Inference on the Returns to Schooling in the Presence of Peer Effects

Marcelo J. Moreira *   Geert Ridder †

December, 2018

*FGV EPGE, Rio de Janeiro, Brasil.
†Department of Economics and USC Dornsife INET, University of Southern California, USA

Overview

- We consider inference in regression models with an endogenous covariate and weak instruments.
- The random errors of the structural equation and of the first stage can be heteroskedastic.
- The random errors of the structural equation and those of the first stage can be correlated between observations.
- The random errors of the structural equation and those can be correlated for each observation (endogenous covariate), but also across observations.
- In summary: the errors of the first-stage and structural equation are heteroskedastic and autocorrelated (HAC).

- Inference for the regression coefficient of an endogenous variable using weak instruments is fundamentally different in the HAC case.
- Current tests for this case can result in low power c.q. wide confidence intervals, because tests available in statistical software as STATA ignore important information in the HAC case.
- We propose a new test that has high power in cases where current tests fail.
- In a simple model for earnings with endogenous education and peer effects we show that errors are HAC with a rather complicated variance matrix.

Heteroskedastic Errors in the Return to Education

- Model of earnings function with endogenous education as in Card (2001).
- Individual maximizes lifetime utility $\log c(t)$ with $c(t)$ consumption subject to a lifetime budget constraint.
- Net income while at school is 0, earnings grows at constant rate g, and $\phi(t)$ the disutility of attending school.
- FOC for education beyond compulsory level is

  $$\frac{f'(S)}{f(S)} = R - g - \rho e^{-\rho S}\phi(S)$$

  with R the interest rate at which the individual can borrow/lend and ρ the subjective discount factor.
- The LHS is the relative return to education and the RHS is the marginal cost of education $d(S)$.

- The marginal return and the marginal cost depend of (un)observed characteristics of the individual

$$\frac{f'(S_i)}{f(S_i)} = b_i + \beta' X_i \qquad d(S_i) = r_i + \rho X_i + k_2 S_i$$

  with b_i, r_i the unobserved heterogeneity in the marginal return and the marginal cost of education.

- Integration of this expression and assuming that the log earnings y_i of i with work experience $E_i = t - S_i$ is $\log y_i = \log f(S_i) + \lambda_i E_i$, we find for the earnings function

$$\log y_i = a_i + \tilde{\alpha}' X_i + b_i S_i + \beta' X_i S_i + \lambda_i E_i$$

  with $a_i + \tilde{\alpha}' X_i$ the integration constant.

- The optimal level of education is

$$S_i = \frac{b_i - r_i}{k_2} + \frac{(\beta - \rho)'}{k_2} X_i$$

- This demand function for education can be identified if we have an exogenous shock to the marginal cost of education

$$r_i = c_i + \gamma' Z_i$$

  Card (2001), Table 2 lists instruments used in 11 studies.

- Substitution results in the first-stage model

$$S_i = \pi' Z_i + \delta' X_i + \eta_i \qquad \pi = -\gamma / k_2 \quad \delta = (\beta - \rho)/k_2 \qquad \eta_i = (b_i - b - c_i)/k_2$$

- S_i depends on b_i and is therefore endogenous in the earnings function. Also a_i, b_i may be correlated.

- The reduced form of this model is

$$\log y_i = a + b\pi' Z_i + \alpha' X_i + (\beta \otimes \pi)'(X_i \otimes Z_i) + (\beta \otimes \delta)'(X_i \otimes X_i) + \lambda E_i + \zeta_i \tag{1}$$

$$S_i = Z_i'\pi + \delta' X_i + \eta_i \tag{2}$$

with $b = E(b_i)$, $\lambda = E(\lambda_i)$

$$a = E(a_i) + E((b_i - b)\eta_i) \qquad \alpha = \tilde{\alpha} + b\delta$$

- The error that reflects the unobserved heterogeneity in the model is

$$\zeta_i = a_i - E(a_i) + b\eta_i + (b_i - b)\eta_i - E((b_i - b)\eta_i) + \eta_i \beta' X_i + (b_i - b)\pi' Z_i + (b_i - b)\delta' X_i +$$

$$+(\lambda_i - \lambda)E_i$$

- This model can be used to estimate the average return to education b. Mean independence of η_i, $a_i - E(a_i))$ and $b_i - E(b_i))$ of Z_i, X_i, E_i is not sufficient and we need full independence because $E[(b_i - b)\eta_i | Z_i]$ may change with Z_i (Card(2001)).

