MODEL-SELECTION TESTS FOR COMPLEX SURVEY SAMPLES

Iraj Rahmani[a] and Jeffrey M. Wooldridge[b]
[a] *Department of Economics, Nazarbayev University, Kazakhstan*
[b] *Department of Economics, Michigan State University, USA*

ABSTRACT

We extend Vuong's (1989) model-selection statistic to allow for complex survey samples. As a further extension, we use an M-estimation setting so that the tests apply to general estimation problems – such as linear and nonlinear least squares, Poisson regression and fractional response models, to name just a few – and not only to maximum likelihood settings. With stratified sampling, we show how the difference in objective functions should be weighted in order to obtain a suitable test statistic. Interestingly, the weights are needed in computing the model-selection statistic even in cases where stratification is appropriately exogenous, in which case the usual unweighted estimators for the parameters are consistent. With cluster samples and panel data, we show how to combine the weighted objective function with a cluster-robust variance estimator in order to expand the scope of the model-selection tests. A small simulation study shows that the weighted test is promising.

Keywords: Survey sampling; weighted estimation; cluster sampling; nonnested models; model selection test; m-estimation

The Econometrics of Complex Survey Data: Theory and Applications
Advances in Econometrics, Volume 39, 109–135
Copyright © 2019 by Emerald Publishing Limited
All rights of reproduction in any form reserved
ISSN: 0731-9053/doi:10.1108/S0731-905320190000039009

1. INTRODUCTION

Building on White (1982), who studied the asymptotic properties of maximum likelihood estimators under general misspecification, Vuong (1989) develops a classical approach to model selection based on the Kullback–Leibler Information Criterion (KLIC), which measures the closeness of a specified model to the true model. Vuong proposes a simple test based on comparing the log-likelihood functions from two estimated models. An important aspect of Vuong's approach is that the null hypothesis is taken to be that both models, as measured by the KLIC, are equally close to the true model. The alternative is that one model provides a better approximation (in the underlying population). Vuong's general framework allows for one of the two models to be nested in the other, for the models to be completely nonnested, and for the models to overlap (depending on the values of the population parameters). When one model is nested within another, we obtain the standard likelihood ratio testing principle, although the limiting distribution is often nonstandard if the most general model is misspecified.

The most attractive application of Vuong's approach is when the competing models are nonnested in a sense we make precise in Section 5.2. Essentially, the models are nonnested if we cannot obtain either model as a special case of the other by imposing parameter restrictions. As shown by Vuong (1989), his statistic has a limiting standard normal distribution under the null hypothesis that the models are nonnested and equally close to the true model. Necessarily, both models are misspecified under the null; if one model were correctly specified, it would necessarily be closer to the true model. Under the alternative, one model may be correctly specified, but the test has power against the alternative that one of the models provides a better approximation in the population. Due to its computational simplicity and standard normal limiting distribution, Vuong's test has become popular in applied work and provides an alternative to computationally intensive approaches based on Cox (1961, 1962). Computationally simple alternatives to the Cox approach, such as those in a regression context proposed by Davidson and MacKinnon (1981) and Wooldridge (1990), are limited in scope.

Vuong's original framework is based on the assumption that the sample consists of independent, identically distributed draws from the population. In practice, many large surveys, such as the Current Population Survey, the Panel Survey of Income Dynamics and National Survey of Families and Households (NSFH), to name a few, adopt stratification and clustering schemes. In such cases, Vuong's (1989) original framework is not valid.

In this paper we extend Vuong's model-selection procedure to allow for various survey sampling methods. We are unaware of any previous work that does so. Findley (1990, 1991), Findley and Wei (1993), and Rivers and Vuong (2002) consider time series problems, such as ARMA models and dynamic regression models, but the nature of the adjustment to the test statistics is different from what we address here.

An important feature of our approach, which also extends Vuong's (1989) original framework, is that we study the model-selection problem in the context of general M-estimation. This generality allows us to explicitly cover situations where only a specific feature of a distribution is correctly specified, such as the conditional mean. For example, because of its robustness for estimating the parameters of a conditional mean – see Gourieroux, Monfort, and Trognon (1984) – Poisson regression is commonly used in situations where little if any interest lies in the rest of the distribution. However, we may have two competing models of the conditional mean. While one could use nonlinear least squares estimation, for efficiency reasons, the Poisson quasi-MLE is often preferred. We can apply Vuong's (1989) approach in this situation to obtain a statistic that does not take a stand on the distribution; it is purely a test of the conditional mean function. By contrast, if we use the same mean function – say, an exponential function with the same covariates – and we use two different quasi-MLEs, such as the Poisson and Geometric, then the Vuong approach is a test of which distribution fits betters. Understanding the distinction between a conditional mean test and a full test of the conditional distribution is something easily described in our setting.

The remainder of the paper is organized as follows. In Section 5.2, we define the estimation problems that effectively define the two nonnested competing models. Section 5.3 shows how to modify the Vuong (1989) statistic to accommodate standard stratified (SS) sampling in the context of general M-estimation. We start with SS and variable probability (VP) sampling because they are widely used in practice, and it is then clear what role weighting plays in more complex sampling designs. Section 5.4 shows how to modify the estimated variance to account for clustering in a multistage design. In Section 5.5, we extend the model-selection test to panel data models with standard stratification design. In Section 5.6, we discuss how weighting is desirable even under what is typically called "exogenous" stratification. Section 5.7 provides several examples, and Section 5.8 contains a small simulation study. Section 5.9 contains a brief conclusion.

2. NONNESTED COMPETING MODELS AND THE NULL HYPOTHESIS

Let $\mathbf{W}$ be an $M \times 1$ random vector taking values in $\mathcal{W} \subset \mathbb{R}^M$. Our goal is to learn something about the distribution of $\mathbf{W}$, which represents an underlying population. As discussed in, say, Wooldridge (2010, Chapter 12), many population parameters of interest – those indexing conditional means, conditional variances, conditional quantiles, and so on – can be identified using a minimization problem in the population. For a $P \times 1$ vector of parameters $\boldsymbol{\theta}$ in a parameter space $\boldsymbol{\Theta} \subset \mathbb{R}^P$, let $q : \mathcal{W} \times \boldsymbol{\Theta} \to \mathbb{R}$ be a real-valued function. Typically we assume that the population minimization problem

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{E}\,[q(\mathbf{W}, \boldsymbol{\theta})] \tag{1}$$

has a unique solution – which is required for identification. In standard settings, the solution is often denoted $\boldsymbol{\theta}_o$, and $\boldsymbol{\theta}_o$ is assumed to index the quantity of interest, such as the parameters in a conditional mean function $\mathbb{E}\,(Y|\mathbf{X})$. In a conditional MLE setting, where $q(\mathbf{W}, \boldsymbol{\theta}) = -\log\,[f(Y|\mathbf{X}; \boldsymbol{\theta}]$ is the negative of the log likelihood, $\boldsymbol{\theta}_o$ is the vector of parameters indexing the conditional density of Y given $\mathbf{X}$. There are many other applications where $\mathbf{W}$ is partitioned as $\mathbf{W} = (\mathbf{X}, \mathbf{Y})$, where $\mathbf{X}$ and $\mathbf{Y}$ are, respectively, K and L dimensional vectors with $L + K = M$. We are particularly interested in the case where $q\ (\cdot)$ is the negative of a quasi-log-likelihood function in the linear exponential family.

Rather than estimate, the parameters using a single objective function – underlying which is a parametric model of some feature of a distribution $\mathbb{D}\,(\mathbf{W})$ or, more likely, a conditional distribution – we suppose we have two estimation methods, represented by the objective functions $q_1(\mathbf{W}, \boldsymbol{\theta}_1)$ and $q_2(\mathbf{W}, \boldsymbol{\theta}_2)$, where the parameter vectors may have different dimensions. We need to make precise the sense in which these competing models and estimation methods are nonnested. Let $\boldsymbol{\theta}_1^*$ and $\boldsymbol{\theta}_2^*$ be the unique solutions to the population problems

$$\min_{\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_1} \mathbb{E}\,[q_1(\mathbf{W}, \boldsymbol{\theta}_1)] \text{ and } \min_{\boldsymbol{\theta}_2 \in \boldsymbol{\Theta}_2} \mathbb{E}\,[q_2(\mathbf{W}, \boldsymbol{\theta}_2)]\,. \tag{2}$$

These solutions are often called the “pseudo-true values” or “quasi-true values.” The null hypothesis is that the models evaluated at the pseudo-true values fit equally well *on average*, where fit is measured by the mean of the objective functions in the population. Precisely,

$$\mathrm{H}_0 : \mathbb{E}\left[q_1(\mathbf{W}, \boldsymbol{\theta}_1^*)\right] = \mathbb{E}\left[q_2(\mathbf{W}, \boldsymbol{\theta}_2^*)\right]. \tag{3}$$

In the maximum likelihood setting, (3) states that the KLIC distances of the two models to the true model are the same. In a regression context using nonlinear least squares, the null is that the population sum of squared residuals are the same; equivalently, the two models provide a function that has the same mean squared error relative to the true conditional mean.

