# THRET: Threshold Regression with Endogenous Threshold Variables


**Andros Kourtellos**
Department of Economics
University of Cyprus*

**Thanasis Stengos**
Department of Economics
University of Guelph†

**Chih Ming Tan**
Department of Economics
Tufts University‡



First Draft: June 2006
This Draft: November 13, 2007



**Abstract**

This paper extends the simple threshold regression framework of Hansen (2000) and Caner and Hansen (2004) to allow for endogeneity of the threshold variable. We develop a concentrated two-stage least squares (C2SLS) estimator of the threshold parameter that is based on an inverse Mills ratio bias correction. Our method also allows for the endogeneity of the slope variables. We show that our estimator is consistent and investigate its performance using a Monte Carlo simulation that indicates the applicability of the method in finite samples. We also illustrate its usefulness with an empirical example from economic growth.



JEL Classifications: C13, C51



*P.O. Box 537, CY 1678 Nicosia, Cyprus, e-mail: andros@ucy.ac.cy.
†Guelph, Ontario N1G 2W1, Canada, email: tstengos@uoguelph.ca
‡8 Upper Campus Road, Medford, MA 02155, email: chihming.tan@tufts.edu.

# 1 Introduction

One of the most interesting forms of nonlinear regression models with wide applications in economics is the threshold regression model. The attractiveness of this model stems from the fact that it treats the sample split value (threshold parameter) as unknown. That is, it internally sorts the data, on the basis of some threshold determinants, into groups of observations each of which obeys the same model. While threshold regression is parsimonious it also allows for increased flexibility in functional form and at the same time is not as susceptible to curse of dimensionality problems as nonparametric methods.

While there are several econometric studies on the statistical inference of this model, there is as yet no available inference when the threshold variable itself is endogenous. Chan (1993) showed that the asymptotic distribution of the threshold estimate is a functional of a compound Poisson process. This distribution is too complicated for inference as it depends on nuisance parameters. Hansen (2000) using a concentrated least squares (TR-CLS) approach developed a more useful asymptotic distribution theory for both the threshold parameter estimate and the regression slope coefficients under the assumption that the threshold effect becomes smaller as the sample increases. Using a similar set of assumptions, Caner and Hansen (2004) studied the case of endogeneity in the slope variables. They proposed a concentrated two stage least squares estimator (IVTR-C2SLS) for the threshold parameter and a GMM estimator for the slope parameters. Gonzalo and Wolf (2005) proposed subsampling to conduct inference in the context of threshold autoregressive models. Seo and Linton (2005) allow the threshold variable to be a linear index of observed variables and propose a smoothed least squares estimation strategy based on smoothing the objective function in the sense of Horowitz's smoothed maximum scored estimator. They show that their estimator exhibits asymptotic normality but it depends on the choice of bandwidth.

In all these studies a crucial assumption is that the threshold variable is exogenous. It turns out, however, that in economics many threshold variables depend on their dynamics. In this paper we introduce the Threshold Regression with Endogenous Threshold variables (THRET) and the Threshold Regression with both Endogenous Threshold and Slope variables (THRETS) models and propose an estimation strategy that extends Hansen (2000) and Caner and Hansen (2004). First of all, we show that the naive concentrated 2SLS estimator is an inconsistent estimator of the threshold parameter. Instead, we propose concentrated two-stage least squares estimation (C2SLS) procedure by augmenting the threshold regression with the inverse Mills ratio which resembles the Heckman's selection correction. Under similar assumption as in Caner and Hansen (2004) we show that our estimators are consistent. Our estimation method also allows for endogeneity in the slope variables. To examine the finite sample properties of our estimators we provide a thorough Monte Carlo analysis that shows that for different sample sizes and parameter combinations our proposed estimators for the threshold parameter and the slope coefficients are relatively more efficient than their existing competitors and their distributions have the correct means.