- ζ_i is heteroskedastic and the covariance of η_i and ζ_i depends on Z_i, E_i, X_i. This is typical for a structural model with unobserved heterogeneity. There is no correlation across observations.

Peer Effects and HAC Errors in the Return of Education

- Following Graham (2008) we assume that an individual's return to education depends on the peer group average

$$b_i - b = \nu_p + (\tau - 1)(\bar{b}_p - b) + \xi_i$$

with ν_p peer-group characteristics.

- The first-stage error is

$$\eta_i = \frac{\nu_p}{k_2} + \frac{\tau - 1}{k_2}(\bar{b}_p - b) + \frac{\xi_i - c_i}{k_2}$$

- The error of the earnings function is

$$\zeta_i = (a_i - E(a_i)) + b\eta_i + \eta_i \beta' X_i + (\lambda_i - \lambda)E_i + \eta_i \nu_p + (\tau - 1)(\bar{b}_p - b)\eta_i + \xi_i \eta_i +$$

$$+\nu_p(\pi' Z_i + \delta' X_i) + (\tau - 1)(\bar{b}_p - b)(\pi' Z_i + \delta' X_i) + \xi_i(\pi' Z_i + \delta' X_i)$$

- Note that the peer effect in the return to education induces a peer effect in the choice of the level of education (positive dependence on $\bar{b}_p$ if $\tau > 1, k_2 > 0$).

- The error of the earnings equation is still heteroskedastic. In addition the errors of the reduced form are correlated within, but not across peer groups. Most importantly for inference, the ζ_i and η_j are correlated within peer groups (and the correlation depends X_i, Z_i).

- Conclusion: introducing peer effects in Card's prototypical model of schooling level choice and earnings produces a triangular linear system with HAC errors. We consider inference in such a system.

Inference with HAC errors and weak instruments

- Triangular system with single endogenous variable and k possibly weak instruments and n observations

$$
\begin{aligned}
y_1 &= y_2\beta + u \\
y_2 &= Z\pi + v_2
\end{aligned}
$$

  Goal is to do inference (test, confidence interval) on β.

- The errors in u, v_2 can be correlated, both within (endogeneity) and between observations, and can be heteroskedastic.

- Correlation between observations occurs in time-series data (HAC, see e.g. Newey and West (1987)), in spatial data (spatial HAC, see e.g. Conley (1999)) and in data with a group structure (clustering, see e.g. Cameron and Miller(2015)).

- We consider the implication of correlation of the errors between observations on weak-instrument robust inference.

- With $Y = [y_1 \ y_2]$ and $a = (\beta \ 1)'$ the reduced form is

$$Y = Z\pi a' + V \tag{3}$$

- We pre-multiply the reduced form by $(Z'Z)^{-1/2}Z'$ and define $R = (Z'Z)^{-1/2}Z'Y$ so that

$$R = \mu a' + \widetilde{V} \tag{4}$$

  with $\mu = (Z'Z)^{1/2}\,\pi$ and $\widetilde{V} = (Z'Z)^{-1/2}\,Z'V$.

- The variance matrix of $vec(\widetilde{V})$ is the $2k \times 2k$ matrix Σ

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

  is unrestricted.

- Σ is estimated as in Newey and West (1987) (HAC), Conley (1999) (spatial HAC) or with a White (1980) type estimator.
- The cluster-robust estimator (White (1980)) is

$$\widehat{\Sigma}_{pq} = (Z'Z)^{-1/2} \sum_{g=1}^{G} Z_g' \hat{v}_{pg} \hat{v}_{qg}' Z_g (Z'Z)^{-1/2}$$

- We ignore the details of estimation of Σ but focus on the impact of features of this variance matrix on inference.

