta,\eta}[u(y, x, \theta, \eta) | s]. \quad (4.9)$$

Hence, the logit model provides an example where fully-orthogonal moments (in fact, all of them) can be obtained by de-meaning. However, this does not mean that the resulting moments are automatically informative about θ . Indeed, nontrivial

moments are available if and only if $y \mapsto x_2^\top y$ is a non-injective mapping; that is, if two distinct outcome configurations share the same value of the sufficient statistic. Suppose that x_2 has continuously-distributed entries. Then, the 2^n values $\{x_2^\top y : y \in \{0, 1\}^n\}$ are almost surely distinct, so the map $y \mapsto x_2^\top y$ is injective, and (4.9) always returns $u_\infty^* = 0$ irrespective of the original moment u . It follows that a nontrivial fully-orthogonal moment does not exist.

Consider now the panel logit model (2.8). In this case the nuisance parameter is a set of fixed effects, and so the x_2 matrix in the previous example is a matrix of unit indicators. Here, $T \geq 2$ already suffices to obtain nontrivial fully-orthogonal moments provided x_{it} varies within units. Indeed, $\binom{T}{c}$ distinct outcome sequences all yield $\sum_t y_t = c$. The conditional logit score is one such moment.

Consider next the β -model (2.9) of network formation. In this model, the degree sequence $d_i = \sum_{j \neq i} y_{ij}$, $i = 1, \dots, N$, serves as sufficient statistic for η at given θ . The non-injectivity of the degree map can be shown by the following invariance argument, which underlies the estimators of Graham (2017) and Jochmans (2018): for any four distinct agents $\{i, j, k, l\}$, the configurations

$$\{y_{ij} = y_{kl} = 1, y_{ik} = y_{jl} = 0\} \quad \text{and} \quad \{y_{ij} = y_{kl} = 0, y_{ik} = y_{jl} = 1\}$$

leave the degree sequence unchanged. This delivers informative fully-orthogonal moments based on network tetrads.

4.4 A mixture interpretation

The projection in (4.6) that constructs fully-orthogonal moments can be understood as implementing the construction in Neyman (1959), applied to a different nuisance parameter and with a different reference measure. In turn, this approach is closely related to functional differencing, which was initially proposed for panel data models in Bonhomme (2012).

To describe this interpretation, let us treat η as a random vector, drawn from an unrestricted conditional distribution given covariates, $\rho(\eta | x)$. This perspective is natural in panel data, where it is sometimes referred to as “correlated random effects”, but we do not need to restrict ourselves to the panel data setting. The model then

becomes a semiparametric mixture, with likelihood

$$f^{\text{mixt}}(y \mid x; \theta, \rho) = \int_{\mathcal{H}} f(y \mid x; \theta, \eta) \rho(\eta \mid x) d\eta. \quad (4.10)$$

We leave ρ unrestricted, aside from the requirement that $\mathcal{H}$ be its support. Note that f^{mixt} depends linearly on ρ .

In this mixture model, u^* being fully orthogonal as in (4.3) can be interpreted as $u^*(y, x, \theta)$ having zero mean conditional on x and η , almost surely in x, η . Hence, u^* being fully orthogonal is equivalent to it belonging to the null space of the conditional expectation operator

$$E_{x;\theta} : u(y) \mapsto E_{x;\theta}[u](\eta) = \int u(y) f(y \mid x; \theta, \eta) dy.$$

Finding such functions is the goal of functional differencing. Elements in the null-space of $E_{x;\theta}$ can be found analytically in certain models, as shown by [Honoré and Weidner \(2025\)](#) and [Dano \(2023\)](#) in a variety of discrete choice panel models and in [Bonhomme and Dano \(2024\)](#) in some network models. However, as we have seen in [Example 1](#), fully-orthogonal moments may be uninformative about the parameter of interest. [Chamberlain \(2010\)](#) highlights that most binary choice models only possess trivial fully-orthogonal moments outside the logit case.

We now show that the construction in [Neyman \(1959\)](#), in the mixture model (4.10), provides a general way to construct such functions, which coincides with the functional differencing projection in [Bonhomme \(2012\)](#). To see this, note that in the mixture model the nuisance parameter is ρ , and the score with respect to it is the collection

$$\nabla_{\delta} \ln f^{\text{mixt}}(y \mid x; \theta, \rho_{\delta}) \Big|_{\delta=0} = \frac{\int f(y \mid x; \theta, \eta) \nabla_{\delta} \ln \rho_{\delta}(\eta \mid x) \Big|_{\delta=0} \rho(\eta \mid x) d\eta}{\int f(y \mid x; \theta, \eta) \rho(\eta \mid x) d\eta}, \quad (4.11)$$

for all sufficiently regular one-dimensional submodels $\{\rho_{\delta}\}$ passing through ρ (see Chapter 25 in [Van der Vaart, 2000](#)). Neyman's orthogonal moment, see (3.7), is then the projection of u onto the orthogonal complement to the span of these scores, the *orthocomplement of the tangent set* of the semiparametric model.

Now, since the quantities $\nabla_{\delta} \ln \rho_{\delta}(\eta \mid x) \Big|_{\delta=0}$ in (4.11) are essentially unrestricted, the zero-mean functions in the orthocomplement of the tangent set at given x and θ

are

$$\left\{ u^* \in L^2 \left(f^{\text{mixt}}(\cdot \mid x; \theta, \rho) \right) : \int u^*(y) f(y \mid x; \theta, \eta) dy = 0 \quad \text{for all } \eta \in \mathcal{H} \right\},$$

that is, they are all the fully-orthogonal moments that are square integrable with respect to $f^{\text{mixt}}(\cdot \mid x; \theta, \rho)$.² This shows that Neyman’s strategy applied to the semiparametric mixture model, and the fully-orthogonal projection in (4.6), coincide, except for the fact that the two projections use different reference measures – $f^{\text{mixt}}(\cdot \mid x; \theta, \rho)$ for the former and $f(\cdot \mid x; \theta, \eta)$ for the latter. Furthermore, the projection onto the orthocomplement of the tangent set of the mixture model coincides with the functional differencing projection. Bonhomme (2012) proposes a numerical strategy to approximate this projection for estimation.

Remark 1 (Efficiency). The tangent set interpretation reveals that, when $u = \nabla_{\theta} \ln f^{\text{mixt}}$, the corresponding projection residual is the efficient score of the semiparametric mixture model. While achieving efficiency requires knowledge of the mixing distribution ρ , the characterization remains useful in practice. One may compute the projection at a convenient density $\tilde{\rho}$ rather than at the true ρ_0 : the resulting moment is still fully orthogonal, since full orthogonality does not depend on the reference measure, and it is efficient under $\tilde{\rho}$. The choice of $\tilde{\rho}$ thus plays the role of a working model, guiding the selection among fully-orthogonal moments without affecting validity.

Remark 2 (Dynamics). The mixture formulation extends to richer conditioning structures. Bonhomme, Dano, and Graham (2023, 2025) apply it to nonlinear panel models in which strict exogeneity of covariates is relaxed to allow for dynamic feedback. There, the nuisance parameter is not only the density of the heterogeneity (given initial conditions), but also the density of covariates given past covariates, past outcomes, and the heterogeneity, also known as the feedback process.

²To see this, note that u^* is orthogonal to the score (4.11) if and only if, by Fubini’s theorem,

$$\int_{\mathcal{H}} \nabla_{\delta} \ln \rho_{\delta}(\eta \mid x) \Big|_{\delta=0} \left(\int u^*(y) f(y \mid x; \theta, \eta) dy \right) \rho(\eta \mid x) d\eta = 0.$$

As the submodel scores $\nabla_{\delta} \ln \rho_{\delta}(\cdot \mid x) \Big|_{\delta=0}$ range over the mean-zero functions of η , this holds for all submodels if and only if $\int u^*(y) f(y \mid x; \theta, \eta) dy$ does not vary with η on the support $\mathcal{H}$ of ρ . Requiring in addition that u^* have mean zero under f^{mixt} sets the constant to zero.

5 Higher-order orthogonality

We now consider a third notion of orthogonality that is intermediate between the previous two.

5.1 Motivation

To motivate (2.5), consider first a moment function u_2^* orthogonal to order $q = 2$. Suppose in this subsection that the nuisance parameter and the moment function are both univariate, so $d_\eta = d_u = 1$. Expanding u_2^* around η_0 (this time to second order) and taking expectations while noting that u_2^* has mean zero at the truth, gives

$$\begin{aligned}\mathbb{E}[u_2^*(y, x, \theta_0, \hat{\eta})] &= \mathbb{E}[\nabla_\eta u_2^*(y, x, \theta_0, \eta_0) (\hat{\eta} - \eta_0)] \\ &\quad + \frac{1}{2} \mathbb{E}[\nabla_\eta^2 u_2^*(y, x, \theta_0, \eta_0) (\hat{\eta} - \eta_0)^2] + (\text{higher-order terms}).\end{aligned}$$

When $\hat{\eta}$ is independent of the data (for example, estimated on a separate sample, as explained in Subsection 3.1), this gives

$$\begin{aligned}\mathbb{E}[u_2^*(y, x, \theta_0, \hat{\eta})] &= \underbrace{\mathbb{E}[\nabla_\eta u_2^*(y, x, \theta_0, \eta_0)]}_{=0 \text{ under (2.5) at } j=1} \mathbb{E}[(\hat{\eta} - \eta_0)] \\ &\quad + \frac{1}{2} \underbrace{\mathbb{E}[\nabla_\eta^2 u_2^*(y, x, \theta_0, \eta_0)]}_{=0 \text{ under (2.5) at } j=2} \mathbb{E}[(\hat{\eta} - \eta_0)^2] + (\text{higher-order terms}).\end{aligned}\tag{5.1}$$

Hence, since u_2^* is orthogonal to order $q = 2$, the two dominant terms in (5.1) are equal to zero. Furthermore, the remainder term will generally be of order $\mathbb{E}[\|\hat{\eta} - \eta_0\|^3]$, implying that the bias induced by estimation noise in the nuisance parameter is asymptotically negligible when $\sqrt{n} \mathbb{E}[\|\hat{\eta} - \eta_0\|^3] \rightarrow 0$ as $n \rightarrow \infty$.

More generally, while the case $q = 1$ recovers the classical Neyman-orthogonality condition, larger values of q yield increased robustness to estimation noise in the nuisance parameter. By the same logic, if u_q^* is orthogonal to order q , then $\mathbb{E}[u_q^*(y, x, \theta_0, \hat{\eta})]$ is of order $\mathbb{E}[\|\hat{\eta} - \eta_0\|^{q+1}]$, implying that the bias induced by estimation noise in the nuisance parameter is asymptotically negligible when $\sqrt{n} \mathbb{E}[\|\hat{\eta} - \eta_0\|^{q+1}] \rightarrow 0$ as $n \rightarrow \infty$. This requires that the first-step estimator $\hat{\eta}$ converge at a rate faster than $n^{-1/2(q+1)}$, thus relaxing Neyman's $n^{-1/4}$ condition when $q > 1$ (Mackey, Syrgkanis, and Zadik,

2018).

5.2 Construction by projection on generalized scores

In the likelihood setting, we require orthogonality to order q to hold conditionally on x and at all values of the parameters, similarly to our definitions of Neyman and full orthogonality. That is, we call u_q^* orthogonal to order q if it is conditionally unbiased,

$$\mathbb{E}_{x;\theta,\eta} [u_q^*(y, x, \theta, \eta)] = 0 \quad \text{for all } \theta \in \Theta, \eta \in \mathcal{H}, \text{ almost every } x, \quad (5.2)$$

and all its first q derivatives in η have zero expectation as well,

$$\mathbb{E}_{x;\theta,\eta} [\nabla_\eta^j u_q^*(y, x, \theta, \eta)] = 0 \quad \text{for all } j = 1, \dots, q, \theta \in \Theta, \eta \in \mathcal{H}, \text{ almost every } x, \quad (5.3)$$

where recall that ∇_η^j collects all derivatives at order j in a vector.