We consider an application of our estimation strategy to a problem that formed our original motivation for thinking about THRET models. We revisit in Section 5 of the paper one of the most important and ongoing debates in the growth empirics literature: the "institutions vs. geography" debate. The key question in this debate is whether geography has direct effects on long-run economic performance or if its influence is limited only to its effects on other growth determinants, such as institutions. Attempts to resolve this debate have centered on the use of linear cross-country regressions where the dependent variable is purchasing-power parity adjusted GDP per capita in 1995 while proxies for institutional quality, climate, disease ecology, macroeconomic policies, and endowments form the set of regressors.

Acemoglu, Johnson, and Robinson (2001), Easterly and Levine (2003), and Rodrik, Subramanian, and Trebbi (2004) conclude that geography's influence on long-run income levels is solely indirect through its effects on institutions, while Sachs (2003) argues that their conclusions are wrong once a measure of malaria transmission is included. Sachs goes further by suggesting that the search for mono-causal effects of fundamental growth determinants on growth may be misdirected. He concludes that, "[t]here is good theoretical and empirical reason to believe that the development process reflects a complex interaction of institutions, policies, and geography [Sachs (2003), p. 9]."

We have explored these points in other papers on the debate. For instance, Tan (2005) employs regression tree methods similar to those used in Durlauf and Johnson (1996) to uncover multiple regimes that classify countries into different convergence clubs. A related but conceptually different approach to modeling parameter heterogeneity and nonlinearities has been taken by Durlauf, Kourtellos, and Minkin (2001) and Mamuneas, Savvides, and Stengos, (2006). These papers have employed varying coefficient models that allow the parameters of the model to vary smoothly as opposed to abruptly in the case of sample splitting methods with a threshold variable. However, these previous studies have assumed that the threshold variable is exogenous. This assumption may be plausible if geography variables or, perhaps, ethnic fractionalization variables were responsible for the threshold effect, but not if institutional quality was the threshold variable since the literature has argued strongly that institutions are endogenous.

In terms of our findings, our results suggest that Sachs' conclusion is only valid for countries with quality of institutions above a threshold level. For low-quality institutions countries, the one factor that appears to have a significant positive impact on economic performance is the degree of trade openness. These results differ from the ones obtained from methods that either ignore the presence of thresholds altogether or ignore the possible endogeneity of the threshold variable.

The paper is organized as follows. Section 2 describes the model and the setup. Section 3 describes the estimator and the main arguments. Section 4 presents our extensive Monte Carlo experiments. Section 5 illustrates our estimator via the empirical example discussed above and section 6 concludes.

## 2 The Threshold Regression with Endogenous Thresholds (THRET) model

We assume that $\{y_i, x_i, q_i, u_i\}_{i=1}^n$ is strictly stationary, ergodic and $\rho-$mixing and that $E(u_i|\mathcal{F}_{i-1})$=0 where $y_i$ is the dependent variable, $x_i$ is a $p \times 1$ vector of covariates and $q_i$ is a threshold variable. Let us first consider the simple case of endogeneity in the threshold alone so that $x_i$ is exogenous and does not include $q_i$. In this case the $l \times 1$ vector of instruments is given by $z_i = (z_{1i}, z_{2i})$, where $z_{2i} = x_i$.

Consider the following THRET model,

$$y_i = x_i'\beta_1 + u_{1i}, \quad q_i \leq \gamma \tag{2.1}$$

$$y_i = x_i'\beta_2 + u_{2i}, \quad q_i > \gamma \tag{2.2}$$

$$q_i = z_i'\pi + v_i \tag{2.3}$$

Equations (2.1) and (2.2) describe the relationship between the variables of interest in each of the two regimes, $q_i$ is the threshold variable with $\gamma$ being the sample split (threshold) value. Equation (2.3) is the selection equation that determines the regime that applies. Note that $q_i$ is observed but the sample split value is unknown.

