Econometric Reviews

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/lecr20

Inference in the presence of unknown rates

Hao Dong, Taisuke Otsu & Luke Taylor

To cite this article: Hao Dong, Taisuke Otsu & Luke Taylor (24 Dec 2024): Inference in the presence of unknown rates, Econometric Reviews, DOI: 10.1080/07474938.2024.2434189

To link to this article: https://doi.org/10.1080/07474938.2024.2434189

ECONOMETRIC REVIEWS
https://doi.org/10.1080/07474938.2024.2434189

Taylor & Francis
Taylor & Francis Group

Inference in the presence of unknown rates

Hao Dong[a], Taisuke Otsu[b], and Luke Taylor[c]

[a]Department of Economics, Southern Methodist University, Dallas, Texas, USA; [b]Department of Economics, London School of Economics, London, UK; [c]Department of Economics and Business Economics, Aarhus University, Aarhus, Denmark

ABSTRACT

The convergence rate of an estimator can vary when applied to datasets from different populations. As the population is unknown in practice, so is the corresponding convergence rate. In this article, we introduce a method to conduct inference on estimators whose convergence rates are unknown. Specifically, we extend the subsampling approach of Bertail, Politis, and Romano (1999) to situations where the convergence rate may include logarithmic components. This extension proves to be particularly relevant in certain statistical inference problems. To illustrate the practical relevance and implementation of our results, we discuss two main examples: (i) non parametric regression with measurement error; and (ii) intercept estimation in binary choice models. In each case, our approach provides robust inference in settings where convergence rates are unknown; simulation results validate our findings.

ARTICLE HISTORY
23 July 2024
20 November 2024

KEYWORDS
Binary choice; convergence rate; measurement error; subsampling

1. Introduction

To conduct valid statistical inference on an estimator of a population, object requires quantifying the estimator's uncertainty. Conventionally, the asymptotic variance of the estimator is used to approximate its finite sample variance. Unfortunately, oftentimes, the asymptotic variance is unknown, difficult to estimate, or inaccurate in finite samples. In such cases, bootstrap methods are employed instead. And when the bootstrap fails, all eyes turn to subsampling (Politis and Romano, 1994).

Subsampling is a robust solution that is valid under minimal assumptions. However, it requires knowledge of the convergence rate of the estimator if one is unwilling — or unable — to standardize the estimator using an estimate of the variance. Bertail, Politis, and Romano (1999) proposed a method to estimate the convergence rate when the rate is polynomial in the sample size n. More concretely, let $\{X_l\}_{l=1}^n$ be a random sample of $X \sim \mathcal{P}_\theta$, and T_n be a consistent estimator of a parameter θ based on the sample with

$$n^{\beta_1}(\log n)^{\beta_2}(T_n - \theta) \xrightarrow{d} L,$$

where β_1 and β_2 depend on $\mathcal{P}_\theta$. Suppose that the limit law L has cumulative distribution function $K(x, \mathcal{P}_\theta)$. If β_1 and β_2 are known, $K(x, \mathcal{P}_\theta)$ can be approximated by subsampling as in Politis and Romano (1994).

If β_1 is unknown and $\beta_2 = 0$, Bertail, Politis, and Romano (1999) proposed a method to estimate β_1 by comparing distributions from varying subsample sizes. The distribution $K(x, \mathcal{P}_\theta)$ can then be approximated via subsampling based on the estimate of β_1. In this article, we extend their method to the

CONTACT Taisuke Otsu t.otsu@lse.ac.uk Department of Economics, London School of Economics, Houghton Street, London, WC2A 2AE, UK.

general case where β_1 and β_2 are both unknown, show the consistency of our estimators of β_1 and β_2, and establish the asymptotic validity of subsampling inference using these estimated convergence rates.

Before presenting our main result in Section 2, we close this section with some motivating examples of our inference method. We provide full details of the examples to be used in our numerical illustrations in Section 3, as well as some briefer comments on further examples. Note that this list of examples is far from exhaustive; additional examples that we do not discuss include Chernozhukov and Fernndez-Val (2011), Armstrong (2013), Sasaki and Ura (2021), Sasaki and Wang (2022), and Dobronyi, Ouyang, and Yang (2024).

1.1. Non parametric regression with errors-in-variables

Consider the non parametric regression model with error-in-variables

$$E[Y|X^*] = g(X^*), \quad X = X^* + \epsilon, \quad W = X^* + \nu,$$

where $(Y, X, W) \in \mathbb{R}^3$ are observable and $(X^*, \epsilon, \nu) \in \mathbb{R}^3$ are unobservable. X and W are noisy measurements of the unobserved X^* with measurement errors ϵ and ν, respectively; ϵ and ν are classical measurement errors in the sense that they are independent of X^*. In this case, to estimate the regression function $g(\cdot)$ using a random sample $\{Y_j, X_j, W_j\}_{j=1}^n$ of (Y, X, W), it is common to use

$$\hat{g}(x) = \frac{\sum_{j=1}^n \hat{\mathbb{K}}\left(\frac{x-W_j}{b_n}\right) Y_j}{\sum_{j=1}^n \hat{\mathbb{K}}\left(\frac{x-W_j}{b_n}\right)},$$

where $\hat{\mathbb{K}}(u) = \frac{1}{2\pi} \int e^{itu} \frac{K^{\mathrm{ft}}(t)}{\hat{f}_\epsilon^{\mathrm{ft}}(t/b_n)} dt$ is known as a deconvolution kernel, K is a conventional kernel function, b_n is a bandwidth parameter, and $\hat{f}_\epsilon$ is an estimate of the characteristic function f_ϵ^{ft} of ϵ, based on $\{X_j, W_j\}_{j=1}^n$. For example, based on 1967's (1967) identity, under mean independence between measurement errors, Schennach (2004) suggested

$$\hat{f}_\epsilon^{\mathrm{ft}}(t) = \frac{\sum_{j=1}^n e^{\mathrm{i}tX_j}}{n \exp\left(\int_0^t \frac{\sum_{j=1}^n X_j e^{\mathrm{i}sW_j}}{\sum_{j=1}^n e^{\mathrm{i}sW_j}} ds\right)}.$$

