Many Terms in a Series Estimator of the Partially Linear Model[1]

Matias D. Cattaneo, Department of Economics, University of Michigan, cattaneo@umich.edu.
Michael Jansson, Department of Economics, UC Berkeley, mjansson@econ.berkeley.edu.
Whitney K. Newey, Department of Economics, MIT, wnewey@mit.edu.

September, 2010

VERY PRELIMINARY AND INCOMPLETE DRAFT

JEL classification: C13, C31.
Keywords: partially linear model, many terms, adjusted variance.

[1] This very preliminary version of the paper was prepared for the Fourth Annual Meeting of the Impact Evaluation Network (IEN), Latin American and the Caribbean Economic Association (LACEA).

Proposed Running Head: Many Terms Asymptotics

Corresponding Author:
Whitney K. Newey
Department of Economics
MIT, E52-262D
Cambridge, MA 02142-1347

Abstract

This papers studies the asymptotic behavior of a series-based semiparametric estimator for the partially linear model, and derives a generalized large sample theory that accommodates (but does not require) a "large" number of terms (or covariates) relative to the sample size. This asymptotic distribution theory covers the classical large sample results based on the asymptotic linear representation of the estimator, and also provides a distributional approximation even when the estimator is not asymptotically linear. Using these large sample results, it is shown that the classical unbiased standard errors estimator from least squares theory is consistent under homoskedasticity, even when the number of regressors grows proportionally to the sample size. On the other hand, the classical Eicker-Huber-White heteroskedasticity-robust standard errors are shown to be inconsistent in general. Two new heteroskedasticity- and many terms-robust standard errors are proposed.

1 Introduction

Semiparametric procedures are popular in econometrics because they reduce misspecification biases while retaining many of the attractive properties of parametric modelling. These procedures typically require choosing a preliminary nonparametric estimator that depends on user-defined tuning and smoothing parameters (e.g., a bandwidth and a kernel, or the number of terms in and a form of series of approximation). Unfortunately, such procedures are considerably less popular among empirical researchers because inference based on classical large sample approximations is known to be highly sensitive to perturbations in the choice of tuning and smoothing parameters, making empirical work unreliable in general. The lack of robustness of semiparametric-based statistical procedures with respect to changes in these parameters is a common problem in many econometric models. As a consequence, inference procedures that are insensitive to changes in the tuning and smoothing parameters are highly desirable, as they will increase substantially the validity of the empirical results obtained in specific empirical applications.

This papers studies the asymptotic behavior of a series-based semiparametric estimator for the partially linear model, and derives a generalized large sample theory based on an alternative asymptotic experiment. Specifically, this paper studies the asymptotic behavior of the corresponding semiparametric t-test under tuning parameter sequences (i.e., the number of terms in the series approximation) that may render asymptotic linearity invalid, and hence capturing features of the semiparametric statistic that are typically assumed away by conventional large sample results (e.g., Newey and McFadden (1994) and Chen (2007)). This type of large sample approximations have been shown to provide a better finite sample characterization of the statistic of interest, when compared to the classical, asymptotically linear distributional approximations. This idea has been employed in a variety of contexts, including matching estimators with fixed number of matches (Abadie and Imbens (2006)), IV estimators with many/weak instruments (Hansen, Hausman, and Newey (2008)), and density-weigthed average derivatives (Cattaneo, Crump, and Jansson (2010)).

The new asymptotic approximation presented here is not only important from a theoretical point of view, but also relevant for applications. The semiparametric linear model is popular among empirical researchers because fits naturally into a "control function" approach, is a commonly used dimension reduction technique, and may be justified in the context of a conditional independence assumption. Moreover, as discussed in detail below, the theoretical results presented here also apply to linear models where the number of regressors is large compared to the sample size, even if there is no approximation bias.

This paper presents four main results. First, it is shown that a generalized central limit theorem may be obtained for the classical series-based partially linear estimator, which is based on an approximate bilinear expansion. This result is shown to cover the classical asymptotic approximation under conventional asymptotics, although in general the estimator has a larger asymptotic variance,

which is not invariant with respect to the tuning and smoothing parameters employed. Second, it is shown that under known homoskedasticity the classical degrees-of-freedom-corrected standard errors estimator from least squares is valid, even when the number of terms in the approximation series is "large". Third, it is shown that the conventional Eicker-Huber-White heteroskedasticity-robust (HR) standard errors estimator is inconsistent in general under the generalized asymptotics. In particular, when the underlying model is homoskedastic, the HR standard errors estimator is biased downwards, leading to liberal inference. Finally, three new HR standard errors estimators are proposed that are also asymptotically valid when the number of terms (or regressors) is "large" relative to the sample size.

2 Model and Classical Results

Let $(y_i, x_i', z_i')'$, $i = 1, \ldots, n$, be a random sample of the random vector $(y, x', z')'$, where $y \in \mathbb{R}$ is a dependent variable, and $x \in \mathbb{R}^{d_x \times 1}$ and $z \in \mathbb{R}^{d_z \times 1}$ are explanatory variables. The so-called partially linear model is given by

$$y_i = x_i'\beta + g(z_i) + \varepsilon_i, \qquad \mathbb{E}\left[\varepsilon_i | x_i, z_i\right] = 0, \qquad \sigma^2(x_i, z_i) = \mathbb{E}\left[\varepsilon_i^2 | x_i, z_i\right],$$

where $v_i = x_i - h(z_i)$, with $h(z_i) = \mathbb{E}\left[x_i | z_i\right]$.[2] Donald and Newey (1994) provide sufficient conditions for $\sqrt{n}$-consistency and asymptotic linearity of a series-based semiparametric estimator of β. Specifically, the estimator $\hat{\beta}$ is constructed by regressing y_i on x_i and $p_K(z_i)$, where $p_K(z) = (p_{K1}(z), \ldots, p_{KK}(z))'$ is an appropriate basis of approximation, such as polynomials or splines, and $K = K(n) \to \infty$.