Condition (3) can hold for nested as well as nonnested models, but the nature of our approach requires us to focus on the latter. The reason is that we supplement

(3) with the assumption that the objective functions, evaluated at the pseudo-true values, differ with positive probability:

$$\mathbb{P}\left[q_1(\mathbf{W},\boldsymbol{\theta}_1^*)\neq q_2(\mathbf{W},\boldsymbol{\theta}_2^*)\right]>0. \tag{4}$$

The requirement in (4) means that the two functions $q_1(\mathbf{W},\boldsymbol{\theta}_1^*)$ and $q_2(\mathbf{W},\boldsymbol{\theta}_2^*)$ must differ for a nontrivial set of outcomes on the support of $\mathbf{W}$. If (4) does not hold then the variance of $q_1(\mathbf{W},\boldsymbol{\theta}_1^*)-q_2(\mathbf{W},\boldsymbol{\theta}_2^*)$ is 0, and that will invalidate the Vuong (1989) approach taken in the paper.

The combination of (3) and (4) effectively rules out nested models, where one model is obtained as a special case of the other and the same objective function – such as the negative of the log-likelihood function – is used. Then, the only way that (3) can be true is when the more general model collapses to the restricted version; otherwise $\mathbb{E}\left[q_1(\mathbf{W},\boldsymbol{\theta}_1^*)\right]>\mathbb{E}\left[q_2(\mathbf{W},\boldsymbol{\theta}_2^*)\right]$. In other words, for nested models, we cannot have (3) and (4) both be true. The two conditions (3) and (4) rule out other forms of degeneracies. For example, assume we have a random variable Y and would like to model $\mathbb{E}(Y|\mathbf{X})$ as a function of the explanatory variables $\mathbf{X}$, a $1\times K$ vector. We specify two competing models and we estimate both by nonlinear least squares. Specifically, $q_1(\mathbf{W},\boldsymbol{\theta}_1)=(Y-\alpha_1-\mathbf{X}\boldsymbol{\beta}_1)^2$ and $q_2(\mathbf{W},\boldsymbol{\theta}_2)=\left[Y-\exp(\alpha_2+\mathbf{X}\boldsymbol{\beta}_2)\right]^2$. If $\mathbb{E}(Y|\mathbf{X})$ actually depends on $\mathbf{X}$, then (4) generally holds. However, if Y is mean independent of $\mathbf{X}$, so that $\mathbb{E}(Y|\mathbf{X})=\mathbb{E}(Y)$, then the two models are simply different parameterizations of a constant conditional mean, and (4) fails. As we will see, this failure causes the standard normal limiting distribution for the Vuong-type statistic to break down. Incidentally, in this example, provided the models satisfy (4), the Vuong test with random sampling would reduce to comparing the R-squareds from the two least squares regressions. We require no additional assumptions – for example, neither homoskedasticity nor normality – for the test to be valid.

The nature of the alternative is inherently one sided, as we wish to determine whether one model can be rejected in favor of the other. Because we have defined the optimization problem to be a minimization problem, the alternative that model one – technically, model one combined with whatever objective function we choose – fits better in the population is

$$\mathrm{H}_{A_{q_1}}:\mathbb{E}\left[q_1(\mathbf{W},\boldsymbol{\theta}_1^*)\right]<\mathbb{E}\left[q_2(\mathbf{W},\boldsymbol{\theta}_2^*)\right].$$

Likewise, the alternative that model two fits better is

$$\mathrm{H}_{A_{q_2}}:\mathbb{E}\left[q_1(\mathbf{W},\boldsymbol{\theta}_1^*)\right]>\mathbb{E}\left[q_2(\mathbf{W},\boldsymbol{\theta}_2^*)\right].$$

In Vuong's setup, these functions are the negative of a log-likelihood function, but we can apply these alternatives much more generally. Naturally, if either $\mathrm{H}_{A_{q_1}}$ or $\mathrm{H}_{A_{q_2}}$ holds then the nondegeneracy condition (4) must hold.

3. TESTING UNDER STRATIFIED SAMPLING

Under random sampling and weak regularity conditions – see, for example, Newey and McFadden (1994) or Wooldridge (2010, Chapter 12) – we can consistently estimate $\boldsymbol{\theta}_g^*$, $g = 1, 2$ by solving the sample counterparts to (2):

$$\min_{\boldsymbol{\theta}_g \in \boldsymbol{\Theta}_g} N^{-1} \sum_{i=1}^{N} q_g(\mathbf{W}, \boldsymbol{\theta}_g).$$

In this section, we are interested in cases where the population has been divided into J strata and then the resulting sample is not necessarily representative of the population.

3.1. Standard Stratified Sampling

The first sampling scheme we consider is SS sampling. As in Wooldridge (2001), for each $j \in \{1, 2, \ldots, J\}$ assume we have a random sample $\{\mathbf{W}_{ij} : i = 1, 2, \ldots, N_j\}$ from the conditional distribution $\mathbb{D}(\mathbf{W}|\mathbf{W} \in \mathcal{W}_j)$, where $\mathcal{W}_j$ is the jth stratum. Let $Q_j = \mathbb{P}(\mathbf{W} \in \mathcal{W}_j)$ be the population share associated with stratum j. Then we choose $\hat{\boldsymbol{\theta}}_g$, $g = 1, 2$ to solve the weighted M-estimation problem

$$\min_{\boldsymbol{\theta}_g \in \Theta_g} \sum_{j=1}^{J} Q_j \left(\frac{1}{N_j} \sum_{i=1}^{N_j} q_g(\mathbf{W}_{ij}, \boldsymbol{\theta}_g) \right). \tag{5}$$

The objective function in Eq. (5) can be rewritten as

$$\frac{1}{N} \sum_{j=1}^{J} \frac{Q_j}{H_j} \sum_{i=1}^{N_j} q_g(\mathbf{W}_{ij}, \boldsymbol{\theta}_g), \tag{6}$$

where $H_j \equiv N_j/N$ is the fraction of observations drawn from stratum j. Further, by letting j_i be the stratum for draw i and dropping the division by N, we can write the objective function in the more familiar form

$$\sum_{i=1}^{N}\left(\frac{Q_{j_i}}{H_{j_i}}\right) q_g(\mathbf{W}_i,\boldsymbol{\theta}_g), \tag{7}$$

where Q_{j_i}/H_{j_i} is the sampling weight for observation i. Thus, the estimator represented in (7) is obtained by weighting the objective function by the sampling weight for each observation i. As discussed in Wooldridge (2001), the representation in (5) is the form used to obtain asymptotic properties of the weighted M-estimator.

We first show that when the models are nonnested in the sense of (4), a properly standardized version of the objective function has a limiting distribution that does not depend on the limiting distribution of $\sqrt{N}(\hat{\boldsymbol{\theta}}_g-\boldsymbol{\theta}_g^*)$ provided $\hat{\boldsymbol{\theta}}_g$ is $\sqrt{N}$-consistent, which is standard. Wooldridge (2001, Theorem 3.2) contains sufficient conditions. The proof relies on fairly standard asymptotics and so we show only its main features.

Theorem 3.1. *Assume that for* $g\in\{1,2\}$,

1. $\{\mathbf{W}_{ij}: i=1,\ldots,N_j; j=1,\ldots,J\}$ *satisfies the SS sampling scheme with* $N_j/N\to a_j>0,\ j=1,\ldots,J$.
2. $\boldsymbol{\Theta}_g$ *is compact.*
3. *The objective function* $\mathbb{E}\left[q_g(\mathbf{W},\boldsymbol{\theta}_g)\right]$ *has a unique minimum on* $\boldsymbol{\Theta}_g$ *at* $\boldsymbol{\theta}_g^*$.
4. $\boldsymbol{\theta}_g^*\in\operatorname{int}\left(\boldsymbol{\Theta}_g\right)$.
5. *For each* $\mathbf{w}\in\mathcal{W}$, $q_g(\mathbf{w},\cdot)$ *is continuous on* $\boldsymbol{\Theta}_g$ *and twice continuously differentiable on* $\operatorname{int}\left(\boldsymbol{\Theta}_g\right)$.
6. *For all* $\boldsymbol{\theta}_g\in\boldsymbol{\Theta}_g$, $\left|q_g(\mathbf{w},\boldsymbol{\theta}_g)\right|\le b\,(\mathbf{w})$, $\left|\partial q_g(\mathbf{w},\boldsymbol{\theta}_g)/\partial\theta_{gk}\right|\le b\,(\mathbf{w})$, $\left|\partial^2 q_g(\mathbf{w},\boldsymbol{\theta}_g)/\partial\theta_{gk}\partial\theta_{gm}\right|\le b\,(\mathbf{w})$ *for a function* $b\,(\mathbf{w})$ *with* $\mathbb{E}\,[b(\mathbf{W})]<\infty$.
7. $\mathbb{E}\left[\nabla_{\boldsymbol{\theta}} q_g(\mathbf{W},\boldsymbol{\theta}_g^*)\nabla_{\boldsymbol{\theta}} q_g(\mathbf{W},\boldsymbol{\theta}_g^*)'\right]<\infty$ *and* $\mathbb{E}\left[\nabla_{\boldsymbol{\theta}} q_g(\mathbf{W},\boldsymbol{\theta}_g^*)\right]=\mathbf{0}$.
   Then

$$\frac{1}{\sqrt{N}}\sum_{j=1}^{J}\frac{Q_j}{H_j}\sum_{i=1}^{N_j} q_g(\mathbf{W}_{ij},\hat{\boldsymbol{\theta}}_g)=\frac{1}{\sqrt{N}}\sum_{j=1}^{J}\frac{Q_j}{H_j}\sum_{i=1}^{N_j} q_g(\mathbf{W}_{ij},\boldsymbol{\theta}_g^*)+o_p(1).$$