Orthogonality to order q can be equivalently written as an uncorrelatedness condition, similarly to the other two notions of orthogonality we have considered. To see this, recall the likelihood ratio, for given x and θ ,

$$r_{\eta'}(y, x, \theta, \eta) = \frac{f(y | x; \theta, \eta')}{f(y | x; \theta, \eta)}, \quad \eta' \in \mathcal{H}.$$

Denote its vector of j th derivatives as

$$w_j(y, x, \theta, \eta) = \nabla_{\eta'}^j r_{\eta'}(y, x, \theta, \eta) \Big|_{\eta'=\eta} = \frac{\nabla_\eta^j f(y | x; \theta, \eta)}{f(y | x; \theta, \eta)}.$$

Observe that, when $j = 1$, w_j coincides with the score for the nuisance parameter, w_1 , from Section 3. When $j = 2$,

$$w_2(y, x, \theta, \eta) = \nabla_\eta^2 \ln f(y | x; \theta, \eta) + \nabla_\eta \ln f(y | x; \theta, \eta) \otimes \nabla_\eta \ln f(y | x; \theta, \eta),$$

which has mean zero by the information identity. The set of functions $\{w_j\}$ is known as the [Bhattacharyya \(1946\)](#) basis.³ Its elements can be interpreted as generalized

³More generally, w_j is a polynomial in $\nabla_\eta \ln f, \dots, \nabla_\eta^j \ln f$. In the scalar case, it is given by the complete Bell polynomial evaluated at these elements. The vector case follows from the multivariate Faà di Bruno formula ([Constantine and Savits, 1996](#)).

score functions, and they satisfy (by differentiating j times $\int r_{\eta'} f_{\eta} dy = 1$ at $\eta' = \eta$):

$$\mathbb{E}_{x;\theta,\eta}[w_j(y, x, \theta, \eta)] = 0 \quad \text{for all } j \geq 1, \theta \in \Theta, \eta \in \mathcal{H}, \text{ almost every } x. \quad (5.4)$$

Now consider again $q = 2$ in the all-scalar case. As in our derivation of Neyman's original device, differentiate the unbiasedness condition (5.2) once with respect to η . This yields the moment condition

$$\underbrace{\mathbb{E}_{x;\theta,\eta} [\nabla_{\eta} u_2^*(y, x, \theta, \eta)]}_{=0 \text{ under (5.3) at } j=1} + \mathbb{E}_{x;\theta,\eta} [u_2^*(y, x, \theta, \eta) w_1(y, x, \theta, \eta)] = 0, \quad (5.5)$$

so

$$\mathbb{E}_{x;\theta,\eta} [u_2^*(y, x, \theta, \eta) w_1(y, x, \theta, \eta)] = 0 \text{ for all } \theta \in \Theta, \eta \in \mathcal{H}, \text{ almost every } x. \quad (5.6)$$

Next, twice differentiate (5.2) to see that

$$\underbrace{\mathbb{E}_{x;\theta,\eta} [\nabla_{\eta}^2 u_2^*(y, x, \theta, \eta)]}_{=0 \text{ under (5.3) at } j=2} + 2 \mathbb{E}_{x;\theta,\eta} [\nabla_{\eta} u_2^*(y, x, \theta, \eta) w_1(y, x, \theta, \eta)] + \mathbb{E}_{x;\theta,\eta} [u_2^*(y, x, \theta, \eta) w_2(y, x, \theta, \eta)] = 0,$$

so

$$2 \mathbb{E}_{x;\theta,\eta} [\nabla_{\eta} u_2^*(y, x, \theta, \eta) w_1(y, x, \theta, \eta)] + \mathbb{E}_{x;\theta,\eta} [u_2^*(y, x, \theta, \eta) w_2(y, x, \theta, \eta)] = 0. \quad (5.7)$$

Furthermore, because (5.6) can be written as $\int u_2^* \nabla_{\eta} f dy = 0$ and needs to hold at all parameter values, we obtain, by differentiating it,

$$\mathbb{E}_{x;\theta,\eta} [\nabla_{\eta} u_2^*(y, x, \theta, \eta) w_1(y, x, \theta, \eta)] + \mathbb{E}_{x;\theta,\eta} [u_2^*(y, x, \theta, \eta) w_2(y, x, \theta, \eta)] = 0. \quad (5.8)$$

Combining (5.7) and (5.8) then implies

$$\mathbb{E}_{x;\theta,\eta} [u_2^*(y, x, \theta, \eta) w_2(y, x, \theta, \eta)] = 0. \quad (5.9)$$

This argument shows that, if u_2^* is orthogonal to order $q = 2$, then it is uncorrelated with the first two Bhattacharyya functions w_1 and w_2 . Moreover, by following the same argument in reverse it is easy to see that the converse is also true. Hence, second-

order orthogonality, for a moment satisfying (5.2), can be equivalently expressed as (5.6) and (5.9).

Iterating the argument, now for the general vector-valued case, reveals that u_q^* is orthogonal to order q if and only if it satisfies the unbiasedness condition (5.2) and the orthogonality conditions

$$\mathbb{E}_{x;\theta,\eta}[u_q^*(y, x, \theta, \eta) w_j(y, x, \theta, \eta)^\top] = 0, \quad j = 1, \dots, q. \quad (5.10)$$

The representation in (5.10) suggests constructing higher-order orthogonal functions by projection, as we did before for the other two notions of orthogonality. Specifically, u_q^* is constructed as the residual of a projection of the original moment function u onto the span of $w_1, \dots, w_q$:

$$u_q^*(\cdot, x, \theta, \eta) = u(\cdot, x, \theta, \eta) - \text{Proj}_{x;\theta,\eta}(u(\cdot, x, \theta, \eta) \parallel \text{span}\{w_1(\cdot, x, \theta, \eta), \dots, w_q(\cdot, x, \theta, \eta)\}), \quad (5.11)$$

where $\text{Proj}_{x;\theta,\eta}[v \parallel \mathcal{S}]$ is again given by (3.6).

Stacking $W_q = (w_1^\top, \dots, w_q^\top)^\top$, we can write the explicit expression

$$u_q^*(y, x, \theta, \eta) = u(y, x, \theta, \eta) - \Lambda_q(x; \theta, \eta) W_q(y, x, \theta, \eta), \quad (5.12)$$

with least-squares projection coefficient

$$\Lambda_q(x; \theta, \eta) = \mathbb{E}_{x;\theta,\eta} [u(y, x, \theta, \eta) W_q(y, x, \theta, \eta)^\top] \mathbb{E}_{x;\theta,\eta} [W_q(y, x, \theta, \eta) W_q(y, x, \theta, \eta)^\top]^{-1}.$$

Note that, by (5.4), u_q^* automatically satisfies (5.2) provided the original moment u is unbiased at all parameter values.

The explicit expression in (5.12) requires nonsingularity of the Gram matrix of the leading basis functions W_q , $\mathbb{E}_{x;\theta,\eta} [W_q(y, x, \theta, \eta) W_q(y, x, \theta, \eta)^\top]$, as Subsection 5.4 below illustrates. For the special case where $u = \nabla_\theta \ln f$, (5.12) reduces to the projected score of Small and McLeish (1989) and Waterman and Lindsay (1996). The more general construction is introduced in Bonhomme, Jochmans, and Weidner (2024).

Remark 3 (Nested spaces). Note that, as anticipated in Figure 1, the three notions

of orthogonality correspond to a sequence of nested spaces:

$$\underbrace{\text{span}\{w_1(\cdot, x, \theta, \eta)\}}_{\text{Neyman}} \subseteq \underbrace{\text{span}\{w_1(\cdot, x, \theta, \eta), \dots, w_q(\cdot, x, \theta, \eta)\}}_{\text{order } q} \subseteq \underbrace{\mathcal{S}_{x;\theta,\eta}}_{\text{full}},$$

where $\mathcal{S}_{x;\theta,\eta}$ is the closed span of likelihood ratios defined in (4.5). The second inclusion holds because each w_j is a limit of finite differences of likelihood ratios. A larger space brings more robustness to estimation noise in η , at the cost of less information about θ .

5.3 Examples

Example 1 (continued). Consider again the general logistic regression model (2.7), for a vector of outcomes $y \in \{0, 1\}^n$. Writing the log-likelihood as

$$\ln f(y \mid x; \theta, \eta) = y^\top x_1 \theta + s^\top \eta - \kappa(x, \theta, \eta),$$

for

$$\kappa(x, \theta, \eta) = \sum_{i=1}^n \ln \left(1 + \exp \left(x_{1i}^\top \theta + x_{2i}^\top \eta \right) \right),$$

we have $\nabla_\eta \ln f = s - \nabla_\eta \kappa$, for $s = x_2^\top y$ the sufficient statistic, while $\nabla_\eta^j \ln f = -\nabla_\eta^j \kappa$ for $j \geq 2$ does not depend on y . Hence w_j is a polynomial of degree j in the sufficient statistic s , its degree- j term being $w_1^{\otimes j}$.

The coefficients of this polynomial can be expressed in terms of the cumulants of s . Indeed, κ is the cumulant generating function of the exponential family in η , so that $\nabla_\eta^j \kappa$ is the j th cumulant of s under $f(\cdot \mid x; \theta, \eta)$. In particular $\nabla_\eta \kappa = \mathbb{E}_{x;\theta,\eta}[s]$ and $\nabla_\eta^2 \kappa = \text{vec}(\text{Var}_{x;\theta,\eta}(s))$, so that

$$w_1 = s - \mathbb{E}_{x;\theta,\eta}[s], \quad w_2 = (s - \mathbb{E}_{x;\theta,\eta}[s])^{\otimes 2} - \text{vec}(\text{Var}_{x;\theta,\eta}(s)).$$

Since the w_j have mean zero and are of degree j , they form a basis of the mean-zero polynomials in s of total degree at most q . Orthogonality to order q therefore amounts to projecting the original moment on powers of the sufficient statistic:

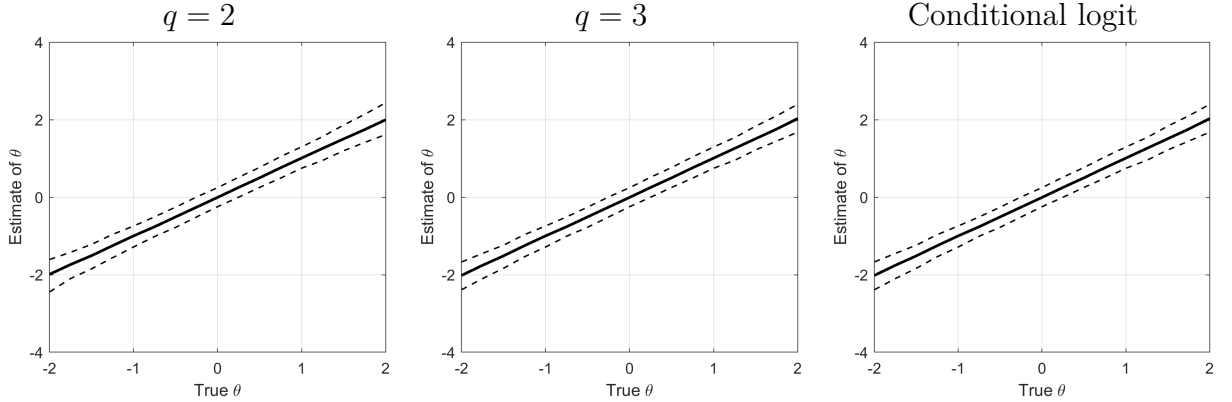
$$u_q^*(\cdot, x, \theta, \eta) = u(\cdot, x, \theta, \eta) - \text{Proj}_{x;\theta,\eta}(u(\cdot, x, \theta, \eta) \parallel \{\text{polynomials in } s \text{ of degree } \leq q\}), \quad (5.13)$$

where the constants may be included in the projection space without affecting the residual, since u has mean zero.