Current practice

- Instead of R we consider equivalent statistics S, T.
- Current practice for testing $H_0 : \beta = 0$ is to use one of the following tests, implemented in STATA (Finlay and Magnusson(2009)).
- LM test

$$LM_1 = \frac{S'T}{(T'T)^{1/2}} \ , \tag{5}$$

- Anderson-Rubin (AR) test

$$AR = S'S$$

- CQLR test

$$CQLR = \frac{AR - T'T + \sqrt{(AR - T'T)^2 + 4LM \cdot T'T}}{2}$$

- Power curves show the performance of these tests with simulated data.
- The LM_1 and CQLR tests are behaving poorly with power equal to size.
- The AR test does better, but it will behave worse if the number of instruments increases (AR is optimal choice if $k = 1$).
- The poor performance is only in the case of weak instruments. If the instruments are strong the LM_1 test dominates the other tests.
- The DGP for which the LM_1 and CQLR tests do poorly only occur if the errors are HAC.
- To be specific the DGP have

$$\mu' \Sigma_{11}^{-1} \Sigma_{12} \Sigma_{11}^{-1} \mu = 0$$

- A necessary and sufficient condition for this is that the eigenvalues of the covariance matrix of the reduced-form and first-stage errors are not all negative or positive.
- With time-series regressions of consumption on asset returns we found that for 9 out of 11 countries the eigenvalues of $\Sigma_{12} + \Sigma_{12}'$ or of opposite signs.

Using additional information

- A further diagnosis shows that the relevant information in the data is in the statistics $S'S, S'T, T'T$ if the errors are homoskedastic and uncorrelated between observations.
- The LM_1, AR, and CQLR tests all depend on these statistics.
- This follows from the fact that the model does not change if the data are transformed in certain way so that the test should not change either.
- The model with HAC errors at first sight changes with the data transformation. However if we consider Σ as a parameter but also as part of the data then there is again a transformation that leaves the model unchanged.
- The statistics $S'S, S'T, T'T$ no longer contain all relevant information.

- To take account of the additional information Moreira and Ridder (2018) propose a new test.

$$CIL = \int_{-\infty}^{\infty} e^{\frac{vec(R_0)'\Sigma_0^{-1}(a_\Delta \otimes I_k)\left((a'_\Delta \otimes I_k)\Sigma_0^{-1}(a_\Delta \otimes I_k)\right)^{-1}(a'_\Delta \otimes I_k)\Sigma_0^{-1}vec(R_0) - T'T}{2}} \times \left|(a'_\Delta \otimes I_k)\Sigma_0^{-1}(a_\Delta \otimes I_k)\right|^{-1/2} .|\Delta|^{k-2} d\Delta \ . \qquad (6)$$

- This is an integrated likelihoood ratio test where we do not maximize over $\Delta = \beta - \beta_0$, but integrate.
- See power curves for performance.

Conclusion

- Inference with weak instruments is different if errors are homoskedastic and serially uncorrelated, and if errors are HAC.
- With weak instruments tests that perform well in the homoskedastic case ignore relevant information in the HAC case.
- That leads to poor performance of these tests for a class of HAC DGP.
- Although an indication that one has such a DGP can be obtained from the data, there is no test of such DGP-s.
- Therefore practitioners should not use the LM, LM_1 and CQLR tests that are currently implemented in STATA. The AR test is preferred over these, but performs poorly if the number of instruments is large.
- The CIL test is a promising alternative.
- This advice affects researchers who use IV with time-series, spatial and grouped data.

Figure 1: Power curves AR, LM, CQLR, and CIL tests for model with HAC errors with $c_{12} = 100$, $c_{11} = 1$ and $c_{22} = c_{12}^2 + c_{12}^{-3}$; varying instrument strength λ, $\alpha = .05$.