Leave-cluster-out and variance estimation

Stanislav Anatolyev*

CERGE-EI and NES

December 2021

Abstract

We introduce the leave-cluster-out (LCO) machinery for clustered samples, a generalization of leave-one-out methods that prove useful for independent data. We use LCO to construct an estimator of the asymptotic variance of the OLS estimator in a linear regression characterized by possibly numerous regressors and arbitrary within-cluster heteroskedasticity. We show consistency of the LCO variance estimator when regressors may be many, regression errors may be heteroskedastic, clusters may be unbalanced and heterogeneous, and cluster sizes may be moderately large. Simulations reveal amazing robustness of the LCO estimator to regressor numerosity and heteroskedasticity.

Keywords: Linear regression, heteroskedasticity, many regressors, leave-out estimation, variance estimation.

JEL codes: C12, C13, C21.

*CERGE-EI, Politických vězňů 7, 11121 Prague 1, Czech Republic. E-mail: stanislav.anatolyev@cerge-ei.cz. This research was supported by the grant 20-28055S from the Czech Science Foundation.

1 Introduction

Given the clustered nature of many microeconomic data sets, as opposed to random samples and hence independent data, modern econometric methods have adapted the regression theory to such structures. This primarily concerns the construction of so called "cluster-robust" asymptotic (co)variance estimates compatible with the block-diagonal structure of the error variance matrix. It is pretty straightforward to obtain consistent estimates also robust to heteroskedasticity (Liang and Zeger, 1986) that generalize White's (1980) "heteroskedasticity-robust" variance estimation. The practical issues in cluster-robust variance estimation (or standard error construction) are described in the surveys by see Cameron and Miller (2015) and Imbens and Kolesár (2016). There has been an effort to improve finite-sample behavior of these estimates in the spirit of HCK modifications (MacKinnon, 2012; Imbens and Kolesár, 2016) to obtain "almost unbiased" variance estimates, in the sense that they are exactly unbiased under homoskedasticity, though not in general.

The formal asymptotic theory for clustered samples is presented in Hansen and Lee (2019) who laid out the conditions under which the large sample theory, including the central limit theorem and consistency of the Liang and Zeger (1986) variance estimates, takes place. In particular, they consider unbalanced clusters and allow the size of the biggest cluster to asymptotically increase with a moderate rate, as the number of observations and the number of clusters go to infinity. Hansen and Lee (2019) also work out several leading econometric models, including the regression setting. Earlier, White (1984) developed asymptotic theory for the case of balanced homogeneous clusters of fixed size, Hansen (2007) derived asymptotics for the case of balanced clusters with asymptotically increasing cluster size, and Carter, Schnepel, and Steigerwald (2017) allowed unbalanced heterogeneous clusters of different size.

In its own right, the regression theory has been moving toward tolerance to the presence of many regressors (or covariates, or controls). Because the classical tools are, in their majority, not robust to regressor numerosity, various modifications have been proposed to robustify the classical estimation and especially inference in a linear regression. For example,

Calhoun (2011) and Anatolyev (2012) provided modifications of the exact F and classical trio of asymptotic tests, respectively, so that the modified tests are valid within the asymptotic framework where the number of regressors and possibly restrictions is proportional to the sample size, though in a conditionally homoskedastic setup. Under heteroskedasticity, Cattaneo, Jansson, and Newey (2018), Kline, Saggio, and Sølvsten (2020) and Jochmans (2021) provided tools for valid inference with a finite number of restrictions, but allowing for conditional heteroskedasticity. Anatolyev and Sølvsten (2021) consider testing of asymptotically many restrictions in a heteroskedastic environment.

One notable idea exploited in construction of some variance matrix estimates compatible with heteroskedasticity is utilization of leave-out estimation – repeated estimation of the same model when some, usually one, observations are removed from the sample. Although the very idea of leave-out estimation has been around in the statistics and econometric literatures for long (for example, in cross-validation methods, in jackknife bias reduction), only recently has it been discovered to be useful in estimation of conditional variances of individual observations in unbiased way. These simple but attractive estimates are due to Kline, Saggio, and Sølvsten (2020), and they have been found their way into variance estimation under heteroskedasticity and many covariates (Kline, Saggio, and Sølvsten, 2020; Jochmans, 2021), valid inference under heteroskedasticity and many restrictions (Anatolyev and Sølvsten, 2021) and model selection (and potentially model averaging) under heteroskedasticity and many predictors (Anatolyev, 2021).

In this paper, we introduce an analog of leave-out estimation for clustered samples, which we call *leave-cluster-out* (LCO). As the name suggests, it entails estimation with the current cluster's observations removed from computations, which yields the *LCO parameter estimator*. The *LCO residuals* are related to the regular OLS residuals by a relationship that involves blocks of the orthogonal projection matrix, which neatly generalizes an analogous celebrated relationship for leave-one-out estimates. On the basis of LCO residuals, we construct unbiased estimates of cluster-wise variance matrices, which are cluster analogues of unbiased individual variance estimates from Kline, Saggio, and Sølvsten (2020) for in-

dependent data, and eventually develop an improved asymptotic variance estimator. This *LCO variance estimator* is thus robust to clustering, to conditional heteroskedasticity, and to many regressors. Like Hansen and Lee (2019), we allow unbalanced clustered sampling, and allow the number of observations in the maximal cluster to slowly grow with the sample size. The inference about linear combinations of parameters are carried out in the usual way, relying on quantiles of the standard normal distribution. In simulations, the LCO estimator exhibits great performance and amazing robustness to regressor numerosity.

The paper is organized as follows. Section 2 describes the setup and introduces the LCO technology. Some important properties of the LCO estimates are derived. In Section 3, we construct the cluster-robust LCO variance estimator and show its consistency. Section 4 discusses the results of simulation experiments, and Section 5 concludes. Proofs of theoretical results are collected in the Appendix. Some notes on notation not explicitly introduced in the body of the paper: $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimal and maximal eigenvalues of square symmetric matrix A; $\operatorname{dg}\{A_g\}_{g=1}^{G}$ denote a square block-diagonal matrix with square blocks A_g; $\|d\|_{L^p}$ denotes the L^p-norm of vector d.