The variance covariance matrix of the errors $(u_{1i}, u_{2i}, v)'$ has the following properties. $E(u_{1i}, u_{2i}) = 0$, $E(u_{1i}v_i) = \sigma_{u_1v} \neq 0$, $E(u_{2i}v_i) = \sigma_{u_2v} \neq 0$, $E(u_{1i}^2) = \sigma_1^2 > 0$, $E(u_{2i}^2) = \sigma_2^2 > 0$, and $E(v_i^2) = \sigma_v^2 = 1$ due to a normalization. Notice that if $\sigma_{u_1v} = \sigma_{u_2v} = 0$ then we get the exogenous threshold model as in Seo and Linton (2005) that allow the threshold variable to be a linear index of observed variables. If, further, $q_i$ is exogenously given then we get the threshold regression model of Hansen (2000) and Caner and Hansen (2004). Estimation in these two latter models is based on TR-CLS and IVTR-C2SLS, respectively.

One may be tempted to use a naive (plug-in) estimator as in the case of endogeneity in the slope and use a naive concentrated two stage least squares method by replacing $q_i$ with the fitted values from a first stage regression, $\widehat{q}_i$ and then minimize the concentrated least squares criterion. However, such a strategy will not work and the resulting estimator will not be consistent $\widehat{\gamma}^*_{NAIVE-CLS} - \gamma = O_p(1)$ because the conditional mean zero property of the errors is not restored due to omitted bias correction terms.

To see this let us define the indicator variable

$$I_i = \begin{cases} 1 & \text{iff } v_i \leq \gamma - z_i'\pi : \text{Regime 1} \\ 0 & \text{iff } v_i > \gamma - z_i'\pi : \text{Regime 2} \end{cases} \tag{2.4}$$

Let us also assume that that joint distribution between $u_{1i}$ and $v_i$ is given as

$$\binom{u_{1i}}{v_i} | x_i, z_i \sim N\left(0, \left(\begin{array}{cc} \sigma^2_{u_1} & \sigma_{u_1 v} \\ \sigma_{u_1 v} & 1 \end{array}\right)\right) \tag{2.5}$$

and using the following standard transformation

$$\binom{\varepsilon_{1i}}{v_i} = \left(\begin{array}{cc} 1 & -\sigma_{u_1 v} \\ 0 & 1 \end{array}\right) \binom{u_{1i}}{v_i} \tag{2.6}$$

we can get that

$$\binom{\varepsilon_{1i}}{v_i} | x_i, z_i \sim N\left(0, \left(\begin{array}{cc} \sigma^2_1 - \sigma^2_{u_1 v} & 0 \\ 0 & 1 \end{array}\right)\right) \tag{2.7}$$

Similarly we can define the joint distribution between $u_{2i}$ and $v_i$ and also introduce $\varepsilon_{2i}$ in the same way as we did $\varepsilon_{1i}$ to be uncorrelated with $v_i$. Let $\kappa_1 = \sigma_{u_1 v} = \rho_1 \sigma_{u_1}$, $\kappa_2 = \sigma_{u_2 v} = \rho_2 \sigma_{u_2}$, and define

$$u_{1i} = \kappa_1 v_i + \varepsilon_{1i} = \kappa_1 \lambda_1 (\gamma - z_i \pi) + e_{1i} \tag{2.8}$$

$$u_{2i} = \kappa_2 v_i + \varepsilon_{2i} = \kappa_2 \lambda_2 (\gamma - z_i \pi) + e_{2i} \tag{2.9}$$

we have the following conditional expectations for each of the regimes

$$E(y | x_1, z_1, v_i \leq \gamma - z_i'\pi) = x_i \beta_1 + \kappa_1 \lambda_{1i}(\gamma - z_i'\pi) \tag{2.10}$$

$$E(y|x_2, z_2, v_i > \gamma - z_i'\pi) = x_i\beta_2 + \kappa_1\lambda_{2i}(\gamma - z_i'\pi) \tag{2.11}$$

where $\lambda_1(\gamma - z_i'\theta) = -\frac{\phi(\gamma - z_i'\theta)}{\Phi(\gamma - z_i'\theta)}$ and $\lambda_2(\gamma - z_i'\theta) = \frac{\phi(\gamma - z_i'\theta)}{1-\Phi(\gamma - z_i'\theta)}$ are the inverse Mills bias correction terms.

