Parameter-invariant unbiased estimates of individual variances

Stanislav Anatolyev* CERGE-EI and NES

Yaroslav Korobka† CERGE-EI

Abstract

Recent literature proposed variants of estimators of conditional error variances that are robust to the presence of many regressors. One of the proposals produces unbiased estimates, which, however, are sensitive to regression function values. We propose its modification based on cross-fitting and averaging over random sample splits, which preserves the unbiasedness but eliminates that sensitivity. We analyze the properties of these estimates and compare them with alternatives in a simulation setup focused on asymptotic variance estimation when there are many covariates.

Keywords: linear regression, many regressors, variance estimation

*Corresponding author. Address: CERGE-EI, Politických vězňů 7, 11121 Prague 1, Czech Republic. E-mail: stanislav.anatolyev@cerge-ei.cz. This research was supported by the grant 24-12720S from the Czech Science Foundation.

†Address: CERGE-EI, Politických vězňů 7, 11121 Prague 1, Czech Republic. E-mail: yaroslav.korobka@cerge-ei.cz.

1 Introduction

We consider the linear regression model

$$y_i = x_i'\theta + e_i, \quad E\left[e_i|x_i\right] = 0,$$

where x_i and θ are $m \times 1$, and m may be comparable with the number of observations n. The pairs of observations $(y_i, x_i)_{i\in\mathcal{H}}$, where $\mathcal{H} = \{1, \cdots, n\}$, constitute a random sample. Let us define individual variances:

$$\sigma_i^2 = E\left[e_i^2|x_i\right],$$

assuming that they exist. Kline, Saggio, and Sølvsten (2020) proposed unbiased estimates

$$\hat{\sigma}_i^2 = \frac{y_i\hat{e}_i}{M_{ii}} \tag{1}$$

for each $i \in \mathcal{H}$, where $\hat{e}_i = y_i - x_i'\hat{\theta}$ is i^{th} OLS residual, M_{ii} is i^{th} diagonal element of the annihilation matrix $M = I_n - X\left(X'X\right)^{-1}X'$, and X is the regressor matrix. We label estimates (1) by KSS. Such individual variance estimates are further utilized in Mikusheva and Sun (2022), Jochmans (2022), Anatolyev and Sølvsten (2023), Boot (2023) and other works related to development of regression theories with many regressors/covariates/instruments.

The KSS estimates originate from leave-one-out OLS estimation and the identity between the OLS residuals corrected for leverage and leave-one-out OLS residuals. The KSS estimates are indeed conditionally unbiased:

$$\begin{aligned} E\left[\hat{\sigma}_i^2|X\right] &= \frac{1}{M_{ii}}E\left[(x_i'\theta + e_i)\Big(\sum_{j\in\mathcal{H}} M_{ij}e_j\Big)|X\right] \\ &= M_{ii}^{-1}\sum_{j\in\mathcal{H}} M_{ij}E\left[e_ie_j|X\right] \\ &= \sigma_i^2. \end{aligned}$$

However, the presence of y_i, which contains $x_i'\theta$, in their construction makes the KSS estimates sensitive to the value of θ and, in particular, prone to generating outliers when $\theta \neq 0$.

The effect of dependence on parameter values can be partially offset by "demeaning" y_i in (1), by subtracting the sample mean $\bar{y}$ from each y_i, as is done in, e.g., Anatolyev and Sølvsten (2023). We label such estimates by KSS+. Clearly, this does not entirely remove the dependence on θ, as $x_i'\theta$ is simply replaced by $(x_i - \bar{x})'\theta$; in addition, exact unbiasedness is lost, as e_i is replaced by $e_i - \bar{e}$, which interacts not only with $M_{ii}e_i$, but also with the rest of $\sum_{j\in\mathcal{H}} M_{ij}e_j$.

