Parameter-invariant unbiased estimation of individual variances and their pairwise products

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 parameter values. We propose its modification based on cross-fitting and averaging over random sample splits, which preserves the unbiasedness but eliminates that sensitivity. Then, we extend the cross-fitting machinery to unbiased estimation of pairwise variance products and to testing for many restrictions. We verify properties of associated estimators and tests in a simulation setup with 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. We thank the Editor and two anonymous expert referees for useful suggestions.

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

$$\hat{\sigma}_i^2 = \hat{e}_{-i}\, y_i, \tag{1}$$

where $\hat{e}_{-i}$ is i^{th} leave-one-out OLS residual computed using sample $\mathcal{H}$. We label estimates (1) by KSS.

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, and, in addition, the exact unbiasedness is lost.

2 Individual variance 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 $\hat{e}_{-i}$. The cross-fit machinery provides such a possibility, although not without a cost. The idea was previously mentioned in Jochmans (2022, Section 3). We extend it further by employing averaging over random sample splits; in the next Section, we provide an extension to pairwise variance products.

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

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

where $\mathring{e}_{\mathcal{H}_i,-i}$ is i^{th} leave-one-out OLS residual computed using subsample $\mathcal{H}_i$, and $\mathring{e}_{\mathcal{H}_{-i},i}$ is i^{th} OLS residual computed using subsample $\mathcal{H}_{-i}$. We label estimates (2) by CF, for "cross-fit."

The sample splitting embedded in CF estimates ensures conditional unbiasedness (see SA1).[1]

In order for both residuals in (2) 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. As for proportions of the split, an equal split intuitively exploits information in the subsamples optimally, on top of allowing higher m. An inspection of the conditional variance of $\mathring{\sigma}_i^2$ reveals (see SA2) that in subsample-average terms it is minimal when the split is indeed equal, at least when m is of a smaller asymptotic order than 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 split of $\mathcal{H}$. The original CF estimator (2) corresponds to $\ell = 1$.

The restriction $|\mathcal{H}_i| = \lfloor n/2 \rfloor$ for all i is reminiscent of the sufficient condition $m/n < 1/2$ figuring in the alternative individual variance estimates of Cattaneo, Jansson, and Newey (2018), CJN for short, derived from different principles:

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

where M is the annihilation matrix and $\hat{e}$ is the OLS residual vector. Further, 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 $(I_n - M) \odot (I_n - M)$ from $M \odot M$ in (3) before taking an inverse. For completeness, we will have a look at properties of these two methods as well.

3 Variance product estimates

Some recent developments in regression analysis require estimation of pairwise products of individual variances

$$\omega_{ij} = \sigma_i^2 \sigma_j^2.$$

Such products arise in the literature on many weak instruments (Mikusheva and Sun, 2022) and inference with many regressors (Anatolyev and Sølvsten 2023, Boot 2023). We next extend the cross-fit idea to estimation of variance products.

[1] Henceforth, SA1 through SA5 refer to sections of the Supplementary Appendix available online at https://pages.nes.ru/sanatoly/Papers/PIUE.htm

Split the sample $\mathcal{H}$ into four (i,j)-specific non-overlapping subsamples,

$$\mathcal{H} = \mathcal{H}_{(i,-j)} \cup \mathcal{H}_{(-i,j)} \cup \mathcal{H}^1_{(-i,-j)} \cup \mathcal{H}^2_{(-i,-j)}, \tag{4}$$

such that $i \in \mathcal{H}_{(i,-j)}$ and $j \notin \mathcal{H}_{(i,-j)}$, $j \in \mathcal{H}_{(-i,j)}$ and $i \notin \mathcal{H}_{(-i,j)}$, and $i,j \notin \mathcal{H}^\varsigma_{(-i,-j)}$ for $\varsigma \in \{1,2\}$. Define the CF estimator by

$$\mathring{\omega}_{ij} = \mathring{e}_{\mathcal{H}_{(i,-j)},-i}\, \mathring{e}_{\mathcal{H}^1_{(-i,-j)},i}\, \mathring{e}_{\mathcal{H}_{(-i,j)},-j}\, \mathring{e}_{\mathcal{H}^2_{(-i,-j)},j}, \tag{5}$$

where leave-one-out and usual OLS residuals are computed using corresponding subsamples. Naturally, each subsample needs to have at least m observations, which restricts applicability to setups when m is at most a quarter of n. The spliting ensures conditional unbiasedness of $\mathring{\omega}_{ij}$, see SA1. Note that $\mathring{\omega}_{ij}$ is permutation invariant with respect to indices, i.e. $\mathring{\omega}_{ij} = \mathring{\omega}_{ji}$ for any pair $i,j \in \mathcal{H}$. Similarly to $\mathring{\sigma}^2_i$, one may average over ℓ random partitions of $\mathcal{H}$:

$$\mathring{\omega}_{ij} = \frac{1}{\ell}\sum_{l=1}^{\ell} (\mathring{\omega}_{ij})_l,$$

where $(\mathring{\omega}_{ij})_l$ is an estimate (5) from l^{th} random split of $\mathcal{H}$.