Specializing to the panel logit model (2.8) makes the construction fully explicit. At the level of a single unit, x_2 is a vector of ones and the sufficient statistic is the scalar $s = \sum_{t=1}^T y_t$, so that w_j is a polynomial of degree j in s whose coefficients are the cumulants of a sum of T independent Bernoulli variables; Appendix A collects the explicit formulas. Orthogonality to order q therefore projects the score on the powers $s, s^2, \dots, s^q$.

Figure 4 illustrates this in Monte Carlo simulations based on the logit model, using the design mentioned in Section 3. The left graph shows the estimator orthogonal to $q = 2$, the middle one shows the estimator orthogonal to $q = 3$, and the right one shows the conditional logit estimator, which is based on a fully-orthogonal moment. We see that the estimators behave very similarly in this simulation. Hence, even if there is no guarantee for the second-order orthogonal moment to be fully orthogonal in the logit model, increasing the orthogonality requirement by a single order (from $q = 1$ to $q = 2$) is sufficient to almost fully remove the bias of the Neyman-orthogonal estimator.

Figure 4: Orthogonal estimators in a logit model



Notes: see notes to Figure 3.

Example 3 (continued). Consider now the nonlinear regression model (2.11). Its log-likelihood is

$$\ln f(y \mid x; \theta, \eta) = \frac{1}{\sigma^2} y^\top m(x, \beta, \eta) - \frac{1}{2\sigma^2} \left(y^\top y + m(x, \beta, \eta)^\top m(x, \beta, \eta) \right) - \frac{n}{2} \ln (2\pi\sigma^2),$$

where recall that $\theta = (\beta^\top, \sigma^2)^\top$. Viewed as a family in η at fixed θ , this is a Gaussian location family whose natural parameter is $m(x, \beta, \eta)/\sigma^2$ and whose sufficient statistic is y itself. Since η is d_η -dimensional while y is n -dimensional, the natural parameter belongs to a d_η -dimensional submanifold of $\mathbb{R}^n$, and unless m is linear in η the model is a *curved* exponential family in the sense of Efron (1975).⁴ This is the difference with Examples 1 and 2: there the sufficient statistic s reduced the data to d_η dimensions, whereas here no reduction takes place.

Writing $z = (y - m)/\sigma$ for the vector of standardized residuals, the score for the nuisance parameter is

$$w_1 = \nabla_\eta \ln f = \frac{1}{\sigma} \nabla_\eta m(x, \beta, \eta) z,$$

and the higher-order elements of the Bhattacharyya basis follow from the chain rule. Indeed, the components of z being independent standard normal, the derivatives of the Gaussian location family with respect to m deliver products of Hermite polynomials: for a multi-index $j = (j_1, \dots, j_n)$,

$$\frac{\nabla_m^j f(y \mid x; \theta, \eta)}{f(y \mid x; \theta, \eta)} = \frac{1}{\sigma^{|j|}} \prod_{i=1}^n H_{j_i}(z_i), \quad |j| = \sum_{i=1}^n j_i,$$

where H_k denotes the (probabilists') Hermite polynomial of degree k and ∇_m^j the corresponding mixed partial derivative with respect to m . Composing with $\eta \mapsto m(x, \beta, \eta)$ by the multivariate Faà di Bruno formula (Constantine and Savits, 1996) then expresses w_j as a combination of such products of total degree at most j , with coefficients built from the derivatives of m up to order j . When $n = 1$ and η is scalar, for instance,

$$w_2 = \frac{\nabla_\eta^2 m}{\sigma} H_1(z) + \frac{(\nabla_\eta m)^2}{\sigma^2} H_2(z).$$

The projection (5.11) is available in closed form despite the absence of a sufficient statistic. The reason is the orthogonality of the Hermite polynomials under the

⁴In our context with covariates and nuisance parameters, a member of the curved exponential family takes the form

$$f(y \mid x; \theta, \eta) = a_\theta(y, x) b_\theta(x, \eta) \exp(s(y, x)^\top \tau_\theta(x, \eta)),$$

where $s(y, x)$ is the sufficient statistic and $\tau_\theta(x, \eta)$ the natural parameter.

Gaussian reference measure,

$$\mathbb{E}_{x;\theta,\eta} \left[\prod_{i=1}^n H_{j_i}(z_i) \prod_{i=1}^n H_{k_i}(z_i) \right] = \mathbf{1}\{j = k\} \prod_{i=1}^n j_i!,$$

which delivers the Gram matrix of W_q and the cross moments with u as explicit functions of the derivatives of m . [Bonhomme, Jochmans, and Weidner \(2024\)](#) implement this approach to estimate the model of team production of [Ahmadpoor and Jones \(2019\)](#) using a sample of researchers in economics.

5.4 The large- q limit: connection to full orthogonality

Full orthogonality implies orthogonality to every order $q \geq 1$ (by differentiating $\mathbb{E}_{x;\theta,\eta}[u^*(y, x, \theta, \eta')] = 0$ in η' under the integral sign). We now show that the converse is also true, provided that an analyticity condition holds.⁵ In this sense, order- q orthogonality interpolates continuously between Neyman and full orthogonality, with the limit recovering the latter under analyticity.

To develop the argument, suppose $\mathcal{H}$ is open and connected, $u^*(y, x, \theta, \eta)$ is orthogonal to every order $q \geq 1$, and the function

$$G_{x;\theta,\eta} : \eta' \mapsto \mathbb{E}_{x;\theta,\eta}[u^*(y, x, \theta, \eta')]$$

is real-analytic on $\mathcal{H}$, for all θ, η , almost every x . Then, the entire Taylor expansion of $G_{x;\theta,\eta}$ around η has vanishing coefficients, so $G_{x;\theta,\eta}$ is identically zero on $\mathcal{H}$. This implies that, for all θ, η , and almost every x ,

$$G_{x;\theta,\eta}(\eta') = \int u^*(y, x, \theta, \eta') f(y \mid x; \theta, \eta) dy = 0 \quad \text{for all } \eta' \in \mathcal{H}.$$

It follows that u^* is fully orthogonal.

Example 1 (continued). In the panel logit model, because s takes only the $T + 1$ values $\{0, 1, \dots, T\}$, polynomials of degree at most T already exhaust all functions of s . Since w_j has degree j in s , the functions $w_1, \dots, w_T$ are linearly independent while w_{T+1} is a combination of them. The construction thus terminates at $q = T$: the

⁵[Small and McLeish \(2011\)](#) similarly invoke analyticity to show that the Bhattacharyya basis spans the set of conditional-mean zero functions of the sufficient statistic in exponential family models.

Gram matrix of W_q is singular for $q > T$, and at $q = T$ the projection space contains all mean-zero functions of s , so that, by (4.9),

$$u_T^* = u - \mathbb{E}_{x;\theta,\eta}[u \mid s] = u_\infty^*.$$

When $u = \nabla_\theta \ln f$, the factorization $\ln f(y \mid x; \theta, \eta) = \ln f(y \mid s, x; \theta) + \ln f(s \mid x; \theta, \eta)$ gives $\mathbb{E}_{x;\theta,\eta}[u \mid s] = \nabla_\theta \ln f(s \mid x; \theta, \eta)$, so that u_T^* is the conditional-logit score. The sequence $u_1^*, \dots, u_T^*$ thus interpolates between the variance-weighted demeaning of (3.12) and the fully-orthogonal moment u_∞^* .

This property is consistent with the Monte Carlo exercise reported in Figure 4. In that setting, we used $T_{\text{est}} = 3$ periods for estimation after having used two periods for sample splitting. In the simulations, u_3^* and the conditional-logit score u_∞^* coincide up to the tolerance of the root-finding algorithm.

The same argument applies to the general form (2.7). There $s = x_2^\top y$ is multivariate and the basis consists of the centered monomials in s of degree j . Since s has finite support, polynomials of sufficiently high degree exhaust the functions on that support (a bound on the required degree is the number of support points of s minus 1). However, when the map $y \mapsto x_2^\top y$ is injective, the limiting moment is $u_\infty^* = 0$.

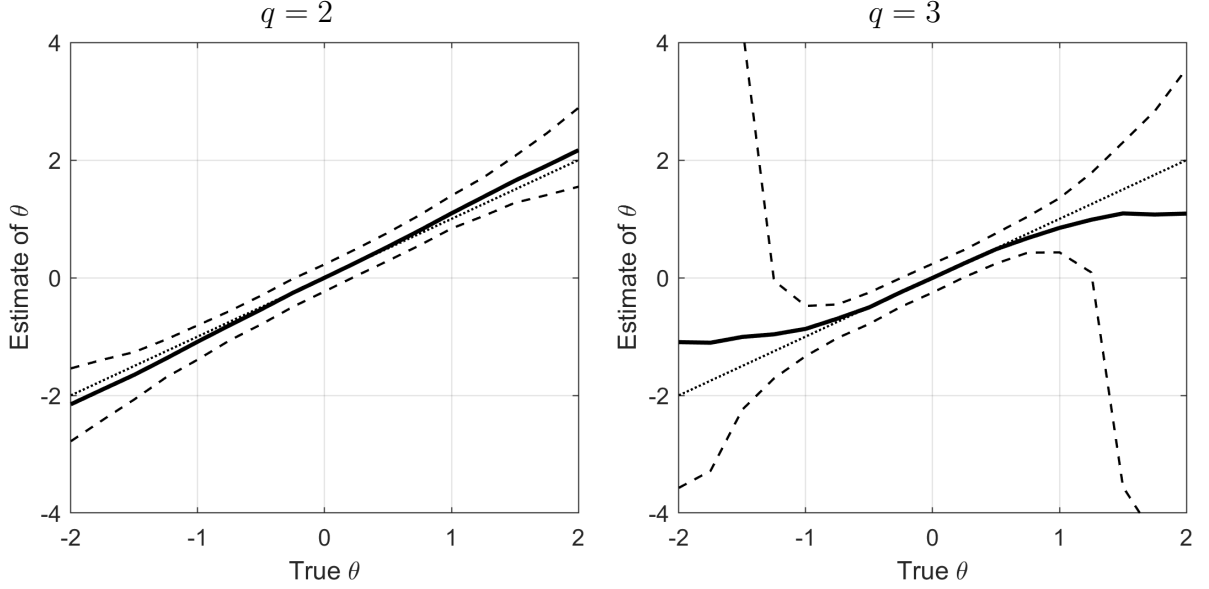
We illustrate this last point with simulations based on a panel logit model with two scalar covariates and a heterogeneous coefficient,

$$y_{it} = \mathbf{1}\{\theta_0 x_{it} + \eta_{i0} z_{it} + \varepsilon_{it} \geq 0\}, \quad \varepsilon_{it} \mid x_{i1}, \dots, x_{iT}, z_{i1}, \dots, z_{iT} \sim \text{i.i.d. Logistic}, \quad (5.14)$$

where x_{it} is binary and z_{it} is continuously distributed.

