Standard Errors in Shift-Share Regressions

Michal Kolesár

August 19, 2024

Contents

1 Summary

The package `ShiftShareSE` implements confidence intervals proposed by Adão et al. [2019] for inference in shift-share least squares and instrumental variables regressions, in which the regressor of interest (or the instrument) has a shift-share structure, as in Bartik [1991]. A shift-share variable has the structure $X_i = \sum_{s=1}^{S} w_{is}\mathcal{X}_s$, where i indexes regions, s indexes sectors, $\mathcal{X}_s$ are sectoral shifters (or shocks), and w_{is} are shares, such as initial share of region i's employment in sector s.

This vignette illustrates the use of the package using a dataset from Autor et al. [2013] (ADH hereafter). The dataset is included in the package as the list `ADH`. The first element of the list, `ADH$reg` is a data-frame with regional variables, the second element, `ADH$sic` is a vector of SIC codes for the sectors, and `ADH$W` is a matrix of shares. See `?ADH` for a description of the dataset.

2 Examples

We now replicate column (1) of Table V in Adão et al. [2019]. First we load the package, define the vector of controls, and define a vector of 3-digit SIC codes:

```
library("ShiftShareSE")
ctrls <- paste("t2 + l_shind_manuf_cbp + l_sh_popedu_c +",
```

```
    "l_sh_popfborn + l_sh_empl_f + l_sh_routine33", " + l_task_outsource + division")
sic <- floor(ADH$sic/10)
```

We cluster the standard errors at the 3-digit SIC code (using the option `sector_cvar`), and, following ADH, weight the data using the weights `ADH$reg$weights`. See `?reg_ss` and `?ivreg_ss` for full description of the options.

The first-stage regression:

```
reg_ss(as.formula(paste("shock ~ ", ctrls)), W = ADH$W,
    X = IV, data = ADH$reg, weights = weights, region_cvar = statefip,
    sector_cvar = sic, method = "all")
#> Estimate: 0.6310409
#>
#> Inference:
#>                Std. Error      p-value  Lower CI  Upper CI
#> Homoscedastic 0.02732516 0.000000e+00 0.5774846 0.6845973
#> EHW           0.08700719 4.083400e-13 0.4605100 0.8015719
#> Reg. cluster  0.09142372 5.113909e-12 0.4518537 0.8102281
#> AKM           0.05296055 0.000000e+00 0.5272402 0.7348417
#> AKM0          0.07671358 1.282891e-03 0.5375710 0.8382827
```

Note that for `"AKM0"`, `"Std. Error"` corresponds to the normalized standard error, i.e. the length of the confidence interval divided by $2z_{1-\alpha/2}$.

The reduced-form and IV regressions:

```
reg_ss(as.formula(paste("d_sh_empl ~", ctrls)), W = ADH$W,
    X = IV, data = ADH$reg, region_cvar = statefip, weights = weights,
    sector_cvar = sic, method = "all")
#> Estimate: -0.4885687
#>
#> Inference:
#>                Std. Error      p-value   Lower CI   Upper CI
#> Homoscedastic 0.06332778 1.221245e-14 -0.6126889 -0.3644485
#> EHW           0.11244360 1.392685e-05 -0.7089541 -0.2681833
#> Reg. cluster  0.07578147 1.140306e-10 -0.6370977 -0.3400398
#> AKM           0.16419445 2.924641e-03 -0.8103839 -0.1667535
#> AKM0          0.25437489 4.218033e-04 -1.2368853 -0.2397541
ivreg_ss(as.formula(paste("d_sh_empl ~", ctrls, "| shock")),
    W = ADH$W, X = IV, data = ADH$reg, region_cvar = statefip,
    weights = weights, sector_cvar = sic, method = "all")
#> Estimate: -0.7742267
#>
#> Inference:
#>                Std. Error      p-value   Lower CI   Upper CI
#> Homoscedastic  0.1069532 4.523049e-13 -0.9838511 -0.5646022
#> EHW            0.1647892 2.623532e-06 -1.0972075 -0.4512459
#> Reg. cluster   0.1758096 1.063809e-05 -1.1188071 -0.4296462
#> AKM            0.2403730 1.277718e-03 -1.2453492 -0.3031041
```

```
#> AKM0          0.3318966 4.218033e-04 -1.6903240 -0.3893132
```

3 Collinear share matrix

Let W denote the share matrix with the (i, s) element given by w_{is} and sth column w_s. Suppose that columns of W are collinear, so that it has rank $S_0 < S$. Without loss of generality, suppose that the first S_0 columns of the matrix are full rank, so that the collinearity is caused by the last $S - S_0$ sectors. In this case, it is not possible to recover, $\tilde{\mathcal{X}}_s$, the sectoral shifters with the controls partialled without further assumptions. The `reg_ss` and `ivreg_ss` functions will return a warning message `"Share matrix is collinear"`. To compute the standard errors, the commands implement a default solution to this issue based on aggregating the shocks to the collinear sectors, which we describe in Section 3.1 below. However, there are other ways of dealing with collinearity in the share matrix, as we describe in 3.2 below. Depending on the the setting, researchers may wish to instead use one of these alternatives.