Anatolyev and Sølvsten (2023) develop a test, hereby labeled by AS, for many restrictions in a heteroskedastic linear regression setup, using unbiased estimation of individual variances and pairwise variance products. The test statistic has the form

$$\frac{\mathcal{F} - \hat{E}_{\mathcal{F}}}{\hat{V}^{1/2}_{\mathcal{F}}},$$

where $\mathcal{F}$ is the conventional F-statistic, $\hat{E}_{\mathcal{F}}$ estimates the conditional mean of $\mathcal{F}$ using KSS estimates, and $\hat{V}_{\mathcal{F}}$ is an unbiased estimate of the conditional variance of the difference $\mathcal{F} - \hat{E}_{\mathcal{F}}$, which uses leave-three-out machinery to estimate variance products (see Anatolyev and Sølvsten 2023, formula (12)). In the end, neither $\hat{E}_{\mathcal{F}}$ nor $\hat{V}_{\mathcal{F}}$ is parameter-invariant.

We derive a modification of the AS test, hereby labeled by AS+, that uses only parameter-invariant elements. The test statistic has a similar form, but uses, in place of $\hat{E}_{\mathcal{F}}$ and $\hat{V}_{\mathcal{F}}$, the analogs based on an arbitrary sample split $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2$ and estimates (2) and (5):

$$\mathring{E}_{\mathcal{F}} = \sum_{i=1}^{n} B_{ii}\mathring{\sigma}^2_i$$

and

$$\mathring{V}_{\mathcal{F}} = 2\sum_{i\neq j\in\mathcal{H}_1} C^2_{1,ij}\mathring{\omega}_{ij} + 2\sum_{i\neq j\in\mathcal{H}_2} C^2_{2,ij}\mathring{\omega}_{ij} + \sum_{i\in\mathcal{H}_1}\sum_{j\in\mathcal{H}_2} \left(D_{2,ij} + D_{1,ji}\right)^2 \mathring{\omega}_{ij}$$

for observable B_{ii}, $C_{\varsigma,ij}$ and $D_{\varsigma,ij}$, $\varsigma \in \{1,2\}$. See SA4 for a derivation, implementation details, and discussion of averaging over sample splits.

4 Computational aspects

In terms of computational burden, sample splitting techniques do not pose overwhelming difficulties. Generally, running OLS on an $n \times m$ regressor matrix takes $\mathcal{O}(nm^2)$ operations. For CF, averaging over ℓ splits then yields the $\mathcal{O}(\ell nm^2)$ complexity, which grows linearly with ℓ. For CJN-type estimators, the complexity is dominated by inversion of the matrix $M \in \mathbb{R}^{n\times n}$, which leads to the total $\mathcal{O}(n^3)$ complexity. Thus, the computational complexity of CJN increases with n significantly faster than for CF. When m is really large and comparable to n, computation of CF becomes comparable to that of CJN if the number of splits is not large. However, typically, m is a relatively small fraction of n, which leaves a large leverage for ℓ to be big so that computations of CF and CJN are comparable. CF is also less demanding with respect to computer memory, since storing large dense matrices is costly.

The four-split estimator, in turn, requires four random subsamples subject to restrictions (4) placed on all (i, j) pairs. In SA3, we describe a simple algorithm that enforces the restriction, at the same time avoiding expensive computation of matrix inverses for each pair (i, j) separately. This algorithm can also be readily exploited by the AS+ test.

As the proposed estimators are constructed using some form of a random sample split, they are not unique for a given sample. The approach mitigating this issue is recently proposed by Ritzwoller and Romano (2026), who develop a simple procedure for sequentially aggregating statistics constructed with multiple splits of the same sample, which guarantees reproducible results up to a specified bound and nominal error rate. In our setting, this corresponds to selecting some large value of ℓ when averaging over sample splits.

5 Simulation evidence

We consider a regression setup 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.[2]

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

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 $n = 200$. The number of simulations varies from 5,000 to 50,000.

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

	σ_i^2	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%	0.000	0.007	0.007	0.008	0.008	0.006
10%	0.006	0.077	0.075	0.134	0.134	0.068
50%	0.185	0.670	0.663	1.66	1.66	0.641
90%	2.07	4.13	4.11	12.4	12.3	4.11
99%	13.9	19.4	19.3	53.3	53.0	20.0

First, we compare the properties of CJN, CJN+, KSS, KSS+, and CF (with $\ell = 10$) 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 positive estimates when $q = 41$, along with the quantiles of population variances. 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 such corrections to KSS that its distribution gets very close to CJN's. Thus, 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 $q \in \{21, 41, 71, 81\}$. 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.