It is known that the convergence rate of $\hat{g}$ depends on the smoothness of the density f_ϵ of ϵ, the density f of X^*, and the regression function g. In particular, following Schennach (2004), suppose $|\{gf\}^{\mathrm{ft}}(t)| \leq d_1(1+|t|)^{-\gamma}$ for $t \in \mathbb{R}$ and positive constant γ, if $d_0(1+|t|)^{-\gamma_{x,0}} \leq |f^{\mathrm{ft}}(t)| \leq d_1(1+|t|)^{-\gamma_{x,0}}$ and $d_0(1+|t|)^{-\gamma_\epsilon} \leq |f_\epsilon^{\mathrm{ft}}(t)| \leq d_1(1+|t|)^{-\gamma_\epsilon}$ for $t \in \mathbb{R}$ and some positive constants $d_0, d_1, \gamma_\epsilon$ and $\gamma_{x,0}$, using a bandwidth of order $O\left(n^{\frac{1}{-2\gamma_{x,0}-\gamma_\epsilon+\gamma}}\right)$, we have

$$n^{\beta_1}\{\hat{g}(x) - g(x)\} \xrightarrow{d} N(b_1(x), v_1(x)),$$

where $\beta_1 = \frac{\gamma+1}{-2\gamma_{x,0}-\gamma_\epsilon+\gamma}$. If $d_0 \exp(a|t|^{\gamma_{x,1}}) \leq |f^{\mathrm{ft}}(t)| \leq d_1 \exp(a|t|^{\gamma_{x,1}})$ and $d_0(1+|t|)^{-\gamma_\epsilon} \leq |f_\epsilon^{\mathrm{ft}}(t)| \leq d_1(1+|t|)^{-\gamma_\epsilon}$ for $t \in \mathbb{R}$ and some positive constants $a, d_0, d_1, \gamma_\epsilon$, and $\gamma_{x,1}$, using a bandwidth of order $O\left(\{\log n\}^{1/\gamma_{x,1}}\right)$, we have

$$(\log n)^{\beta_2}\{\hat{g}(x) - g(x)\} \xrightarrow{d} N(b_2(x), v_2(x)),$$

where $\beta_2 = \frac{\gamma+1}{\gamma_{x,1}}$. These results can be summarized as

$$n^{\beta_1}(\log n)^{\beta_2}\{\hat{g}(x) - g(x)\} \xrightarrow{d} L,$$

where L is a normally distributed random variable with potentially non zero mean and non unit variance. Note that these derivations require knowledge of the optimal bandwidth, in which case, we would also

know the values of β_1 and β_2, so estimating them using our method would be unnecessary. However, in practical cases, the optimal bandwidth is typically unknown because f_ϵ, f, and g are unknown. Thus, the convergence rate will differ from the above expression and, more importantly, will be unknown.

Self-normalizing $\hat{g}$ prior to subsampling could be an option to allow inference; however, since $\hat{g}$ is a non linear function of multiple analog estimators, it must be linearized, resulting in an influence function that depends on unknown densities and conditional expectations in a complex manner; see equations (24)–(27) of Schennach (2004). Estimating such a complex function is challenging, making self-normalization of $\hat{g}$ infeasible in practice. Furthermore, as Kato and Sasaki (2019) discussed, no valid bootstrap procedure has been found for this setting.

1.2. Estimating the intercept of a binary choice model

Consider the binary choice model

$$Y = \mathbb{I}\{\alpha + Z - U \geq 0\},$$

where $(Y, Z) \in \mathbb{R}^2$ are observable, and $U \in \mathbb{R}$ is unobservable. U has zero mean, a strictly increasing cumulative distribution function, and is independent of Z. To estimate the intercept α, following Lewbel (1997), we can use

$$\hat{\alpha} = \frac{1}{n}\sum_{j=1}^{n} \frac{Y_j - \mathbb{I}\{Z_j > 0\}}{\hat{h}(Z_j)} \mathbb{I}\{|Z_j| \leq \tau_n\},$$

where $\hat{h}$ is an estimator of the density of Z (e.g., the kernel density estimator) and τ_n is a trimming sequence. Khan and Tamer (2010) showed that

$$n^{\beta_1}(\log n)^{\beta_2}(\hat{\alpha} - \alpha) \xrightarrow{d} L,$$

where L is a normally distributed random variable with potentially non zero mean and non unit variance, and values of β_1 and β_2 depend on the tail behavior of the densities of Z and U. In particular, when both Z and U have a standard logistic distribution, $\beta_1 = 0.5$ and $\beta_2 = -0.5$; when Z has a standard normal distribution and U has a standard logistic distribution, $\beta_1 = 0.25$ and $\beta_2 = -0.25$; and when Z has a Cauchy distribution and U has a standard logistic distribution, $\beta_1 = 0.5$ and $\beta_2 = 0$, i.e., the regular parametric rate. In practice, we do not know the values of β_1 and β_2 because the distributions of Z and U are unknown so conventional subsampling is infeasible. While the asymptotic variance of this estimator is simple to compute, as show in Lewbel (1997) is can be inaccurate in finite samples and its validity requires strong conditions on the tail index of the error distribution, casting doubt on using a self-normalized statistic. Furthermore, as shown in Khan and Nekipelov (2022) and Heiler and Kazak (2021), estimators of irregularly identified objects (such as the intercept in this model), do not admit valid bootstrap inference.