We can also rewrite the THRET model (2.1), (2.2), and (2.3) as a single equation. Let $\lambda_i = \lambda_i(\gamma - z_i'\pi) = d(\gamma)\lambda_{1i} + (1 - d(\gamma))\lambda_{2i}$, $\widetilde{\lambda}_{1i} = d(\gamma)\lambda_{1i}$, $e_i = d(\gamma)e_{1i} + (1 - d(\gamma))\, e_{2i}$, $\delta_\kappa = (\kappa_1 - \kappa_2)$, $\beta = \beta_2$, and $\kappa = \kappa_2$ then we get

$$y_i = x_i'\beta + x_i(\gamma)'\delta_\beta + \kappa\lambda_i(\gamma - z_i'\pi) + \delta_\kappa\widetilde{\lambda}_{1i}\left(\gamma - z_i'\pi\right) + e_i \tag{2.12}$$

where $d_i(\gamma) = I(q_i \leq \gamma)$ and $x_i(\gamma) = x_i d_i(\gamma)$.

It is easy to see that in the case when the two regimes enjoy the same error structure $u_1 = u_2$, or when there is no regime dependent heteroskedasticity, we simply get

$$y_i = x_i\beta + x_i(\gamma)'\delta_\beta + \kappa\lambda_i(\gamma - z_i'\pi) + e_i \tag{2.13}$$

and when $\rho = 0$ and hence $\kappa = 0$ we get Hansen's (2000) Threshold Regression for exogenous threshold and slope variables model,

$$y_i = x_i\beta + x_i(\gamma)'\delta_\beta + e_i \tag{2.14}$$

It is also apparent that THRET is similar in nature to the case of the error interdependence that exists in limited dependent variable models between the equation of interest and the sample selection equation, see Heckman (1979). However, there is one important difference. While in sample selection models, we observe the assignment of observations into regimes but the variable that drives this assignment is taken to be latent, here, it is the opposite; we do not know which observations belong to which regime (we do not know the threshold value), but we can observe the threshold variable.

## 2.1 Estimation

Our estimation procedure proceeds in three steps. First, we estimate the parameter vector $\pi$ in the threshold equation (2.3) by Least Squares (LS). Second, we estimate the threshold estimate by minimizing a concentrated two stage least squares (THRET-C2SLS) criterion using the estimates of $\widehat{\pi}$ from the first stage

$$S_n^{C2SLS}(\beta(\gamma),\delta_\beta(\gamma),\delta_\kappa(\gamma),\kappa(\gamma),\gamma)=\sum_{i=1}^{n}(y_i-x_i'\beta-x_i'(\gamma)\delta_\beta-\kappa\lambda_i(\gamma-z_i'\widehat{\pi})-\delta_\kappa\widetilde{\lambda}_{1i}\left(\gamma-z_i'\widehat{\pi}\right))^2 \tag{2.15}$$

Third, we estimate the LS estimates of the slope parameters based on the split samples implied by $\widehat{\gamma}_{THRET-C2SLS}$.

This sum of squared errors criterion (2.15) implies that Hansen's TR-CLS criterion which is used for estimation of (2.14) will yield an inconsistent estimator for the THRET model given by equations (2.1), (2.2), and (2.3), where

$$S_n^{CLS}(\beta(\gamma),\delta_\beta(\gamma),\delta_\kappa(\gamma),\gamma)=\sum_{i=1}^{n}(y_i-x_i'\beta-x_i'(\gamma)\delta_\beta)^2 \tag{2.16}$$

since it be can be shown that $S_n^{C2SLS}(\beta,\delta_\beta,\delta_\kappa,\kappa,\gamma)=S_n^{CLS}(\beta,\delta,\gamma)+O_p(1)$.

**Proposition 1: Consistency of C2SLS Estimator in THRET** For the C2SLS estimator in the case of endogenous threshold but exogenous slope variables defined as $\widehat{\gamma}_{C2SLS}=\arg\min\left(S_n^{C2SLS}(\gamma)-e'e\right)$ we have that $\widehat{\gamma}_{THRET-C2SLS}\xrightarrow{p}\gamma_0$.