To formally describe this estimator, let $Y = [y_1, \cdots, y_n]' \in \mathbb{R}^{n \times 1}$, $X = [x_1, \cdots, x_n]' \in \mathbb{R}^{n \times d_x}$, $Z = [z_1, \cdots, z_n]' \in \mathbb{R}^{n \times d_z}$, $\varepsilon = [\varepsilon_1, \cdots, \varepsilon_n]' \in \mathbb{R}^{n \times 1}$, $V = [v_1, \cdots, v_n]' \in \mathbb{R}^{n \times d_x}$, $G = [g(z_1), \cdots, g(z_n)]' \in \mathbb{R}^{n \times 1}$, $H = [h(z_1), \cdots, h(z_n)]' \in \mathbb{R}^{n \times d_x}$, and $P = [p^K(z_1), \ldots, p^K(z_n)]'$. The series-based estimator of β is given by

$$\hat{\beta} = (X'MX)^- X'MY, \qquad M = I - Q, \qquad Q = P(P'P)^- P',$$

where A^- denotes a generalized inverse of a matrix A (satisfying $AA^-A = A$). For fixed n, this estimator coincides with a "partial-out" regression estimator $\hat{\beta} = (\tilde{X}'\tilde{X})^- \tilde{X}'\tilde{Y}$, where $\tilde{X} = MX = [\tilde{x}_1, \cdots, \tilde{x}_n]'$ and $\tilde{Y} = MY = [\tilde{y}_1, \cdots, \tilde{y}_n]'$. (Similarly, denote $\tilde{\varepsilon} = M\varepsilon = [\tilde{\varepsilon}_1, \cdots, \tilde{\varepsilon}_n]'$ and $\tilde{V} = MV = [\tilde{v}_1, \cdots, \tilde{v}_n]'$.)

The estimator $\hat{\beta}$ may be intuitively interpreted as a two-step semiparametric estimator with smoothing parameter $p_K(\cdot)$ and tuning parameter K, because the unknown (regression) functions

[2] See Robinson (1988) for the analysis of this model when using kernel regression, and Linton (1995) for the corresponding higher-order properties.

$g(\cdot)$ and $h(\cdot)$ are non-parametrically estimated in a preliminary step by the series estimator. In particular, the following assumption characterizes the rate at which the approximation error of series estimator should vanish.

Assumption B. (i) For some $\alpha_h > 0$, there exists η_h so that

$$\frac{1}{n} H'MH = \frac{1}{n} \min_{\eta} \|H - P\eta\|^2 = \frac{1}{n} \sum_{i=1}^{n} \left\| h(z_i) - p_K(z_i)' \eta_h \right\|^2 = O_{as}(K^{-2\alpha_h}).$$

(ii) For some $\alpha_g > 0$, there exists η_g so that

$$\frac{1}{n} G'MG = \frac{1}{n} \min_{\eta} \|G - P\eta\|^2 = \frac{1}{n} \sum_{i=1}^{n} \left[g(z_i) - p_K(z_i)' \eta_g \right]^2 = O_{as}(K^{-2\alpha_g}).$$

The conditions required in Assumption B are implied by conventional assumptions from the series-based nonparametric literature (e.g., Newey (1997, Assumption 3)). Thus, under appropriate assumptions, commonly used basis of approximation such as polynomials or splines will satisfy Assumption B with $\alpha_h = d_z/s_h$ and $\alpha_g = d_z/s_g$, where s_h and s_g denotes the number of continuous derivatives of h and g, respectively.

Under regularity conditions (given in Section 3) and Assumption B, Donald and Newey (1994) obtained the following (infeasible) classical asymptotic approximation for $\hat{\beta}$: if

$$nK^{-2(\alpha_h + \alpha_g)} \to 0 \qquad \text{and} \qquad \frac{K}{n} \to 0 \tag{1}$$

then

$$\sqrt{n}(\hat{\beta} - \beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_i + o_p(1) \to_d \mathcal{N}(0, \Omega), \qquad \Omega = \Gamma^{-1} \Sigma \Gamma^{-1}, \tag{2}$$

where

$$\psi_i = \Gamma^{-1} v_i \varepsilon_i, \qquad \Gamma = \mathbb{E}\left[v_i v_i' \right], \qquad \Sigma = \mathbb{E}\left[v_i \mathbb{V}\left[\varepsilon | X, Z \right] v_i' \right].$$

The classical asymptotic linear representation of $\hat{\beta}$, given in (2), is established by analyzing the first-order stochastic properties of the "numerator" and "denominator" of the estimator. Intuitively, in each case, the analysis proceeds by first finding conditions so that $QX \approx H$ and $Q(Y - X\beta) \approx G$, which captures the bias introduced by the series approximation, and then finding conditions so that the corresponding reminders are well-behaved asymptotically. Specifically, for the "denominator" of $\hat{\beta}$, it can be shown (see Lemma 1 below) that if Condition (1) holds, then the (Hessian) matrix

satisfies (recall that $M = I - Q$ and $QP = P$)

$$\hat{\Gamma}_n = \frac{1}{n} X'MX \approx \frac{1}{n} V'MV = \frac{1}{n} V'V - \frac{1}{n} V'QV \approx \frac{1}{n} \sum_{i=1}^{n} v_i v_i' \to_p \Gamma,$$

where the first approximation captures the bias introduced by the series estimator (Assumption B(i)), and the second approximation requires the contribution of $V'QV$ to vanish asymptotically. Similarly, for the "numerator" of $\hat{\beta}$, it can be shown (see Lemmas 2–4 below) that

$$\frac{1}{\sqrt{n}} X'M\left(Y - X\beta\right) \approx \frac{1}{\sqrt{n}} V'M\varepsilon = \frac{1}{\sqrt{n}} V'\varepsilon - \frac{1}{\sqrt{n}} V'Q\varepsilon \approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} v_i \varepsilon_i \to_d \mathcal{N}\left(0, \Sigma\right),$$

where the first approximation is again related to the bias introduced by the nonparametric estimator and the second approximation requires $V'Q\varepsilon$ to be asymptotically negligible.

In both cases, the approximation error associated with the bias is controlled by the condition $nK^{-2(\alpha_h + \alpha_g)} \to 0$, which requires K to be "large" (provided the underlying functions g and h are smooth enough). On the other hand, condition $K/n \to 0$ guarantees that both $V'QV$ and $V'Q\varepsilon$ are asymptotically negligible, as required for the classical, asymptotically linear, approximation to be valid. The latter condition controls the variance of the estimator, and it is directly related to the behavior of the nonparametric estimator, which in this case is described by Q.