1.3. Other examples

1.3.1. Inverse probability weighting

Sasaki and Ura (2022) considered estimation of moments of the form $E[B/A]$. A common example is the mean potential outcome given by $E[Y(1)] = E[DY/p(X)]$, where D is a binary treatment, $Y(1)$ is a potential outcome for $D = 1$, Y is an observable outcome, X is a vector of covariates, and $p(X) = P(D = 1|X)$ is the propensity score. When dividing by a probability, care must be taken to deal with cases where the probability is close to zero; typically, this is achieved by trimming away observations where the denominator is below some threshold. Unlike Khan and Tamer (2010), the trimming bias is explicitly corrected by Sasaki and Ura (2022). They then derive the asymptotic normality for the standardized/self-normalized bias-corrected trimmed estimator, and use the normal asymptotic approximation for inference.

Ma and Wang (2020) also considered a trimmed version of the inverse probability weighting (IPW) estimator, and derive the asymptotic distribution when the propensity score is known. Unlike Sasaki and Ura (2022), inference is conducted using subsampling for the standardized estimator (i.e., the variance has been estimated before subsampling). Heiler and Kazak (2021) also derived the asymptotic distribution of the IPW estimator, but do not trim and allow the propensity score to be estimated. They showed that the convergence rate is $[n^{\alpha-1}/l(n)]^{\frac{1}{\alpha}}$, where $l(\cdot)$ is a slowly varying function (e.g., $\log(\cdot)$) and α is the — unknown — tail index of the asymptotic distribution of the estimator. For inference, they used the m out of n bootstrap (with replacement) for the self-normalized estimator. According to Bickel, GÖtze, and van Zwet (2012), it is possible to obtain a similar result under weaker assumptions using our subsampling procedures.

1.3.2. Sample selection model

Khan and Nekipelov (2022) proposed a closed-form estimator for the intercept of the outcome equation in a sample selection model; see, e.g., Heckman (1990) and Andrews and Schafgans (1998) for the practical importance of the intercept in such models. While this estimator is consistent over large classes of error distributions, it will have a rate of convergence that depends on the tail behavior of the instrument and the joint distribution of the error terms — something that is inherently unobservable. The convergence rate ranges from $\sqrt{n}$ to $\log n^{\kappa}$ for some $\kappa > 0$. Khan and Nekipelov (2022) went on to show that any intercept estimator for this model that is uniformly consistent over such a class of error distributions is not compatible with inference using pivotal statistics or the bootstrap. In answer to this, they developed a novel form of inference termed locally uniform inference based on drifting parameter asymptotics. We note, however, that the subsampling approach of this article is applicable under weaker assumptions than they impose.

1.3.3. Conditional moment inequality models

Armstrong (2015) proposed a Kolmogorov–Smirnov-style test for conditional moment inequality models when the parameters may be on the boundary of the parameter space. To determine critical values for his test, he noted that the convergence rate of the statistic depends on unknown quantities; thus, he first estimates the convergence rate. However, in Theorem 5.1, he showed that if the moment function is not smooth enough, the rate is at least as slow as $n^{(1+p)/(1+2p)}$, where p is the number of bounded derivatives of the moment function, so that logarithmic rates cannot be ruled out. Consequently, he adjusted the approach of Bertail, Politis, and Romano (1999) by truncating the convergence rate from above whenever the polynomial rate requirement of the approach appears to be violated. Consequently, although the tests proposed by Armstrong (2015) are exact when the convergence rate is polynomial, when truncation is applied (i.e., when the rate is logarithmic), the test is conservative. By using our approach, it is likely that his test could be exact in all cases with no unnecessary loss in power for logarithmic settings.

2. Main result

We first present our estimation method for the convergence rates, considering the general case when θ belongs to a normed linear space with norm $\|\cdot\|$. Let $\{X_l\}_{l=1}^n$ be a random sample of $X \sim \mathcal{P}_\theta$, and T_n be a consistent estimator of θ based on $\{X_l\}_{l=1}^n$ at rate $\tau(n) = n^{\beta_1}(\log n)^{\beta_2}$ with

$$P(\tau(n)\|T_n - \theta\| \le x) \to K(x, \mathcal{P}_\theta)$$

for some $K(x, \mathcal{P}_\theta)$ continuous in x, where β_1 and β_2 are unknown parameters depending on $\mathcal{P}_\theta$. As in Bertail, Politis, and Romano (1999), our rate estimator is constructed using the empirical distribution function of subsampled statistics of T_n, that is

$$K_{b_n}(x;\tau) = \frac{1}{q}\sum_{s=1}^{q} \mathbb{I}\{\tau(b_n)\|T_{b_n,s} - T_n\| \le x\},$$

where $\{T_{b_n,s}\}_{s=1}^{q}$ are values of the statistic T_n applied to subsamples with a subsample size b_n. Note that 'subsampling' here refers to the random draw of b_n observations from the full sample. Let $K_{b_n}(x) = K_{b_n}(x;1)$, and $K^{-1}(t,\mathcal{P}_\theta)$ and $K_{b_n}^{-1}(t)$ be the t-th quantiles of $K(x,\mathcal{P}_\theta)$ and $K_{b_n}(x)$, respectively. Under mild conditions presented below, an application of Bertail, Politis, and Romano (1999, Lemma 1) implies the following relationship for these quantiles:

$$\log K_{b_n}^{-1}(t) = \log K^{-1}(t,\mathcal{P}_\theta) - \beta_1 \log b_n - \beta_2 \log\log b_n + o_p(1).$$