2 Regression model and LCO estimation

Consider a linear regression model where the n observations belong to G clusters, g^{th} cluster having n_g observations:

$$y_{g,i} = x'_{g,i}\beta + e_{g,i}, \quad E[e_{g,i}|X_g] = 0,$$

where β is $m \times 1$ regression parameter vector and $X_g = \left(x_{g,1}, .., x_{g,n_g}\right)'$ is $n_g \times m$ matrix of cluster g's regressors. Define, for each $g = 1, ..., G$, $n_g \times 1$ vectors $[y]_g = (y_{g,1}, ..., y_{g,n_g})'$ and $[e]_g = (e_{g,1}, ..., e_{g,n_g})$. The same notation for clusterized vectors will be sustained throughout. The collections $\{[y]_g, X_g\}_{g=1}^{G}$ are assumed to be independent across g. The dimension m of the regressors may be large and comparable to sample size with $m \leq n - \max_{1\leq g\leq G} n_g$. One can rewrite the model in the matrix form

$$y = X\beta + e, \quad E[y|X] = 0,$$

and also in terms of cluster-wise data:

$$\begin{bmatrix} [y]_1 \\ [y]_2 \\ \vdots \\ [y]_g \\ \vdots \\ [y]_G \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_g \\ \vdots \\ X_G \end{bmatrix} \beta + \begin{bmatrix} [e]_1 \\ [e]_2 \\ \vdots \\ [e]_g \\ \vdots \\ [e]_G \end{bmatrix},$$

where $E\left[[y]_g | X_g\right] = 0$ for each $g = 1, ..., G$. Denote $\Omega_g = var\left(e_g | X_g\right)$, so that $\Omega(X) = var(e|X) = \mathrm{dg}\left\{\Omega_g\right\}_{g=1}^{G}$. No structure on any Ω_g is imposed, beyond requiring all of them being finite, symmetric and positive semidefinite, nor is any functional form of dependence on X_g. In turn, the dependence structure within any X_g is not restricted either, beyond invertibility of cluster-wise submatrices of the orthogonal projection matrix (see below).

Let $P = X\left(X'X\right)^{-1}X'$, and let P_{gg} be the $n_g \times n_g$ matrix corresponding to the g^{th} diagonal block of P. Define $M_{gg} = I_{n_g} - P_{gg}$. The vector of OLS residuals is $\hat{e} = My$, and $[\hat{e}]_g$ is an $n_g \times 1$ vector corresponding to its g^{th} cluster. Further, let X_{-g} be matrix X with the rows corresponding to the g^{th} cluster removed, and $[y]_{-g}$ be vector y with the elements corresponding to the g^{th} cluster removed.

Denote by $\hat{\beta}_{-g}$ the *leave-cluster-out (LCO) estimator*

$$\hat{\beta}_{-g} = \left(X'_{-g}X_{-g}\right)^{-1} X'_{-g}[y]_{-g},$$

and the *leave-cluster-out (LCO) residuals* for g^{th} cluster

$$[\hat{e}_{-g}]_g = [y]_g - X_g\hat{\beta}_{-g}.$$

These LCO objects are an extension of leave-one-out estimator and leave-one-out residuals when there is no clustering, i.e. when $n_g = 1$ for all $g = 1, ..., G$ and $G = n$.

Assumption 1. For all $g = 1, 2, ..., G$, the matrix M_{gg} is non-singular.

As we will see later from Lemma 2, Assumption 1 makes sure that the LCO estimates and residuals exist for all the clusters. This is a condition on sufficiently non-atomic structure

of the distribution of regressors within each cluster. On the one extreme, if all regressors are continuously distributed, Assumption 1 must be satisfied provided that the biggest cluster is of size no larger than $n - m$, which we have already imposed. On the other extreme, if group dummies for the same groups are used as regressors, all matrices M_{gg} are singular. As long as the configuration in the presence of discrete regressors is not that unreasonable, Assumption 1 is expected to hold.

Lemma 1. Under Assumption 1, the product $[y]_g [\hat{e}_{-g}]_g'$ is conditionally unbiased for Ω_g, $g = 1, 2, ..., G$.

Lemma 1 extends the idea of constructing unbiased estimates of individual variance components of Kline, Saggio, and Sølvsten (2020) to a regression on clustered data. Lemma 2 below is an analog of a celebrated relation between the OLS residuals and leave-one-out residuals, $\hat{e}_{-i} = \hat{e}_i / M_{ii}$, $i = 1, ..., n$.

Lemma 2. Under Assumption 1, we have for all $g = 1, 2, ..., G$,

$$[\hat{e}_{-g}]_g = M_{gg}^{-1} [\hat{e}]_g .$$

The result in Lemma 2, in particular, allows one to compute the LCO residuals for all clusters without running G LCO regressions, and instead compute all of them from one set of OLS estimates. This requires though additional G square matrix inversions of submatrices of the orthogonal projection matrix. In addition, this relation greatly helps with proving theoretical results.

3 Cluster-robust variance estimation

On the basis of LCO residuals, we can construct an unbiased cluster-wise variance estimate for each $\Omega_g (X_g)$, $g = 1, 2, ..., G$, based on LCO residuals:

$$\hat{\Omega}_g^{LCO} = \frac{[y]_g [\hat{e}_{-g}]_g' + [\hat{e}_{-g}]_g [y]_g'}{2}. \tag{1}$$

We have used the machinery of symmetrization in order for the variance matrix estimate to be symmetric, as the product in Lemma 1 need not be symmetric by construction.

Having been equipped with the unbiased cluster-wise variance estimates (1), we construct the OLS variance estimate as

$$\hat{V}^{LCO} = (X'X)^{-1} \left(\sum_{g=1}^{G} X_g' \hat{\Omega}_g^{LCO} X_g \right) (X'X)^{-1} . \tag{2}$$

This is an analog of the variance estimate for independent data that is robust to conditional heteroskedasticity and many regressors (or covariates) proposed in Kline, Saggio, and Sølvsten (2020) and implemented in Jochmans (2021).[1] Because $\hat{\Omega}_g^{LCO}$ is conditionally unbiased, $\hat{V}^{LCO}$ is too. This unbiasedness holds for arbitrary within-cluster heteroskedasticity and for any number of regressors, in contrast to the members of a class of HCK variance estimators (see, e.g., Bell and McCaffrey, 2002; Imbens and Kolesár, 2016) – those that are adjusted to be exactly unbiased in the special circumstance of conditional homoskedasticity.