The classical approach to form a confidence interval for β_0 is to use the asymptotic distributional result coupled with a consistent standard errors estimator. A plug-in approach employs the (asymptotically) pivotal test statistic $T_{0,n}(K) = \Omega_0^{-1/2}\sqrt{n}(\hat{\beta} - \beta)$, together with a plug-in consistent estimator for Ω_0. Under heteroskedasticity, the feasible test statistic is given by

$$\hat{T}_{0,n}(K_n) = \hat{\Omega}_n^{-1/2}\sqrt{n}(\hat{\beta} - \beta), \qquad \hat{\Omega}_n = \hat{\Gamma}_n^{-1}\hat{\Sigma}_n\hat{\Gamma}_n^{-1},$$

where

$$\hat{\Gamma}_n = \frac{X'MX}{n}, \qquad \hat{\Sigma}_n = \frac{1}{n - K - d_x} \sum_{i=1}^{n} \tilde{x}_i \tilde{x}_i' \hat{\varepsilon}_i^2, \qquad \hat{\varepsilon} = \tilde{Y} - \tilde{X}'\hat{\beta} = [\hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_n]'.$$

In this case, the standard errors estimator is the classical Heteroskedasticity-Robust (HR) standard errors estimator commonly used in regression analysis. Under Condition (1) it is not difficult to show that $\hat{\Omega}_n^{-1}\Omega_0 \to_p I_{d_x}$.

3 Generalized Asymptotic Distribution

This section derives a generalized asymptotic distribution for $\sqrt{n}(\hat{\beta} - \beta)$, which relaxes the condition $K_n/n \to 0$. This non-standard asymptotic theory encompasses the classical result discussed in the

previous section, and also captures the effect of the "quadratic" term of the expansion, which is assumed away by Condition (1). Intuitively, this asymptotic experiment captures the effect of K "large" (relative to n) by breaking down the asymptotic linearity of the estimator.

To characterize the generalized central limit theory, it is natural to study the stochastic behavior of the estimator $\sqrt{n}(\hat{\beta}-\beta)$ as a "ratio" of two bilinear forms:

$$\sqrt{n}(\hat{\beta}-\beta)=\left(\frac{1}{n}X'MX\right)^{-}\frac{1}{\sqrt{n}}X'M\left(Y-X\beta\right)=\hat{\Gamma}_n^{-}\frac{1}{\sqrt{n}}X'M\left(Y-X\beta\right).$$

The following lemma characterizes the behavior of the Hessian matrix $\hat{\Gamma}_n$ under quite weak conditions. (Throughout this paper A_{ij} denotes the (i,j)-th element of a matrix A.)

Lemma 1. Suppose that $\mathbb{E}[\|v_i\|^4|z_i]\leq C_v<\infty$ (a.s.) and Assumption B(i) holds. Then,

$$\hat{\Gamma}_n=\frac{1}{n}X'MX=\Gamma_n+o_p(K/n),\qquad \Gamma_n=\frac{1}{n}\sum_{i=1}^{n}M_{ii}\mathbb{E}\left[v_iv_i'|z_i\right]=O_{as}(1+K/n).$$

This lemma characterizes the stochastic behavior of the Hessian matrix under conditions that are weaker than those entertained by the classical, asymptotically linear, distribution theory. Specifically, because $M=I-Q$ is an idempotent symmetric matrix, $M_{ii}\in(0,1)$ and $\sum_{i=1}^{n}M_{ii}\leq n-K$, Lemma 1 implies that $\hat{\Gamma}_n$ remains asymptotically bounded even when $K/n\not\to 0$. In particular, $\Gamma_n=\mathbb{E}\left[v_iv_i'\right]+o_p\left(1\right)$ when $K/n\to 0$. Moreover, in the case of homoskedasticity of v_i, that is, if $\mathbb{E}\left[v_iv_i'|z_i\right]=\mathbb{E}\left[v_iv_i'\right]$ (and $\text{rank}\left(Q\right)=K$), then $\Gamma_n=\left(1-K/n\right)\mathbb{E}\left[v_iv_i'\right]$. Finally, if $\lambda_{\min}\left(\mathbb{E}\left[v_iv_i'|z_i\right]\right)\geq F_V>0$ and $M_{ii}=1-Q_{ii}\geq F_Q>0$ (a.s.) then $\lambda_{\min}\left(\Gamma_n\right)\geq F_QF_V>0$. ($\lambda_{\min}\left(A\right)$ denotes the minimum eigenvalue of a matrix A.)

To fully characterize the asymptotic behavior of the "numerator" of $\sqrt{n}(\hat{\beta}-\beta)$ it is convenient to proceed in two steps. First, under appropriate bias assumptions, it is possible to show that the numerator is asymptotically equivalent to quadratic form based on mean-zero random variables.

Lemma 2. Suppose the assumptions of Lemma 1 hold, and $\mathbb{E}[\varepsilon_i^4|z_i]\leq C_\varepsilon<\infty$ (a.s.) and Assumption B(ii) holds. Then,

$$\frac{1}{\sqrt{n}}X'M\left(Y-X\beta\right)=\frac{1}{\sqrt{n}}V'M\varepsilon+O_p(\sqrt{n}K^{-(\alpha_h+\alpha_g)}+K^{-\alpha_h}+K^{-\alpha_g}).$$

As in Lemma 1, this result only requires bounded moments and a bias condition. In this case, the bias arises from both the approximation of the unknown functions h and g. As mentioned above,

the high-level Assumption B is implied by the standard assumption of best approximation from the sieve literature. Interestingly, in this model there is a trade-off in terms of curse of dimensionality: provided that $\min\{\alpha_h, \alpha_g\} > 0$, the bias condition is given by $\sqrt{n}K^{-(\alpha_h+\alpha_g)} \to 0$, which implies a trade-off between smoothness and dimensionality between h and g.