3.1 Default way of dealing with collinear sectors

We use a QR factorization of W with column pivoting (see Chapter 5.4.2 in Golub and Van Loan [2013]) to drop the collinear columns in W. That is, we decompose $W = QRP'$, where Q is an $N \times S$ orthogonal matrix, the matrix R takes the form $R = \left(\begin{smallmatrix} R_1 & R_2 \\ 0 & 0 \end{smallmatrix}\right)$, where R_1 is an $S_0 \times S_0$ upper triangular matrix, R_2 has dimensions $S_0 \times (S - S_0)$, and P is a permutation matrix such that the diagonal elements of R are decreasing. We then drop $S_0 - S$ columns of W that correspond to the last $S - S_0$ columns of QR, as indicated by the permutation matrix, obtaining a new share matrix W_{new}. Most software implementations of ordinary least squares, including LAPACK used by `R`, use this algorithm to drop collinear columns of the regressor matrix.

This solution keeps the regional shocks X_i the same, so that the point estimates do not change, while implicitly redefining the sectoral shocks $\mathcal{X}_s$. In particular, by definition of collinearity, each column w_s of W that we drop can be written as a linear combination of the new share matrix W_{new}. We can determine the coefficients γ_s in this linear combination by regressing w_s onto W_{new}. Observe that since

$$X = W\mathcal{X} = W_{new}\mathcal{X}_0 + \sum_{s=S_0+1}^{S} (W_0\gamma_s)X_s = W_{new}\left[\mathcal{X}_0 + \sum_{s=S_0+1}^{S} \gamma_s X_s\right] = W_{new}\mathcal{X}_{new},$$

dropping the collinear columns of W doesn't change the regional shocks X_i if we implicitly define a new sectoral shock vector $\mathcal{X}_{new}$ as

$$\mathcal{X}_{new} = \mathcal{X}_0 + \sum_{s=S_0+1}^{S} \gamma_s x_s.$$

Here $\mathcal{X}_0$ corresponds to the first S_0 entries of the S-vector of shocks $\mathcal{X}$.

Note that re-ordering the columns of W will generally result in different columns being dropped, so that the standard errors will generally depend on the order of the sectors.

3.2 Other solutions

There are alternative ways of dealing with collinearity, including:

1. Drop the collinear sectors, defining $X_i = \sum_{s=1}^{S_0} w_{is}\mathcal{X}_s$, and defining the share matrix W to only have S_0 columns, as in the default solution. This effectively puts shocks to the collinear sectors into the residual (which is analogous to letting say the shock to non-manufacturing sectors be part of the residual), and changes the point estimate as well as the estimand.
2. Aggregate the sectors. For instance, if originally the sectors correspond to 4-digit SIC industries, we may wish to work with 3-digit industries. This solution will change the point estimate, as well as the estimand. Alternatively, we may only aggregate the collinear sectors.
3. If the only controls are those with shift-share structure, and we have data on $\mathcal{Z}_s$, we can estimate $\tilde{\mathcal{X}}_s$ by running a sector-level regression of $\mathcal{X}_s$ onto $\mathcal{Z}_s$, and taking the residual. This solution doesn't affect the point estimate or the definition of the estimand.

4 Extensions to multiple shifters and multiple endogenous variables

We now discuss how the methods in Adão et al. [2019] extend to the case where there are multiple shifters, or, in the case of an IV regression, multiple endogenous variables. Currently, these extensions are not implemented in the package.

4.1 OLS

Suppose that we're interested in the effect of a k-vector of shift-share regressors, $X_i = \sum_s w_{is}\mathcal{X}_s$, where $\mathcal{X}_s$ is a vector of length k. For inference on the coefficient on the jth element of X_i, we proceed as if this was the only shift-share regressor, treating the remaining shifters as part of the controls.