Assumption 2. As $n \to \infty$, we have $G \to \infty$ and $\max_{1 \leq g \leq G} n_g = o\left(n^{1/2}\right)$.

Assumption 2 restricts asymptotic growth of maximally sized cluster and allows moderately large clusters.

Take c to be an $m \times 1$ constant vector with $\|c\|_{L^2} = O(1)$, and take the parameter of interest to be $c'\beta$, so that the restriction being tested involves an asymptotically finite subset of parameters.

Assumption 3 below lists various technical regularity conditions.

[1] Jochmans (2021) follows the setup of Cattaneo, Jansson, and Newey (2018) who consider inference for a finite number of parameters of interest but allows moderate regressions misspecification due to imperfect approximation of the true regression by a linear function of many covariates. We stick to the standard setup of a correctly specified regression. This allows us, in particular, to be a bit more flexible in formulating the parameter of interest. We conjecture that a moderate amount of misspecification can also be allowed without jeopardizing consistency of the LCO estimator.

Assumption 3.

(i) There exists $C_\Omega > 0$ such that $\lambda_{\max}(\Omega(X)) \leq C_\Omega$, and there exists $C_\kappa > 0$ such that $\max_{1\leq g\leq G} \max_{1\leq i\leq n_g} E\left[[e]_{g,i}^4 | X_g\right] \leq C_\kappa$.

(ii) There are $c_{XX} > 0$ and $C_{Xc} > 0$ such that $\Pr\{\lambda_{\min}(n^{-1}X'X) \geq c_{XX}\} \to 1$ as $n \to \infty$ and $\Pr\left\{\max_{1\leq g\leq G} \|X_g c\|_{L^2}^2 > C_{Xc} \max_{1\leq g\leq G} n_g\right\} \to 0$ as $G \to \infty$.

(iii) The vector of coefficients β is such that $\|X\beta\|_{L^2}^2 = O_P(n)$ and $\|\beta\|_{L^2} = O(1)$.

(iv) There is $c_M > 0$ such that $\Pr\{\min_{1\leq g\leq G} \lambda_{\min}(M_{gg}) \geq c_M\} \to 1$ as $n \to \infty$.

The first condition in Assumption 3(i) restricts eigenvalues of the conditional variance matrix, and automatically restricts eigenvalues of all cluster-wise blocks; its independent-sample analog would be $\max_{1\leq i\leq n} E[e_i^2|x_i] \leq C_\Omega$. The second condition in Assumption 3(i) restricts individual fourth moments. The first condition in Assumption 3(ii) restrrules out near-collinearity of the regressors, while the second condition restricts within-cluster growth of regressors whose coefficients participate in the null restriction. The conditions in Assumption 3(iii) preclude growth of regressors with the sample size and restrict the explanatory power in the regression. They are needed because the LCO variance estimator, or more precisely, the estimates of cluster-wise variances, involve levels of the dependent variable. On the one extreme, the number of regressors may be finite, then all of these may have non-zero bounded coefficients; on the other extreme, the number of regressors may be proportional to the sample size, then the non-zero coefficients may be asymptotically few but fixed, many but $\sqrt{1/n}$-local-to-zero, or take a suitable in-between configuration. The condition in Assumption 3(iv) strengthens Assumption 1. It is an analog of the condition of leverages $\min_{1\leq i\leq n} M_{ii} \geq c_M$ typically imposed in many-regressor literature for independent data.

The main result is Theorem 1 below, which established consistency of the LCO variance estimator $\hat{V}^{LCO}$ in (2). We presume that $\sqrt{n}\left(c'\hat{\beta} - c'\beta\right)$ is asymptotically normal with

mean zero and variance $c'Vc$,where

$$V = (X'X)^{-1}\left(\sum_{g=1}^{G} X_g'\Omega_g X_g\right)(X'X)^{-1}.$$

Theorem 1. Suppose Assumptions 2-3 hold. Then, for the LCO variance estimator $\hat{V}^{LCO}$, we have

$$n\left(c'\hat{V}^{LCO}c - c'Vc\right) = o_p(1),$$

as $n \to \infty$.

The significance level ϕ asymptotic two-sided test of the null $H_0 : c'\beta = \omega$, where ω is a precified value, is then performed in the usual way by comparing the value of the t statistic

$$t_{c'\beta} = \frac{c'\hat{\beta} - \omega}{\sqrt{c'\hat{V}^{LCO}c}}$$

with the right $(1-\phi/2)$-quantile $z_{1-\phi/2}$ of the standard normal distribution, and the confidence level $1-\phi$ asymptotic confidence interval for $c'\beta$ is constructed as $c'\hat{\beta} \mp z_{1-\phi/2}\sqrt{c'\hat{V}^{LCO}c}$.

Remark. As follows from the proof of Theorem 1, the condition $\|c\|_{L^2} = O(1)$ may be slightly relaxed as long as it is still consistent with Assumption 3(i), i.e. participation of a moderately large number of coefficients may be allowed in the combination, at the expense of a smaller growth rate of $\max_g n_g$ stated in Assumption 1. For example, if $\max_g n_g$ is asymptotically fixed, one may allow c to grow up to $\|c\|_{L^2} = o\left(n^{1/2}\right)$.