Lemma 2 justifies focusing on the bilinear form

$$\theta_n = \frac{1}{\sqrt{n}} V' M \varepsilon = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij} v_i \varepsilon_j,$$

where $\mathbb{E}[\theta_n | Z, X] = 0$. Moreover, under the assumptions imposed in Lemma 2, a simple variance calculation yields

$$\Sigma_n = \mathbb{V}[\theta_n | Z] = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij}^2 \mathbb{E}[v_i v_i' \varepsilon_j^2 | z_i, z_j] = O_{as}\left(1 + \frac{K}{n}\right).$$

In particular, if $K/n \to 0$ then

$$\Sigma_n = \frac{1}{n} \sum_{i=1}^{n} M_{ii}^2 \mathbb{E}\left[v_i v_i' \varepsilon_i^2 | z_i\right] + o_p(1) = \mathbb{E}\left[v_i v_i' \varepsilon_i^2\right] + o_p(1) = \Sigma + o_p(1),$$

as given in Section 2. Moreover, under homoskedasticity, that is, if $\mathbb{E}\left[\varepsilon_i^2 | x_i, z_i\right] = \sigma^2$, then

$$\mathbb{V}\left[\frac{1}{\sqrt{n}} V' M \varepsilon \middle| X, Z\right] = \frac{\sigma^2}{n} V' M V,$$

and

$$\Sigma_n = \frac{\sigma^2}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij}^2 \mathbb{E}\left[v_i v_i' | z_i\right] = \sigma^2 \Gamma_n,$$

because $\sum_{j=1}^{n} M_{ij}^2 = M_{ii}$. Furthermore, if $\mathbb{E}\left[\varepsilon_i^2 | x_i, z_i\right] = \sigma^2$ and $\mathbb{E}[v_i v_i' | z_i] = \mathbb{E}[v_i v_i']$ (and $\operatorname{rank}(Q) = K$ (a.s.)), then $\Sigma_n = \sigma^2 (1 - K/n) \mathbb{E}[v_i v_i']$. Finally, if $\lambda_{\min}\left(\mathbb{E}\left[v_i v_i' \varepsilon^2 | z_i\right]\right) \geq F_{V\varepsilon} > 0$ and $M_{ii} = 1 - Q_{ii} \geq F_Q > 0$ (a.s.), then $\lambda_{\min}(\Sigma_n) \geq F_Q^2 F_{V\varepsilon} > 0$.

The following Lemma characterizes the asymptotic distribution of the bilinear form θ_n.

Lemma 3. Suppose the assumptions of Lemma 2 hold, and $M_{ii} > F_M > 0$ (a.s.) and $\lambda_{\min}(\Sigma_n) > F_\Sigma > 0$ (a.s.). Then,

$$\Sigma_n^{-1/2} \frac{1}{\sqrt{n}} V' M \varepsilon \to_d \mathcal{N}(0, I_{d_x}).$$

The following theorem is a direct consequence of the previous lemmas and Slutsky Theorem,

and constitutes the main result of this section.

Theorem 1. Suppose the assumptions Lemma 3 hold and suppose $\lambda_{\min}(\Gamma_n) > F_H > 0$. Then, if

$$nK^{-2(\alpha_x+\alpha_g)} \to 0 \qquad \text{and} \qquad \frac{K}{n} \to \alpha \in [0,1) \tag{3}$$

then

$$\hat{\Omega}_n^{-1/2}(\hat{\beta}-\beta) \to_d \mathcal{N}(0, I_{d_x}). \tag{4}$$

If, moreover, $(\Gamma_n, \Sigma_n) = (\Gamma_\infty, \Sigma_\infty) + o_p(1)$ for some $(\Gamma_\infty, \Sigma_\infty)$, then

$$\sqrt{n}(\hat{\beta}-\beta) \to_d \mathcal{N}(0, \Omega_\infty), \qquad \Omega_\infty = \Gamma_\infty^{-1}\Sigma_\infty\Gamma_\infty^{-1}.$$

If, moreover, $\mathbb{E}\left[\varepsilon_i^2 | x_i, z_i\right] = \sigma^2$ for some σ^2, then

$$\sqrt{n}(\hat{\beta}-\beta) \to_d \mathcal{N}\left(0, \sigma^2\Gamma_\infty^{-1}\right).$$

Theorem 1 shows that the central limit theorem for $\hat{\beta}$ holds under the weaker condition (3). (Compare to Condition (1).) This result does not rely on asymptotic linearity, nor on the actual convergence of the matrices Γ_n and Σ_n. However, if $K/n \to 0$, then $(\Gamma_n, \Sigma_n) = (\Gamma, \Sigma) + o_p(1)$ with $\Gamma = \mathbb{E}[v_i v_i']$ and $\Sigma = \mathbb{E}\left[v_i v_i' \varepsilon_i^2\right]$, and the resulting large sample distribution theory does rely on the asymptotically linear representation of $\hat{\beta}$, as given in (2).

Importantly, if $K/n \nrightarrow 0$ and $(\Gamma_n, \Sigma_n) = (\Gamma, \Sigma) + o_p(1)$, then $\Gamma \neq \mathbb{E}[v_i v_i']$ and $\Sigma \neq \mathbb{E}\left[v_i v_i' \varepsilon_i^2\right]$, in general. For instance, if (v_i, ε_i) is independent of z_i, then $\Gamma_n = (1 - K/n)\,\mathbb{E}[v_i v_i']$ and

$$\Sigma_n = \left(1 - \frac{K}{n}\right)\mathbb{E}\left[v_i v_i' \varepsilon_i^2\right] + \left(\frac{1}{n}\sum_{i=1}^n Q_{ii}^2 - \frac{K}{n}\right)\left(\mathbb{E}\left[v_i v_i' \varepsilon_i^2\right] - \mathbb{E}\left[v_i v_i'\right]\mathbb{E}\left[\varepsilon_i^2\right]\right).$$

4 Standard Errors

This section discusses different homoskedastic- and heteroskedastic- standard errors estimators, and their properties under the generalized asymptotics studied in this paper.

4.1 Homoskedasticity

If $\mathbb{E}\left[\varepsilon_i^2 | x_i, z_i\right] = \sigma^2$ for all $i = 1, 2, \cdots, n$, then $\Sigma_n = \sigma^2\Gamma_n^-$. Thus, a natural plug-in estimator is given by $\hat{V}_{HOM} = \hat{\sigma}^2\hat{\Gamma}^-$, where $\hat{\sigma}^2$ is chosen so that $\hat{\sigma}^2 = \sigma^2 + o_p(1)$. The usual OLS estimator is

$$\hat{\sigma}_n^2 = \frac{1}{n - K - d_x}\sum_{i=1}^n \hat{\varepsilon}_i^2 = \frac{1}{n - K - d_x}\hat{\varepsilon}'\hat{\varepsilon}.$$

As shown in Lemma 1, $\hat{\Gamma}_n = \Gamma_n + o_p(1)$ under the many terms asymptotics discussed in this paper (Condition (3)), and therefore it remains to verify that $\hat{\sigma}_n^2$ is also consistent under this generalized asymptotic experiment. Heuristically, this estimator is consistent because

$$\hat{\varepsilon}'\hat{\varepsilon} = (Y - X\hat{\beta})'M(Y - X\hat{\beta}) \approx \varepsilon' M \varepsilon \approx (n - K)\,\sigma^2,$$

where the first approximation is based on the $\sqrt{n}$-consistency of $\hat{\beta}$ and the approximation bias of the series estimator, while the second approximation is based on the fact that the bilinear form $\varepsilon' M \varepsilon = \varepsilon'(I - Q)\varepsilon$ is dominated by its diagonal. These heuristics are formalized in the following theorem.