Proof. Under the assumptions, $q(\cdot)$ is continuously differentiable with respect to $\boldsymbol{\theta}_g$ and $\boldsymbol{\theta}_g^*$ is in the interior of Θ_g. Therefore, from a Taylor expansion of $\sum_{i=1}^{J} q_g(\mathbf{W}_{ij},\boldsymbol{\theta}_g)$ about $\boldsymbol{\theta}_g^*$, and then dividing both side by N_j, we obtain

$$\frac{1}{N_j}\sum_{i=1}^{N_j} q_g(\mathbf{W}_{ij},\hat{\boldsymbol{\theta}}_g)=\frac{1}{N_j}\sum_{i=1}^{N_j} q_g(\mathbf{W}_{ij},\boldsymbol{\theta}_g^*)+\frac{1}{N_j}\sum_{i=1}^{N_j}\nabla_{\boldsymbol{\theta}} q_g(\mathbf{W}_{ij},\ddot{\boldsymbol{\theta}}_g^j)(\hat{\boldsymbol{\theta}}_g-\boldsymbol{\theta}_g^*),$$

where $\ddot{\boldsymbol{\theta}}_g^j$ is a mean value between $\hat{\boldsymbol{\theta}}_g$ and $\boldsymbol{\theta}_g^*$. Because $\hat{\boldsymbol{\theta}}_g \xrightarrow{p} \boldsymbol{\theta}_g^*$, $\ddot{\boldsymbol{\theta}}_g \xrightarrow{p} \boldsymbol{\theta}_g^*$. Now, by a corollary of the uniform law of large numbers, for example, Wooldridge (2010, Lemma 12.1),

$$\frac{1}{N_j}\sum_{i=1}^{N_j} \nabla_{\boldsymbol{\theta}} q_g(\mathbf{W}_{ij}, \ddot{\boldsymbol{\theta}}_g^j) = \frac{1}{N_j}\sum_{i=1}^{N_j} \nabla_{\boldsymbol{\theta}} q_g(\mathbf{W}_{ij}, \boldsymbol{\theta}_g^*) + o_p(1).$$

As shown in Wooldridge (2001, Theorem 3.2), the assumptions ensure that $\sqrt{N}(\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g^*) = O_p(1)$, and so

$$\begin{aligned}&\left[\frac{1}{N_j}\sum_{i=1}^{N_j} \nabla_{\boldsymbol{\theta}} q_g(\mathbf{W}_{ij}, \ddot{\boldsymbol{\theta}}_g^j)\right]\sqrt{N}(\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g^*)\\ &\quad = \left[\frac{1}{N_j}\sum_{i=1}^{N_j} \nabla_{\boldsymbol{\theta}} q_g(\mathbf{W}_{ij}, \boldsymbol{\theta}_g^*)\right]\sqrt{N}(\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g^*) + o_p(1).\end{aligned}$$

Therefore, if we multiply by $\sqrt{N}Q_j$ and sum across j we get

$$\begin{aligned}\sqrt{N}\sum_{j=1}^{J}\frac{Q_j}{N_j}\sum_{i=1}^{N_j} q_g(\mathbf{W}_{ij}, \hat{\boldsymbol{\theta}}_g) &= \sqrt{N}\sum_{j=1}^{J}\frac{Q_j}{N_j}\sum_{i=1}^{N_j} q_g(\mathbf{W}_{ij}, \boldsymbol{\theta}_g^*)\\ &\quad + \left[\sum_{j=1}^{J}\frac{Q_j}{N_j}\sum_{i=1}^{N_j} \nabla_{\boldsymbol{\theta}} q_g(\mathbf{W}_{ij}, \boldsymbol{\theta}_g^*)\right]\sqrt{N}(\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g^*).\end{aligned} \tag{8}$$

As in Wooldridge (2001, Theorem 3.2),

$$\begin{aligned}\operatorname*{plim}_{N\to\infty}\sum_{j=1}^{J} Q_j\left(\frac{1}{N_j}\sum_{i=1}^{N_j} \nabla_{\boldsymbol{\theta}} q_g(\mathbf{W}_{ij}, \boldsymbol{\theta}_g^*)\right) &= \sum_{j=1}^{J} Q_j\left(\operatorname*{plim}_{N\to\infty}\frac{1}{N_j}\sum_{i=1}^{N_j} \nabla_{\boldsymbol{\theta}} q_g(\mathbf{W}_{ij}, \boldsymbol{\theta}_g^*)\right)\\ &= \sum_{j=1}^{J} Q_j \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} q_g(\mathbf{W}_i, \boldsymbol{\theta}_g^*)|\mathbf{W}_i \in \mathcal{W}_j\right]\\ &= \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} q_g(\mathbf{W}_i, \boldsymbol{\theta}_g^*)\right] = \mathbf{0},\end{aligned} \tag{9}$$

where the last equality holds from the population first order condition for $\boldsymbol{\theta}_g^*$. Therefore

$$\sum_{j=1}^{J} Q_j \left(\frac{1}{N_j} \sum_{i=1}^{N_j} \nabla_{\boldsymbol{\theta}} q_g(\mathbf{W}_{ij}, \boldsymbol{\theta}_g^*) \right) = o_p(1),$$

and since $\sqrt{N}(\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g^*) = O_p(1)$, the second term product in (8) is $o_p(1)$. We have shown

$$\sqrt{N} \sum_{j=1}^{J} \frac{Q_j}{N_j} \sum_{i=1}^{N_j} q_g(\mathbf{W}_{ij}, \hat{\boldsymbol{\theta}}_g) = \sqrt{N} \sum_{j=1}^{J} \frac{Q_j}{N_j} \sum_{i=1}^{N_j} q_g(\mathbf{W}_{ij}, \boldsymbol{\theta}_g^*) + o_p(1),$$

and using $H_j = N_j/N$, we can rewrite this as

$$\frac{1}{\sqrt{N}} \sum_{j=1}^{J} \frac{Q_j}{H_j} \sum_{i=1}^{N_j} q_g(\mathbf{W}_{ij}, \hat{\boldsymbol{\theta}}_g) = \frac{1}{\sqrt{N}} \sum_{j=1}^{J} \frac{Q_j}{H_j} \sum_{i=1}^{N_j} q_g(\mathbf{W}_{ij}, \boldsymbol{\theta}_g^*) + o_p(1).$$

This complete the proof. ■

We can use Theorem 3.1 to construct a simple test statistic that allow us to discriminate between two competing models. Let $r(\mathbf{w}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \equiv q_1(\mathbf{w}, \boldsymbol{\theta}_1) - q_2(\mathbf{w}, \boldsymbol{\theta}_2)$ be the difference in the two objective functions evaluated at $\mathbf{w} \in \mathcal{W}$ and generic values of the parameters. The null hypothesis is

$$\mathrm{H}_0 : \mathbb{E}\left[r(\mathbf{W}, \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*)\right] = 0.$$

By the assumption that the models are nonnested, $\mathbb{V}\left[r(\mathbf{W}, \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*)\right] > 0$. Applied to both estimation problems, and under the null hypothesis, Theorem 3.1 implies

$$\frac{1}{\sqrt{N}} \sum_{j=1}^{J} \frac{Q_j}{H_j} \sum_{i=1}^{N_j} r(\mathbf{W}_{ij}, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) = \frac{1}{\sqrt{N}} \sum_{j=1}^{J} \frac{Q_j}{H_j} \sum_{i=1}^{N_j} r(\mathbf{W}_{ij}, \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*) + o_p(1). \quad (10)$$

Eq. (10) is the key result in the paper. It extends Vuong (1989) by allowing for stratified sampling, and also any objective functions – not just MLE. Because the right hand side of (10) does not depend on the limiting distributions of $\sqrt{N}(\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g^*)$, its distribution is easy to study by now applying, again, the results in Wooldridge (2001) on SS sampling.