Figure 5 shows two orthogonal estimators, to order $q = 2$ (left) and $q = 3$ (right). The $q = 2$ estimator is well-behaved and shows little bias. In contrast, the $q = 3$ estimator exhibits bias and much wider confidence bands. In fact, the estimator fails to exist in approximately a quarter of the simulations. This contrasts sharply with Figure 4, in the standard panel logit case, where the $q = 3$ estimator was numerically identical to the fully-orthogonal conditional logit estimator. The model with a heterogeneous coefficient of a continuous covariate is a setting where the map $y_i \mapsto \sum_t z_{it} y_{it}$ is injective, and there is thus no non-trivial fully-orthogonal moment in this model. The simulations show that, in such cases, increasing the orthogonality order can have a high cost in terms of variability.

Figure 5: Orthogonal estimators in a logit model with a heterogeneous coefficient



Notes: see notes to Figure 3. The results are computed based on the instances where the estimator exists (100% of cases for $q = 2$, approximately 75% of cases for $q = 3$).

6 Other target parameters

In many applications the researcher is interested in a parameter $\mu_0 \in \mathcal{M}$ that satisfies the moment condition

$$\mathbb{E}[u(y, x, \theta_0, \eta_0, \mu_0)] = 0, \quad (6.1)$$

and thus is a functional of θ_0 , η_0 , and the distribution of x . Leading examples are average partial effects, average elasticities, and counterfactual policy effects in structural models. The three notions of orthogonality can be adapted to handle these quantities.

6.1 Orthogonality

The key difference with the case discussed in the earlier sections is that u is not unbiased at all parameter values, that is,

$$\mathbb{E}_{x;\theta,\eta}[u(y, x, \theta, \eta, \mu)]$$

is not identically zero: only its expectation over x is zero when evaluated at the true values $\theta = \theta_0$, $\eta = \eta_0$, $\mu = \mu_0$.

For all three notions, we replace the unbiasedness requirement on the orthogonal moment by

$$\begin{aligned} \mathbb{E}_{x;\theta,\eta} [u^*(y, x, \theta, \eta, \mu)] &= \mathbb{E}_{x;\theta,\eta} [u(y, x, \theta, \eta, \mu)] \\ \text{for all } \theta \in \Theta, \eta \in \mathcal{H}, \mu \in \mathcal{M}, \text{ almost every } x, \end{aligned} \quad (6.2)$$

and we will require (6.2) for all $u^* \in \{u_1^*, u_q^*, u_\infty^*\}$.

With this change, Neyman orthogonality is as before, except for the presence of μ and for the replacement of unbiasedness by (6.2). That is, we say that u_1^* is Neyman orthogonal if

$$\mathbb{E}_{x;\theta,\eta} [\nabla_\eta u_1^*(y, x, \theta, \eta, \mu)] = 0 \text{ for all } \theta \in \Theta, \eta \in \mathcal{H}, \mu \in \mathcal{M}, \text{ almost every } x, \quad (6.3)$$

and $u^* = u_1^*$ satisfies (6.2).

Likewise, for all $q \geq 1$, we say that u_q^* is orthogonal to order q if

$$\begin{aligned} \mathbb{E}_{x;\theta,\eta} [\nabla_\eta^j u_q^*(y, x, \theta, \eta, \mu)] &= 0 \\ \text{for all } j = 1, \dots, q, \theta \in \Theta, \eta \in \mathcal{H}, \mu \in \mathcal{M}, \text{ almost every } x, \end{aligned} \quad (6.4)$$

and $u^* = u_q^*$ satisfies (6.2).

Finally, we say that u_∞^* is fully orthogonal if

$$\begin{aligned} \mathbb{E}_{x;\theta,\eta} [u_\infty^*(y, x, \theta, \eta', \mu)] &\text{ does not depend on } \eta' \in \mathcal{H} \\ \text{for all } \theta \in \Theta, \eta \in \mathcal{H}, \mu \in \mathcal{M}, \text{ almost every } x, \end{aligned} \quad (6.5)$$

and $u^* = u_\infty^*$ satisfies (6.2).

Under full orthogonality, plugging in any arbitrary value for η' , using (6.5) and (6.2), and noting that u is unbiased at the truth, we obtain

$$\begin{aligned} \mathbb{E} [u_\infty^*(y, x, \theta_0, \eta', \mu_0)] &= \mathbb{E} [\mathbb{E}_{x;\theta_0,\eta_0} [u_\infty^*(y, x, \theta_0, \eta', \mu_0)]] \\ &= \mathbb{E} [\mathbb{E}_{x;\theta_0,\eta_0} [u_\infty^*(y, x, \theta_0, \eta_0, \mu_0)]] \\ &= \mathbb{E} [\mathbb{E}_{x;\theta_0,\eta_0} [u(y, x, \theta_0, \eta_0, \mu_0)]] = 0, \end{aligned}$$

where the outer expectations are taken with respect to the distribution of x . Given a consistent estimate of θ_0 , μ_0 can then be consistently estimated using standard GMM.

6.2 The Neyman projection

The orthogonal constructions take a different form in the case of the target parameter μ_0 . Starting with Neyman orthogonality, note that differentiating (6.2) in η for $u^* = u_1^*$ implies

$$\begin{aligned} \mathbb{E}_{x;\theta,\eta} [\nabla_\eta u_1^*(y, x, \theta, \eta, \mu)] + \mathbb{E}_{x;\theta,\eta} [w_1(y, x, \theta, \eta) u_1^*(y, x, \theta, \eta, \mu)^\top] \\ = \nabla_\eta \mathbb{E}_{x;\theta,\eta} [u(y, x, \theta, \eta, \mu)], \end{aligned}$$

where w_1 is the score with respect to η , and recall that, with our definition, the matrix $\nabla_\eta \mathbb{E}_{x;\theta,\eta} [u(y, x, \theta, \eta, \mu)]$ is $d_\eta \times d_u$.

Hence, Neyman orthogonality is equivalent to (6.2) at $u^* = u_1^*$, and

$$\begin{aligned} \mathbb{E}_{x;\theta,\eta} [w_1(y, x, \theta, \eta) u_1^*(y, x, \theta, \eta, \mu)^\top] = \nabla_\eta \mathbb{E}_{x;\theta,\eta} [u(y, x, \theta, \eta, \mu)] \\ \text{for all } \theta \in \Theta, \eta \in \mathcal{H}, \mu \in \mathcal{M}, \text{ almost every } x. \end{aligned} \quad (6.6)$$

Conditions (6.2) and (6.6) define a set of affine restrictions on u_1^* . Denote this set by $\mathcal{T}_{x;\theta,\eta,\mu}^1$. Given an original moment function u , we construct a Neyman-orthogonal u_1^* as

$$u_1^*(\cdot, x, \theta, \eta, \mu) = \text{Proj}_{x;\theta,\eta} [u(\cdot, x, \theta, \eta, \mu) \parallel \mathcal{T}_{x;\theta,\eta,\mu}^1], \quad (6.7)$$

where $\text{Proj}_{x;\theta,\eta} [v \parallel \mathcal{S}]$ is given by (3.6).

Since η is finite-dimensional, the projection is finite-dimensional as well, and u_1^* in (6.7) takes the explicit form

$$u_1^*(y, x, \theta, \eta, \mu) = u(y, x, \theta, \eta, \mu) - \Lambda_1(x; \theta, \eta, \mu) w_1(y, x, \theta, \eta), \quad (6.8)$$

with projection coefficient

$$\begin{aligned} \Lambda_1(x; \theta, \eta, \mu) = \left(\mathbb{E}_{x;\theta,\eta} [u(y, x, \theta, \eta, \mu) w_1(y, x, \theta, \eta)^\top] - b_1(x; \theta, \eta, \mu)^\top \right) \\ \times \mathbb{E}_{x;\theta,\eta} [w_1(y, x, \theta, \eta) w_1(y, x, \theta, \eta)^\top]^{-1}, \end{aligned}$$

where

$$b_1(x; \theta, \eta, \mu) = \nabla_\eta \mathbb{E}_{x; \theta, \eta} [u(y, x, \theta, \eta, \mu)]$$

is the sensitivity of the expected moment to the nuisance parameter. Indeed, u_1^* in (6.8) satisfies (6.2), since w_1 has mean zero, and it satisfies (6.6) by construction of Λ_1 . Comparing with (3.8), the only difference is the subtraction of b_1 from the covariance between u and the nuisance score.

6.3 The order- q projection

The same argument applies at higher orders. Proceeding as in Section 5.2, but starting from (6.2) in place of unbiasedness at all parameter values, orthogonality to order q is equivalent to (6.2) at $u^* = u_q^*$ and

$$\begin{aligned} \mathbb{E}_{x; \theta, \eta} [w_j(y, x, \theta, \eta) u_q^*(y, x, \theta, \eta, \mu)^\top] &= \nabla_\eta^j \mathbb{E}_{x; \theta, \eta} [u(y, x, \theta, \eta, \mu)] \\ \text{for all } j = 1, \dots, q, \theta \in \Theta, \eta \in \mathcal{H}, \mu \in \mathcal{M}, \text{ almost every } x, \end{aligned} \quad (6.9)$$

where the right-hand side is a total derivative, with η entering both the expectation and u .

Conditions (6.2) and (6.9) again define a set of affine restrictions on u_q^* , which we denote by $\mathcal{T}_{x; \theta, \eta, \mu}^q$, and we construct

$$u_q^*(\cdot, x, \theta, \eta, \mu) = \text{Proj}_{x; \theta, \eta} [u(\cdot, x, \theta, \eta, \mu) \parallel \mathcal{T}_{x; \theta, \eta, \mu}^q]. \quad (6.10)$$

Stacking $W_q = (w_1^\top, \dots, w_q^\top)^\top$ and $b_q = (b_1^\top, \dots, b_q^\top)^\top$, where

$$b_j(x; \theta, \eta, \mu) = \nabla_\eta^j \mathbb{E}_{x; \theta, \eta} [u(y, x, \theta, \eta, \mu)],$$

the projection (6.10) takes the explicit form

$$u_q^*(y, x, \theta, \eta, \mu) = u(y, x, \theta, \eta, \mu) - \Lambda_q(x; \theta, \eta, \mu) W_q(y, x, \theta, \eta), \quad (6.11)$$

with projection coefficient

$$\begin{aligned}\Lambda_q(x; \theta, \eta, \mu) &= \left(\mathbb{E}_{x; \theta, \eta} \left[u(y, x, \theta, \eta, \mu) W_q(y, x, \theta, \eta)^\top \right] - b_q(x; \theta, \eta, \mu)^\top \right) \\ &\quad \times \mathbb{E}_{x; \theta, \eta} \left[W_q(y, x, \theta, \eta) W_q(y, x, \theta, \eta)^\top \right]^{-1}.\end{aligned}$$

Similarly to the case $q = 1$, (6.2) holds at $u^* = u_q^*$ because the w_j have mean zero, and (6.9) holds by construction of Λ_q . Comparing with (5.12), the construction differs only through the subtraction of b_q . Expression (6.11) is Theorem 1 in [Bonhomme, Jochmans, and Weidner \(2024\)](#).