Theorem 2. Suppose the assumptions of Theorem 1 hold. Then, $\hat{\sigma}_n^2 = \sigma^2 + o_p\,(1)$.

It follows by Lemma 1, Theorem 2 and Slutsky Theorem that

$$\hat{V}_{HOM} = \sigma^2 \Gamma_n^- + o_p\,(1) = \Sigma_n + o_p\,(1)\,,$$

and therefore

$$\hat{V}_{HOM}^{-1/2}(\hat{\beta} - \beta_0) \to_d \mathcal{N}\,(0, I_d)\,,$$

under Condition (3). Thus, under known homoskedasticity, the usual finite sample standard errors estimator from least squares theory turns out to be valid even when K is large. However, the "consistent" but biased standard errors estimator $(\hat{\varepsilon}'\hat{\varepsilon}/n)\Gamma_n^-$ will not be consistent unless $K/n \to 0$, which implies that using the finite sample, unbiased standard errors estimator under homoskedasticity is important even in large samples when the degrees of freedom is small (i.e., when K is large).

4.2 Heteroskedasticity

Under heteroskedasticity of unknown form, a natural candidate for standard errors estimator is the (family of) Eicker-Huber-White estimators given by

$$\hat{V}_{HR} = (\tilde{X}'\tilde{X})^- \tilde{X}'\Upsilon\tilde{X}(\tilde{X}'\tilde{X})^- = \frac{1}{n}\hat{\Gamma}_n^- \hat{\Sigma}_n \hat{\Gamma}_n^-,$$

$$\hat{\Sigma}_n = \frac{1}{n} X'M\Upsilon MX, \qquad \Upsilon = \operatorname{diag}\left(\omega_1\hat{\varepsilon}_1^2, \cdots, \omega_n\hat{\varepsilon}_n^2\right) = \begin{bmatrix} \omega_1\hat{\varepsilon}_1^2 & & \\ & \ddots & \\ & & \omega_n\hat{\varepsilon}_n^2 \end{bmatrix}$$

where $\{\omega_i : i = 1, \cdots, n\}$ are appropriate weights. Classical choices of weights include $\omega_i = 1$, $\omega_i = n/(n - K - d_x)$, $\omega_i = M_{ii}^{-1}$, etc. Since $\hat{\Gamma}_n = \Gamma_n + o_p(1)$ according to Lemma 1, it only remains

to characterized the middle matrix of this classical sandwich formula. Heuristically, the asymptotic properties of $\hat{\Sigma}_n$ are given by

$$\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n} \omega_i \tilde{x}_i \tilde{x}_i' \hat{\varepsilon}_i^2 \approx \frac{1}{n}\sum_{i=1}^{n} \omega_i \tilde{v}_i \tilde{v}_i' \tilde{\varepsilon}_i^2 \approx \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\omega_i \tilde{v}_i \tilde{v}_i' \tilde{\varepsilon}_i^2 \middle| Z\right],$$

where, as before, the first approximation captures the bias of the series estimator and removes the estimation error of β, and the second approximation shows that a (conditional) law of large numbers holds in this case (i.e., a variance condition). This idea is summarized in the following theorem.

Theorem 3. Suppose the assumptions of Theorem 1 hold with $\alpha_h > 1/2$, and $\omega_i \in \sigma(Z)$ for all i. Then, $\hat{\Sigma}_n = \tilde{\Sigma}_n + o_p(1)$, where

$$\tilde{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\sum_{\ell=1}^{n} \omega_i M_{i\ell}^2 M_{j\ell}^2\right) \mathbb{E}\left[v_i v_i' \varepsilon_j^2 \middle| z_i, z_j\right].$$

Recall that the population asymptotic middle matrix Σ_n is given by

$$\Sigma_n = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij}^2 \mathbb{E}[v_i v_i' \varepsilon_j^2 | z_i, z_j],$$

which implies that the classical HR standard errors will not be consistent in general when Condition (3). For example, the bias may be characterized under homoskedasticity: assuming $\mathbb{E}\left[\varepsilon_i^2 | x_i, z_i\right] = \sigma^2$ and $\omega_i = 1$, a simple calculation yields

$$\Sigma_n = \frac{\sigma^2}{n}\sum_{i=1}^{n} M_{ii} \mathbb{E}\left[v_i v_i' | z_i\right],$$

while, using basic properties of idempotent matrices, it is easy to verify that

$$\tilde{\Sigma}_n = \Sigma_n - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij}^2 (1 - M_{jj}) \mathbb{E}\left[v_i v_i' | z_i\right] < \Sigma_n.$$

As a consequence, the classical Eicker-Huber-White estimator is downward biased when K is "large". It is important to note that this result continue to hold in a simple linear model where the number of regressors is "large" when compared to the sample size. Therefore, there is an important sense in which the classical HR standard errors estimator is not robust, even in a simple linear model.

On the other hand, if $K/n \to 0$, then the asymptotic results presented above imply that $\tilde{\Sigma}_n = \Sigma_n + o_p(1)$, which verifies that the classical HR standard errors estimator is indeed consistent under heteroskedasticity of unknown form.

4.3 HR and Many Terms Robust Standard Errors

Intuitively, the failure of the classical HR standard errors estimator is due to the fact that both $\tilde{x}_i$ and $\hat{\varepsilon}_i$ are estimated with too much error when $K/n \not\to 0$. Thus, it is possible to fix this problem by considering alternative (consistent) estimators for either $\tilde{x}_i$ or $\hat{\varepsilon}_i$. To describe the new estimators, let K_g and K_h be two choices of truncation for an approximation series, and let

$$\tilde{X} = (I - Q_h)X, \qquad Q_h = P_{K_h}(P_{K_h}'P_{K_h})^{-}P_{K_h}', \qquad M_h = I - Q_h,$$

$$\hat{\varepsilon} = (I - Q_g)\varepsilon, \qquad Q_g = P_{K_g}(P_{K_g}'P_{K_g})^{-}P_{K_g}', \qquad M_g = I - Q_g.$$

Using this notation, Theorem 3 may be extended to the following result.