4.2 IV with a single endogenous regressor and multiple shift-share instruments

Now suppose that the k-vector X_i defined in section 4.1 is a k-vector of instruments. Let X denote the $N \times k$ matrix with rows given by X_i'. Consider the setup in Section IV.C of Adão et al. [2019], with the first-stage coefficients β_{is} in eq. (31) now a k-vector, and α being the scalar treatment effect of Y_2 on Y_1 as in eq. (30). Letting $\ddot{X} = X - Z(Z'Z)^{-1}Z'X$ denote the $N \times k$ matrix of instruments with the covariates partialled out, the two-stage least squares estimator is given by

$$\hat{\alpha} = \frac{Y_2'\ddot{X}(\ddot{X}'\ddot{X})^{-1}\ddot{X}'Y_1}{Y_2'\ddot{X}(\ddot{X}'\ddot{X})^{-1}\ddot{X}'Y_2} = \frac{\hat{\beta}'\ddot{X}'Y_1}{\hat{\beta}'\ddot{X}'\ddot{X}\hat{\beta}},$$

where $\hat{\beta} = (\ddot{X}'\ddot{X})^{-1}\ddot{X}'Y_2$ is a k-vector of first-stage coefficients.

Thus,

$$\hat{\alpha} - \alpha = \frac{Y_2'\ddot{X}(\ddot{X}'\ddot{X})^{-1}\ddot{X}'Y_1(0)}{Y_2'\ddot{X}(\ddot{X}'\ddot{X})^{-1}\ddot{X}'Y_2}.$$

Now, letting $Y_1(0) = Z'\delta + \epsilon$, we have, as in the proof of Proposition 4 in the paper,

$$r_N^{1/2}\ddot{X}'Y_1(0) = r_N^{1/2}X'(I - Z(Z'Z)^{-1}Z)\epsilon = r_N^{1/2}\tilde{\mathcal{X}}'W'\epsilon + o_p(1).$$

Thus, using arguments in Proposition 4 in the paper, we obtain the infeasible standard error formula

$$\text{se}(\hat{\alpha}) = \frac{\sqrt{\sum_s (\hat{\beta}'\tilde{\mathcal{X}}_s)^2 R_s^2}}{\hat{\beta}'\ddot{X}'\ddot{X}\hat{\beta}}, \qquad R_s = \sum_i w_{is}\epsilon_i,$$

where $\tilde{\mathcal{X}}_s$ is a (vector) residual from the population regression of the vector $\mathcal{X}_s$ onto the controls.

This suggests the feasible standard error formula

$$\widehat{\text{se}}(\hat{\alpha}) = \frac{\sqrt{\sum_s (\hat{\beta}'\widehat{\mathcal{X}}_s)^2 \hat{R}_s^2}}{\hat{\beta}'\ddot{X}'\ddot{X}\hat{\beta}}, \qquad \hat{R}_s = \sum_i w_{is}\hat{\epsilon}_i,$$

where $\widehat{\mathcal{X}} = (W'W)^{-1}W'\ddot{X}$ are the regression coefficients from the regression of $\ddot{X}$ onto W (as in Remark 6, except now a $\widehat{\mathcal{X}}$ is an $S \times k$ matrix), and $\hat{\epsilon}_i$ are estimates of structural residual. For AKM, $\hat{\epsilon} = Y_1 - Y_2\hat{\alpha} - Z(Z'Z)^{-1}Z'(Y_1 - Y_2\hat{\alpha})$.

For AKM0, the construction is more complicated. Let $\hat{\gamma} = (\ddot{X}'\ddot{X})^{-1}\ddot{X}'Y_1$ denote the reduced-form coefficient. Let $\hat{R}_{s,\alpha_0} = \sum_i w_{is}\hat{\epsilon}_{\alpha_0}$, where $\hat{\epsilon}_{\alpha_0} = (I - Z(Z'Z)^{-1}Z')(Y_1 - Y_2\alpha_0)$. Then

$$Q(\alpha_0) = (\hat{\gamma} - \hat{\beta}\alpha_0)'(\ddot{X}'\ddot{X})\left(\sum_s \widehat{\mathcal{X}}_s\widehat{\mathcal{X}}_s'\hat{R}_{s,\alpha_0}^2\right)^{-1}(\ddot{X}'\ddot{X})(\hat{\gamma} - \hat{\beta}\alpha_0)$$

will be distributed χ_k^2 in large samples, because $(\ddot{X}'\ddot{X})^{-1}\sum_s \widehat{\mathcal{X}}_s\widehat{\mathcal{X}}_s'\hat{R}_{s,\alpha_0}^2(\ddot{X}'\ddot{X})^{-1}$ consistently estimates the asymptotic variance of $\hat{\gamma} - \hat{\beta}\alpha_0$ under the null. Therefore, we reject the null $H_0\colon \alpha = \alpha_0$ if $Q(\alpha_0) > \chi_{k,1-\alpha}^2$, where $\chi_{k,1-\alpha}^2$ is the $1-\alpha$ quantile of a χ_k^2. A confidence set is collected by all nulls that are not rejected,

$$\text{AKM0 confidence set} = \{\alpha \in \mathbb{R}\colon Q(\alpha) \leq \chi_{k,1-\alpha}^2\},$$