Taking sequences of quantile points $\{t_j\}_{j=1}^{J}$ and subsample sizes $\{b_{in}\}_{i=1}^{I}$ and averaging over j, this relation becomes

$$\underbrace{\frac{1}{J}\sum_{j=1}^{J}\log K_{b_{in}}^{-1}(t_j)}_{y_i} = \underbrace{\frac{1}{J}\sum_{j=1}^{J}\log K^{-1}(t_j,\mathcal{P}_\theta)}_{\beta_0} - \beta_1 \underbrace{\log b_{in}}_{x_i} - \beta_2 \underbrace{\log\log b_{in}}_{z_i} + o_p(1), \tag{1}$$

for $i = 1,\ldots,I$. Based on this, we propose to estimate β_1 and β_2 by the OLS estimator for the regression of y_i on $(1, x_i, z_i)$, denoted by $\hat{\beta}_1(\{b_{in}\}_{i=1}^{I})$ and $\hat{\beta}_2(\{b_{in}\}_{i=1}^{I})$, respectively. Compared to Bertail, Politis, and Romano (1999), we introduce an additional regressor $z_i = \log\log b_{in}$ to estimate the logarithmic convergence rate β_2.

To derive the convergence rates of these estimators, we impose the following assumptions.

Assumption.

(i): $\{X_l\}_{l=1}^{n}$ *is a random sample of* $X \sim \mathcal{P}_\theta$.
(ii): $P(n^{\beta_1}(\log n)^{\beta_2}\|T_n - \theta\| \le x) \to K(x,\mathcal{P}_\theta)$ *for constants* (β_1,β_2) *such that* $n^{\beta_1}(\log n)^{\beta_2} \to \infty$.
(iii): $K(x,\mathcal{P}_\theta)$ *is continuous in* x *and strictly increasing on* (k_0,k_1) *as a function of* x, *where* $k_0 = \sup\{x : K(x,\mathcal{P}_\theta) = 0\}$ *and* $k_1 = \inf\{x : K(x,\mathcal{P}_\theta) = 1\}$.
(iv): $b_n \to \infty$ *and* $b_n/n \to 0$ *as* $n \to \infty$.

Assumptions (i) and (iii) are taken from Politis and Romano (1994). Assumption (i) may be relaxed to allow weakly dependent data by constructing subsamples for consecutive observations. Assumption (iii) is a standard condition to establish the validity of subsampling approximations. Assumption (ii) is new in that both β_1 and β_2 are considered unknown, a crucial characteristic shared by the examples discussed in Section 1. Assumption (iv) also originates from Politis and Romano (1994), but its interpretation differs slightly in our setting. In particular, to estimate β_1 and β_2 accurately, we need $b_{in} \to \infty$ and $b_{in}/n \to 0$ for Eq. (1) to hold for all $i = 1,\ldots,I$. Theorem 1 below demonstrates that, based on our choice of b_{in}, Assumption (iv) remains sufficient. This indicates that we do not require additional assumptions on the subsample size b_n compared to those given in Politis and Romano (1994), even though the convergence rate is estimated. A similar observation was made by Bertail, Politis, and Romano (1999) in a simpler scenario where β_2 is known to be 0.

Under these assumptions, the consistency and convergence rates of our rate estimators are obtained as follows.

Theorem 1. *Under Assumptions (i)–(iv), it holds that for* $0 < \gamma_1 < \cdots < \gamma_I < 1$,

$$\hat{\beta}_1(\{n^{\gamma_i}\}_{i=1}^{I}) - \beta_1 = o_p((\log n)^{-1}),$$
$$\hat{\beta}_2(\{\exp((\log n)^{\gamma_i})\}_{i=1}^{I}) - \beta_2 = o_p((\log\log n)^{-1}).$$

The proof of this theorem is presented in Appendix A.1. This theorem is a generalization of Bertail, Politis, and Romano (1999, Theorem 1) to the case where β_2 may be non zero. In accordance with Bertail, Politis, and Romano (1999), we employ a series of subsamples of varying size to estimate the convergence rate of T_n. Unlike Bertail, Politis, and Romano (1999), however, here we require different sets of

subsamples to estimate β_1 and β_2. In particular, we use $\{n^{\gamma_i}\}_{i=1}^I$ to estimate β_1 and $\{\exp((\log n)^{\gamma_i})\}_{i=1}^I$ for β_2. These different choices are necessary to ensure the fastest convergence rate for each of the estimators of β_1 and β_2. Consequently, separate regressions are required for β_1 and β_2. We first run a regression of y_i on $(1, x_i, z_i)$ as defined in Eq. (1) using $\{n^{\gamma_i}\}_{i=1}^I$, where the coefficient on x_i gives the estimate $\hat{\beta}_1$; the regression is repeated using $\{\exp((\log n)^{\gamma_i})\}_{i=1}^I$ and the coefficient on z_i gives the estimate $\hat{\beta}_2$. Although the estimate of β_2 from the first regression is consistent, the convergence rate for the estimate of β_2 is faster in the second regression; the same is true for β_1 in the first regression versus the second.

For the remainder of the article, we suppress the dependence on the subsample sizes and let $\hat{\beta}_1 = \hat{\beta}_1(\{n^{\gamma_i}\}_{i=1}^I)$, $\hat{\beta}_2 = \hat{\beta}_2(\{\exp((\log n)^{\gamma_i})\}_{i=1}^I)$, and $\hat{\tau}(n) = n^{\hat{\beta}_1}(\log n)^{\hat{\beta}_2}$ to economize on notation. By plugging in these rate estimators, our subsampling estimator for the distribution function $K(x, \mathcal{P}_\theta)$ of L is defined as

$$K_{b_n}(x|\hat{\tau}) = \frac{1}{q}\sum_{s=1}^{q} \mathbb{I}\{\hat{\tau}(b_n)\|T_{b_n,s} - T_n\| \leq x\}.$$

The following theorem establishes the asymptotic validity of the proposed subsampling procedures.