Theorem 3.2. *Under the conditions of Theorem 3.1, assume that*

$$\mathrm{H}_0 : \mathbb{E}\left[q_1(\mathbf{W}_i, \boldsymbol{\theta}_1^*)\right] = \mathbb{E}\left[q_2(\mathbf{W}_i, \boldsymbol{\theta}_2^*)\right].$$

Then

$$\frac{1}{\sqrt{N}} \sum_{j=1}^{J} \frac{Q_j}{H_j} \sum_{i=1}^{N_j} r(\mathbf{W}_{ij}, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) \xrightarrow{d} \text{Normal}(0, \eta^2),$$

where

$$\eta^2 = \sum_{j=1}^{J} \left(\frac{Q_j^2}{H_j}\right) \mathbb{V}\left[r\left(\mathbf{W}, \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*\right) | \mathbf{W} \in \mathcal{W}_j\right].$$

Proof. From Theorem 3.1 and the asymptotic equivalence lemma, we must argue that

$$\frac{1}{\sqrt{N}} \sum_{j=1}^{J} \frac{Q_j}{H_j} \sum_{i=1}^{N_j} r(\mathbf{W}_{ij}, \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*) \xrightarrow{d} \text{Normal}(0, \eta^2).$$

But this holds from Wooldridge (2001, Theorem 3.2). Namely, we apply the asymptotic variance formula for a weighted objective function under SS sampling to the sequence $\{R_{ij} : i = 1, \ldots, N_j; j = 1, \ldots, J\}$, where

$$R_{ij} \equiv r(\mathbf{W}_{ij}, \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*). \qquad \blacksquare$$

Consistent estimation of η^2 is straightforward. Let

$$\bar{R}_j = N_j^{-1} \sum_{i=1}^{N_j} r(\mathbf{W}_{ij}, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2), \tag{11}$$

be the within-stratum mean of the difference in objective functions. Then

$$\begin{aligned}\hat{\eta}^2 &\equiv \sum_{j=1}^{J} \left(\frac{Q_j^2}{H_j}\right) \left(\frac{1}{N_j} \sum_{i=1}^{N_j} \left[r(\mathbf{W}_{ij}, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) - \bar{R}_j\right]^2\right) \\ &= \frac{1}{N} \sum_{j=1}^{J} \frac{Q_j^2}{H_j^2} \sum_{i=1}^{N_j} \left[r(\mathbf{W}_{ij}, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) - \bar{R}_j\right]^2, \qquad (12)\end{aligned}$$

which is the asymptotic variance estimator in Wooldridge (2001) applied to the scalar case. The model selection t statistic can be written as

$$t_{\text{MS}} = \frac{N^{-1/2}\sum_{i=1}^{N}(Q_{j_i}/H_{j_i})r(\mathbf{W}_{ij},\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_2)}{\hat{\eta}}. \tag{13}$$

Under the null hypothesis, $t_{\text{MS}} \xrightarrow{d}$ Normal(0, 1).

The model selection t statistic is very easy to obtain in practice. For each i, define the difference in objective functions as

$$\hat{R}_i = r(\mathbf{W}_i,\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_2) = q_1(\mathbf{W}_i,\hat{\boldsymbol{\theta}}_1) - q_2(\mathbf{W}_i,\hat{\boldsymbol{\theta}}_2).$$

Then we treat the $\hat{R}_i$ as data obtained from an SS sampling scheme, so we simply need to specify the stratum for each observation, j_i, and the weight, Q_{j_i}/H_{j_i}. For example, in Stata, one applies the “svyset” option – to specify the stratum identifier and weights – and runs the regression $\hat{R}_i$ on a constant. The usual t statistic on the constant is the model-selection test statistic. The sign of the constant indicates which model fits better.

In using t_{MS} as a model-selection test, it is important to understand its properties compared to an approach where weighting is not used but SS sampling has been employed. Because of the nature of the null and alternative, it is not true that the weighted version of the test will always reject the null more often than the unweighted version of the test. To see this, consider what happens when, say, model one is correctly specified with parameters $\boldsymbol{\theta}_{o1}$. Then, generally, we need to use the weights to consistently estimate $\boldsymbol{\theta}_{o1}$; the unweighted estimator converges to some other quantity, say $\boldsymbol{\theta}_1^+$. For model two, we can write the probability limits as $\boldsymbol{\theta}_2^*$ and $\boldsymbol{\theta}_2^+$ for the weighted and unweighted problems, respectively. Now, there is no guarantee that $q_1(\mathbf{W}_i,\boldsymbol{\theta}_{o1})$ is further from $\mathbb{E}\left[q_2(\mathbf{W}_i,\boldsymbol{\theta}_2^*)\right]$, on average, than $q_1(\mathbf{W}_i,\boldsymbol{\theta}_1^+)$ is from $\mathbb{E}\left[q_2(\mathbf{W}_i,\boldsymbol{\theta}_2^+)\right]$, and so the test based on the unweighted objective function may reject more often in favor of model one than the weighted version of the test. This turns out not to be a good thing, for two reasons. First, even though the unweighted estimator might point to the correct model/estimation method, the estimator of $\boldsymbol{\theta}_{o1}$ is generally inconsistent. In other words, the unweighted approach may choose the correct model but with parameter estimators that are essentially useless! In fact, for computing quantities of interest, such as average partial effects, there is no telling that it would be better to use model one with inconsistent parameter estimators or model two, with estimators converging to $\boldsymbol{\theta}_2^+$.

A second important shortcoming of the unweighted test is that it may systematically opt for model two when model one is correctly specified. And the problem would generally be worse as the sample size grows. This cannot happen with the weighted version of the test, provided we have chosen our model and objective function in a way that generates consistent estimators under correct specification of the feature of interest. The reason is that, if model one is correctly specified and the objective function is chosen appropriately, $\mathbb{E}\left[q_1(\mathbf{W},\boldsymbol{\theta}_{o1})\right] < \mathbb{E}\left[q_2(\mathbf{W},\boldsymbol{\theta}_2^*)\right]$. In other words, the weighted test is a consistent test for choosing the true model when one of the models is correctly specified. With the unweighted test, it could easily be that $\mathbb{E}\left[q_2(\mathbf{W},\boldsymbol{\theta}_2^+)\right] < \mathbb{E}\left[q_1(\mathbf{W},\boldsymbol{\theta}_1^+)\right]$, in which case the unweighted test will systematically select the wrong model. And this will happen with probability approaching one as the sample size grows. We will see this phenomenon in the simulations in Section 8.

An analogy that does not require thinking about weighting versus not weighting might be helpful. In fact, assume random sampling, as in the original Vuong (1989) work. Now suppose we specify two nonnested conditional mean models, $m_1\left(\mathbf{x}_1,\boldsymbol{\theta}_1\right)$ and $m_2\left(\mathbf{x}_1,\boldsymbol{\theta}_2\right)$, and model one is correctly specified. If we use an objective function that identifies conditional means – say, the squared residual function – then the Vuong test will detect that model one is correct with probability approaching one. Suppose we use another objective function, such as the least absolute deviations (LAD). In general, neither model one nor model two is correctly specified for the conditional median. Consequently, using LAD in the Vuong statistic has essentially unknown properties. It could incorrectly choose in favor of model two because model two is closest to the conditional median. But it could also frequently reject model two in favor of model one; in fact, nothing says the rejection frequency could not be higher than when using the squared residual objective function. In other words, using the wrong objective function may actually lead to a more powerful test. The problem is that when this occurs, it is essentially a fluke. And the LAD estimators are not generally consistent for conditional mean parameters, and so it is difficult to know how it helps to choose model one: we have the correct model but inconsistent estimators of the parameters.

When the competing models contain different numbers of parameters, the finite sample performance of t_{MS} may suffer. As in Vuong (1989), we can penalize the objective functions for the number of parameters. Since we are minimizing the objective function, we add a penalty that is a function of the number of parameters. The resulting statistic is

$$\tilde{t}_{\mathrm{MS}} = \frac{N^{-1/2}\sum_{i=1}^{N}(Q_{j_i}/H_{j_i})r(\mathbf{W}_{ij},\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_2) + N^{-1/2}\left[K(P_1)-K(P_2)\right]}{\hat{\eta}}. \quad (14)$$

where P_1 and P_2 are the number of parameters in the different models and $K(\cdot)$ is the penalty function. For example, $K(P)=P$ gives the Akaike (1973) criterion and $K(P)=(P/2)\log(N)$ gives the Schwarz (1978) criterion. In both cases, $N^{-1/2}K(P)\to 0$ for fixed P, and so the penalty does not affect the asymptotic distribution of the test statistic: $\tilde{t}_{\mathrm{MS}}$ and t_{MS} have the same asymptotic distributions under H_0. The statistic $\tilde{t}_{\mathrm{MS}}$ has the feature of penalizing models that are not parsimonious in the number of parameters. One could instead simply add $N^{-1/2}\,[K(P_1) - K(P_2)]$ to t_{MS}, which means the penalty would not be divided by $\hat{\eta}$. Again, the resulting statistic is asymptotically equivalent. In what follows, we drop the penalty function for notational convenience.