Theorem 4. Suppose the assumptions of Theorem 3 hold, then

$$\check{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\sum_{\ell=1}^{n}\omega_i M_{h,i\ell}^2 M_{g,j\ell}^2\right)\mathbb{E}\left[v_i v_i' \varepsilon_j^2 | z_i, z_j\right].$$

This theorem leads naturally to two alternative recipes for heteroskedasticity and many terms robust estimators. Specifically, for $\omega_i = 1$, if $\min\{K_h, K_g\} = o(n)$ and $\max\{K_h, K_g\} = K$, then $\check{\Sigma}_n = \Sigma_n + o_p(1)$. This result implies that

$$\hat{V}_{HET}^{-1/2}(\hat{\beta} - \beta_0) \to_d \mathcal{N}(0, I_d), \qquad \hat{V}_{HET} = \frac{1}{n}\hat{\Gamma}_n^{-}\check{\Sigma}_n\hat{\Gamma}_n^{-}.$$

5 Simulations

To explore the consequences of using many terms in the partially linear model, or alternatively using many covariates in a linear model, this section reports preliminary results from a Monte Carlo experiment. Specifically, the simulations consider the following model:

$$\begin{aligned} y_i &= x_i'\beta + g(z_i) + \varepsilon_i, & \varepsilon_i &= \sigma_\varepsilon(x_i, z_i)\, u_{1i}, \\ x_i &= h(z_i) + v_i, & v_i &= \sigma_v(z_i)\, u_{2i}, \end{aligned}$$

with $d_x = 1$, $d_z = 10$, $g(z) = 1$, $h(z) = 0$, and $u = (u_{1i}, u_{1i})' \sim \mathcal{N}(0, I_2)$ and $z_i \sim \mathcal{U}(-1, 1)$ independently of u. Note that this data generating process does not have smoothing bias. Four

models of heteroskedasticity are considered, as given in Table 1 (with $\iota = (1, 1, \cdots, 1)' \in \mathbb{R}^{d_z}$).

Table 1: Simulation Models ()

	$\sigma_v^2(z_i) = 1$	$\sigma_v^2(z_i) = (z_i'\iota)^2$
$\sigma_\varepsilon^2(x_i, z_i) = 1$	Model 1	Model 3
$\sigma_\varepsilon^2(x_i, z_i) = (z_i'\iota + x_i)^2$	Model 2	Model 4

For simplicity, the simulations consider additive-separable power series, that is, the unknown function $g(z_i)$ is assumed to satisfy $g(z_i) = 1 + g_1(z_{1i}) + \cdots + g_{d_z}(z_{d_z i})$ and each component is estimated by $g_j(z_{ji}) \approx p_K(z_{ji})'\gamma_j$, $j = 1, 2, \cdots, d_z$, with $p_K(z_{ji}) = (0, z_{ji}, z_{ji}^2, \cdots, z_{ji}^{K-1})'$.

We consider the classical least squares homoskedasticity-consistent standard errors estimators

$$V_{HO1} = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n}\hat{\Gamma}^- \qquad \text{and} \qquad V_{HO2} = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n - K - d_x}\hat{\Gamma}^-,$$

and the classical heteroskedasticity-consistent standard errors estimators

$$V_{HR1} = \hat{\Gamma}^-\tilde{X}'\Upsilon\tilde{X}\hat{\Gamma}^-, \qquad \Upsilon = \operatorname{diag}(\hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_n)/n,$$

$$V_{HR2} = \hat{\Gamma}^-\tilde{X}'\Upsilon\tilde{X}\hat{\Gamma}^-, \qquad \Upsilon = \operatorname{diag}(\hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_n)/(n - K - d_x).$$

Also, we report the two alternative heteroskedasticity- and many terms- robust standard errors estimators proposed in Theorem 4. These estimators are given by

$$V_{CJN1} = \hat{\Gamma}^-\tilde{X}'\Upsilon\tilde{X}\hat{\Gamma}^-, \qquad \Upsilon = \operatorname{diag}(\hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_n)/n,\ K_h = K_{CV},\ K_g = K,$$

$$V_{CJN2} = \hat{\Gamma}^-\tilde{X}'\Upsilon\tilde{X}\hat{\Gamma}^-, \qquad \Upsilon = \operatorname{diag}(\hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_n)/n,\ K_h = K,\ K_g = K_{CV},$$

where K_{CV} represents a cross-validation estimate of the optimal K.

The results are presented in Figure 1, for a grid of $K = (0, 1, 2, \cdots, 20)$. The effective degrees of freedom is determined by the choice of K and d_z.

6 Technical Appendix

All statements involving conditional expectations are understood to hold almost surely. Recall that $M = I - Q$ is symmetric and idempotent, and therefore $|M_{ii}| \leq 1$, $n - K = \sum_{i=1}^{n} M_{ii}$ and $M_{ij} = \sum_{\ell=1}^{n} M_{i\ell} M_{\ell j}$.

Proof of Lemma 1. It follows from $H'MH/n = o_p(1)$ and the Cauchy-Schwarz inequality that $X'MX/n = (V+H)' M (V+H)/n = V'MV/n + H'MH/n + 2H'MV/n = V'MV/n + o_p(1)$, provided that

$$\frac{1}{n} V'MV = \frac{1}{n} \sum_{i=1}^{n} M_{ii} v_i v_i' + \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} M_{ij} v_i v_j' = O_p(1).$$

First, using $|M_{ii}| \leq 1$ and the Markov inequality,

$$\frac{1}{n} \sum_{i=1}^{n} M_{ii} v_i v_i' = \frac{1}{n} \sum_{i=1}^{n} M_{ii} \mathbb{E}\left[v_i v_i' | z_i\right] + o_p(1),$$

because

$$\mathbb{V}\left[\frac{1}{n} \sum_{i=1}^{n} M_{ii} \|v_i\|^2 \middle| Z \right] = \frac{1}{n^2} \sum_{i=1}^{n} M_{ii}^2 \mathbb{V}[\|v_i\|^2 | z_i] = O_{as}\left(n^{-1}\right).$$