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%

Figure 1: Rejection rates of t-test with different variance estimators

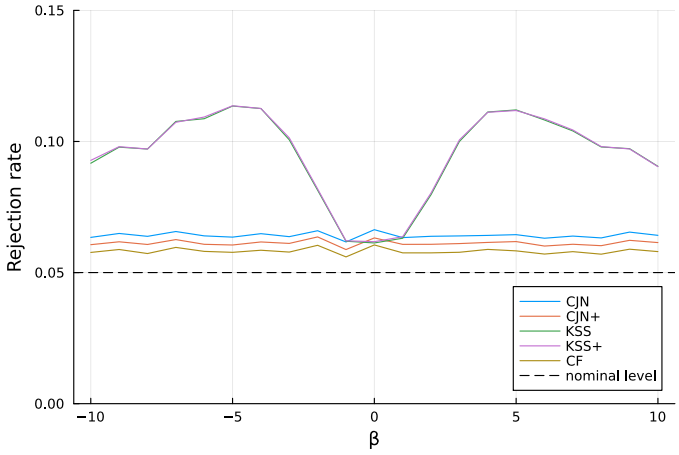


Table 2 also contains (below rejection rates in a small font) percentages of negative definite asymptotic variance estimates, in which cases they are replaced by White estimates. One can see that while the CJN and CJN+ estimates are practically free of such a phenomenon, the KSS/KSS+ estimates are susceptible 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. See also SA5 for simulation experiments with other covariate distributions.

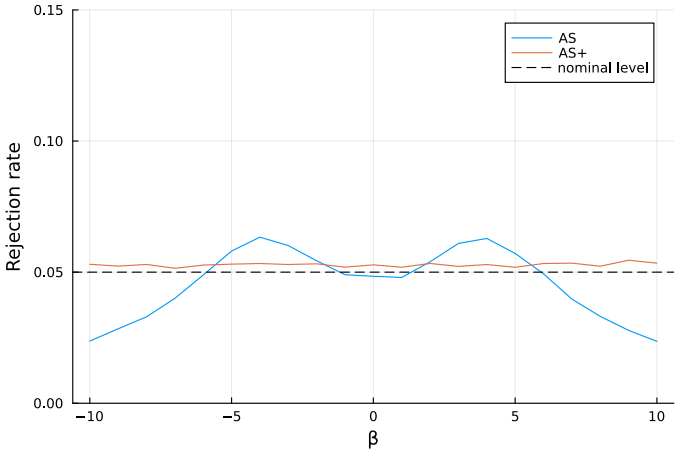
The large distortions from the use of KSS and KSS+ estimates documented in Table 2 occur primarily because $\beta = 5$ is substantially far from zero. Figure 1 shows explicitly the dependence of rejection rates on the value of β in the range from -10 to 10 when $q = 41$. One can see strong dependence 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 free KSS from the scale effect, so that the two curves are almost identical, while CF does the job perfectly.

Next, we analyze the behavior of CF variance product estimates. The unbiasedness is demonstrated in Table 3, which shows actual biases when $\beta = 5$; additionally, AS labels the leave-three-out estimates of Anatolyev and Sølvsten (2023), and CJN labels products of CJN individual variance estimates. The CF biases are tiny and do not grow with covariate dimensionality.

Table 3: Biases of product estimates

q	11	21	31	41
CF	-0.03	-0.03	-0.03	-0.02
AS	-0.03	-0.04	-0.03	0.02
CJN	0.04	0.05	0.05	0.06

Figure 2: Rejection rates of AS and AS+ tests



Finally, we compare the actual sizes of the AS and AS+ tests of the null hypothesis that

all m regression parameters, in both β and γ, equal their true values. The variance estimator $\mathring{V}_{\mathcal{F}}$ in AS+ is averaged over $\ell = 30$ sample splits. Figure 2 shows the dependence of rejection rates on the value of β when $q = 41$. The AS+ rejection rate oscillates close to the nominal level of 5%, and is clearly insensitive to change in the slope coefficient. The original AS test has an excellent coverage in the vicinity of zero, but elsewhere may depends on β in either direction. In addition, AS produces negative variance estimators in up to 30% of cases, while the AS+ test, due to the averaging over sample splits, effectively produces none.

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.

Ritzwoller, D.M. and J.P. Romano (2026). Reproducible aggregation of sample-split statistics. *American Economic Review*, conditionally accepted.