2 Cross-fit estimates

It would be desirable to completely eliminate the dependence on θ, by using some proxies for e_i in place of y_i, that would be independent of $\sum_j M_{ij}e_j$. The cross-fit machinery provides such a possibility, although not without a cost. The idea was previously mentioned in Jochmans (2022, end of Section 3), and we extend it further by employing averaging over random sample splits.

Split the sample into two i-specific non-overlapping subsamples, $\mathcal{H} = \mathcal{H}_i \cup \mathcal{H}_{-i}$ such that $i \in \mathcal{H}_i$. Then, define

$$\mathring{\sigma}_i^2 = \frac{\mathring{e}_{\mathcal{H}_i,i}\mathring{e}_{\mathcal{H}_{-i},i}}{M_{\mathcal{H}_i,ii}}, \tag{2}$$

where $M_{\mathcal{H}_i} = I_{|\mathcal{H}_i|} - X_{\mathcal{H}_i}\left(X_{\mathcal{H}_i}'X_{\mathcal{H}_i}\right)^{-1}X_{\mathcal{H}_i}'$ is the annihilation matrix based only on regressors $X_{\mathcal{H}_i}$ included in $\mathcal{H}_i$, $M_{\mathcal{H}_i,ij}$ is its $(i,j)^{th}$ element, $\mathring{e}_{\mathcal{H}_i,i}$ is i^{th} residual based on the OLS estimator using subsample $\mathcal{H}_i$, and, similarly, $\mathring{e}_{\mathcal{H}_{-i},i}$ is i^{th} residual based on the OLS estimator using subsample $\mathcal{H}_{-i}$. We label estimates (2) by CF, for "cross-fit." The sample splitting embedded in CF estimates ensures conditional unbiasedness:

$$\begin{aligned}
E\left[\mathring{\sigma}_i^2|X\right] &= \frac{1}{M_{\mathcal{H}_i,ii}}E\left[\Big(\sum_{j\in\mathcal{H}_i}M_{\mathcal{H}_i,ij}e_j\Big)\Big(e_i+\sum_{k\in\mathcal{H}_{-i}}M_{\mathcal{H}_{-i},ik}e_k\Big)|X\right] \\
&= M_{\mathcal{H}_i,ii}^{-1}\sum_{j\in\mathcal{H}_i}M_{\mathcal{H}_i,ij}E\left[e_je_i|X\right] \\
&\quad + M_{\mathcal{H}_i,ii}^{-1}\sum_{j\in\mathcal{H}_i}\sum_{k\in\mathcal{H}_{-i}}M_{\mathcal{H}_i,ij}M_{\mathcal{H}_{-i},ik}E\left[e_je_k|X\right] \\
&= M_{\mathcal{H}_i,ii}^{-1}M_{\mathcal{H}_i,ii}E\left[e_i^2|X\right] + 0 \\
&= \sigma_i^2.
\end{aligned}$$

Note that in order for both OLS residuals $\mathring{e}_{\mathcal{H}_i,i}$ and $\mathring{e}_{\mathcal{H}_{-i},i}$ to be defined, both subsamples $\mathcal{H}_i$ and $\mathcal{H}_{-i}$ need to have at least m observations, which restricts applicability of CF to setups when m is at most half of the sample size n.

Now, the split of $\mathcal{H}$ into $\mathcal{H}_i$ and $\mathcal{H}_{-i}$ is arbitrary, even subject to conditions $i \in \mathcal{H}_i$ and $|\mathcal{H}_i| = \lfloor n/2 \rfloor$. This gives an additional possibility of generating a number of sets of CF estimates based on $\ell \geq 2$ random splits for each fixed $i \in \mathcal{H}$, and further averaging them:

$$\mathring{\sigma}_i^2 = \frac{1}{\ell}\sum_{i=1}^{\ell}(\mathring{\sigma}_i^2)_l,$$

where $(\mathring{\sigma}_i^2)_l$'s are estimates (2) from l^{th} random sample split. The original CF estimator (2) corresponds to $\ell = 1$.