3.2. Variable Probability Sampling

When observations in the strata are difficult to identify prior to sampling or when collecting information on the variable determining stratification is cheap relative to the cost of collecting the remaining information, VP sampling is convenient. In VP sampling, a unit is first drawn at random from the population. If the unit falls into stratum j, it is kept with probability p_j. For example, if we define stratification in terms of individual income, we draw a person randomly from the population, determine the person's income and then keep that person with probability that depends on income class that is set by the researcher. As discussed in Wooldridge (1999), consistent estimation of the population parameters generally requires weighting the objective function by the inverse of the probability of being kept in the sample. With J strata, these probabilities are $\{p_j : j = 1, \ldots, J\}$. It is straightforward to show the analog of Lemma 3.1 carries over, and so, under the null hypothesis that the models are nonnested and fit the population equally well, estimation of the parameters does not affect the limiting distribution. This leads to the test statistic

$$\frac{N^{-1/2}\sum_{i=1}^{N} p_{j_i}^{-1} r(\mathbf{W}_i, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)}{\left(N^{-1}\sum_{i=1}^{N} p_{j_i}^{-2}\left[r(\mathbf{W}_i, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)\right]^2\right)^{1/2}} = \frac{\sum_{i=1}^{N} p_{j_i}^{-1} r(\mathbf{W}_i, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)}{\left(\sum_{i=1}^{N} p_{j_i}^{-2}\left[r(\mathbf{W}_i, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)\right]^2\right)^{1/2}}, \quad (15)$$

where again j_i is the stratum for observation i. (Remember that under VP sampling, we do not always observe a draw from the population; this statistic necessarily depends only on the draws we keep.)

One way that the denominator of (15) differs from that of (13) is that the within-stratum means are not removed in (15). Wooldridge (1999) shows that if the known sampling probabilities, p_j, are replaced with the observed frequencies, then it is proper to remove the means, $\bar{R}_j$, in (15). Using the sample frequencies means that we know how many times each stratum was drawn – call this M_j. Then $\hat{p}_j = N_j/M_j$, where N_j is the kept number of draws in stratum j (and which we always observe). We replace p_j in (15) with $\hat{p}_j$ and then replace $r(\mathbf{W}_i, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)$ with $r(\mathbf{W}_i, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) - \bar{R}_j$ in the denominator for all i in stratum j. In many cases, the number of times each stratum was drawn is not available, and so one must use the p_j rather than the $\hat{p}_j$. If one uses the $\hat{p}_j$ directly in (15), then the statistic is conservative in the sense that, asymptotically, its size will be less than the nominal size (because the estimated standard deviation is systematically too large).

4. TESTS STATISTICS UNDER MULTISTAGE SAMPLING

The model-selection statistic can also be modified to account for complex surveys that feature cluster sampling, stratified sampling and variable probability sampling. Clusters are groups of families, households or individuals positioned or occurring in relatively close association. For example, in a school, students in each class form a cluster. In rural areas the villages and in urban areas the neighborhoods are clusters. Stratification often occurs at a larger geographical level, such as county or state in the United States.

Complex survey methods are commonly used. For example, the NSFH is a complex survey sample. It has multistage design that involves clustering, stratification, and VP sampling.

The formal design structure we use here is closely related to Bhattacharya (2005). In the first stage, the population of interest is divided into S subpopulations or strata. These could be states, counties, prefectures, and other geographic entities. They are exhaustive and mutually exclusive within the designated population. Within stratum s, there are C_s clusters in the population. In stratum s, N_s clusters are drawn randomly. Since the asymptotic analysis is based on number of clusters going to infinity, we assume that in each stratum, a "large" number of clusters is sampled. Because of the cluster sampling, units (e.g., households) within each cluster can be arbitrarily correlated. Each sampled cluster c in stratum s contains a finite population of M_{sc} units (e.g., households). At the final sampling stage, for each sampled cluster c in stratum s, we randomly sample K_{sc} households. (The following formula assumes sampling is with replacement.) Define a (nonrandom) weight for each stratum-cluster pair as

$$v_{sc} = \frac{C_s}{N_s}\frac{M_{sc}}{K_{sc}},$$

where we require information on the number of clusters in the population and the number of units per cluster. The weighted objective function

$$\sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}} v_{sc} q_g\left(\mathbf{W}_{scm}, \boldsymbol{\theta}_g\right)$$

identifies the population pseudo-true parameters $\boldsymbol{\theta}_g^*$, $g = 1,2$. See Bhattacharya (2005) for linear regression and Rahmani (2018) for the general M-estimation case.

Using reasoning similar to Lemma 3.1, it can be shown that the asymptotic distribution of

$$\frac{1}{\sqrt{N}} \sum_{s=1}^{S} \sum_{c=1}^{N_s} \sum_{m=1}^{K_{sc}} v_{sc} \cdot r\left(\mathbf{W}_{scm}, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2\right)$$

is not affected by the estimation of $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$, where

$$r\left(\mathbf{W}_{scm}, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2\right) = q_1\left(\mathbf{W}_{scm}, \hat{\boldsymbol{\theta}}_1\right) - q_2\left(\mathbf{W}_{scm}, \hat{\boldsymbol{\theta}}_2\right)$$

is the difference between the two objective functions for each unit m, in cluster c, in stratum s. In other words,

$$\frac{1}{\sqrt{N}} \sum_{s=1}^{S} \sum_{c=1}^{N_s} \sum_{m=1}^{K_{sc}} v_{sc} \cdot r\left(\mathbf{W}_{scm}, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2\right) = \frac{1}{\sqrt{N}} \sum_{s=1}^{S} \sum_{c=1}^{N_s} \sum_{m=1}^{K_{sc}} v_{sc} \cdot r\left(\mathbf{W}_{scm}, \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*\right) + o_p(1). \quad (16)$$

Consequently, as in the case of simple stratification, the problem reduces to obtaining a valid standard error for a population mean, namely, $\mathbb{E}\left[q_1\left(\mathbf{W}, \boldsymbol{\theta}_1^*\right) - q_2\left(\mathbf{W}, \boldsymbol{\theta}_2^*\right)\right]$. Under the null hypothesis, this mean is zero, and (16) means we can directly apply the results of Bhattacharya (2005) directly:

$$\frac{1}{\sqrt{N}} \sum_{s=1}^{S} \sum_{c=1}^{N_s} \sum_{m=1}^{K_{sc}} v_{sc} \cdot r\left(\mathbf{W}_{scm}, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2\right) \xrightarrow{d} N(0, \xi^2)$$

and a consistent estimator of ξ^2 is

$$\begin{aligned} \hat{\xi}^2 = N^{-1} \Bigg\{ & \sum_{s=1}^{S} \sum_{c=1}^{N_s} \sum_{m=1}^{K_{sc}} v_{sc}^2 r_{scm}^2 \left(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2\right) \\ & + \sum_{s=1}^{S} \sum_{c=1}^{N_s} \sum_{m=1}^{K_{sc}} \sum_{m' \neq m}^{K_{sc}} v_{sc}^2 r_{scm} \left(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2\right) r_{scm'} \left(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2\right) \\ & - \sum_{s=1}^{S} \frac{1}{N_s} \left(\sum_{c=1}^{N_s} \sum_{m=1}^{K_{sc}} v_{sc} r_{scm} \left(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2\right) \right)^2 \Bigg\}. \end{aligned} \quad (17)$$

The first term in (17) would be a consistent estimator of the variance under simple random sampling. The second term accounts for within-cluster correlation, and the third term properly subtracts off the within-strata means. Typically, the second term is positive, reflecting the positive correlation within cluster. The third term, without the minus sign, is always nonnegative. Therefore, the second and third terms tend so work in opposite directions. In any case, the resulting test statistic,

$$\frac{N^{-1/2}\sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}} v_{sc}\cdot r_{scm}\left(\mathbf{W}_{scm},\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_2\right)}{\hat{\xi}}, \tag{18}$$

is easy to compute once the difference in objective functions is obtained for each observation. We simply compute the standard error for the sample mean of $r_{scm}\left(\mathbf{W}_{scm},\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_2\right)$ under the complex sampling scheme. In Stata, the "svyset" command can be used to set the sample as including stratification and clustering, where the sampling weights are also specified.

Extending Bhattacharya's (2005) model to more complex sampling designs, Rahmani (2018) considers a sampling design with VP sampling in the final stage. The framework closely resembles several complex surveys, including the NSFH. The asymptotic variance becomes more complicated due to the final sampling stage. The formula for the asymptotic variance becomes even more complicated, mostly due to notation, but the adjustments are straightforward to describe. The weights are simply adjusted to reflect the final probability sampling stage, where an observation is further weighted by $1/p_j$, where j represents a stratum for the final VP sampling stage. The statistic still has the same general form as in (18) and is easily computed using standard software once the weights are properly adjusted for the VP sampling.

5. MODEL-SELECTION TESTS WITH PANEL DATA

Model-selection tests in panel data models with complex sampling designs are similar to the tests in the cross-sectional cases, but in using standard software we must make sure to account for serial correlation in the difference in objective functions when using a pooled estimation method. Here we cover the case where stratified sampling is done in an initial time period, as is very common. Consequently, the sampling weights, Q_j/H_j for the strata $j = 1, \ldots, J$, do not change over time.