In the appendix we provide a proof that uses similar regularity conditions as Hansen (2000).

# 3 The Threshold Regression with Endogenous Threshold and Slope model (THRETS)

In this section we generalize THRET to the more realistic case of a Threshold Regression with Endogenous Threshold and Slope (THRETS) variables. THRETS takes the form of

$$y_i=x_i'\beta+x_i(\gamma)'\delta_\beta+\kappa\lambda_i(\gamma-z_i'\pi)+\delta_\kappa\widetilde{\lambda}_{1i}\left(\gamma-z_i'\pi\right)+e_i \tag{3.17}$$

and

$$x_i=\Pi'z_i+\eta_i \tag{3.18}$$

where the $l\times 1$ vector $z_i=(z_{1i},z_{2i})$ with $z_{2i}=x_{2i}$ and $E(\eta_i|z_i)=0$, and where $l\geq p$. $\pi_1$ is the parameter vector of the regression of $q_i$ on $z_i$ such that $\Pi=(\pi_1,\Pi_2)$.

Again we propose an estimation procedure based on three steps. First, we estimate the parameter vector $\Pi$ in the threshold equation (3.18) by LS . Second, we estimate the sample split (threshold) value by minimizing a Concentrated Two Stage Least Squares (C2SLS) criterion using the estimates of $\widehat{\Pi}$ from the first stage

$$S_n^{C2SLS}(\beta(\gamma), \delta_\beta(\gamma), \delta_\kappa(\gamma), \kappa(\gamma), \gamma) = \sum_{i=1}^{n}(y_i - \widehat{x}_i'\beta - \widehat{x}_i'(\gamma)\delta_\beta - \kappa\lambda_i(\gamma - z_i'\widehat{\pi}_1) - \delta_\kappa\widetilde{\lambda}_{1i}\left(\gamma - z_i'\widehat{\pi}_1\right))^2 \tag{3.19}$$

Third, we estimate the slope parameters using 2SLS or GMM on the split samples implied by the estimate of $\gamma$.

Using a similar framework as in Caner and Hansen (2004) it can be shown that $\widehat{\gamma}_{THRETS-C2SLS} = \arg\min\left(S_n^{C2SLS}(\gamma) - e'e\right)$ is consistent.

# 4 Monte Carlo

We proceed below with an exhaustive simulation study that compares the small sample performance of our estimator against existing estimators. In particular, when we allow for the endogeneity of the threshold alone we compare THRET-C2SLS estimates of the threshold parameter against estimates based on TR-CLS (Hansen, 2000) and a naive C2SLS estimator (NAIVE-C2SLS) that simply uses the fitted values from a first stage as a threshold variable. We also compare the LS estimates of the slope coefficients that are based on the subsamples implied by $\widehat{\gamma}$. Likewise when we allow for the endogeneity of both the slope and the threshold variable we compare our estimator against the IVTR-C2SLS (Hansen, 2004), and the naive C2SLS estimator (NAIVE-C2SLS) that replaces both the threshold and the slope variables with the fitted values from a first stage and then minimizes a concentrated least squares criterion. In this case we compare the GMM estimates of the slope coefficients for the various estimators.

The Monte Carlo design is based on the following threshold regression

$$y_i = \begin{cases} x_i'\beta_1 + u_i, & q_i \leq 2 \\ x_i'\beta_2 + u_i, & q_i > 2 \end{cases} \tag{4.20}$$