Theorem 2. *Under Assumptions (i)–(iv), it holds that for* $0 < \gamma_1 < \cdots < \gamma_I < 1$,

$$\sup_x |K_{b_n}(x|\hat{\tau}) - K(x, \mathcal{P}_\theta)| \xrightarrow{p} 0.$$

The proof of this theorem is presented in Appendix A.2. This theorem implies that the t-th quantile $K_{b_n}^{-1}(t|\hat{\tau})$ of $K_{b_n}(x|\hat{\tau})$ is also consistent for the t-th quantile $K^{-1}(t, \mathcal{P}_\theta)$ of $K(x, \mathcal{P}_\theta)$. Thus, the asymptotic coverage probability of the interval $[T_n - n^{-\hat{\beta}_1}(\log n)^{-\hat{\beta}_2}K_{b_n}^{-1}(t|\hat{\tau}), T_n + n^{-\hat{\beta}_1}(\log n)^{-\hat{\beta}_2}K_{b_n}^{-1}(t|\hat{\tau})]$ is the nominal level t. This indicates that Theorems 1 and 2 together provide a method to construct valid confidence intervals based on subsampling in the case where the convergence rate is unknown. Given that we do not assume a particular model, this method is widely applicable.

3. Simulation

We evaluate the finite sample performance of our procedure in two canonical settings: (i) non parametric regression with a mismeasured regressor, and (ii) estimation of the intercept in a binary choice model.

As discussed in the previous section, we must choose a sequence of subsample sizes, $\{b_{in}\}_{i=1}^I$, to estimate β_1 and β_2. In that section, we show that different sequences for β_1 and β_2 are required, and must take the form $\{n^{\gamma_i}\}_{i=1}^I$ and $\{\exp((\log n)^{\gamma_i})\}_{i=1}^I$, for β_1 and β_2, respectively. Here, we propose a method to determine a suitable choice for $(\gamma_1, \ldots, \gamma_I)$. First, we choose an overly wide range from which we can then search for a more appropriate range in a data-driven way. The overly wide range is chosen as a grid from 0.5 to 0.9 with increments of 0.025, with a lower bound for the grid such that $\min_{1\leq i\leq I} b_{in} > \log(n)^2$, as in Heiler and Kazak (2021). Next, we estimate the OLS regression to determine (β_1, β_2) using this initial $(\gamma_1, \ldots, \gamma_I)$, and save the R-squared from this regression. We then re-estimate the OLS regression but now use $(\gamma_2, \ldots, \gamma_I)$, and again save the R-squared. Following this, we estimate the OLS regression using $(\gamma_1, \ldots, \gamma_{(I-1)})$ and save the R-squared. Now estimate the OLS regression with $(\gamma_2, \ldots, \gamma_{(I-1)})$; save the R-squared. Continue this process alternating between removing the smallest and largest γ. The regression that produces the largest R-squared is used as the sequence of choice, giving the estimated (β_1, β_2). Intuitively, this procedure finds the 'correct' range of subsample sizes that can most accurately estimate the convergence rate.[1]

To avoid excessive computational cost, we can use the same subsamples to determine the optimal subsample size for estimating the distribution of the deconvolution estimator. We do this using the

[1] Simulation results (not presented) suggest that the results are insensitive to the initial choice of $(\gamma_1, ..., \gamma_I)$ due to the second data-driven search step.

method of Bickel and Sakov (2008). This constitutes choosing the optimal subsample size as the one whose distribution is closest to the distribution of the next consecutive candidate subsample distribution; we use the Kolmogorov distance to measure this. In other words, the optimal subsample size b_n^*, is chosen as $b_n^* = \text{argmin}_{b_{in}} \sup_x |L_{b_{in}}(x) - L_{b_{(i+1)n}}(x)|$, where $L_{b_{in}}$ denotes the distribution using subsamples of size b_{in}. Bickel and Sakov (2008, Section 5.1) extensively study the patterns for the choice of the subsample size using this Kolmogorov distance. Throughout, we use 2000 subsamples to approximate the distribution of each subsample distribution.

For setting (a), we use the deconvolution estimator of Schennach (2004), as detailed above in Section 1.1, and follow the simulation setting of that article. In particular, we use the regression function

$$E[Y|X^*] = \begin{cases} -1 & \text{if } X^* < 1, \\ X^* & \text{if } X^* \in [-1, 1], \\ 1 & \text{if } X^* > 1, \end{cases}$$

with $X = X^* + \epsilon$ and $W = X^* + \nu$, where only (Y, X, W) are observable; so (X, W) are repeated noisy measurements of the unobserved true regressor X^*. The regression error term is independent of (X^*, ϵ, ν) and drawn from $N(0, 1/4)$, and X^* is independent of (ϵ, ν). As in Schennach (2004), our object of interest is $E[Y|X = 1]$.

To showcase the ability of our subsampling method to adapt to varying convergence rates, we consider two cases based on the smoothness of the distributions of X^*, ϵ, and ν. First, we take X^*, ϵ, and ν to be normally distributed. Schennach (2004) showed that in this case, and when the regression function is ordinary smooth of order 2 (as in this simulation), the deconvolution estimator converges at rate $O_p((\ln n)^{-1/2})$ to a Gaussian distribution. In the second case, we take X^*, ϵ, and ν to follow a Laplace distribution. In this case, Schennach (2004) showed convergence to a Gaussian distribution at rate $O_p(n^{-1/4})$.