When a probability density function for the joint distribution $\mathbb{D}(\mathbf{Y}_{i1}, \ldots, \mathbf{Y}_{iT}|\mathbf{X}_{i1}, \ldots, \mathbf{X}_{iT})$ is fully specified, the methods in Sections 5.3 and 5.4 apply directly: the objective function is the joint log likelihood conditional on $(\mathbf{X}_{i1}, \ldots, \mathbf{X}_{iT})$.

For many reasons, one often wants to compare models estimated using pooled methods. Pooled estimation methods are computationally simpler, often much more so. More importantly, we are often interested in a feature of $\mathbb{D}(\mathbf{Y}_{it}|\mathbf{X}_{it})$ or even $\mathbb{D}(\mathbf{Y}_{it}|\mathbf{X}_{i1},\ldots,\mathbf{X}_{iT})$, and we do not wish to take a stand on how the $\{\mathbf{Y}_{it}: t=1,...,T\}$ are related to each other. For example, we might be interesting in estimating $\mathbb{E}(\mathbf{Y}_{it}|\mathbf{X}_{it})$ or $\mathbb{E}(\mathbf{Y}_{it}|\mathbf{X}_{i1},\ldots,\mathbf{X}_{iT})$ using pooled quasi-MLE in the linear exponential family. Such an approach is robust to other distributional misspecification and to arbitrary serial correlation. Therefore, any model-selection statistic should be robust to arbitrary serial dependence, too.

As an example, suppose we use pooled nonlinear least squares to estimate two models of the conditional mean. The difference in objective functions at time t, evaluated at the pseudo-true values, is

$$\left[Y_{it} - m_1(\mathbf{X}_{it},\boldsymbol{\theta}_1^*)\right]^2 - \left[Y_{it} - m_2(\mathbf{X}_{it},\boldsymbol{\theta}_2^*)\right]^2.$$

There are essentially no interesting cases where this difference would be serially uncorrelated over time. We would have to assume that $\{(\mathbf{X}_{it},Y_{it}): t=1,\ldots,T\}$ is an independent sequence, and this is very unlikely in a panel data setting.

In models with unobserved heterogeneity, say $\mathbf{C}_i$, we can take a correlated random effects approach (as in Wooldridge, 2010, Section 13.9) and propose a model for

$$\mathbb{D}(\mathbf{C}_i|\mathbf{X}_{i1},\ldots,\mathbf{X}_{iT}) = \mathbb{D}(\mathbf{C}_i|\bar{\mathbf{X}}_i)$$

Then, if we assume strict exogeneity of $\{\mathbf{X}_{it}\}$ conditional on $\mathbf{C}_i$,

$$\mathbb{D}(\mathbf{Y}_{it}|\mathbf{X}_{i1},\ldots,\mathbf{X}_{iT},\mathbf{C}_i) = \mathbb{D}(\mathbf{Y}_{it}|\mathbf{X}_{it},\mathbf{C}_i)$$

then we obtain a model for

$$\mathbb{D}(\mathbf{Y}_{it}|\mathbf{X}_{i1},\ldots,\mathbf{X}_{iT}) = \mathbb{D}(\mathbf{Y}_{it}|\mathbf{X}_{it},\bar{\mathbf{X}}_i).$$

The outcome $\{\mathbf{Y}_{it}: t=1,\ldots,T\}$ would essentially never be independent conditionally on $\{\mathbf{X}_{i1},\ldots,\mathbf{X}_{iT}\}$, but we can still apply pooled MLE, pooled quasi-MLE or some other pooled estimation procedure.

With a time dimension, there is a subtle issue about the nature of the null hypothesis. One could imagine wanting to test the models against each other for each time period. Even if this is desirable, it would be tricky because the nature of the alternative – that one model fits at least as well in all time periods – would imply multiple inequality restrictions. Plus, what would we do if one model fits better in three time periods but the other model fits much better in two time periods? In the end, we would very likely average across time. Therefore, here we take the null to be that the two models fit the same when we average (or sum) across time.

To be precise, define pooled objective functions

$$q_g(\mathbf{W}_i,\boldsymbol{\theta}_g) = \sum_{t=1}^{T} q_{gt}(\mathbf{W}_{it},\boldsymbol{\theta}_g), \quad g = 1,2.$$

Then the null hypothesis is

$$\mathbb{E}\left[q_1(\mathbf{W}_i,\boldsymbol{\theta}_1^*)\right] = \mathbb{E}\left[q_2(\mathbf{W}_i,\boldsymbol{\theta}_2^*)\right], \tag{19}$$

where $\boldsymbol{\theta}_1^*$ and $\boldsymbol{\theta}_1^*$ are the pseudo-true values. A sufficient but not necessary condition is that the models fit equally well for each t:

$$\mathbb{E}\left[q_{1t}(\mathbf{W}_{it},\boldsymbol{\theta}_1^*)\right] = \mathbb{E}\left[q_{2t}(\mathbf{W}_{it},\boldsymbol{\theta}_2^*)\right], \quad t = 1,\ldots,T. \tag{20}$$

We will take (19) as the null hypothesis in what follows.

Under stratified sampling, it is easy to obtain a valid model-selection statistic that adds no additional assumptions to (19). In fact, we can simply apply the same statistics in (12) and (13). The estimate $\hat{\eta}^2$ automatically accounts for the serial correlation in

$$r(\mathbf{W}_{ij},\boldsymbol{\theta}_1^*,\boldsymbol{\theta}_2^*) = \sum_{t=1}^{T}\left[q_{1t}(\mathbf{W}_{itj},\boldsymbol{\theta}_1^*) - q_{2t}(\mathbf{W}_{itj},\boldsymbol{\theta}_2^*)\right].$$

This can be seen by expanding the term

$$\frac{1}{N_j}\sum_{i=1}^{N_j}\left[r(\mathbf{W}_{ij},\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_2) - \bar{R}_j\right]^2$$

and noting that it includes cross products between time periods t and s, $t \neq s$.

Standard software can be tricked into computing the model-selection statistic by specifying the strata, j, the sampling weights, Q_{j_i}/H_{j_i}, and specifying each cross-sectional unit i as a cluster. As is well known – see, for example, Arellano (1987) – the form of the robust variance estimator for small-T panel data estimators is the same as for cluster correlation.

6. EXOGENOUS STRATIFICATION

In most applications, we partition $\mathbf{W}$ as $\mathbf{W} = (\mathbf{X},\mathbf{Y})$ and we are interested in some feature of the distribution of $\mathbf{Y}$ given $\mathbf{X}$. If the feature is correctly specified, and we choose a suitable objective function, then the population (true) value of the parameters, $\boldsymbol{\theta}_o$, solves

$$\min_{\boldsymbol{\theta}\in\Theta} \mathbb{E}\,[q(\mathbf{W},\boldsymbol{\theta})|\mathbf{X}=\mathbf{x}]$$

for all $\mathbf{x}\in\mathcal{X}$, the support of $\mathbf{X}$. For example, in the case of estimating the conditional mean $\mathbb{E}(Y|\mathbf{X})$, one suitable choice of the objective function is the squared residual function:

$$q(\mathbf{W},\boldsymbol{\theta})=[Y-m(\mathbf{X},\boldsymbol{\theta})]^{2.}$$

When the conditional mean is correctly specified, that is,

$$\mathbb{E}\,(Y|\mathbf{X}=\mathbf{x})=m(\mathbf{x},\boldsymbol{\theta}_o),\ \mathbf{x}\in\mathcal{X},$$

it is easily shown – see, for example, Wooldridge (2010, Chapter 12) – that $\boldsymbol{\theta}_o$ solves

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbb{E}\{[Y-m(\mathbf{X},\boldsymbol{\theta})]^2\,|\mathbf{X}=\mathbf{x}\}$$

for all $\mathbf{x}$.

Now consider a situation where stratification is based entirely on $\mathbf{X}$, and so $\{\mathcal{X}_j : j=1,\ldots,J\}$ represents the mutually exclusive and exhaustive strata. Then, as discussed in Wooldridge (1999, 2001), $\boldsymbol{\theta}_o$ solves

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbb{E}\{[Y-m(\mathbf{X},\boldsymbol{\theta})]^2\,|\mathbf{X}\in\mathcal{X}_j\}$$

for each j. Wooldridge (1999, 2001) shows that this feature of $\boldsymbol{\theta}_o$ implies that the unweighted M-estimator is generally consistent for $\boldsymbol{\theta}_o$.

Given consistency of the unweighted estimator under correct specification, it may be tempting to ignore stratification when it is based on $\mathbf{X}$ and to simply apply Vuong's (1989) statistic in the M-estimation context. But the unweighted statistic does not achieve our objectives because the null hypothesis is that each model is misspecified. Even when stratification is based on $\mathbf{X}$, we need to use the weights to uncover $\boldsymbol{\theta}_1^*$ and $\boldsymbol{\theta}_2^*$. In particular, under the null hypothesis of interest, $\boldsymbol{\theta}_g^*$ does not generally solve

$$\min_{\boldsymbol{\theta}_g\in\boldsymbol{\Theta}_g} \mathbb{E}\left[q_g(\mathbf{W},\boldsymbol{\theta}_g)|\mathbf{X}=\mathbf{x}\right]$$

for all $\mathbf{x}\in\mathcal{X}$, and therefore the unweighted estimator is inconsistent for $\boldsymbol{\theta}_g^*$. Our goal is to compare the models in the population, and the weighted estimator always consistently estimates $\boldsymbol{\theta}_g^*$ under the null and alternatives – including if model g is correctly specified – whether or not stratification is based on $\mathbf{X}$, $\mathbf{Y}$, or both. To summarize, this observation argues in favor of weighting for both estimation and model selection.