The necessary restriction $|\mathcal{H}_i| = \lfloor n/2 \rfloor$ is reminiscent of the sufficient condition figuring in the alternative individual variance estimates of Cattaneo, Jansson, and Newey (2018), CJN for short, derived from different principles:

$$\breve{\sigma}_i^2 = \left((M \odot M)^{-1} \hat{e} \odot \hat{e}\right)_i. \tag{3}$$

Additionally, Anatolyev (2018) proposed a finite sample improvement of estimates (3) based on ensuring estimation unbiasedness under homoskedasticity. This modification, which we label by CJN+, requires subtracting $P \odot P$ from $M \odot M$ in (3), where $P = I_n - M$, before taking an inverse. Note that in both cases, one needs to invert an $n \times n$ matrix, while KSS/KSS+ and CF do not require that. For completeness, we will have a look at properties of these two methods as well.

3 Simulation evidence

We compare the properties of CJN, CJN+, KSS, KSS+, and CF (with $\ell = 10$) estimates in the context of a regression with one regressor and many covariates, in which case $x_i' = (z_i, w_i')$, $\theta' = (\beta, \gamma')$, and w and γ are $q \times 1$, where $q = m - 1$. The OLS estimate of the parameter of interest β equals $\left(\sum_{i\in\mathcal{H}} \hat{v}_i^2\right)^{-1} \sum_{i\in\mathcal{H}} \hat{v}_i y_i$, and the asymptotic variance estimate is computed as $\left(\sum_{i\in\mathcal{H}} \hat{v}_i^2\right)^{-2} \sum_{i\in\mathcal{H}} \hat{v}_i^2 \tilde{\sigma}_i^2$, where $\hat{v}_i = \sum_{j\in\mathcal{H}} \left(\mathbb{I}_{\{i=j\}} - w_i'(\sum_{k\in\mathcal{H}} w_k w_k')^{-1} w_j\right) x_j$ and $\tilde{\sigma}_i^2$'s are one of the five sets of individual variance estimates. For CJN/CJN+ and KSS/KSS+, if the asymptotic variance estimator turns out negative, we replace it with the White variance estimate; for CF, we dispose of it and compute it anew with another set of random splits, and continue until the estimate turns out positive.[1]

In simulations, $z_i \sim \mathcal{N}(0,1)$, w_i contains unity and a collection of $q-1$ independent standard normal random variables. The error term is heteroskedastic and generated as $e_i = \vartheta(z_i^2 + q^{-1} w_i' w_i) \cdot \mathcal{N}(0,1)$, where ϑ is such that $var(e_i) = 1$. The nuisance parameter is $\gamma = (g, ..., g)'$, with g such that the partial R^2 in a regression of z_i on w_i equals 0.7. The true parameter of interest is $\beta = 5$, but later we also analyze the dependence on the value of β. To focus on small sample properties, we set the sample size to $n = 200$.

First, we look at some features of the distributions of individual variance estimates, in particular, how often the estimates are negative and how often they generate outliers. Table 1 shows the percentages of negative estimates and selected quantiles of the distribution of the

[1] In our simulation experiments, we usually had to invoke this replacement less than five times per asymptotic variance estimate, if at all.

Table 1: Percentages of negative and empirical quantiles of positive individual variance estimates

	CJN	CJN+	KSS	KSS+	CF
	percentages of negative estimates				
$\% < 0$	42.9%	41.8%	48.2%	48.1%	41.2%
	quantiles of positive estimates				
1%	0.007	0.007	0.008	0.008	0.006
2%	0.014	0.014	0.018	0.018	0.012
10%	0.077	0.075	0.134	0.134	0.068
50%	0.670	0.663	1.66	1.66	0.641
90%	4.13	4.11	12.4	12.3	4.11
98%	13.0	13.0	37.2	37.0	13.3
99%	19.4	19.3	53.3	53.0	20.0

positive estimates, collected from 20,000 simulations when $q = 41$, and Figure 1 depicts their smoothed censored densities. The percentage of negative estimates is comparable across all estimators, though is slightly higher for KSS and KSS+, while averaging embedded in CF reduces this fraction to that of CJN/CJN+. The left tails of the KSS and KSS+ censored densities are slightly thinner than the others', but their medians exceed the others' by almost threefold, and their right tails, correspondingly, contain many more outliers. In all parts of the distribution, CF makes corrections to KSS, so that the distribution gets so close to CJN's that in Figure 1 they are visually indistinguishable. To conclude, the averaging over random sample splits helps reduce the presence of negative estimates and outliers.