We keep the signal-to-noise ratio fixed for both designs: X^* has unit variance, and both measurement errors have a variance of $1/4$ (this again follows the simulation design in Schennach, 2004). We use the infinite-order flat-top kernel proposed by McMurry and Politis (2004), which is defined by its Fourier transform

$$K^{\text{ft}}(t) = \begin{cases} 1 & \text{if } |t| \leq 0.05, \\ \exp\left\{\frac{-\exp(-(|t|-0.05)^2)}{(|t|-1)^2}\right\} & \text{if } 0.05 < |t| < 1, \\ 0 & \text{if } |t| \geq 1. \end{cases}$$

and the bandwidth is selected using the leave-one-out method of Dong, Otsu, and Taylor (2023), which exhibits good finite sample performance for this deconvolution estimator (as evidenced in the original paper). As noted in Section 1.1, we only know the convergence rate of the estimator if we know — and use — the optimal bandwidth. In this case, we assume we do not know the data-generating process, hence do not know the optimal bandwidth nor the convergence rate.

Table 1 reports coverage probabilities (deviations from the nominal level) for a range of sample sizes $n = \{500, 1000, 2000\}$ based on 1000 Monte Carlo replications. We report results for our method and the existing subsampling method of Bertail, Politis, and Romano (1999) using an estimated polynomial rate. Overall, the coverage of our subsampling confidence intervals are accurate for moderate sample sizes ($n \geq 1000$). They perform favorably compared to the approach of Bertail, Politis, and Romano (1999) when the data is normally distributed, i.e., when the convergence rate is logarithmic. This result is expected since Bertail, Politis, and Romano (1999) does not consider logarithmic rates. For data from a Laplace distribution, the rate is polynomial, so the approach of Bertail, Politis, and Romano (1999) is valid; it is encouraging to see that the two methods produce very similar results.

Table 2 reports the (size-adjusted) power of a two-sided t-test with 5% nominal level for a test of the null hypothesis $E[Y|X = 1] = 0.5$, where the true value is 1. As expected based on the convergence rates, the test shows greater power in the logistic setting than the normal. It is encouraging to see the power increases with the sample size, and our method shows slightly greater power than the existing method.

Table 1. Coverage probabilities for setting (a).

Distribution	Normal						Laplace					
Sample size	500		1000		2000		500		1000		2000	
Nominal prob.	90	95	90	95	90	95	90	95	90	95	90	95
New method	-2.3	-2.0	-2.3	-0.6	-0.6	0.3	3.3	1.3	2.4	1.1	-0.7	0.6
Existing method	-5.2	-2.2	-2.6	-0.9	-1.7	-0.8	3.3	1.1	2.2	1.2	0.7	0.5

Table 2. Power for setting (a).

Distributions	Normal			Laplace		
Sample size	500	1000	2000	500	1000	2000
New method	19.8	39.1	98.8	37.4	53.3	81.7
Existing method	19.4	35.6	98.5	34.5	49.3	78.4

Table 3. Coverage probabilities for setting (b).

Distribution	Logistic						Normal						Cauchy					
Sample size	250		500		1000		250		500		1000		250		500		1000	
Nominal prob.	90	95	90	95	90	95	90	95	90	95	90	95	90	95	90	95	90	95
New method	1.1	0.5	0.8	0.6	0.0	-0.1	1.3	0.7	1.1	1.6	0.1	0.9	1.0	1.1	-0.9	-1.0	-0.3	-0.1
Existing method	1.9	0.7	1.8	1.5	0.7	0.5	1.9	1.5	1.2	0.5	-0.4	-0.4	2.3	1.4	-1.2	-1.0	0.1	-0.1

Table 4. Power for setting (b).

Distributions	Logistic			Normal			Cauchy		
Sample size	250	500	1000	250	500	1000	250	500	1000
New method	66.1	89.4	99.1	54.9	75.5	95.5	78.9	91.4	98.8
Existing method	65.7	89.6	98.9	51.7	72.3	93.6	77.8	89.1	98.8

For setting (b), we use the estimator of Lewbel (1997), as detailed above in Section 1.2. The trimming parameter τ_n is fixed at 0.001 (as in Lewbel, 1997), but results for $\tau_n = 0.01$ and $\tau_n = 0.0001$ are almost identical. We use a kernel density estimator with a Gaussian kernel and bandwidth chosen using likelihood cross-validation (Silverman, 1986).

We use the data generating process

$$Y = \mathbb{I}\{\alpha + \delta Z + U \geq 0\},$$

with $\alpha = 0$ as the parameter of interest, $\delta = 1$, and Z independent of U, where U follows a logistic distribution with unit variance. We consider three cases based on the distribution of Z. As shown by Khan and Tamer (2010), when Z has a logistic distribution, the estimator converges to a Gaussian distribution at rate $O_p(n^{1/2}(\log n)^{-1/2})$; when Z has a normal distribution, convergence to the Gaussian distributions occurs at rate $O_p(n^{1/4}(\log n)^{-1/4})$; and when Z has a Cauchy distribution, the convergence rate to a Gaussian distribution is $O_p(n^{1/2})$. In each case, we set Z to have unit variance.

Table 3 reports coverage probabilities (deviations from the nominal level) for a range of sample sizes $n = \{250, 500, 1000\}$ based on 1000 Monte Carlo replications. Again, with moderate sample sizes, the proposed subsampling procedure with estimated convergence rates exhibits accurate coverage properties across all cases. In the Cauchy setting, the existing method of Bertail, Politis, and Romano (1999) is valid, so it is encouraging to see that our method performs equally well here.

Table 4 reports the (size-adjusted) power of a two-sided t-test with 5% nominal level for a test of the null hypothesis $\alpha = 0.5$, where the true value is 0. Again, as the theoretical convergence rates would suggest, the Cauchy setting exhibits the greatest power and the normal setting has the lowest power. The power of our approach is comparable to that of Bertail, Politis, and Romano (1999), and all settings see an increase in power with the sample size.