Similarly,

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} M_{ij} v_i v_j' = O_p(n^{-1} K^{1/2}) = o_p(1),$$

because

$$\begin{aligned} &\mathbb{E}\left[\left(\sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} M_{ij} c' v_i v_j' c \right) \left(\sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} M_{ij} c' v_i v_j' c \right)' \middle| Z \right] \\ &\leq 2 \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} M_{ij}^2 \mathbb{E}[\|v_i\|^2 | z_i] \mathbb{E}[\|v_j\|^2 | z_j] = O_{as}(K) \end{aligned}$$

for every d-vector c with $c'c = 1$, where the inequality uses basic moment calculations for quadratic forms. ■

Proof of Lemma 2. First note that

$$\frac{1}{\sqrt{n}} X'M(Y - X\beta_0) = \frac{1}{\sqrt{n}} V'M\varepsilon + \frac{1}{\sqrt{n}} H'M\varepsilon + \frac{1}{\sqrt{n}} X'MG,$$

where $H'M\varepsilon/\sqrt{n} = O_p\left(K^{-\alpha_h}\right)$ because

$$\mathbb{E}\left[\left\| \frac{1}{\sqrt{n}} H'M\varepsilon \right\|^2 \middle| Z \right] = \frac{1}{n} \operatorname{tr}\left(H'M\mathbb{E}[\varepsilon\varepsilon'|Z]MH\right) \leq \frac{C}{n} \operatorname{tr}\left(H'MH\right) = O\left(K_n^{-2\alpha_h}\right).$$

Next,

$$\frac{1}{\sqrt{n}} X'Mg = \frac{1}{\sqrt{n}} V'MG + \frac{1}{\sqrt{n}} \bar{X}'MG = O_p\left(K^{-\alpha_g} + \sqrt{n} K^{-(\alpha_X + \alpha_g)}\right),$$

because

$$\mathbb{E}\left[\left\|\frac{1}{\sqrt{n}}V'MG\right\|^2\middle|Z\right]=\frac{1}{n}G'M\mathbb{E}\left[VV'|Z\right]MG\leq\frac{C}{n}G'MG=O_{as}\left(K^{-2\alpha_g}\right),$$

and

$$\frac{1}{\sqrt{n}}H'MG\leq\sqrt{n}\sqrt{\frac{1}{n}H'MH}\sqrt{\frac{1}{n}G'MG}=O_{as}\left(\sqrt{n}K^{-(\alpha_X+\alpha_g)}\right),$$

which completes the proof. ■

Lemma 4. Suppose $\mathbb{E}[\left\|V_i\right\|^4|Z_i]<C_V<\infty$ and $\mathbb{E}\left[\varepsilon_i^4|Z_i\right]<C_\varepsilon<\infty$ (a.s.). If $1-Q_{ii}>F_Q>0$ (a.s.) and if $\lambda_{\min}\left(\Sigma_n\right)>F_\Sigma>0$, then

$$\Sigma_n^{-1/2}\frac{1}{\sqrt{n}}V'(I-Q)\varepsilon\rightarrow_d\mathcal{N}\left(0,I_d\right).$$

Proof of Lemma 4. Use Lemma A2 in Chao, Swanson, Hausman, Newey, and Woutersen (2009). ■

Proof of Theorem 2. First, it follows from $G'MG/n=o_p\left(1\right)$ and the Cauchy-Schwarz inequality that

$$\begin{aligned}\frac{1}{n}\hat{\varepsilon}'\hat{\varepsilon}&=\frac{1}{n}(Y-X\hat{\beta})'M(Y-X\hat{\beta})\\&=\frac{1}{n}(Y-X\hat{\beta}-G)'M(Y-X\hat{\beta}-G)+\frac{1}{n}G'MG-\frac{2}{n}(Y-X\hat{\beta}-G)'MG\\&=\frac{1}{n}(Y-X\hat{\beta}-G)'M(Y-X\hat{\beta}-G)+o_p(1),\end{aligned}$$

provided $(Y-X\hat{\beta}-G)'M(Y-X\hat{\beta}-G)/n=O_p\left(1\right)$. Next, note that Lemma 1 and $\hat{\beta}-\beta=o_p(1)$ imply $(\hat{\beta}-\beta)'X'MX(\hat{\beta}-\beta)/n=o_p\left(1\right)$, which together with the Cauchy-Schwarz inequality gives

$$\begin{aligned}&\frac{1}{n}(Y-X\hat{\beta}-G)'M(Y-X\hat{\beta}-G)\\&=\frac{1}{n}\varepsilon'M\varepsilon+\frac{1}{n}(\hat{\beta}-\beta)'X'MX(\hat{\beta}-\beta)-\frac{2}{n}(Y-X(\hat{\beta}-\beta)-G)'M(\hat{\beta}-\beta)\\&=\frac{1}{n}\varepsilon'M\varepsilon+o_p(1)=\frac{1}{n}\tilde{\varepsilon}'\tilde{\varepsilon}+o_p(1).\end{aligned}$$

Finally, consider the bilinear form

$$\frac{1}{n}\tilde{\varepsilon}'\tilde{\varepsilon}=\frac{1}{n}\varepsilon'M\varepsilon=\frac{1}{n}\sum_{i=1}^n M_{ii}\varepsilon_i^2+\frac{1}{n}\sum_{i=1}^n\sum_{j=1,j\neq i}^n\varepsilon_iM_{ij}\varepsilon_j.$$

First, using $|M_{ii}|\leq1$ and the fact that $S_n=\mathbb{E}\left[S_n|Z\right]+O_p\left(\left(\mathbb{E}\left[\mathbb{V}\left[S_n|Z\right]\right]\right)^{1/2}\right)$,

$$\frac{1}{n}\sum_{i=1}^n M_{ii}\varepsilon_i^2=\frac{1}{n}\sum_{i=1}^n M_{ii}\mathbb{E}\left[\varepsilon_i^2|z_i\right]+o_p\left(1\right)=\frac{n-K}{n}\sigma^2+o_p(1),$$

because

$$\mathbb{E}\left[\mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} M_{ii}\varepsilon_i^2 \middle| Z\right]\right] = \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n} M_{ii}^2 \mathbb{V}[\varepsilon_i^2|z_i]\right] \leq \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{E}\left[\mathbb{V}[\varepsilon_i^2|z_i]\right] \leq \frac{C_V}{n} = o\left(1\right).$$