4 Simulation evidence

We borrow elements of the simulation setup in Carter, Schnepel, and Steigerwald (2017):

$$y_{g,i} = \beta_0 + \sum_{j=1}^{d} \beta_j x_{g,i,j} + u_{g,i},$$

where the regressors are generated by

$$x_{g,i} = z_g + z_{g,i},$$

with z_g and $z_{g,i}$ IID standard normal. Note that $m = d + 1$. The error $u_{g,i}$ follows the error component structure

$$u_{g,i} = \sigma_g \varepsilon_g + \sigma_{g,i} \eta_{g,i},$$

with ε_g and $\eta_{g,i}$ IID standard normal. The conditional standard deviations σ_g and $\sigma_{g,i}$ are generated by

$$\sigma_g = \gamma \left(\sum_{j=1}^{d} z_g^2 \right)^{1/2}, \ \sigma_{g,i} = \gamma \left(\sum_{j=1}^{d} z_{g,i}^2 \right)^{1/2},$$

with $\gamma = 5$, which induces pretty strong heteroskedasticity. There are $n = 2,500$ observations divided into $G = 100$ clusters. In the first, 'balanced', design, all clusters are equally sized, with $n_g = 25$. In the second, 'unbalanced', design, there is a big dispersion in cluster sizes: $n_1 = n_2 = n_3 = 1$, $n_4 = n_5 = 2$, ..., $n_{96} = n_{97} = 48$. $n_{98} = n_{99} = n_{100} = 49$, so that the average number is the same as in the balanced design. Note that $\max_{1 \leq g \leq G} n_g$ is either 25 or 49, which are a bit higher figures than what can be expected from the order $o\left(n^{1/2}\right)$ of Assumption 2.

The true values of parameters are zero, $\beta_0 = \beta_2 = ... = \beta_d = 0$, except $\beta_1 = 1$. Then, $\|\beta\|_{L^2} = 1$. We are looking at the actual sizes for the null $H_0 : \beta_1 = 1$, which is its true value, so that $\|c\|_{L^2} = 1$. Two variance estimators are compared: one is the proposed LCO estimator $\hat{V}^{LCO}$, and the other is the benchmark clustered estimator (labeled LZ, for Liang and Zeger, 1986)

$$\hat{V}^{LZ} = (X'X)^{-1} \left(\sum_{g=1}^{G} X_g' \left[\hat{e}\right]_g \left[\hat{e}\right]_g' X_g \right) (X'X)^{-1}.$$

Figure 1 shows the actual rejection rates, corresponding to the nominal size of 5%, obtained from 20,000 simulations, the upper panel for the balanced design, and the lower panel for the unbalanced design. The graphs are drawn for d running from $d = 5$ to $d = 100$ with a step of 5. One can see that the LZ estimator leads to overrejection that is uniformly higher than that from the LCO estimator, and reaches the value of 'additional' 5% for d as small as $20 \div 40$. The size distortions from the LZ estimator seem to be increasing roughly linearly with d, and reach $10 \div 15\%$ on top of the nominal size when d reaches 100, which is not that large with $n = 2,500$. Interestingly, these distortions are consistently higher for

the unbalanced design than for the balanced design, by approximately 20%. In contrast, the LCO estimator leads to uniformly small size distortions, also of the overrejection type, of no higher than just 1%, some part of which is certainly the simulation noise. The size distortions do not rise with d at all, for either design, even for the unbalanced design, whose $\max_{1\leq g\leq G} n_g$ is pretty big for the $o\left(n^{1/2}\right)$ rule.

Note that in the previous simulation exercise, with $n = 2,500$, the ratio of regressor numerosity even with the maximal d is quite small, only ≈ 0.04, yet the size distortions resulted from the use of the LZ estimator, are large. We have also run experiments with a really big number of non-constant regressors, $d = 500$ for the unbalanced design and $d = 1,000$ for the balanced design, with perceptible dimensionality ratios of 0.2 and 0.4, respectively. The actual rejection rates from 5,000 simulations turned out to be around $15 \div 16\%$ when the LZ variance estimator is used, while with the LCO variance estimation, they are around $5.1 \div 5.3\%$ keeping up with those in Figure 1 in much smaller dimensional situations.

5 Conclusion

In this paper, we have used the LCO method, an extension of leave-one-out machinery adapted for clustered data, for constructing an unbiased asymptotic OLS variance estimation in a linear regression model with many regressors. The proposed LCO technology may turn to be useful in other regression setups with many regressors and heteroskedasticity – for example, in adaptation of the Mallows criterion (Anatolyev, 2021) for model selection or model averaging, or in adaptation of testing for many restrictions (Anatolyev and Sølvsten, 2021), in case the regression errors are clustered. Potentially, it can find its way into other clustered data situations where cross-validation methods are used – for example, bandwidth selection in nonparametric models. This is an interesting agenda for future research.

References

Anatolyev, S. (2012). Inference in regression models with many regressors. *Journal of Econometrics*, 170(2), 368-382.

Anatolyev, S. (2021). Mallows criterion for heteroskedastic linear regressions with many regressors. *Economics Letters*, 203, 109864.

Anatolyev, S. and M. Sølvsten (2021). Testing many restrictions under heteroskedasticity. arXiv preprint arXiv:2003.07320.

Bell, R.M. and D.F. McCaffrey (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, 28(2), 169-181.

Calhoun, G. (2011). Hypothesis testing in linear regression when k/n is large. *Journal of Econometrics*, 165 (2), 163-174.

Cameron, A.C. and D.L. Miller (2015). A practitioner's guide to cluster robust inference. *Journal of Human Resources*, 50, 317-372.

Carter, A.V., K.T. Schnepel, and D.G. Steigerwald (2017). Asymptotic behavior of a t-test robust to cluster heterogeneity. *Review of Economics and Statistics*, 99(4), 698-709.

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.

Hansen, B.E. and S. Lee (2019). Asymptotic theory for clustered samples. *Journal of Econometrics*, 210, 268-290.

Hansen, C.B. (2007). Asymptotic properties of a robust variance matrix estimator for panel data when T is large. *Journal of Econometrics*, 141, 597-620.

Imbens, G.W. and M. Kolesár (2016). Robust standard errors in small samples: Some practical advice. *Review of Economics and Statistics,* 98, 701-712.

Jochmans, K. (2021). Heteroscedasticity-robust inference in linear regression models with many covariates. *Journal of the American Statistical Association* (forthcoming).

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

Liang, K.-Y. and S.L. Zeger. (1986). Longitudinal data analysis for generalized linear models. *Biometrika*, 73(1), 13-22.

MacKinnon, J. G. (2012). Thirty years of heteroskedasticity-robust inference. In: Chen, X. and N. R. Swanson, eds., *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, 437–461, Springer, New York.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4), 817-838.