6.4 The fully-orthogonal projection

Since (6.5) holds at every value of η' , we may fix it at any convenient value and simply write $u_\infty^*(y, x, \theta, \mu)$, as in Section 4.1. The condition then reads, for all θ , η , and x ,

$$\mathbb{E}_{x; \theta, \eta} [u_\infty^*(y, x, \theta, \mu)] = b_0(x; \theta, \eta, \mu)^\top \quad \text{for all } \eta \in \mathcal{H},$$

where $b_0(x; \theta, \eta, \mu) = \mathbb{E}_{x; \theta, \eta} [u(y, x, \theta, \eta, \mu)^\top]$. Fixing a reference value $\eta \in \mathcal{H}$ and applying the change of measure of Section 4.2, we obtain that (6.5) is equivalent to

$$\begin{aligned}\mathbb{E}_{x; \theta, \eta} [u_\infty^*(y, x, \theta, \mu) r_{\tilde{\eta}}(y, x, \theta, \eta)] &= b_0(x; \theta, \tilde{\eta}, \mu)^\top \\ \text{for all } \theta \in \Theta, \eta \in \mathcal{H}, \tilde{\eta} \in \mathcal{H}, \mu \in \mathcal{M}, \text{ almost every } x.\end{aligned}\tag{6.12}$$

Condition (6.12) defines an affine subset of $L^2(f(\cdot \mid x; \theta, \eta))$, which we denote by $\mathcal{T}_{x; \theta, \eta, \mu}^\infty$. As before, we construct

$$u_\infty^*(\cdot, x, \theta, \eta, \mu) = \text{Proj}_{x; \theta, \eta} [u(\cdot, x, \theta, \eta, \mu) \parallel \mathcal{T}_{x; \theta, \eta, \mu}^\infty], \tag{6.13}$$

the closest fully-orthogonal moment to u in $L^2(f(\cdot \mid x; \theta, \eta))$. As in Section 4.2, the feasible set consists of functions that do not involve η , whereas the projected moment does. When u is unbiased at all parameter values we have $b_0 = 0$, and (6.13) coincides with the residual construction (4.6) of Section 4.2.

Two differences with the other projections are worth noting. First, the fully-orthogonal projection is again infinite-dimensional; Appendix B gives the corresponding operator equation. Second, $\mathcal{T}_{x; \theta, \eta, \mu}^\infty$ may be empty, in which case no fully-orthogonal

moment exists for the target μ_0 , as we illustrate in Example 1 next.

6.5 Examples

Example 1 (continued). Consider the panel logit model (2.8) with a scalar continuous covariate, suppress the unit index as before, and consider the average partial effect of x_t

$$\mu_0 = \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \nabla_x \Pr(y_t = 1 \mid x_t = x) \right] = \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \theta_0 \pi_t (1 - \pi_t) \right]. \quad (6.14)$$

A unit's contribution to the moment condition is

$$u(y, x, \theta, \eta, \mu) = \frac{\theta}{T} \sum_{t=1}^T v_t(\theta, \eta) - \mu, \quad v_t = \pi_t(1 - \pi_t).$$

This moment does not depend on the outcome, so that $\mathbb{E}_{x;\theta,\eta}[u w_j^\top] = 0$ for every j and the construction only depends on the sensitivities b_j . Those have a compact form. Recall from Section 5.3 that $\nabla_\eta^j \kappa$ is the j th cumulant κ_j of the sufficient statistic $s = \sum_t y_t$, and note that $\sum_t v_t = \kappa_2$. Hence

$$b_0 = \frac{\theta}{T} \kappa_2 - \mu, \quad b_j = \frac{\theta}{T} \kappa_{j+2}, \quad j \geq 1.$$

The orthogonal moments follow. With $\mathbb{E}_{x;\theta,\eta}[w_1^2] = \kappa_2$, the Neyman construction (6.8) gives

$$u_1^* = u + \frac{\theta}{T} \frac{\kappa_3}{\kappa_2} w_1, \quad w_1 = s - \kappa_1,$$

and, using $w_2 = (s - \kappa_1)^2 - \kappa_2$ together with $\mathbb{E}[w_1 w_2] = \kappa_3$ and $\mathbb{E}[w_2^2] = \kappa_4 + 2\kappa_2^2$, the order-2 construction (6.11) gives

$$u_2^* = u + \frac{\theta}{T} (\kappa_3, \kappa_4) \begin{pmatrix} \kappa_2 & \kappa_3 \\ \kappa_3 & \kappa_4 + 2\kappa_2^2 \end{pmatrix}^{-1} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}.$$

Fully-orthogonal moments, in contrast, do not exist here. Fix x and θ , and write a candidate as $u_\infty^* = g(s, x, \theta, \mu) + h$, with $\mathbb{E}_{x;\theta,\eta}[h \mid s] = 0$. By sufficiency, the conditional distribution of y given s does not depend on η , so h contributes nothing

at any value of the fixed effect and

$$\mathbb{E}_{x;\theta,\eta} [u_{\infty}^*(y, x, \theta, \mu)] = \sum_{k=0}^T g(k, x, \theta, \mu) \Pr(s = k \mid x; \theta, \eta).$$

Full orthogonality thus requires the $T + 1$ numbers $g(0, x, \theta, \mu)$, ..., $g(T, x, \theta, \mu)$ to satisfy

$$\sum_{k=0}^T g(k, x, \theta, \mu) \Pr(s = k \mid x; \theta, \eta) = b_0(x; \theta, \eta, \mu), \quad \text{for all } \eta \in \mathcal{H},$$

which is a finite-dimensional system facing a continuum of restrictions. For $T = 2$, clearing denominators and matching powers of e^η leaves the restriction

$$\frac{\theta}{2} (e^{x_1\theta} - e^{x_2\theta})^2 e^{2\eta} = 0,$$

which holds only when $x_1 = x_2$ or $\theta = 0$.

Example 3 (continued). As a special case of model (2.11), consider the linear model

$$y = x\eta_0 + \sigma_0\varepsilon, \quad \varepsilon \mid x \sim \mathcal{N}(0, I_n). \quad (6.15)$$

A leading application is the two-way model of [Abowd, Kramarz, and Margolis \(1999\)](#), in which y collects log wages, x encodes worker and firm identities, and η_0 the corresponding effects. Besides σ_0^2 , interest centers on variance components: quadratic forms $\mu_0 = \eta_0^\top Q \eta_0$ for a symmetric Q , measuring the dispersion of worker and firm effects and their covariance (see [Bonhomme, Lamadon, and Manresa, 2026](#) for a survey). Here we take σ^2 as known, and start with the original moment function $u(y, x, \theta, \eta, \mu) = \eta^\top Q \eta - \mu$. We provide detailed derivations in [Appendix D](#).

Since u does not depend on the outcome, the construction is governed by the sensitivities alone. Here the construction terminates at $q = 2$, since the w_j are uncorrelated and

$$b_0 = \eta^\top Q \eta - \mu, \quad b_1 = 2Q\eta, \quad b_2 = 2 \text{vec}(Q), \quad b_j = 0 \text{ for } j \geq 3.$$

With $w_1 = \sigma^{-2}x^\top(y - x\eta)$ and $\mathbb{E}_{x;\theta,\eta}[w_1 w_1^\top] = \sigma^{-2}x^\top x$, the Neyman construction

gives

$$u_1^* = \eta^\top Q\eta - \mu + 2\eta^\top Q(x^\top x)^{-1}x^\top(y - x\eta), \quad (6.16)$$

which still involves η . This moment is moreover of no practical use when $\hat{\eta}$ is the least-squares estimator computed on the same sample: then $x^\top(y - x\hat{\eta}) = 0$, the correction vanishes, and u_1^* reduces to the uncorrected plug-in $\hat{\eta}^\top Q\hat{\eta}$.

Imposing second-order orthogonality changes the picture. Adding

$$w_2 = \text{vec} \left(-\sigma^{-2}x^\top x + \sigma^{-4}x^\top(y - x\eta)(y - x\eta)^\top x \right)$$

to the basis, the order-two moment is

$$u_2^* = y^\top x(x^\top x)^{-1}Q(x^\top x)^{-1}x^\top y - \sigma^2 \text{tr} \left(Q(x^\top x)^{-1} \right) - \mu, \quad (6.17)$$

which no longer depends on η and is unbiased at every value of it. Moreover, u_2^* is fully orthogonal. Note the contrast with Example 1: while no fully-orthogonal moment exists for the average effect in the panel logit model with a continuous covariate, in model (6.15) quadratic forms in η admit fully-orthogonal moments. We recognize the trace-corrected variance-component expression of Andrews, Gill, Schank, and Upward (2008), generalized by Kline, Saggio, and Sølvesten (2020) to the non-Gaussian, heteroskedastic case.⁶

A second special case is the nonlinear model

$$y = x_1\eta_0 + \beta_0 x_2(\eta_0 \otimes \eta_0) + \sigma_0\varepsilon, \quad \varepsilon \mid x \sim \text{iid } \mathcal{N}(0, I_n), \quad (6.18)$$

where $x = (x_1, x_2)$, $\otimes$ denotes the Kronecker product, and β_0 is scalar. This model contains as a special case the Tukey generalization of Abowd, Kramarz, and Margolis (1999) studied in Crippa (2025), here for Gaussian errors, and reduces to (6.15) when the complementarity parameter β_0 equals zero.

Let us focus on the variance component $\mu_0 = \eta_0^\top Q\eta_0$. Since the sensitivities b_j

⁶A second-order projection also provides an unbiased estimator of σ_0^2 itself, which we have taken as known. Writing $M_x = I_n - x(x^\top x)^{-1}x^\top$ and taking $u = \nabla_{\sigma^2} \ln f = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(y - x\eta)^\top(y - x\eta)$, the moment

$$u_2^* = \frac{1}{2\sigma^4} (y^\top M_x y - (n - d_\eta) \sigma^2)$$

has mean zero at every $\eta \in \mathcal{H}$, since $M_x y = \sigma M_x \varepsilon$ does not involve η and $\mathbb{E}[\varepsilon^\top M_x \varepsilon] = \text{tr}(M_x) = n - d_\eta$. It is thus fully orthogonal. Solving $u_2^* = 0$ gives the degrees-of-freedom corrected variance estimator $y^\top M_x y / (n - d_\eta)$. See Appendix D.

depend on the target alone, they are as above. However, the Bhattacharyya basis changes relative to the linear model. Writing $m(\eta) = x_1\eta + \beta x_2(\eta \otimes \eta)$ and $z = (y - m(\eta))/\sigma$, the Jacobian

$$M(\eta) = (\nabla_\eta m(\eta))^\top = x_1 + \beta x_2 (I_{d_\eta} \otimes \eta + \eta \otimes I_{d_\eta})$$

now depends on η , and $\nabla_\eta^2 m$ is a nonzero constant proportional to β . Following Example 3 in Subsection 5.3, $w_1 = \sigma^{-1} M(\eta)^\top z$ and w_2 is the corresponding combination of first- and second-order Hermite polynomials in z , with coefficients built from $M(\eta)$ and $\nabla_\eta^2 m$. The Gram matrix of W_2 and the cross moments follow in closed form from the orthogonality of the Hermite polynomials, so that the order-two moment u_2^* in (6.11) is available in closed form, and the same holds at higher orders. Unlike (6.17), however, u_2^* still depends on η through $M(\eta)$, so that estimation requires a first-step estimate and sample splitting. Crippa (2025) constructs fully-orthogonal moments for β_0 in the special case of the Tukey model. Whether informative fully-orthogonal moments exist for quadratic forms μ_0 in (6.18) is left for future work.

7 Conclusion

We have described three constructions that make a moment function insensitive to the presence of nuisance parameters in a nonlinear model. The first is Neyman orthogonality, the second is orthogonality to order q , and the third is full orthogonality. All three apply the same principle. In each case the moment function is projected on a subspace of functions of the data determined by the likelihood, and the residual is retained. The three differ only in the choice of subspace, which is spanned by the nuisance score, by the first q generalized scores, and by all the likelihood ratios, respectively. These subspaces are nested and, under an analyticity condition, orthogonality to every order implies full orthogonality, so that the three notions form a single increasing sequence of projections. We have also shown that the projection delivering full orthogonality is Neyman's own construction, applied to a model in which the distribution of the nuisance parameter is itself the nuisance parameter. Finally, we have shown that the projection approach extends naturally to target parameters that are functionals of the nuisance parameters and of the distribution of the covariates.