7. EXAMPLES

The previous framework has many applications. Here we describe a few that are not completely standard.

Example 7.1: (Binary and Fractional Responses) Let Y_i be either a binary response or a fractional response, so $Y_i \in [0, 1]$. In either case, sensible models are of the form

$$\mathbb{E}\left(Y_i | \mathbf{X}_i\right) = F(\mathbf{X}_i \boldsymbol{\theta}_o)$$

where $F(\cdot) \in (0, 1)$ is typically a smooth, strictly increasing function. Common examples are $F(z) = \Phi(z)$ (the standard normal CDF), $F(z) = \Lambda(z)$ (the logistic CDF) and $F(z) = 1 - \exp[-\exp(z)]$ (complementary-log-log). The index form is not important but common. For example, we could include models that have heteroskedasticity in an underlying latent variable formulation, such as a heteroskedastic probit.

Whether Y_i is binary or fractional, maximizing the Bernoulli log likelihood is a common estimation method. In the case of a binary response, our hope is that we are using true MLE. In the case of a fractional response, estimation is clearly quasi-MLE, but we are interested in testing specification of the conditional mean. In the previous setup, the objective functions are the negative of the quasi-log likelihood functions. In particular

$$q_g(\mathbf{W}_i, \boldsymbol{\theta}_g) = -\left\{(1 - Y_i)\log\left[1 - F_g(\mathbf{X}_i \boldsymbol{\theta}_g)\right] + Y_i \log\left[F_g(\mathbf{X}_i \boldsymbol{\theta}_g)\right]\right\}.$$

Many software packages, such as Stata, allow for estimation of binary and fractional response models with survey sampling. After the estimates $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ have been obtained, compute

$$r(\mathbf{W}_i, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) = q_1(\mathbf{W}_i, \hat{\boldsymbol{\theta}}_1) - q_2(\mathbf{W}_i, \hat{\boldsymbol{\theta}}_2).$$

Then, using a survey option, run a weighted regression of $r(\mathbf{W}_i, \hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)$ on a constant and use the appropriate variance estimate to account for the stratified sampling or more complex forms of survey sampling.

In a panel data context where we use the correlated random effects approach, the models for comparison could be of the form

$$m_g(\mathbf{X}_{it}, \bar{\mathbf{X}}_i, \boldsymbol{\theta}_g) = F_g\left(\alpha_{gt} + \mathbf{X}_{it}\boldsymbol{\beta}_g + \mathbf{Z}_i \boldsymbol{\gamma}_g + \bar{\mathbf{X}}_i \boldsymbol{\xi}_g\right),$$

where $\mathbf{Z}_i$ are time constant variables and $\bar{\mathbf{X}}_i = T^{-1}\sum_{t=1}^{T} \mathbf{X}_{it}$ are the time averages of the time-varying covariates. Whether Y_{it} is binary or fractional, one can use a pooled QMLE based on the Bernoulli quasi-LLF. If there is no cluster sampling, we only need to specify the PSUs, the sampling weights, and taking each unit be its own cluster to account for serial dependence in the objective function.

Example 7.2 (Gamma versus Lognormal) If $Y_i > 0$ is continuous, and we wish to fit its entire distribution, then it makes sense to compare to fully specified PDFs. Two leading cases are the gamma and lognormal distributions (although one could use others, such as the Weibull). Both are two-parameter families and can be parameterized in terms of the mean and an additional dispersion parameter. Typically, the mean would be parameterized as $\exp(\mathbf{X}_i \boldsymbol{\beta}_g)$ in both cases, with dispersion parameters $\tau_g^2 > 0$. Once the MLEs have been computed, the model-selection statistic, appropriately adjusted for weights, stratification, and clustering, is easily obtained.

Example 7.3 (Nonlinear Least Squares) Suppose we want to model $\mathbb{E}(Y_i|\mathbf{X}_i)$ using functions $m_1(\mathbf{X}_i, \boldsymbol{\theta}_1)$ and $m_2(\mathbf{X}_i, \boldsymbol{\theta}_2)$. Let $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ be the weighted nonlinear least squares estimators, obtained using the sampling weights (if appropriate). For each unit i, the difference in objective functions is

$$\hat{R}_i = \left[Y_i - m_1(\mathbf{X}_i, \hat{\boldsymbol{\theta}}_1)\right]^2 - \left[Y_i - m_2(\mathbf{X}_i, \hat{\boldsymbol{\theta}}_2)\right]^2 = \hat{U}_{i1}^2 - \hat{U}_{i2}^2,$$

the difference in squared residuals. The model-selection statistic is based on the difference in SSRs, obtained from the regression

$$\hat{U}_{i1}^2 - \hat{U}_{i2}^2 \text{ on } 1, \quad i = 1, \ldots, N$$

using weights, if necessary, and adjusting the standard error of the constant for the sampling scheme. For example, if $Y_i \geq 0$, possibly taking on the value zero, it is somewhat common to start with a linear model estimated by OLS. That can be compared with an exponential model estimated by OLS.

8. A SMALL SIMULATION STUDY

In this section, we present findings of a small simulation study. Our goal is to show that the weighted version of Vuong's statistic can help choose a correctly specified model when one of the two models is correctly specified. We also find evidence that the weighted statistic often chooses the model that is the better approximation to the truth. Just as importantly, we will see that using the unweighted statistic can be very misleading.

In the first set of simulations, we create a population of 100,000 units, where the outcome variable, Y, follows a Poisson distribution conditional on a set of covariates. We consider two different conditional mean functions. In the first case, we generate five covariates, all normally distributed, such that

$$\mathbb{E}\left(Y|X_1, X_2, X_3, X_4, X_5\right) = \exp\left(0.5 + 0.5X_1 + 0.5X_2 + 0.2X_3\right), \tag{21}$$

so that the true model includes X_3 but excludes X_4 and X_5. We call this model one. We chose the parameter values so that the test does not choose the correct model with probability one but still has substantial power in its direction. As competing models, we replace X_3 first with X_4 (model two) and then with X_5 (model three):

$$\begin{aligned} m_2(x_1, x_2, x_3, x_4, x_5) &= \exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_4\right) \\ m_3(x_1, x_2, x_3, x_4, x_5) &= \exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_5\right) \end{aligned}$$

We generate X_4 and X_5 to make it somewhat difficult to distinguish among the models by setting

$$\begin{aligned} X_4 &= X_3 + R_4 \\ X_5 &= X_3 + R_5, \end{aligned}$$

where R_4 and R_5 are independent Normal$(0, 1/9)$ random variables. Models two and three fit equally well using any objective function that suitably identifies a conditional mean function, such as a quasi-log likelihood from the linear exponential family.

In order to study the performance of the test in a quasi-MLE framework, we also generated the conditional distribution of Y to be exponential in the population, with the same mean function (21). We must emphasize that we are still using the Poisson log-likelihood function, so the estimator, in this case, is properly called QMLE. As in the Poisson case, there is no guarantee that the weighted estimator will choose the correct model one more frequently than the unweighted test. But we know it will not systematically choose an incorrect conditional mean over a correct one.

Rather than drawing a random sample, we stratify the sample on the basis of Y. In particular, for the Poisson distribution, we take samples of $1,000$ from the stratum with $Y = 0$ and $1,000$ from the stratum with $Y > 0$. In the population, $\mathbb{P}(Y = 0) = 0.19061$. Therefore, we oversample the stratum with $Y = 0$. There are only two strata, and the sampling weights are $Q_1 = 0.38122$ and $Q_2 = 1.61878$. For

Table 1. Rejection Frequencies, First Specification.

DGP	**Poisson**		**Exponential**	
Estimation Method	**Poisson MLE**		**Poisson QMLE**	
Test Version	**Weighted**	**Unweighted**	**Weighted**	**Unweighted**
$m_1 > m_2$	0.782	0.689	0.264	0.229
$m_2 > m_1$	0.000	0.000	0.004	0.003
$m_1 > m_3$	0.730	0.658	0.189	0.180
$m_3 > m_1$	0.000	0.000	0.006	0.004
$m_2 > m_3$	0.040	0.039	0.033	0.042
$m_3 > m_2$	0.048	0.051	0.055	0.040

the exponential distribution, we choose the strata so that the population frequencies are about 0.727 and 0.273, using a cutoff $Y \leq 2$. Therefore, we oversample units with larger outcomes, rather than small ones.