White, H. (1984). *Asymptotic Theory for Econometricians*. New York: Academic Press.

Woodbury, M. A. (1949). The stability of out-input matrices. *Chicago*, IL 9.

Appendix: Proofs

Proof of Lemma 1. Observe that the conditional expectation is

$$\begin{aligned} E\left(\left[y\right]_g\left[\hat{e}_{-g}\right]_g'|X\right) &= E\left(\left(X_g\beta+\left[e\right]_g\right)\left[X\left(\beta-\hat{\beta}_{-g}\right)+e\right]_g'|X\right) \\ &= \left(X_g\beta\right)E\left(\beta-\hat{\beta}_{-g}|X\right)'X_g' \\ &\quad +E\left(\left[e\right]_g|X\right)E\left(\beta-\hat{\beta}_{-g}|X\right)'X_g' \\ &\quad +\left(X_g\beta\right)E\left(\left[e\right]_g'|X\right) \\ &\quad +E\left(\left[e\right]_g\left[e\right]_g'|X\right) \\ &= \Omega_g. \end{aligned}$$

The first term here is zero as the OLS estimator is conditionally unbiased; in the second term, expectation of the product factorizes because $\hat{\beta}_{-g}$ does not use data from cluster g; the third term is zero because of the regression assumption. □

Proof of Lemma 2. Denote also $P_{\circ g}=X\left(X'X\right)^{-1}X_g'$. Let us look at

$$\left[\hat{e}_{-g}\right]_g=\left[y-X\hat{\beta}_{-g}\right]_g=\left[y-X\left(X_{-g}'X_{-g}\right)^{-1}X_{-g}'y_{-g}\right]_g.$$

Using the Woodbury matrix identity (Woodbury, 1949),

$$\begin{aligned} \left(X_{-g}'X_{-g}\right)^{-1} &= \left(X'X-X_g'X_g\right)^{-1} \\ &= \left(X'X\right)^{-1}+\left(X'X\right)^{-1}X_g'\left(I_{n_g}-X_g\left(X'X\right)^{-1}X_g'\right)^{-1}X_g\left(X'X\right)^{-1} \\ &= \left(X'X\right)^{-1}+\left(X'X\right)^{-1}X_g'M_{gg}^{-1}X_g\left(X'X\right)^{-1}. \end{aligned}$$

Then

$$\left[\hat{e}_{-g}\right]_g=\left[y-X\left(X'X\right)^{-1}\left(I_k+X_g'M_{gg}^{-1}X_g\left(X'X\right)^{-1}\right)\left(X'y-X_g'y_g\right)\right]_g.$$

Note that

$$y-X\left(X'X\right)^{-1}\left(I_k+X_g'M_{gg}^{-1}X_g\left(X'X\right)^{-1}\right)X'y=My-P_{\circ g}M_{gg}^{-1}P_{\circ g}'y$$

and

$$X\left(X'X\right)^{-1}\left(I_k + X_g'M_{gg}^{-1}X_g\left(X'X\right)^{-1}\right)X_g'y_g = P_{\circ g}\left(I + M_{gg}^{-1}P_{gg}\right)y_g = P_{\circ g}M_{gg}^{-1}y_g,$$

Therefore,

$$\begin{aligned}
\left[\hat{e}_{-g}\right]_g &= \left[My - P_{\circ g}M_{gg}^{-1}P_{\circ g}'y + P_{\circ g}M_{gg}^{-1}y_g\right]_g \\
&= \left[My\right]_g - P_{gg}M_{gg}^{-1}\left(P_{\circ g}'y - y_g\right) \\
&= \left[My\right]_g - \left(I_{n_g} - M_{gg}\right)M_{gg}^{-1}\left[-My\right]_g \\
&= \left[My\right]_g - \left(I_{n_g} - M_{gg}^{-1}\right)\left[My\right]_g \\
&= M_{gg}^{-1}\left[\hat{e}\right]_g.
\end{aligned}$$

□

Proof of Theorem 1. We will denote by C a positive generic constant, which may be different in different instances. Define $n_g \times 1$ vector $a_g = X_g\left(X'X\right)^{-1}c$. First, note that

$$\begin{aligned}
\sum_{g=1}^{G}\left\|a_g\right\|_{L^2}^2 &= c'\left(X'X\right)^{-1}\sum_{g=1}^{G}X_g'X_g\left(X'X\right)^{-1}c = c'\left(X'X\right)^{-1}c \\
&\leq n^{-1}\lambda_{\min}\left(n^{-1}X'X\right)^{-1}\left\|c\right\|^2 \leq C\left\|c\right\|^2 n^{-1}.
\end{aligned}$$

Second, note that

$$\begin{aligned}
\left\|a_g\right\|_{L^2}^2 &= c'\left(X'X\right)^{-1}X_g'X_g\left(X'X\right)^{-1}c \\
&\leq n^{-2}\lambda_{\min}\left(n^{-1}X'X\right)^{-2}\left\|X_g c\right\|^2,
\end{aligned}$$

hence, using Assumption 3, with probability approaching one,

$$\max_{1\leq g\leq G}\left\|a_g\right\|_{L^2}^2 \leq n^{-2}\cdot c_{XX}^{-2}\cdot C_{Xc}\max_{1\leq g\leq G}n_g \leq Cn^{-2}\max_{1\leq g\leq G}n_g.$$

Similarly,

$$\max_{1\leq g\leq G}\left\|M_{gg}^{-1}a_g\right\|_{L^2}^2 \leq \min_{1\leq g\leq G}\lambda_{\min}\left(M_{gg}\right)^{-2}\max_{1\leq g\leq G}\left\|a_g\right\|_{L^2}^2 \leq Cn^{-2}\max_{1\leq g\leq G}n_g.$$

Then, also using Assumptions 2 and 3, we have:

(i) $\sum_{g=1}^{G} \|a_g\|_{L^2}^4 \leq \left(\max_{1\leq g\leq G} \|a_g\|_{L^2}^2\right) \sum_{g=1}^{G} \|a_g\|_{L^2}^2 \leq C \|c\|^2 n^{-3} \max_{1\leq g\leq G} n_g$,

(ii) $\left(a_g' (X_g\beta)\right)^2 \leq \max_g \|a_g\|_{L^2}^2 \beta' \sum_{g=1}^{G} X_g' X_g \beta \leq Cn^{-2} \|X\beta\|_{L^2} \max_{1\leq g\leq G} n_g$.