4. Conclusion

In this article, we introduce a robust inference method to handle situations where the convergence rate of estimators is unknown. In particular, we provide a practical solution to this challenge by extending the subsampling techniques to account for both polynomial and logarithmic convergence components. The method is validated through simulation experiments designed for two settings: non parametric regression with measurement errors and binary choice models. These results demonstrate that our method can deliver accurate inference across different data-generating processes, while also highlighting its flexibility and potentially broad applicability across different statistical problems.

Appendix A

A.1. Proof of Theorem 1

Recall the definitions of y_i, x_i, z_i, and β_0 in (1), and let $u_i = y_i - \beta_0 - \beta_1 x_i - \beta_2 z_i$. Define $S_{ab} = \sum_{i=1}^{I}(a_i - \bar{a})(b_i - \bar{b})$ for $\bar{a} = I^{-1}\sum_{i=1}^{I} a_i$ and $\bar{b} = I^{-1}\sum_{i=1}^{I} b_i$. Note that the rate estimators are explicitly written as

$$\hat{\beta}_1(\{b_{in}\}_{i=1}^{I}) = \frac{S_{zz}S_{xy} - S_{xz}S_{zy}}{S_{xx}S_{zz} - (S_{xz})^2},$$
$$\hat{\beta}_2(\{b_{in}\}_{i=1}^{I}) = \frac{-S_{xz}S_{xy} + S_{xx}S_{zy}}{S_{xx}S_{zz} - (S_{xz})^2}.$$

For $\hat{\beta}_1(\{n^{\gamma_i}\}_{i=1}^{I})$, we have

$$\begin{aligned}\hat{\beta}_1(\{n^{\gamma_i}\}_{i=1}^{I}) - \beta_1 &= \frac{S_{zz}S_{xu} - S_{xz}S_{zu}}{S_{xx}S_{zz} - (S_{xz})^2} \\ &= \frac{1}{1-\hat{\rho}_{x,z}^2}\left(\frac{S_{xu}}{S_{xx}} - \hat{\rho}_{x,z}\frac{S_{zu}}{\sqrt{S_{xx}S_{zz}}}\right).\end{aligned}$$

where $\hat{\rho}_{x,z}$ denotes the sample correlation coefficient between $\{x_i\}_{i=1}^{I}$ and $\{z_i\}_{i=1}^{I}$. Since $|\hat{\rho}_{x,z}| \le 1$, $S_{uu} = o_p(1)$, and

$$\max\left\{\left|\frac{S_{xu}}{S_{xx}}\right|, \left|\frac{S_{zu}}{\sqrt{S_{xx}S_{zz}}}\right|\right\} \le \sqrt{\frac{S_{uu}}{S_{xx}}},$$

it is sufficient to check the magnitude of S_{xx}. Since $x_i = -\gamma_i \log n$, we have $\bar{x} = -\bar{\gamma}\log n$ with $\bar{\gamma} = I^{-1}\sum_{i=1}^{I}\gamma_i$ and the first statement of the theorem follows from

$$S_{xx} = \sum_{i=1}^{I}(\bar{\gamma}\log n - \gamma_i \log n)^2 = A(\log n)^2,$$

for a positive constant $A = \sum_{i=1}^{I}(\bar{\gamma} - \gamma_i)^2$.

By a similar argument, for $\hat{\beta}_2(\{\exp((\log n)^{\gamma_i})\}_{i=1}^{I})$, we have

$$\begin{aligned}\hat{\beta}_2(\{\exp((\log n)^{\gamma_i})\}_{i=1}^{I}) - \beta_2 &= \frac{S_{xx}S_{zu} - S_{xz}S_{xu}}{S_{xx}S_{zz} - (S_{xz})^2} \\ &= \frac{1}{1-\hat{\rho}_{x,z}^2}\left(\frac{S_{zu}}{S_{zz}} - \hat{\rho}_{x,z}\frac{S_{xu}}{\sqrt{S_{xx}S_{zz}}}\right).\end{aligned}$$

Since $|\hat{\rho}_{x,z}| \le 1$, $S_{uu} = o_p(1)$, and

$$\max\left\{\left|\frac{S_{zu}}{S_{zz}}\right|, \left|\frac{S_{xu}}{\sqrt{S_{xx}S_{zz}}}\right|\right\} \le \sqrt{\frac{S_{uu}}{S_{zz}}},$$

it is sufficient to check the magnitude of S_{zz}. Since $z_i = -\gamma_i \log\log n$, the second statement of the theorem follows from

$$S_{zz} = \sum_{i=1}^{I} (-\gamma_i \log\log n + \bar{\gamma} \log\log n)^2 = A(\log\log n)^2.$$

A.2. Proof of Theorem 2

First, note that

$$K_{b_n}(x|\hat{\tau}) = \frac{1}{q}\sum_{s=1}^{q} \mathbb{I}\{b_n^{\hat{\beta}_1} (\log b_n)^{\hat{\beta}_2} \|(T_{b_n,s} - \theta) - (T_n - \theta)\| \le x\},$$

and the triangular inequality implies

$$K_{b_n}(x|\hat{\tau}) \le \frac{1}{q}\sum_{s=1}^{q} \mathbb{I}\{b_n^{\hat{\beta}_1} (\log b_n)^{\hat{\beta}_2} \|T_{b_n,s} - \theta\| \le x + b_n^{\hat{\beta}_1} (\log b_n)^{\hat{\beta}_2} \|T_n - \theta\|\},$$

$$K_{b_n}(x|\hat{\tau}) \ge \frac{1}{q}\sum_{s=1}^{q} \mathbb{I}\{b_n^{\hat{\beta}_1} (\log b_n)^{\hat{\beta}_2} \|T_{b_n,s} - \theta\| \le x - b_n^{\hat{\beta}_1} (\log b_n)^{\hat{\beta}_2} \|T_n - \theta\|\}.$$