Second, using similar arguments,

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} M_{ij}\varepsilon_i\varepsilon_j = O_p\left(n^{-1}K^{1/2}\right) = o_p\left(1\right),$$

because

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} M_{ij}\varepsilon_i\varepsilon_j\right)^2 \middle| Z\right] = \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} M_{ij}^2 \mathbb{E}[\varepsilon_i^2|z_i]\mathbb{E}[\varepsilon_j^2|z_j] \leq \frac{2C_V K}{n^2}.$$

Therefore, because $(n-K)/(n-K-d) \to 1$,

$$\hat{\sigma}^2 = \frac{1}{n-K-d}\hat{\varepsilon}'\hat{\varepsilon} = \sigma^2 + o_p(1),$$

which completes the proof. ■

Proof of Theorem 3. Special case of Theorem 4. ■

Proof of Theorem 4. Available upon request. ■

References

Abadie, A., and G. W. Imbens (2006): "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74(1), 235–267.

Cattaneo, M. D., R. K. Crump, and M. Jansson (2010): "Small Bandwidth Asymptotics for Density-Weighted Average Derivatives," working paper.

Chao, J. C., N. R. Swanson, J. A. Hausman, W. K. Newey, and T. Woutersen (2009): "Asymptotic Distribution of JIVE in a Heteroskedastic IV Regression with Many Instruments," working paper.

Chen, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics, Volume VI*, ed. by J. Heckman, and E. Leamer, pp. 5550–5632. Elsevier Science B.V.

Donald, S. G., and W. K. Newey (1994): "Series Estimation of Semilinear Models," *Journal of Multivariate Analysis*, 50(1), 30–40.

Hansen, C., J. Hausman, and W. K. Newey (2008): "Estimation with Many Instrumental Variables," *Journal of Business and Economic Statistics*, 26(4), 398–422.

Linton, O. (1995): "Second Order Approximation in the Partialy Linear Regression Model," *Econometrica*, 63(5), 1079–1112.

Newey, W. K. (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79, 147–168.

Newey, W. K., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics, Volume IV*, ed. by R. F. Engle, and D. L. McFadden, pp. 2112–2245. Elsevier Science B.V.

Robinson, P. M. (1988): "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56(4), 931–954.

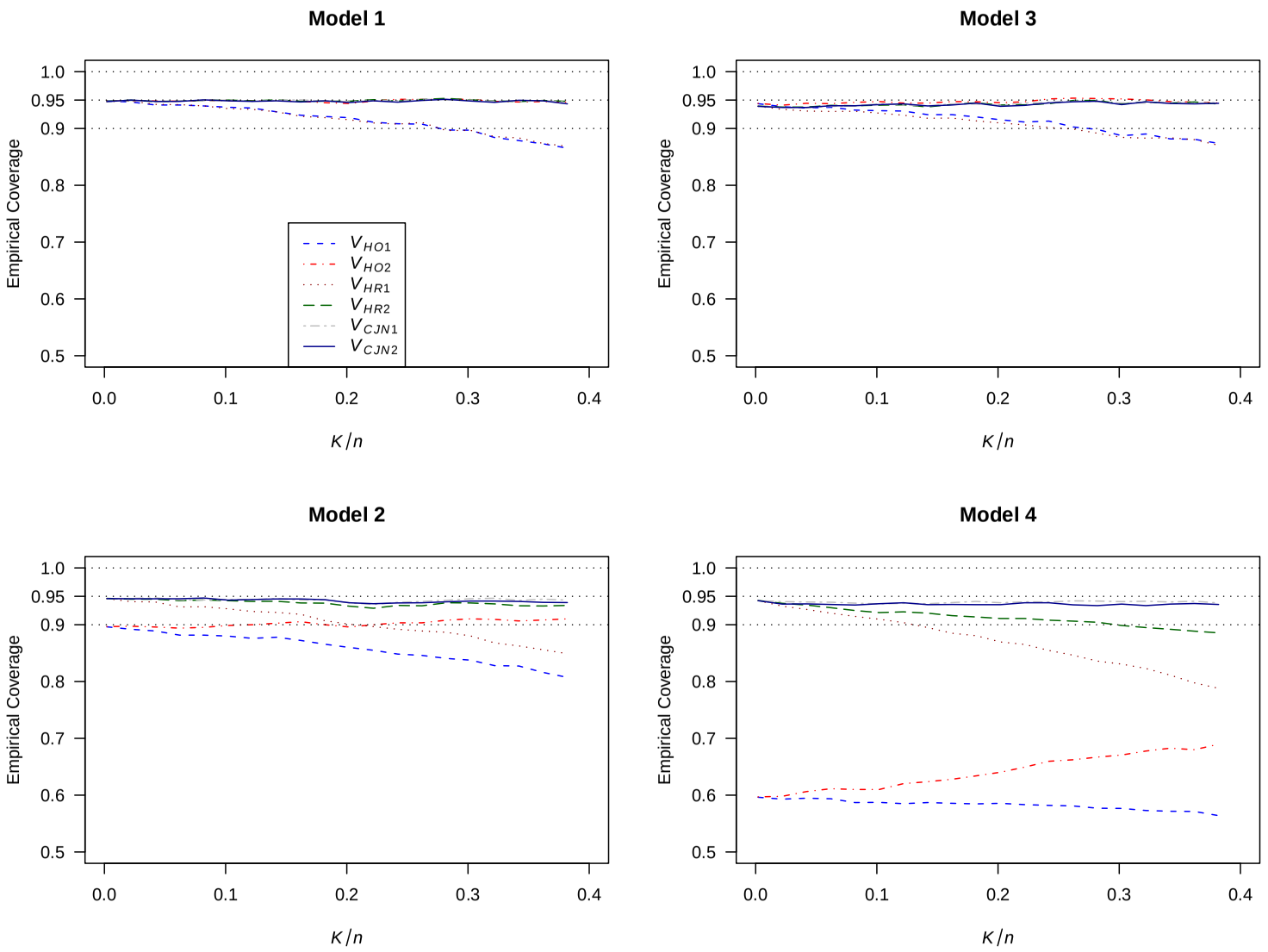

Figure 1: Empirical Coverage Rates for 95% Confidence Intervals: $n = 500$, $S = 3,000$