Let $\hat{V}_0$ be non-symmetrized version of $\hat{V}^{LCO}$, then

$$c'\hat{V}_0 c - c'Vc = \sum_{g=1}^{G} a_g' \left([y]_g [\hat{e}]_g' M_{gg}^{-1} - \Omega_g \right) a_g = A_1 + A_2 + A_3,$$

where

$$\begin{aligned} A_1 &= \sum_{g=1}^{G} a_g' \left([e]_g [e]_g' - \Omega_g \right) a_g, \\ A_2 &= \sum_{g=1}^{G} a_g' (X_g\beta) [Me]_g' M_{gg}^{-1} a_g, \\ A_3 &= \sum_{g=1}^{G} a_g' [e]_g \sum_{h=1, h\neq g}^{G} [e]_h' M_{gh}' M_{gg}^{-1} a_g. \end{aligned}$$

We now consider each of these three terms.

Take the first term, A_1. The expectation is zero by the definition of Ω_g. Next,

$$\begin{aligned} var\left[A_1|X\right] &= var\left[\sum_{g=1}^{G} a_g' \left([e]_g [e]_g' - \Omega_g \right) a_g | X\right] = \sum_{g=1}^{G} var\left[a_g' \left([e]_g [e]_g' - \Omega_g \right) a_g | X_g\right] \\ &< \sum_{g=1}^{G} E\left[\left(a_g' \left([e]_g [e]_g' - \Omega_g \right) a_g\right)^2 | X_g\right] \\ &\leq 2 \max_{1\leq g\leq G} \lambda_{\max}\left(E\left[\left([e]_g [e]_g'\right)^2 | X_g\right]\right) \sum_{g=1}^{G} \|a_g\|_{L^2}^4 \\ &\leq 2 \cdot C_\kappa \max_{1\leq g\leq G} n_g \cdot \|c\|^2 n^{-3} \max_{1\leq g\leq G} n_g \\ &\leq O\left(\|c\|^2 n^{-3} \max_{1\leq g\leq G} n_g^2\right), \end{aligned}$$

also using that

$$\begin{aligned} \max_{1\leq g\leq G} \lambda_{\max}\left(E\left[\left([e]_g [e]_g'\right)^2 | X_g\right]\right) &\leq \max_{1\leq g\leq G} \left\| E\left[\left([e]_g [e]_g'\right)^2 | X_g\right] \right\|_F \\ &\leq \max_{1\leq g\leq G} \sqrt{n_g^2 \max_{1\leq i\leq n_g} E\left[[e]_{g,i}^4 | X_g\right]^2} \\ &\leq C_\kappa \max_{1\leq g\leq G} n_g. \end{aligned}$$

For the second term A_2, the expectation is zero by the conditional mean zero assumption. Let $\iota_{g,i}$ denote a $n \times 1$ unit vector with unity only in the position corresponding to observation i in cluster g. Note that

$$\sum_{h=1}^{G} \iota'_{g_1,i_1} M_{\circ h} E\left[e_h e'_h | X\right] M'_{\circ h} \iota_{g_2,i_2} = \iota'_{g_1,i_1} \left(\sum_{h=1}^{G} M_{\circ h} \Omega_h M'_{\circ h} \right) \iota_{g_2,i_2} = \left[(M\Omega M)_{g_1 g_2} \right]_{i_1 i_2}.$$

Then,

$$\begin{aligned}
var\left[A_2|X\right] &= var\left[\sum_{g=1}^{G} a'_g \left(X_g\beta\right) \left[Me\right]'_g M_{gg}^{-1} a_g | X \right] \\
&= var\left[\sum_{g=1}^{G} a'_g \left(X_g\beta\right) \sum_{i=1}^{n_g} \left(M_{gg}^{-1} a_g\right)_i \sum_{h=1}^{G} \left[M_{\circ h} e_h\right]_{g,i} | X \right] \\
&= \sum_{h=1}^{G} var\left[\sum_{g=1}^{G} \sum_{i=1}^{n_g} a'_g \left(X_g\beta\right) \left(M_{gg}^{-1} a_g\right)_i \left(\iota'_{g,i} M_{\circ h} e_h\right) | X \right] \\
&= \sum_{h=1}^{G} \sum_{g_1=1}^{G} \sum_{i_1=1}^{n_g} a'_{g_1} \left(X_{g_1}\beta\right) \left(M_{g_1 g_1}^{-1} a_{g_1}\right)_{i_1} \\
&\qquad \sum_{g_2=1}^{G} \sum_{i_2=1}^{n_g} a'_{g_2} \left(X_{g_2}\beta\right) \left(M_{g_2 g_2}^{-1} a_{g_2}\right)_{i_2} \iota'_{g_1,i_1} M_{\circ h} E\left[e_h e'_h | X\right] M'_{\circ h} \iota_{g_2,i_2} \\
&= \sum_{g_1=1}^{G} \sum_{i_1=1}^{n_g} a'_{g_1} \left(X_{g_1}\beta\right) \left(M_{g_1 g_1}^{-1} a_{g_1}\right)_{i_1} \\
&\qquad \sum_{g_2=1}^{G} \sum_{i_2=1}^{n_g} a'_{g_2} \left(X_{g_2}\beta\right) \left(M_{g_2 g_2}^{-1} a_{g_2}\right)_{i_2} \left[(M\Omega M)_{g_1 g_2} \right]_{i_1 i_2} \\
&= \sum_{g_1=1}^{G} a'_{g_1} \left(X_{g_1}\beta\right) \sum_{g_2=1}^{G} a'_{g_2} \left(X_{g_2}\beta\right) \\
&\qquad \sum_{i_1=1}^{n_g} \sum_{i_2=1}^{n_g} \left[(M\Omega M)_{g_1 g_2} \right]_{i_1 i_2} \left(M_{g_1 g_1}^{-1} a_{g_1}\right)_{i_1} \left(M_{g_2 g_2}^{-1} a_{g_2}\right)_{i_2} \\
&= \sum_{g_1=1}^{G} \sum_{g_2=1}^{G} \left(a'_{g_1} \left(X_{g_1}\beta\right) M_{g_1 g_1}^{-1} a_{g_1}\right)' (M\Omega M)_{g_1 g_2} \left(a'_{g_2} \left(X_{g_2}\beta\right) M_{g_2 g_2}^{-1} a_{g_2}\right) \\
&= A'_{Mg} M\Omega M A_{Mg},
\end{aligned}$$