Our examples illustrate when each notion is available. In the panel Poisson model,

Neyman orthogonality already delivers a fully-orthogonal moment. In the panel logit, taking $q = T$ returns the conditional-likelihood score. For average effects of a continuous covariate in that model no fully-orthogonal moment exists, and orthogonality to a finite order is the only route available among these constructions. Choosing the order q in practice remains an open question.

Developing methods to construct orthogonal moments outside likelihood models is an important direction for future research. Neyman orthogonality extends naturally to this case, as shown by an active recent literature on double/debiased machine learning. Recent research has extended higher-order orthogonality to moment condition models without a likelihood structure ([Chetverikov, Sørensen, and Tsyvinski, 2026](#); [Bonhomme, Jochmans, Newey, and Weidner, 2026](#)). More work in this direction is needed to better understand the properties of estimators based on orthogonal moments in settings such as panels, networks, and non- and semiparametric estimation problems where nuisance parameters are high-dimensional.

References

- ABOWD, J. M., F. KRAMARZ, AND D. N. MARGOLIS (1999): “High wage workers and high wage firms,” *Econometrica*, 67, 251–333.
- AHMADPOOR, M., AND B. F. JONES (2019): “Decoding team and individual impact in science and invention,” *Proceedings of the National Academy of Sciences*, 116, 13885–13890.
- ANDERSEN, E. B. (1970): “Asymptotic properties of conditional maximum-likelihood estimators,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 32(2), 283–301.
- ANDREWS, M. J., L. GILL, T. SCHANK, AND R. UPWARD (2008): “High wage workers and low wage firms: negative assortative matching or limited mobility bias?,” *Journal of the Royal Statistical Society: Series A*, 171, 673–697.
- ARELLANO, M. (2003a): “Discrete choices with panel data,” *Investigaciones Económicas*, XXVII, 423–458.
- ARELLANO, M. (2003b): *Panel data econometrics*. OUP Oxford.
- ARELLANO, M., AND B. HONORÉ (2001): “Panel data models: some recent developments,” in *Handbook of econometrics*, vol. 5, pp. 3229–3296. Elsevier.
- BHATTACHARYYA, A. (1946): “On some analogues of the amount of information and their use in statistical estimation,” *Sankhyā*, 8, 1–14.
- BLUNDELL, R., R. GRIFFITH, AND F. WINDMEIJER (2002): “Individual effects and dynamics in count data models,” *Journal of econometrics*, 108(1), 113–131.
- BONHOMME, S. (2012): “Functional differencing,” *Econometrica*, 80(4), 1337–1385.
- BONHOMME, S., AND K. DANO (2024): “Functional Differencing in Networks,” *Revue économique*, 75(1), 147–175.
- BONHOMME, S., K. DANO, AND B. S. GRAHAM (2023): “Identification in a binary choice panel data model with a predetermined covariate,” *SERIEs*, 14(3), 315–351.

- (2025): “Moment restrictions for nonlinear panel data models with feedback,” Discussion paper, National Bureau of Economic Research.
- BONHOMME, S., K. JOCHMANS, W. K. NEWHEY, AND M. WEIDNER (2026): “Higher-Order Neyman Orthogonality in Moment-Condition Models,” *arXiv preprint arXiv:2605.10842*.
- BONHOMME, S., K. JOCHMANS, AND M. WEIDNER (2024): “A neyman-orthogonalization approach to the incidental parameter problem,” *arXiv preprint arXiv:2412.10304*.
- BONHOMME, S., T. LAMADON, AND E. MANRESA (2026): “A Users’ Guide to Uncovering Worker and Firm Effects: The ABC of AKM,” *Journal of Economic Perspectives*, 40(2), 195–214.
- CHAMBERLAIN, G. (1980): “Analysis of covariance with qualitative data,” *The review of economic studies*, 47(1), 225–238.
- (1992): “Efficiency bounds for semiparametric regression,” *Econometrica: Journal of the Econometric Society*, pp. 567–596.
- (2010): “Binary response models for panel data: Identification and information,” *Econometrica*, 78(1), 159–168.
- CHATTERJEE, S., P. DIACONIS, AND A. SLY (2011): “Random graphs with a given degree sequence,” *Annals of Applied Probability*, 21, 1400–1435.
- CHERNOZHUKOV, V., D. CHETVERIKOV, M. DEMIRER, E. DUFLO, C. HANSEN, W. NEWHEY, AND J. ROBINS (2018): “Double/debiased machine learning for treatment and structural parameters,” *Econometrics Journal*, 21, C1–C68.
- CHETVERIKOV, D., J. R.-V. SØRENSEN, AND A. TSYVINSKI (2026): “Triple/Double-Debiased Lasso,” *arXiv preprint arXiv:2603.20134*.
- CONSTANTINE, G., AND T. SAVITS (1996): “A multivariate Faà di Bruno formula with applications,” *Transactions of the American Mathematical Society*, 348, 503–520.
- CRIPPA, F. (2025): “Identification, Estimation, and Inference in Two-Sided Interaction Models,” *arXiv preprint arXiv:2510.22884*.

- DANO, K. (2023): “Transition probabilities and moment restrictions in dynamic fixed effects logit models,” *arXiv preprint arXiv:2303.00083*.
- DHAENE, G., AND K. JOCHMANS (2015a): “Profile-score adjustments for incidental-parameter problems,” *Mimeo*.
- DHAENE, G., AND K. JOCHMANS (2015b): “Split-panel jackknife estimation of fixed-effect models,” *The Review of Economic Studies*, 82, 991–1030.
- EFRON, B. (1975): “Defining the curvature of a statistical problem (with applications to second order efficiency),” *The Annals of Statistics*, pp. 1189–1242.
- ENGL, H. W., M. HANKE, AND A. NEUBAUER (1996): *Regularization of inverse problems*, vol. 375. Springer Science & Business Media.
- GHAZAL, G. A., AND H. NEUDECKER (2000): “On second-order and fourth-order moments of jointly distributed random matrices: a survey,” *Linear Algebra and its Applications*, 321, 61–93.
- GRAHAM, B. S. (2017): “An econometric model of network formation with degree heterogeneity,” *Econometrica*, 85, 1033–1063.
- HAHN, J., AND W. K. NEWKEY (2004): “Jackknife and analytical bias reduction for nonlinear panel models,” *Econometrica*, 72, 1295–1319.
- HANSEN, L. P. (1982): “Large sample properties of generalized method of moments estimators,” *Econometrica: Journal of the econometric society*, pp. 1029–1054.
- HAUSMAN, J., B. H. HALL, AND Z. GRILICHES (1984): “Econometric Models for Count Data with an Application to the Patents-R&D Relationship,” *Econometrica*, 52(4), 909–938.
- HONORÉ, B. E., AND M. WEIDNER (2025): “Moment conditions for dynamic panel logit models with fixed effects,” *Review of Economic Studies*, 92(5), 3112–3137.
- JOCHMANS, K. (2018): “Semiparametric analysis of network formation,” *Journal of Business & Economic Statistics*, 36, 705–713.
- KLINE, P., R. SAGGIO, AND M. SØLVSTEN (2020): “Leave-out estimation of variance components,” *Econometrica*, 88, 1859–1898.

- LANCASTER, T. (2002): “Orthogonal parameters and panel data,” *Review of Economic Studies*, 69, 647–666.
- MACKEY, L., V. SYRGKANIS, AND I. ZADIK (2018): “Orthogonal machine learning: Power and limitations,” in *International Conference on Machine Learning*, pp. 3375–3383. PMLR.
- MAGNUS, J. R., AND H. NEUDECKER (1979): “The commutation matrix: some properties and applications,” *Annals of Statistics*, 7, 381–394.
- (1980): “The elimination matrix: some lemmas and applications,” *SIAM Journal on Algebraic Discrete Methods*, 1(4), 422–449.
- NEWBY, W. K. (1994): “The asymptotic variance of semiparametric estimators,” *Econometrica*, 62, 1349–1382.
- NEYMAN, J. (1959): “Optimal asymptotic tests of composite hypotheses,” in *Probability and Statistics*, ed. by U. Grenander, pp. 416–444. Wiley NY.
- RASCH, G. (1961): “On general laws and the meaning of measurement in psychology,” in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, vol. 4, pp. 321–333.
- ROBINS, J., L. LI, E. TCHETGEN TCHETGEN, AND A. VAN DER VAART (2008): “Higher order influence functions and minimax estimation of nonlinear functionals,” in *Probability and Statistics: Essays in Honor of David A. Freedman*, pp. 335–421. Institute of Mathematical Statistics.
- SMALL, C. G., AND D. L. MCLEISH (1989): “Projection as a method for increasing sensitivity and eliminating nuisance parameters,” *Biometrika*, 76, 693–703.
- SMALL, C. G., AND D. L. MCLEISH (2011): *Hilbert space methods in probability and statistical inference*. John Wiley & Sons.
- SUR, P., AND E. J. CANDÈS (2019): “A modern maximum-likelihood theory for high-dimensional logistic regression,” *Proceedings of the National Academy of Sciences*, 116(29), 14516–14525.
- VAN DER VAART, A. W. (2000): *Asymptotic statistics*, vol. 3. Cambridge university press.

WATERMAN, R. P., AND B. G. LINDSAY (1996): “Projected score methods for approximating conditional scores,” *Biometrika*, 83, 1–13.

APPENDIX

A Details for the panel data logit model

A.1 The second derivative of the expected Neyman-orthogonal moment

To show (3.14), note that $\partial_{\eta'} \pi_t = v_t$ and $\partial_{\eta'} v_t = v_t(1 - 2\pi_t)$. Applying the quotient rule to $\bar{x}(\eta') = \sum_t v_t x_t / \sum_t v_t$ and collecting terms gives

$$\partial_{\eta'} \bar{x}(\eta') = \frac{\sum_{t=1}^T v_t (1 - 2\pi_t) (x_t - \bar{x})}{\sum_{t=1}^T v_t}.$$

Differentiating (3.13) twice with respect to η' produces three terms. The first one carries $\pi_t(\eta) - \pi_t(\eta')$ as a factor and so vanishes at $\eta' = \eta$. The second one is $2 \partial_{\eta'} \bar{x}(\eta') \sum_t v_t$, which by the previous display equals twice the right-hand side of (3.14). The third one is $-\sum_t (x_t - \bar{x}) \partial_{\eta'} v_t$, which equals minus the right-hand side of (3.14). Adding the last two terms gives the result. In turn, (3.15) follows on substituting

$$x_1 - \bar{x} = \frac{v_2 (x_1 - x_2)}{v_1 + v_2}, \quad x_2 - \bar{x} = -\frac{v_1 (x_1 - x_2)}{v_1 + v_2}$$

into (3.14) and using $(1 - 2\pi_1)v_1v_2 - (1 - 2\pi_2)v_1v_2 = 2v_1v_2(\pi_2 - \pi_1)$.

A.2 Moments of the sufficient statistic

Carrying out the projection (5.13) in the panel logit model requires the moments of $s = \sum_{t=1}^T y_t$, which supply the coefficients of the w_j , together with the cross moments $\mathbb{E}_{x;\theta,\eta}[u s^j]$.