Second, we analyze the ability to control the size of a two-sided t-test at the 5% level for a hypothesis that β equals its true value. Table 2 contains actual rejection rates for different values of covariate numerosity $q \in \{21, 41, 71, 81\}$, computed from 5,000 simulations. Clearly, the use of KSS and KSS+ estimates lead to quite perceptible size distortions, while the use of CF eliminates most of distortions taking them to an even smaller level than that of CJN+, which slightly improves on the original CJN. Moreover, the remaining distortions from the use of CF (as well as of CJN and CJN+) do not tend to vary with covariate dimensionality, while the distortions from KSS and KSS+ tend to increase with it.

Figure 1: Distribution of individual variance estimates (abscissa axis in log-scale)

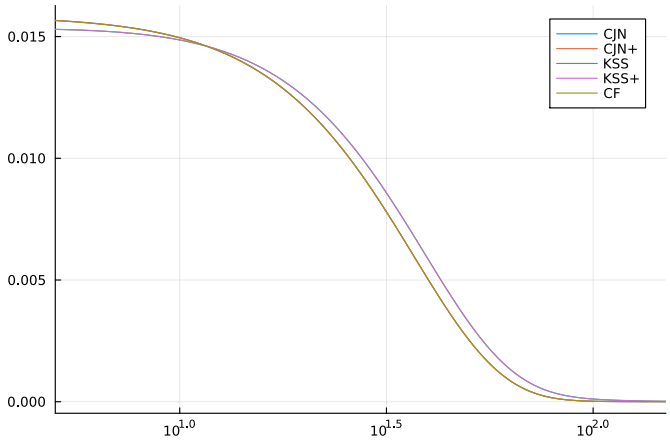


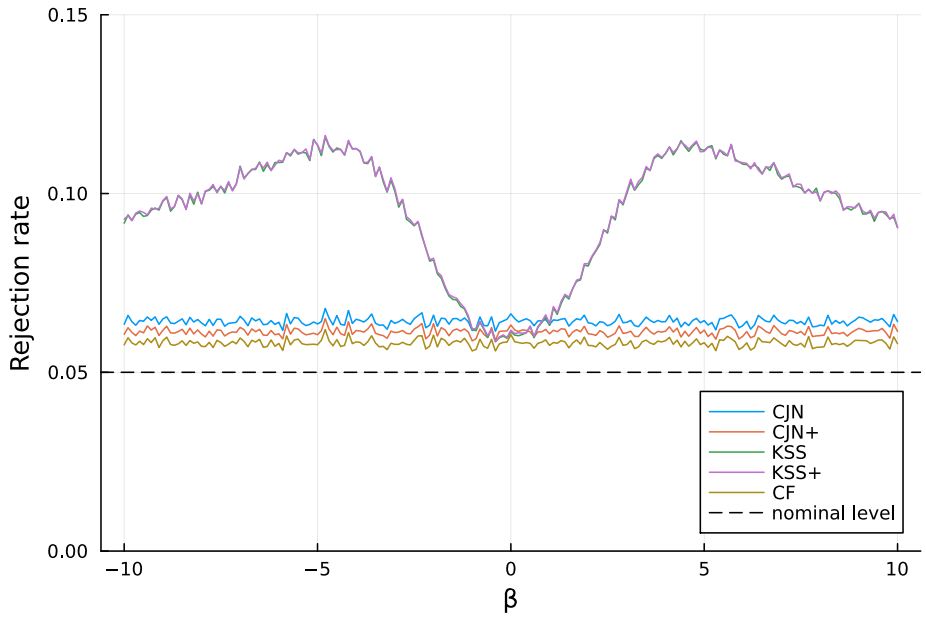
Table 2: Actual rejection rates and percentages of negative asymptotic variance estimates