where A'_{Mg} is $1 \times \sum_{g=1}^{G} n_g = 1 \times n$ row vector with subvectors $a'_g \left(X_g\beta\right) M_{gg}^{-1} a_g$ as g runs

from 1 to G. Note that

$$\begin{aligned}
\|A_{Mg}\|_{L^2}^2 &= \sum_{g=1}^{G} \left(a_g'\left(X_g\beta\right)\right)^2 \left\|M_{gg}^{-1}a_g\right\|_{L^2}^2 \\
&= \max_{1\leq g\leq G} \left\|M_{gg}^{-1}a_g\right\|_{L^2}^2 \sum_{g=1}^{G} \left(a_g'\left(X_g\beta\right)\right)^2 \\
&\leq c_M^{-2} \max_{1\leq g\leq G} \|a_g\|_{L^2}^4 \left(\beta' \sum_{g=1}^{G} X_g'X_g\beta\right) \\
&\leq c_M^{-2} \left(Cn^{-2} \max_{1\leq g\leq G} n_g\right)^2 \|X\beta\|_{L^2}^2 \\
&\leq O(n^{-3} \max_{1\leq g\leq G} n_g^2).
\end{aligned}$$

Hence,

$$\begin{aligned}
var\left[A_2|X\right] &\leq \|A_{Mg}\|_{L^2}^2 \lambda_{\max}\left(\Omega\right) \lambda_{\max}\left(M\right)^2 \\
&\leq O(n^{-3} \max_{1\leq g\leq G} n_g^2) \cdot C_\Omega \cdot 1^2 \\
&= O(n^{-3} \max_{1\leq g\leq G} n_g^2).
\end{aligned}$$

Finally, take the third term, A_3. The expectation is zero because $[e]_g$ and $[e]_h$ do not correlate when $g \neq h$. Now, going for the variance,

$$\begin{aligned}
var\left[A_3|X\right] &= var\left[\sum_{g=1}^{G} a_g'\left[e\right]_g \sum_{h=1,h\neq g}^{G} \left[e\right]_h' M_{gh}'M_{gg}^{-1}a_g|X\right] \\
&= E\left[\sum_{g_1=1}^{G} a_{g_1}'\left[e\right]_{g_1} \sum_{h_1=1,h_1\neq g_1}^{G} \left[e\right]_{h_1}' M_{g_1h_1}'M_{g_1g_1}^{-1}a_{g_1} \right. \\
&\qquad \left. \sum_{g_2=1}^{G} a_{g_2}'\left[e\right]_{g_2} \sum_{h_2=1,h_2\neq g_2}^{G} \left[e\right]_{h_2}' M_{g_2h_2}'M_{g_2g_2}^{-1}a_{g_2}|X\right] \\
&= \sum_{g=1}^{G} \sum_{h=1,h\neq g}^{G} E\left[\left(a_g'\left[e\right]_g\right)^2|X\right] E\left[\left(\left[e\right]_h' M_{gh}'M_{gg}^{-1}a_g\right)^2|X\right] \\
&\qquad + \sum_{g=1}^{G} \sum_{h=1,h\neq g}^{G} E\left[a_g'\left[e\right]_g\left[e\right]_g' M_{hg}'M_{hh}^{-1}a_h|X\right] E\left[a_h'\left[e\right]_h\left[e\right]_h' M_{gh}'M_{gg}^{-1}a_g|X\right].
\end{aligned}$$

The first term equals to and is bounded by

$$\begin{aligned}
& \sum_{g=1}^{G} a_g' \Omega_g a_g a_g' \left(M_{gg}^{-1} \left(M \Omega M \right)_{gg} M_{gg}^{-1} - \Omega_g \right) a_g \\
< & \sum_{g=1}^{G} \lambda_{\max} \left(\Omega_g \right)^2 \left\| a_g \right\|_{L^2}^2 \left\| M_{gg}^{-1} a_g \right\|_{L^2}^2 \lambda_{\max} \left(M_g \right)^2 \\
\leq & \; c_M^{-2} \max_g \lambda_{\max} \left(\Omega_g \right)^2 \max_g \lambda_{\max} \left(M_g \right)^2 \max_g \left\| a_g \right\|_{L^2}^2 \sum_{g=1}^{G} \left\| a_g \right\|_{L^2}^2 \\
\leq & \; C O_P(1)^2 \cdot O_P(1)^2 \cdot C n^{-2} \max_{1 \leq g \leq G} n_g \cdot C \left\| c \right\|^2 n^{-1} \\
= & \; O\left(\left\| c \right\|^2 n^{-3} \max_{1 \leq g \leq G} n_g \right).
\end{aligned}$$

The second term equals to and is bounded by

$$\begin{aligned}
& \left| \sum_{g=1}^{G} \sum_{h=1, h \neq g}^{G} \left(a_g' \Omega_g M_{hg}' M_{hh}^{-1} a_h \right) \left(a_h' \Omega_h M_{gh}' M_{gg}^{-1} a_g \right) \right| \\
< & \sum_{g=1}^{G} \sum_{h=1}^{G} \left\| M_{hg} \Omega_g a_g \right\|_{L^2}^2 \left\| M_{hh}^{-1} a_h \right\|_{L^2}^2 \\
\leq & \sum_{h=1}^{G} \left\| M_{hh}^{-1} a_h \right\|_{L^2}^2 \sum_{g=1}^{G} \left\| M_{hg} \right\|_{L^2}^2 \left\| \Omega_g a_g \right\|_{L^2}^2 \\
\leq & \; c_M^{-2} \max_h \left\| a_h \right\|_{L^2}^2 \sum_{g=1}^{G} \sum_{h=1}^{G} \left\| M_{hg} \Omega_g a_g \right\|_{L^2}^2 \\
\leq & \; c_M^{-2} \max_h \left\| a_h \right\|_{L^2}^2 \max_g \lambda_{\max} \left(M_g \right) \max_g \lambda_{\max} \left(\Omega_g \right)^2 \sum_{g=1}^{G} \left\| a_g \right\|_{L^2}^2 \\
\leq & \; C n^{-2} \max_{1 \leq g \leq G} n_g \cdot O_P(1) \cdot O_P(1)^2 \cdot C \left\| c \right\|^2 n^{-1} \\
= & \; O\left(\left\| c \right\|^2 n^{-3} \max_{1 \leq g \leq G} n_g \right),
\end{aligned}$$