The cumulants of s are obtained by adding up the per-period Bernoulli cumulants, and coincide with the derivatives $\nabla_{\eta}^j \kappa$ of the cumulant generating function of Section 5.3. Writing $v_t = \pi_t(1 - \pi_t)$, the four lowest are

$$\kappa_1 = \sum_{t=1}^T \pi_t, \quad \kappa_2 = \sum_{t=1}^T v_t, \quad \kappa_3 = \sum_{t=1}^T v_t(1 - 2\pi_t), \quad \kappa_4 = \sum_{t=1}^T v_t(1 - 6v_t),$$

and the standard moment-cumulant relations turn these into the moments of s .

For the cross moments, decompose $s = y_t + s_{-t}$, where $s_{-t} = \sum_{r \neq t} y_r$ is independent of y_t , and use $y_t^k = y_t$ for $k \geq 1$ to obtain

$$\mathbb{E}_{x;\theta,\eta} \left[(y_t - \pi_t) s^j \right] = v_t \sum_{k=1}^j \binom{j}{k} \mathbb{E}_{x;\theta,\eta} \left[s_{-t}^{j-k} \right],$$

the moments of s_{-t} following from its cumulants in the same way. Summing over t against x_t delivers $\mathbb{E}_{x;\theta,\eta} [u s^j]$ for the score $u = \sum_t x_t (y_t - \pi_t)$.

B Construction of fully-orthogonal moments

B.1 The target parameter θ

Given a moment function u , and for x, θ, η fixed, the projection residual u_∞^* in (4.6) solves

$$u_\infty^*(\cdot, x, \theta, \eta) = \underset{u^*}{\operatorname{argmin}} \mathbb{E}_{x;\theta,\eta} \left[\|u(y, x, \theta, \eta) - u^*(y)\|^2 \right], \quad (\text{B.1})$$

$$\text{s.t. } \mathbb{E}_{x;\theta,\tilde{\eta}} [u^*(y)] = 0 \text{ for all } \tilde{\eta} \in \mathcal{H}. \quad (\text{B.2})$$

Components of u decouple in the objective and in the constraints alike, so nothing is lost by treating a scalar component. The constraints are associated to a Lagrange multiplier $\lambda(\tilde{\eta})$. We then define the Lagrangian

$$\mathcal{L}(u^*, \lambda) = \frac{1}{2} \mathbb{E}_{x;\theta,\eta} [(u - u^*)^2] - \int_{\mathcal{H}} \lambda(\tilde{\eta}) \mathbb{E}_{x;\theta,\eta} [u^* r_{\tilde{\eta}}] d\tilde{\eta}, \quad (\text{B.3})$$

each constraint having been expressed under the reference distribution using the change of measure underlying (4.4).

Taken pointwise in y , the first-order condition delivers

$$u_\infty^*(y, x, \theta, \eta) = u(y, x, \theta, \eta) + \int_{\mathcal{H}} \lambda(\tilde{\eta}) \frac{f(y \mid x; \theta, \tilde{\eta})}{f(y \mid x; \theta, \eta)} d\tilde{\eta}, \quad (\text{B.4})$$

so the program is solved by the original moment plus a correction drawn from $\mathcal{S}_{x;\theta,\eta}$.

Finally, imposing the constraints, the Lagrange multiplier solves the following

integral equation

$$\int_{\mathcal{H}} K(\tilde{\eta}', \tilde{\eta}) \lambda(\tilde{\eta}) d\tilde{\eta} = -\mathbb{E}_{x;\theta,\tilde{\eta}'} [u(y, x, \theta, \eta)], \quad K(\tilde{\eta}', \tilde{\eta}) = \mathbb{E}_{x;\theta,\eta} [r_{\tilde{\eta}'} r_{\tilde{\eta}}]. \quad (\text{B.5})$$

Remark 4 (Existence). Note that (B.5) is a Fredholm integral equation of the first kind. Its kernel K is positive semi-definite and, in general, compact, so that the inverse operator is unbounded and a solution λ need not exist (see, e.g., Engl, Hanke, and Neubauer, 1996). This does not affect the existence of u_{∞}^* itself. The feasible set in (B.2) is a closed affine subspace of $L^2(f(\cdot | x; \theta, \eta))$ and the objective in (B.1) is strictly convex, so the program has a unique solution, given by the projection residual (4.6). What may fail is only the representation (B.4) of the correction $u_{\infty}^* - u$ as an integral against a multiplier: the projection belongs to the closed span $\mathcal{S}_{x;\theta,\eta}$, whereas (B.4) requires it to lie in the span itself.

B.2 Other target parameters

We now turn to the targets of Section 6. The relevant program is

$$u_{\infty}^*(\cdot, x, \theta, \eta, \mu) = \underset{u^*}{\operatorname{argmin}} \mathbb{E}_{x;\theta,\eta} [(u(y, x, \theta, \eta, \mu) - u^*(y))^2], \quad (\text{B.6})$$

$$\text{s.t. } \mathbb{E}_{x;\theta,\tilde{\eta}} [u^*(y)] = b_0(x; \theta, \tilde{\eta}, \mu) \text{ for all } \tilde{\eta} \in \mathcal{H}, \quad (\text{B.7})$$

which reduces to (B.1)–(B.2) when $b_0 \equiv 0$.

The Lagrangian and the first-order condition are as before, so that the solution again takes the form

$$u_{\infty}^*(y, x, \theta, \eta, \mu) = u(y, x, \theta, \eta, \mu) + \int_{\mathcal{H}} \lambda(\tilde{\eta}) \frac{f(y | x; \theta, \tilde{\eta})}{f(y | x; \theta, \eta)} d\tilde{\eta}. \quad (\text{B.8})$$

Imposing the constraints, the multiplier now solves

$$\int_{\mathcal{H}} K(\tilde{\eta}', \tilde{\eta}) \lambda(\tilde{\eta}) d\tilde{\eta} = \mathbb{E}_{x;\theta,\tilde{\eta}'} [u(y, x, \theta, \tilde{\eta}', \mu) - u(y, x, \theta, \eta, \mu)], \quad (\text{B.9})$$

with K the Gram kernel of (B.5). The right-hand side measures how much the nuisance argument of the original moment matters, relative to the reference value η at which the projection is computed; it vanishes when u does not depend on η , and (B.9) reduces to (B.5) when $b_0 \equiv 0$.

The solution is

$$u_{\infty}^* = g_0 + \left(u - \text{Proj}_{x;\theta,\eta} (u \parallel \mathcal{S}_{x;\theta,\eta}) \right), \quad (\text{B.10})$$

where $g_0 \in \mathcal{S}_{x;\theta,\eta}$ is the unique element (when it exists) satisfying $\mathbb{E}_{x;\theta,\eta}[g_0 r_{\tilde{\eta}}] = b_0(x; \theta, \tilde{\eta}, \mu)$ for all $\tilde{\eta} \in \mathcal{H}$. The second term in (B.10) is the fully-orthogonal projection residual of Section 4.2.

Remark 5 (Existence, other targets). Unlike in (B.1)–(B.2), the feasible set may now be empty: the zero function satisfies (B.7) only when $b_0 \equiv 0$, and g_0 exists only if b_0 lies in the range of the map $g \mapsto (\tilde{\eta} \mapsto \mathbb{E}_{x;\theta,\eta}[g r_{\tilde{\eta}}])$. Existence of a fully-orthogonal moment for the target is thus a genuine restriction on the model, as Example 1 illustrates, and it is distinct from the question of whether the multiplier λ in (B.8) admits an integral representation.

C Monte Carlo simulations

This appendix collects the designs and estimators underlying the simulation evidence reported in the text.

Common design. In each experiment there are $N = 1000$ units and $T = 5$ periods. The covariate of interest is $x_{it} \sim \text{Bernoulli}(0.5)$ drawn independently across i and t , and the individual effects are drawn as $\eta_i \sim \mathcal{N}(0, 1)$, independently of the covariates. The parameter θ_0 moves across a grid from -2 to 2 in increments of 0.25 , with 1000 replications at each point. Sample splitting follows Section 3.1, without cross-fitting: each unit contributes its first $T_{\text{split}} = \lfloor T/2 \rfloor = 2$ periods to the estimation of η_i , and the estimating equation uses the $T_{\text{est}} = T - T_{\text{split}} = 3$ left-out periods. All estimating equations for θ are solved on $[-5, 5]$, and the first-step equations on $[-10, 10]$. An estimate is recorded as missing when its estimating equation has no sign change on that interval; means and percentiles are computed over the remaining replications.

Poisson. The data are generated from (2.10). First-step estimates are obtained by maximizing the split-sample likelihood jointly in $(\theta, \eta_1, \dots, \eta_N)$. Given θ , the individual effects have the closed form

$$\hat{\eta}_i(\theta) = \ln \left(\sum_{t \leq T_{\text{split}}} y_{it} \right) - \ln \left(\sum_{t \leq T_{\text{split}}} \exp(x_{it}\theta) \right),$$

and the split-sample estimate of θ solves the resulting concentrated score, which is the Poisson conditional-likelihood score. We keep $\hat{\eta}_i$ evaluated at the split-sample estimate of θ and discard the latter; the estimating equations are then solved holding $\hat{\eta}_i$ fixed. The closed form is finite provided the unit has a positive count in at least one split-sample period, and units with $\sum_{t \leq T_{\text{split}}} y_{it} = 0$ are dropped; between about 60% ($\theta_0 = -2$) and 93% ($\theta_0 = 2$) of units are retained. We report the plug-in estimator, solving

$$\sum_i \sum_{t > T_{\text{split}}} x_{it} (y_{it} - \mu_{it}(\theta, \hat{\eta}_i)) = 0,$$

and the Neyman-orthogonal estimator based on (3.9), with the weighted covariate mean computed over the left-out periods. By (3.10), the latter does not involve $\hat{\eta}_i$; it is computed on the same retained units.

Logit. The data are generated from (2.8), with the same covariate and effect distributions. Split-sample maximum likelihood again provides $\hat{\eta}_i$: for each θ , $\hat{\eta}_i(\theta)$ solves $\sum_{t \leq T_{\text{split}}} (y_{it} - \pi_{it}) = 0$ by Newton's method, and the split-sample estimate of θ solves the resulting profile score. As in the Poisson design, we keep $\hat{\eta}_i$ evaluated at the split-sample estimate of θ and discard the latter. For a unit whose split-sample outcomes are all zero or all one the effect is unbounded, and such units are dropped; roughly 40% of units are retained under this design. All estimators are computed on the retained units. Besides the plug-in estimator, we report the order- q orthogonal estimators for $q = 1, 2, 3$ and the conditional-logit estimator. For $q = 1$ the moment is (3.12). For $q \geq 2$ we project the estimation-sample score on the centered powers of $S_i = \sum_{t > T_{\text{split}}} (y_{it} - \pi_{it})$, computing the moments $\mathbb{E}[S_i^j]$ and $\mathbb{E}[u_i S_i^j]$ exactly by enumeration over the $2^{T_{\text{est}}}$ outcome configurations of a unit, and solving the resulting $q \times q$ system unit by unit. The conditional-logit score conditions on the choice count $\sum_{t > T_{\text{split}}} y_{it}$.

A heterogeneous coefficient. The third design is based on model (5.14), where the individual effect is multiplied by a continuously distributed regressor, and the choice probabilities are

$$\Pr(y_{it} = 1 \mid x_{it}, z_{it}, \eta_i) = \frac{\exp(x_{it}\theta_0 + z_{it}\eta_i)}{1 + \exp(x_{it}\theta_0 + z_{it}\eta_i)}, \quad (\text{C.1})$$

with $x_{it} \sim \text{Bernoulli}(0.5)$ and $z_{it} \sim \text{Uniform}[1/2, 3/2]$ drawn independently across i and t , $\eta_i \sim \mathcal{N}(0, 1)$, and no intercept. First-step estimation and the drop rule are as in the logit design, with the individual-effect equation $\sum_{t \leq T_{\text{split}}} z_{it}(y_{it} - \pi_{it}) = 0$. We report the order- q orthogonal estimators for $q = 2, 3$, obtained by projecting the estimation-sample score on the centered powers of $\sum_{t > T_{\text{split}}} z_{it}(y_{it} - \pi_{it})$, with the Gram and cross moments again computed exactly by enumeration over the $2^{T_{\text{est}}}$ outcome configurations.