Note that (i) unlike the case with a single instrument (Remark 6, step (iv)), there is not a closed form solution to the confidence set anymore: one needs to do a grid search over the real line, collecting all values of α for which the test doesn't reject, and (ii) the confidence set will be valid even if the instruments are weak; however, if the instruments are strong, the AKM0 test is less powerful than the AKM test, and consequently the AKM0 confidence set will tend to be bigger than the AKM confidence interval.

Not that properties (i) and (ii) are inherited from the properties of the heteroskedasticity-robust version of the Anderson-Rubin test when there is more than one instrument (see, for example, Section 5.1 in Andrews et al. [2019] for a discussion). The AKM0 method adapts this test to the current setting with shift-share instruments, inheriting these properties.

If we do not require validity under weak instruments, we can also use a different version of AKM0, namely computing the confidence set as

$$\text{Alternative AKM0 confidence set} = \left\{\alpha \in \mathbb{R}\colon \frac{(\hat{\alpha} - \alpha)^2}{\frac{\sum_s (\hat{\beta}'\widehat{\mathcal{X}}_s)^2\hat{R}_{s,\alpha}^2}{(\hat{\beta}'\ddot{X}'\ddot{X}\hat{\beta})^2}} \leq z_{1-\alpha/2}^2\right\}.$$

This form of the confidence can be thought of as the analog to the Lagrange multiplier confidence set in likelihood models, rather than the analog of the Anderson-Rubin test. In the case with a single instrument, these concepts coincide, but they are different in general. In this case, the inequality defining the set is just a quadratic inequality in α, and we can solve it explicitly as in Remark 6 in the paper to obtain a closed-form solution. If the instruments are strong, it will take the form of an interval.

4.3 IV with multiple endogenous variables

Consider a general setup with eqs. (30) and (31) in the paper replaced by

$$Y_{1i}(y_2) = Y_{1i}(0) + y_2'\alpha \qquad Y_{2i}(x_1, \ldots, x_S) = Y_{2i}(0) + \sum_s w_{is} B_{is}' x_s$$

with $\mathcal{X}$ and Y_2 now both vectors, and B_{is} has dimensions $\dim(\mathcal{X}) \times \dim(Y_2)$. If $\mathcal{X} = Y_2$, the setup reduces to that in section 4.1. If Y_2 is scalar, the setup reduces to that in section 4.2. The two-stage least squares estimator of α is given by

$$\hat{\alpha} = (Y_2'\ddot{X}(\ddot{X}'\ddot{X})^{-1}\ddot{X}'Y_2)^{-1}Y_2'\ddot{X}(\ddot{X}'\ddot{X})^{-1}\ddot{X}'Y_1.$$

With scalar X_i and Y_{2i}, this expression reduces to eq. (33) in the paper. Now,

$$\hat{\alpha} - \alpha = (Y_2'\ddot{X}(\ddot{X}'\ddot{X})^{-1}\ddot{X}'Y_2)^{-1}Y_2'\ddot{X}(\ddot{X}'\ddot{X})^{-1} \cdot \ddot{X}'(Y_1 - Y_2\alpha)$$

Suppose that

$$E[\mathcal{X}_s \mid \mathcal{F}_0] = \Gamma'\mathcal{Z}_s,$$

where $\mathcal{F}_0 = (Y_1(0), Y_2(0), W, \mathcal{Z}, U, B)$. Let δ be the coefficient on Z in the regression of $Y_{1i} - Y_{2i}'\alpha$ onto Z_i, and let $\epsilon_i = Y_{1i} - Y_{2i}'\alpha - Z_i'\delta = Y_{1i}(0) - Z_i'\delta$. Then, as in proof of Proposition 4 in the paper,

$$\begin{aligned} r_N^{1/2}\ddot{X}'(Y_1 - Y_2\alpha) &= r_N^{1/2}\ddot{X}'(Z\delta + \epsilon) = r_N^{1/2}\tilde{\mathcal{X}}'W'\epsilon + r_N^{1/2}\Gamma'U'\epsilon - r_N^{1/2}\epsilon'Z(\hat{\Gamma} - \Gamma), \\ &= r_N^{1/2}\tilde{\mathcal{X}}'W'\epsilon + o_p(1), \end{aligned}$$