because

$$\begin{aligned}
\sum_{h=1}^{G} \left\| M_{hg} \Omega_g a_g \right\|_{L^2}^2 &= \sum_{h=1}^{G} a_g' \Omega_g M_{hg}' M_{hg} \Omega_g a_g = a_g' \Omega_g M_{gg} \Omega_g a_g \\
&\leq \lambda_{\max} \left(M_g \right) \lambda_{\max} \left(\Omega_g \right)^2 \left\| a_g \right\|_{L^2}^2 .
\end{aligned}$$

So, $var\left[A_3 | X \right] = O\left(\left\| c \right\|^2 n^{-3} \max_{1 \leq g \leq G} n_g \right)$.

To summarize, we have $E\left[c'\hat{V}_0 c - c'Vc|X\right] = 0$ and

$$var\left[c'\hat{V}_0 c - c'Vc|X\right] = O_P\left(\|c\|^2 n^{-3} \max_{1\leq g\leq G} n_g^2\right),$$

and hence

$$n\left(c'\hat{V}_0 c - c'Vc\right) = O_P\left(\|c\| n^{-1/2} \max_{1\leq g\leq G} n_g\right),$$

which is $o_p(1)$ by Assumption 2. Because $c'\hat{V}_0' c = c'\hat{V}_0 c$, the same relation will hold with for the symmetrized version $\hat{V}^{LCO}$. □

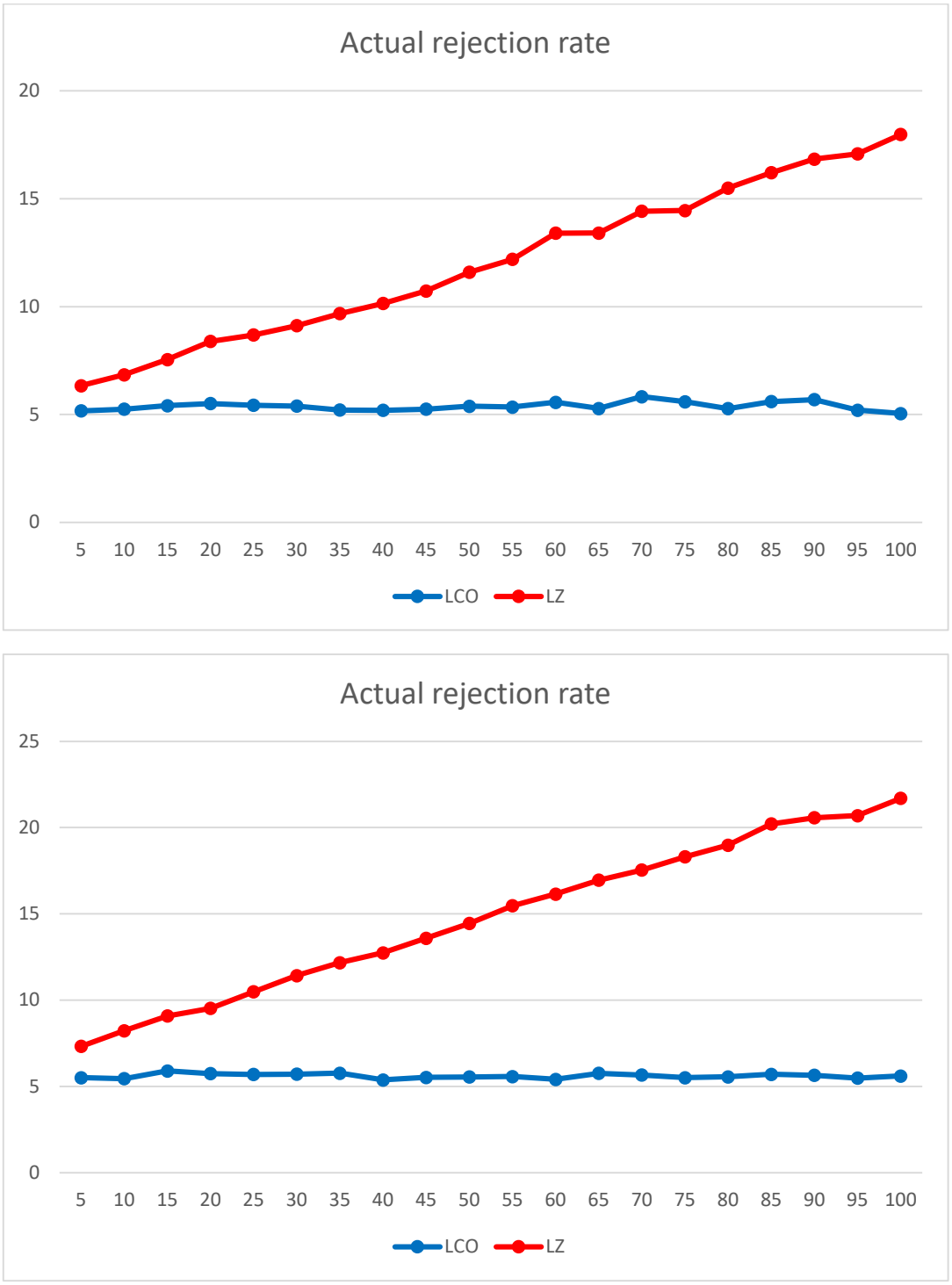

Figure 1. Actual rejection rates corresponding to nominal size of 5%, against number of non-constant regressors. Upper panel: balanced design, lower panel: unbalanced design