Also note that for any $\epsilon > 0$, we have

$$\begin{aligned} &P\left(b_n^{\hat{\beta}_1} (\log b_n)^{\hat{\beta}_2} \|T_n - \theta\| < \epsilon\right) \\ &= P\left(n^{\beta_1} (\log n)^{\beta_2} \|T_n - \theta\| < \epsilon b_n^{\beta_1 - \hat{\beta}_1} (\log b_n)^{\beta_2 - \hat{\beta}_2} \frac{n^{\beta_1} (\log n)^{\beta_2}}{b_n^{\beta_1} (\log b_n)^{\beta_2}}\right) \to 1, \end{aligned} \tag{A.1}$$

where the convergence follows from Theorem 1 and Assumptions (ii) and (iv). Define

$$U_{b_n}(x|\hat{\tau}) = \frac{1}{q}\sum_{s=1}^{q} \mathbb{I}\{b_n^{\hat{\beta}_1} (\log b_n)^{\hat{\beta}_2} \|T_{b_n,s} - \theta\| \le x\}.$$

Then Eq. (A.1) implies

$$U_{b_n}(x - \epsilon|\hat{\tau}) \le K_{b_n}(x|\hat{\tau}) \le U_{b_n}(x + \epsilon|\hat{\tau}),$$

with probability approaching one. The conclusion follows from

$$U_{b_n}(x|\hat{\tau}) = U_n(x b_n^{\beta_1 - \hat{\beta}_1} (\log b_n)^{\beta_2 - \hat{\beta}_2} | \beta) \xrightarrow{p} K(x, \mathcal{P}_\theta)$$

for each x, where the convergence follows from Theorem 1 and the validity of standard subsampling as in Politis and Romano (1994).

References

Andrews, D. W. K., Schafgans, M. M. A. (1998). Semiparametric estimation of the intercept of a sample selection model. *Review of Economic Studies* 65(3):497–517. doi: 10.1111/1467-937X.00055

Armstrong, T. B. (2013). Bounds in auctions with unobserved heterogeneity. *Quantitative Economics* 4(3):377–415. doi: 10.3982/QE167

Armstrong, T. B. (2015). Asymptotically exact inference in conditional moment inequality models. *Journal of Econometrics* 186(1):51–65. doi: 10.1016/j.jeconom.2015.01.002

Bickel, P. J., Götze, F., Van Zwet, W. R. (2012). *Resampling Fewer than n Observations: Gains, Losses, and Remedies for Losses.* New York: Springer, pp. 267–297.

Bickel, P. J., Sakov, A. (2008). On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statistica Sinica* 18:967–985.

Bertail, P., Politis, D. N., Romano, J. P. (1999). On subsampling estimators with unknown rate of convergence. *Journal of the American Statistical Association* 94(446):569–579. doi: 10.1080/01621459.1999.10474151

Chernozhukov, V., Fernandez-Val, I. (2011). Inference for extremal conditional quantile models, with an application to market and birthweight risks. *The Review of Economic Studies* 78(2):559–589. doi: 10.1093/restud/rdq020

Dobronyi, C. R., Ouyang, F., Yang, T. T. (2024). Revisiting panel data discrete choice models with lagged dependent variables, forthcoming in *Journal of Business & Economic Statistics*.

Dong, H., Otsu, T., Taylor, L. (2023). Bandwidth selection for nonparametric regression with errors-in-variables. *Econometric Reviews* 42(4):393–419. doi: 10.1080/07474938.2023.2191105

Heckman, J. (1990). Varieties of selection bias. *American Economic Review* 80(2):313–318.

Heiler, P., Kazak, E. (2021). Valid inference for treatment effect parameters under irregular identification and many extreme propensity scores. *Journal of Econometrics* 222(2):1083–1108. doi: 10.1016/j.jeconom.2020.03.025

Kato, K., Sasaki, Y. (2019). Uniform confidence bands for nonparametric errors-in-variables regression. *Journal of Econometrics* 213(2):516–555. doi: 10.1016/j.jeconom.2019.05.021

Khan, S., Nekipelov, D. (2022). On uniform inference in nonlinear models with endogeneity. *Journal of Econometrics* 240(2):105261.

Khan, S., Tamer, E. (2010). Irregular identification, support conditions, and inverse weight estimation. *Econometrica* 78(6):2021–2042.

Kotlarski, I. (1967). On characterizing the gamma and the normal distribution. *Pacific Journal of Mathematics* 20(1):69–76. doi: 10.2140/pjm.1967.20.69

Lewbel, A. (1997). Semiparametric estimation of location and other discrete choice moments. *Econometric Theory* 13(1):32–51. doi: 10.1017/S0266466600005636

Ma, X., Wang, J. (2020). Robust inference using inverse probability weighting. *Journal of the American Statistical Association* 115(532):1851–1860. doi: 10.1080/01621459.2019.1660173

McMurry, T. L., Politis, D. N. (2004). Nonparametric regression with infinite order flat-top kernels. *Journal of Nonparametric Statistics* 16(3-4):549–562. doi: 10.1080/10485250310001622596

Politis, D. N., Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics* 22:2031–2050.

Sasaki, Y., Ura, T. (2022). Estimation and inference for moments of ratios with robustness against large trimming bias. *Econometric Theory* 38(1):66–112. doi: 10.1017/S0266466621000025

Sasaki, Y., Ura, T. (2021). Slow movers in panel data, arXiv preprint, arXiv:2110.12041.

Sasaki, Y., Y., Wang, (2022). Non-robustness of the cluster-robust inference: with a proposal of a new robust method, arXiv preprint, arXiv:2210.16991.

Schennach, S. M. (2004). Nonparametric regression in the presence of measurement error. *Econometric Theory* 20:1046–1093.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Boca Raton, FL: CRC Press, p. 26.