where the second line follows by arguments in that proof. Now, since $\mathcal{X}_s$ is independent across s conditional on $\mathcal{F}_0$, it follows that conditional on $\mathcal{F}_0$,

$$r_N^{1/2}\tilde{\mathcal{X}}'W'\epsilon = r_N^{1/2}\sum_s \tilde{\mathcal{X}}_s R_s = \mathcal{N}(0, \sum_s R_s^2 E[\tilde{\mathcal{X}}_s\tilde{\mathcal{X}}_s' \mid \mathcal{F}_0]) + o_p(1),$$

where $R_s = \sum_i w_{is}\epsilon_i$. This leads to variance formula

$$\begin{aligned} \widehat{\text{var}}(\hat{\alpha}) &= (Y_2'\ddot{X}(\ddot{X}'\ddot{X})^{-1}\ddot{X}'Y_2)^{-1}Y_2'\ddot{X}(\ddot{X}'\ddot{X})^{-1} \cdot \sum_s \hat{R}_s^2\widehat{\mathcal{X}}_s\widehat{\mathcal{X}}_s' \cdot (\ddot{X}'\ddot{X})^{-1}\ddot{X}'Y_2(Y_2'\ddot{X}(\ddot{X}'\ddot{X})^{-1}\ddot{X}'Y_2)^{-1} \\ &= (\hat{B}'\ddot{X}'\ddot{X}\hat{B})^{-1} \cdot \sum_s \hat{R}_s^2\hat{B}'\widehat{\mathcal{X}}_s\widehat{\mathcal{X}}_s'\hat{B} \cdot (\hat{B}'\ddot{X}'\ddot{X}\hat{B})^{-1}, \end{aligned}$$

where $\hat{R}_s = \sum_i w_{is}\hat{\epsilon}_i$, $\widehat{\mathcal{X}} = (W'W)^{-1}W'\ddot{X}$ as in eq. (36) in the paper, with rows $\mathcal{X}_s'$, and $\hat{B} = (\ddot{X}'\ddot{X})^{-1}\ddot{X}'Y_2$ is a matrix of the first-stage coefficients. Here $\hat{\epsilon}_i$ is an estimate of the structural residual, such as

$$\hat{\epsilon} = (I - Z(Z'Z)^{-1}Z')(Y_1 - Y_{21}'\hat{\alpha}) \tag{1}$$

For standard errors, take square root of the appropriate diagonal element.

The AKM0 version is a little tricky here if $\dim(\alpha) > 1$ and we're only interested in inference on one element of α, say the first: this is analogous to issues with using the Anderson-Rubin test in a setting with multiple endogenous variables.

If we do not require validity under weak instruments, then the analog of the 'alternative AKM0' procedure from the preceding subsection uses the estimate $(\alpha_{10}, \hat{\alpha}_{-1}(\alpha_{10}))$ in place of $\hat{\alpha}$ in (1), where α_{10} is the null hypothesized value, and

$$\hat{\alpha}_{-1}(\alpha_{10}) = (Y_{2,-1}'\ddot{X}(\ddot{X}'\ddot{X})^{-1}\ddot{X}'Y_{2,-1})^{-1}Y_{2,-1}'\ddot{X}(\ddot{X}'\ddot{X})^{-1}\ddot{X}'(Y_1 - Y_{2,1}\alpha_{10}).$$

is the estimate of the remaining elements of α with the null $H_0\colon \alpha_1 = \alpha_{10}$ imposed.

References

Rodrigo Adão, Michal Kolesár, and Eduardo Morales. Inference in shift-share designs: Theory and inference. *Quarterly Journal of Economics*, 134(4):1949–2010, November 2019. doi: 10.1093/qje/qjz025.

Isaiah Andrews, James H. Stock, and Liyang Sun. Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, 11(1):727–753, August 2019. doi: 10.1146/annurev-economics-080218-025643.

David H. Autor, David Dorn, and Gordon H. Hanson. The China syndrome: Local labor market effects of import competition in the United States. *American Economic Review*, 103(6):2121–2168, October 2013. doi: 10.1257/aer.103.6.2121.

Timothy J. Bartik. *Who Benefits from State and Local Economic Development Policies?* W.E. Upjohn Institute for Employment Research, Kalamazoo, MI, 1991.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. The Johns Hopkins University Press, Baltimore, MD, fourth edition, 2013.