q	21	41	71	81
CJN	0.062 0.0%	0.067 0.0%	0.064 0.0%	0.067 0.0%
CJN+	0.060 0.0%	0.064 0.0%	0.060 0.0%	0.063 0.0%
KSS	0.101 8.5%	0.116 9.0%	0.122 10.6%	0.134 10.9%
KSS+	0.102 8.2%	0.117 9.0%	0.124 10.6%	0.132 10.9%
CF	0.057 0.0%	0.061 0.0%	0.057 0.0%	0.058 0.0%

Table 2 also contains (below rejection rates in a small font) percentages of negative definite asymptotic variance estimates, in which cases they are replaced by White estimates. One can see that while the CJN and CJN+ estimates are practically free of such a phenomenon, the KSS/KSS+ estimates are susceptible to it in a non-trivial number of times, around 10% of cases. The recursive replacement embedded in the cross-fitting algorithm, however, is able to free the CF asymptotic variance estimates from the phenomenon as well.

We have also experimented with some non-normal covariate distributions such as Bernoulli, lognormal, and Pareto. The rejection rate patterns stay similar to those reported in Table 2, but the proportions of negative definite variance estimators from the use of KSS increases when the covariate distribution is lognormal (to up to 40%) or Pareto (to up to 45%), though demeaning in KSS+ decreases these figures (to about 20% and about 30%, respectively). The corresponding figures in all cases are virtually zero for CJN/CJN+, and zero by construction

Figure 2: Rejection rates of t-test with different variance estimators



for CF.

The large distortions from the use of KSS and KSS+ estimates documented in Table 2 occur primarily because $\beta = 5$ is substantially far from zero. Figure 2 shows explicitly the dependence of rejection rates on the value of β in the range from -10 to 10 when $q = 41$, computed from 50,000 simulations for each β on a grid. One can see strong dependence of size distortions on values of β outside a small vicinity of zero for KSS/KSS+ and total insensitiveness for the other methods. Notably, the demeaning in KSS+ does not help to free KSS from the scale effect, so that the two curves are almost identical, while CF does this job perfectly.

4 Concluding remarks

We have proposed a cross-fitting modification of the Kline, Saggio, and Sølvsten (2020) individual variance estimates that relieves them from the scale effect and retains exact unbiasedness. In simulations, the cross-fit estimates do exhibit insensitivity to parameter values and other properties comparable to the alternative Cattaneo, Jansson, and Newey (2018) estimates. The cross-fit modification could replace the baseline in new econometric theories that utilize such estimates or their extensions.

5 References

Anatolyev, S. (2018). Almost unbiased variance estimation in linear regressions with many covariates. *Economics Letters*, 169, 20–23.

Anatolyev, S. and M. Sølvsten (2023). Testing many restrictions under heteroskedasticity. *Journal of Econometrics*, 236(1), 105473.

Boot, T. (2023). Joint inference based on Stein-type averaging estimators in the linear regression model. *Journal of Econometrics*, 235(2), 1542–1563.

Cattaneo, M., M. Jansson, and W.K. Newey (2018). Inference in linear regression models with many covariates and heteroscedasticity. *Journal of the American Statistical Association*, 113(523), 1350-1361.

Jochmans, K. (2022). Heteroscedasticity-robust inference in linear regression models with many covariates. *Journal of the American Statistical Association*, 117(538), 887–896.

Kline, P., R. Saggio, and M. Sølvsten (2020). Leave-out estimation of variance components. *Econometrica*, 88(5), 1859–1898.

Mikusheva, A. and L. Sun (2022). Inference with many weak instruments. Review of Economic Studies, 89(5), 2663–2686.