D Orthogonal moments in the linear model

This appendix derives the expressions stated in Example 3 of Section 6. We work throughout with $y = x\eta + \sigma\varepsilon$, $\varepsilon \mid x \sim \text{iid } \mathcal{N}(0, I_n)$, taking $x^\top x$ nonsingular and σ^2 known unless stated otherwise; η has dimension $r = d_\eta$. Three matrices will prove useful for the derivations. The elimination matrix L_r satisfies $\text{vech}(C) = L_r \text{vec}(C)$; the commutation matrix K_r satisfies $K_r \text{vec}(C) = \text{vec}(C^\top)$; and for symmetric C the duplication matrix D_r satisfies $D_r \text{vech}(C) = \text{vec}(C)$. On these, see Magnus and Neudecker (1979, 1980).

D.1 Generalized scores and two lemmas

The log-likelihood is $\ell = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y - x\eta)^\top(y - x\eta)$, with derivatives $\nabla_\eta \ell = x^\top(y - x\eta)/\sigma^2$ and $\nabla_\eta^2 \ell = -x^\top x/\sigma^2$. The first- and second-order generalized scores are therefore

$$w_1 = \frac{1}{\sigma^2} x^\top(y - x\eta), \quad w_2 = \text{vech}(V), \quad V := -\frac{1}{\sigma^2} x^\top x + \frac{1}{\sigma^4} x^\top(y - x\eta)(y - x\eta)^\top x, \quad (\text{D.1})$$

where w_2 collects the distinct entries of the matrix $\nabla_\eta^2 \ell + \nabla_\eta \ell \nabla_\eta \ell^\top$, and we stack $W_2 = (w_1^\top, w_2^\top)^\top$. This differs from the convention used in the main text where w_2 collected all, possibly redundant, entries; the two conventions give the same projection, since the redundant entries are linear combinations of the distinct ones.

Lemma 1. $\mathbb{E}_{x;\theta,\eta} [w_1 w_1^\top] = \frac{1}{\sigma^2} x^\top x$, $\mathbb{E}_{x;\theta,\eta} [w_1 w_2^\top] = 0$, and

$$\mathbb{E}_{x;\theta,\eta} [w_2 w_2^\top] = \frac{1}{\sigma^4} L_r (x^\top \otimes x^\top) (I_{n^2} + K_n) (x \otimes x) L_r^\top. \quad (\text{D.2})$$

Proof of Lemma 1. Substituting $y - x\eta = \sigma\varepsilon$ gives $w_1 = x^\top \varepsilon / \sigma$ and $V = x^\top (\varepsilon \varepsilon^\top - I_n) x / \sigma^2$. The first claim is immediate from $\mathbb{E}[\varepsilon \varepsilon^\top] = I_n$, and the second holds because all third moments of a centered Gaussian vector vanish. For the third claim, $w_2 = \sigma^{-2} L_r (x^\top \otimes x^\top) \text{vec}(\varepsilon \varepsilon^\top - I_n)$, and the Gaussian fourth-moment identity, equation (4.3) of [Ghazal and Neudecker \(2000\)](#),

$$\mathbb{E} [\text{vec}(\varepsilon \varepsilon^\top) \text{vec}(\varepsilon \varepsilon^\top)^\top] = \text{vec}(I_n) \text{vec}(I_n)^\top + I_{n^2} + K_n,$$

implies $\mathbb{E}[\text{vec}(\varepsilon \varepsilon^\top - I_n) \text{vec}(\varepsilon \varepsilon^\top - I_n)^\top] = I_{n^2} + K_n$, which yields (D.2). $\square$

Lemma 2. For symmetric $r \times r$ matrices A and B ,

$$\text{vech}(A)^\top \left[L_r (x^\top \otimes x^\top) (I_{n^2} + K_n) (x \otimes x) L_r^\top \right]^{-1} \text{vech}(B) = \frac{1}{2} \text{tr} \left(A (x^\top x)^{-1} B (x^\top x)^{-1} \right). \quad (\text{D.3})$$

Proof of Lemma 2. We use the identities

$$D_r = (I_{r^2} + K_r) L_r^\top \left(L_r (I_{r^2} + K_r) L_r^\top \right)^{-1}, \quad (\text{D.4})$$

$$(I_{n^2} + K_n) (x \otimes x) = (x \otimes x) (I_{r^2} + K_r), \quad (\text{D.5})$$

$$D_r L_r (I_{r^2} + K_r) = I_{r^2} + K_r, \quad (\text{D.6})$$

$$K_r D_r = D_r, \quad (\text{D.7})$$

from [Magnus and Neudecker \(1980\)](#). By (D.5), the matrix inside the inverse equals $L_r (x^\top x \otimes x^\top x) (I_{r^2} + K_r) L_r^\top$, and combining (D.4)–(D.7) shows that its inverse is $\frac{1}{2} D_r^\top \left((x^\top x)^{-1} \otimes (x^\top x)^{-1} \right) D_r$. Since $D_r \text{vech}(A) = \text{vec}(A)$ for symmetric A , the left-hand side of (D.3) equals

$$\frac{1}{2} \text{vec}(A)^\top \left((x^\top x)^{-1} \otimes (x^\top x)^{-1} \right) \text{vec}(B) = \frac{1}{2} \text{tr}(A (x^\top x)^{-1} B (x^\top x)^{-1}).$$

$\square$

D.2 Variance components

Here $\mu_0 = \eta_0^\top Q \eta_0$ and the base moment is $u = \eta^\top Q \eta - \mu$, which does not involve y , so that $\mathbb{E}[u W_2^\top] = u \mathbb{E}[W_2^\top] = 0$ and the construction (6.11) reduces to

$$u_2^* = u + b_2^\top \left(\mathbb{E} [W_2 W_2^\top] \right)^{-1} W_2,$$

where $b_2 = \left((2Q\eta)^\top, \text{vech}(2Q)^\top\right)^\top$ is the column vector collecting the sensitivities of Section 6 in the coordinates of W_2 : the gradient $\nabla_\eta(\eta^\top Q\eta) = 2Q\eta$ and, in the vech convention of (D.1), the second-derivative block $\text{vech}(2Q)$. By Lemma 1 the Gram matrix is block diagonal, so the two blocks contribute separately. The first-order block contributes

$$(2Q\eta)^\top \left(\frac{x^\top x}{\sigma^2}\right)^{-1} \frac{x^\top(y - x\eta)}{\sigma^2} = 2\eta^\top Q(x^\top x)^{-1}x^\top(y - x\eta),$$

and truncating the construction after this block yields u_1^* in (6.16). The second-order block contributes, by Lemma 2 with $A = 2Q$ and $B = \sigma^4 V$,

$$\begin{aligned} & \sigma^4 \text{vech}(2Q)^\top \left[L_r(x^\top \otimes x^\top)(I_{n^2} + K_n)(x \otimes x)L_r^\top \right]^{-1} \text{vech}(V) \\ &= \sigma^4 \text{tr} \left(Q(x^\top x)^{-1} V (x^\top x)^{-1} \right), \end{aligned}$$

and substituting the definition of V ,

$$\begin{aligned} & \sigma^4 \text{tr} \left(Q(x^\top x)^{-1} V (x^\top x)^{-1} \right) \\ &= (y - x\eta)^\top x (x^\top x)^{-1} Q (x^\top x)^{-1} x^\top (y - x\eta) - \sigma^2 \text{tr} \left(Q (x^\top x)^{-1} \right). \end{aligned}$$

Adding the three pieces and expanding $(y - x\eta)^\top x = y^\top x - \eta^\top x^\top x$, the terms involving η cancel,

$$\begin{aligned} & \eta^\top Q\eta + 2\eta^\top Q(x^\top x)^{-1}x^\top(y - x\eta) + (y - x\eta)^\top x (x^\top x)^{-1}Q(x^\top x)^{-1}x^\top(y - x\eta) \\ &= y^\top x (x^\top x)^{-1}Q(x^\top x)^{-1}x^\top y, \end{aligned}$$

which gives u_2^* in (6.17). As a check, substituting $x^\top y = x^\top x\eta + \sigma x^\top \varepsilon$ shows directly that $\mathbb{E}[y^\top x (x^\top x)^{-1}Q(x^\top x)^{-1}x^\top y] = \eta^\top Q\eta + \sigma^2 \text{tr}(Q(x^\top x)^{-1})$ for every η , confirming that u_2^* is fully orthogonal, and, in this case, exactly unbiased.

D.3 The variance parameter

We now derive the expression given in footnote 6, taking $\theta = \sigma^2$ as the target and $u = \nabla_{\sigma^2} \ell = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(y - x\eta)^\top(y - x\eta)$. Being a score, u is unbiased at all parameter values, so that all sensitivities vanish and the construction of Section 5.2 applies, with $u_2^* = u - \mathbb{E}[u W_2^\top] \left(\mathbb{E}[W_2 W_2^\top] \right)^{-1} W_2$.

Writing $u = (\varepsilon^\top \varepsilon - n)/(2\sigma^2)$, the first-order block vanishes because $\mathbb{E}[u w_1^\top]$ involves third moments of ε . For the second-order block, $\mathbb{E}[\varepsilon_i \varepsilon_j \varepsilon_k \varepsilon_l] = \delta_{ij} \delta_{kl} + \delta_{ik} \delta_{jl} + \delta_{il} \delta_{jk}$ gives $\mathbb{E}[(\varepsilon^\top \varepsilon) \varepsilon \varepsilon^\top] = (n+2)I_n$, hence $\mathbb{E}[(\varepsilon^\top \varepsilon - n) \varepsilon \varepsilon^\top] = 2I_n$ and

$$\mathbb{E}[u V] = \frac{1}{2\sigma^4} x^\top \mathbb{E}[(\varepsilon^\top \varepsilon - n)(\varepsilon \varepsilon^\top - I_n)] x = \frac{1}{\sigma^4} x^\top x.$$

Applying Lemma 2 with $A = x^\top x / \sigma^4$ and $B = \sigma^4 V$, the second-order block contributes $\frac{1}{2} \text{tr}(V(x^\top x)^{-1})$, so that

$$u_2^* = u - \frac{1}{2} \text{tr}(V(x^\top x)^{-1}).$$

Finally, writing $P_x = x(x^\top x)^{-1}x^\top$ and $M_x = I_n - P_x$, the definition of V gives $\text{tr}(V(x^\top x)^{-1}) = (\varepsilon^\top P_x \varepsilon - d_\eta) / \sigma^2$, and since $M_x x = 0$ implies $y^\top M_x y = \sigma^2 \varepsilon^\top M_x \varepsilon$,

$$u_2^* = \frac{\varepsilon^\top \varepsilon - n}{2\sigma^2} - \frac{\varepsilon^\top P_x \varepsilon - d_\eta}{2\sigma^2} = \frac{\varepsilon^\top M_x \varepsilon - (n - d_\eta)}{2\sigma^2} = \frac{1}{2\sigma^4} (y^\top M_x y - (n - d_\eta) \sigma^2),$$

as claimed. It is free of η and has mean zero at every value of it, since $\mathbb{E}[\varepsilon^\top M_x \varepsilon] = \text{tr}(M_x) = n - d_\eta$.