For each draw, we use both the weighted Vuong statistic and the unweighted statistic, both based on the Poisson log-likelihood function with exponential conditional mean. We test each model against the other two, so we have six outcomes. Table 1 reports rejection frequencies obtained from the simulations using $1,000$ replications. Here the alternative is that model m_i is better than m_j, $i \neq j$, or in short $m_i > m_j$.

As can be seen in Table 1, when the population distribution is Poisson, the weighted test does a better job than the unweighted test in detecting that model one provides the best fit – because it is the true model. For model one versus model two, the rejection in favor of model one is almost 10% points higher (78.2% versus 68.9%). Similarly, the weighted test does better in choosing between models one and three (73.0% versus 65.8%). Neither test ever incorrectly chooses model two or model three over model one. Remember, though, that the estimates of the parameters using the unweighted estimator are inconsistent because stratification is based on Y.

Both the weighted and unweighted tests have rejection frequencies close to 0.05 when comparing the two incorrect models, model two and model three, which corresponds to the notion that both models are wrong but fit equally well. The weighted statistic does find a few more "false positives" than the unweighted test, but there are only 1,000 replications. Overall, the weighted test seems clearly preferred, and we must use the weights for consistent parameter estimation, anyway.

When the true conditional distribution is exponential, both tests have a tougher time choosing model one. This possibly is a result that, for the exponential distribution, the variance is the square of the mean, and so there is much more variation

Table 2. Rejection Frequencies, Second Specification.

DGP	**Poisson**		**Exponential**	
Estimation Method	**Poisson MLE**		**Poisson QMLE**	
Test Version	**Weighted**	**Unweighted**	**Weighted**	**Unweighted**
$m_1 > m_2$	0.525	0.000	0.015	0.005
$m_2 > m_1$	0.000	0.333	0.002	0.034
$m_1 > m_3$	0.013	0.006	0.002	0.001
$m_3 > m_1$	0.015	0.012	0.002	0.000
$m_2 > m_3$	0.000	0.297	0.001	0.023
$m_3 > m_2$	0.581	0.001	0.024	0.013

in the outcome Y than when the conditional distribution is Poisson. Plus, in the exponential case, we oversample large outcomes rather than small ones (although the weighted version of the test accounts for that). Overall, the weighted test does somewhat better. For example, it correctly chooses model one over model two 26.4% of the time compared with 22.9% for the unweighted test.

As a second conditional mean specification, we use

$$\mathbb{E}\,(Y|X_1,X_2,X_3,X_4,X_5) = \exp\,(0.5 + 0.5X_1 + 0.5X_2 + 0.3X_3^{-1}), \tag{22}$$

where X_1 and X_2 have the same normal distributions and $X_3 \sim \text{Uniform}(1,3)$. Now we are primarily interested in the ability of the test to detect functional form misspecification. As before, the correct conditional mean function is labeled model one. The alternative models are

$$\begin{aligned} m_2(x_1,x_2,x_3) &= \exp\,(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2) \\ m_3(x_1,x_2,x_3) &= \exp\,(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_3^2). \end{aligned}$$

Model two ignores x_3 entirely; given the simulation findings in Table 1, we would expect the test to do well in choosing model one. Model three misspecifies the functional form in x_3. As a quadratic can mimic the reciprocal function, we expect the test to have a more difficult time telling apart models one and three. In the Poisson population, we still oversample units with $Y = 0$. The strata in the exponential case are defined by $Y \leq 4$ and $Y > 4$.

The findings in Table 2 for the Poisson distribution are very interesting and highlight the danger of using the unweighted version of the test. The weighted test is fairly successfully distinguishing between models one and two: it correctly rejects the null in favor of model one 52.5% of the time, and never chooses model two. By contrast, the unweighted test never correctly picks model one. It even picks model two 33.3% of the time. That means that a researcher is much more likely to think that the correct model is quadratic in X_1 and X_2 and entirely excludes X_3.

Both the weighted and unweighted tests have very little ability to tell the difference between models one and three. This is unlikely to be a bad thing because quantities of interest – such as elasticities, semi-elasticities and average partial effects – are probably pretty similar across the two models. The weighted test shows a clear preference for model three over model two, and this is a good thing: model three is certainly closer to the true model. By contrast, the unweighted test incorrectly shows a clear preference for model two over model three.

Both tests are completely ineffective for selecting among the three models when the data are generated from the exponential distribution. As before, this probably arises from the large variance in an exponential distribution and possibly the oversampling of large outcomes from the population.

9. CONCLUSION

We have extended Vuong's (1989) in several useful directions. First, we allow for general M-estimation rather than maximum likelihood estimation. Second, we allow for complex survey samples rather than assuming random sampling from a population. Third, we allow panel data applications combined with survey sampling.

The key to obtaining computationally simple tests is contained in Theorem 3.1, which shows that when the models are appropriately nonnested and they fit equally well, the limiting distribution of the standardized difference in objective functions is nondegenerate and does not depend on the limiting distributions of the estimators themselves. This means we can apply standard asymptotic variance estimators for stratified samples, cluster samples and combinations of these directly to the differences in the unit-specific objective functions.

Section 5.7 contains just a couple of examples that show how the results can be applied to problems that are explicitly quasi-MLE in nature, including popular fractional response models and models for nonnegative responses.

For the most part, the simulation results in Section 5.8 are promising. In addition to providing consistent estimators of the pseudo-true values, weighting the objective function generally allows us to better choose the best fitting model in cases where the best fitting model is the true model or the best fitting model is "close" to the true model. In one case, the unweighted test systematically selects the worst of the three models while almost never choosing the correct model.

More simulations could be informative. For example, seeing what happens when stratification is based on **X** is something we did not do.

There are several interesting directions for future research. First, it would be helpful to study the finite-sample properties of the version of the test statistic that penalizes the number of parameters – see (14). Second, our setup can be extended to the case where the goodness-of-fit functions are not the same as the objective functions used to obtain the $\hat{\boldsymbol{\theta}}_g$. For example, in a Tobit model, one might maximize the log likelihood but then want to make comparisons based on the conditional mean, in which case we might want to compare a sum of squared residuals from a Tobit to that from, say, an exponential mean estimated using Poisson QMLE. The analog of Theorem 3.1 will not be as simple, but such extensions seem worthwhile.

Finally, as suggested by a reviewer, rather than relying on standard first-order asymptotics, one could possibly bootstrap the test statistic. Given the nature of the null hypothesis and the complex survey sampling, this poses an interesting challenge for the future.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd international symposium on information theory*, Budapest (pp. 267–281).

Arellano, M. (1987). Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics*, *49*, 431–434.

Bhattacharya, D. (2005). Asymptotic inference from multi-stage samples. *Journal of Econometrics*, *126*, 145–171.

Cox, D. R. (1961). Tests of separate families of hypotheses. In L. M. LeCam, J. Neyman, E. L. Scott (Eds.), *Proceedings of the 4th berkeley symposium on mathematical statistics and probability, Vol. 1*, , University of California, Berkeley Press (pp. 105–123).

Cox, D. R. (1962). Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society, Series B*, *24*, 406–424.

Davidson, R., & MacKinnon, J. G. (1981). Several tests for model specification in the presence of alternative hypotheses. *Econometrica*, *49*, 781–793.

Findley, D. F. (1990). *Making difficult model comparisons*. Mimeo, U.S. Bureau of the Census.

Findley, D. F. (1991). Convergence of finite multistep predictors from incorrect models and its role in model selection. *Note di Matematica*, *XI*, 145–155.

Findley, D. F., & Wei, C. Z. (1993). Moment bound for deriving time series CLT's and model selection procedures. *Statistica Sinica*, *3*, 453–480.

Gourieroux, C., Monfort, A., & Trognon, C. (1984). Pseudo-maximum likelihood methods: Applications to Poisson models. *Econometrica*, 52, 701–720.

Newey, W. K., & McFadden, D. (1994). Large sample estimation and hypothesis testing. In R. F. Engle & D. L. McFadden (Eds.), *Handbook of econometrics* (Vol. IV, pp. 2111–2245). Amsterdam: North-Holland Publishing.

Rahmani, I. (2018). Asymptotic Inference of M-Estimator from Multistage Samples with Variable Probability in the Final Stage, Working Paper.

Rivers, D., & Vuong, Q. (2002). Model selection tests for nonlinear dynamic models. *The Econometrics Journal 5*, 1–39.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464.

Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, *57*, 307–333.

White, H. (1982). Maximum likelihood estimation of misspecified models. Econometrica 50, 1-25.

Wooldridge, J. M. (1990). An encompassing approach to conditional mean tests with applications to testing nonnested hypotheses. *Journal of Econometrics*, *45*, 331–350.

Wooldridge, J. M. (1999). Asymptotic properties of weighted M-estimators for variable probability samples. *Econometrica*, *67*, 1385–1406.

Wooldridge, J. M. (2001). Asymptotic properties of weighted M-estimators for standard stratified samples. *Econometric Theory*, *17*, 451–470.

Wooldridge, J. M. (2010). *Econometric analysis of cross-section and panel data* (2nd ed.). Cambridge, MA: MIT Press.