The threshold equation is given by

$$q_i = 2 + 3z_{1i} + 3z_{2i} + v_i \tag{4.21}$$

where $v_i, \varepsilon_i \sim NIID(0,1)$ and $u_i = \sigma_u^2 \left(\rho_0 v_i + (1-\rho_0)\varepsilon_i\right) / \sqrt{\left(\rho_0^2 + (1-\rho_0)^2\right)}$ so that the degree of the endogeneity is controlled via the correlation between $u_i$ and $v_i$ given by $\rho = \rho_0 / \sqrt{\left(\rho_0^2 + (1-\rho_0)^2\right)}$. We specify $\rho_0 = 0.05$, 0.50, and 0.95. We fix $\beta_2 = 1$ and vary $\beta_1$ by examining various $\delta = \beta_1 - \beta_2$, $\delta = (0.01, 0.05, 0.1, 0.25, 0.5, 1)$. First, we examine the case where the threshold variable is the only endogenous variable $x_i = (1, x_{2i})$ and second, we look into the more realistic case that allows for both endogeneity in the threshold and the slope variables $x_i = (1, q_i, x_{2i})$. Furthermore, we consider the implications of the degree of correlation between the (excluded) instrumental variables $z_i$ and the exogenous slope variables $x_{2i}$ through $z_{ij} = \left(\omega_0 x_2 + (1-\omega_0)\xi_{ij}\right) / \sqrt{\left(\omega_0^2 + (1-\omega_0)^2\right)}$, where $\xi_{ij} \sim NIID(0,1)$ and $\omega_0 / \sqrt{\left(\omega_0^2 + (1-\omega_0)^2\right)}$ is the degree of correlation between $z_i$ and $x_{2i}$. Finally, we consider sample sizes of $100, 200$, and 500 using 1000 Monte Carlo simulations.

Tables 1-3 discuss the relative Mean Square Error (MSE) while Figures 1-7 present the Gaussian kernel density estimates using Silverman's bandwidth parameter of the Monte Carlo estimates of the threshold coefficient $\gamma$ and the difference of slope coefficients $\delta = \beta_1 - \beta_2$ of the various estimators.

First we consider the estimation of the threshold $\gamma$ in 2.1, 2.2, and 2.3 in the case of endogeneity in the threshold alone. Table 1(a) presents the relative MSEs of TR-CLS and NAIVE-C2SLS relative to THRET-C2SLS estimator given by $MSE_{TR}/MSE_{THRET}$ and $MSE_{NAIVE}/MSE_{THRET}$, respectively, across different values of $\delta$, different quantiles and sample sizes $n$ when $\rho_0 = 0.95$ and $\omega_0 = 0.95$. For all $\delta$ and all sample sizes the relative MSEs show that THRET is relatively more efficient than CLS and NAIVE-2SCLS. These efficiencies are largest for the right tail as shown by the 95th quantile of standard error. Similarly, Table 1(b) demonstrates the relative efficiency of THRET-C2SLS when there is endogeneity in both the threshold and slope variables using $MSE_{IVTR}/MSE_{THRETS}$ and $MSE_{NAIVE}/MSE_{THRETS}$ across different values of $\delta$, different quantiles and sample sizes $n$ when $\rho_0 = 0.95$ and $\omega_0 = 0.95$. Figures 1-2 show the corresponding kernel density estimates of the threshold estimator for various values of $\delta$. Figures 6-7 show for $\delta = 0.5$ the kernel density estimates of the threshold estimator for various degrees of endogeneity $\rho_0 = 0.05, 0.50, 0.95$.[1] It is evident that the distribution of THRET-C2SLS and THRETS-C2SLS centers around the true value and dominates its competitors in terms of efficiency. Under the assumption of small thresholds effects in the sense that $\delta_{\beta,n} \to 0$ as $n \to \infty$, the asymptotic distribution of the threshold estimator is a suitably modified version of the non-standard distribution derived by Hansen (2000) and by Caner and Hansen (2000) for the case of exogenous

[1] We have conducted experiments across different degrees of threshold endogeneity (different values of $\rho$) and a broad range of values of $\delta$. Although these experiments are not reported in detail to conserve space, they are available from the authors on request.

and endogenous regressors, respectively. This is verified by the figures that we obtained for the different values of $\delta$.

In terms of slope coefficients our estimator performs at least as well as the respective competitors. Tables 2(a) presents the relative MSEs of the LS estimates and Table 2(b) presents the relative MSEs of the GMM estimates of the slope coefficient of the exogenous covariate of THRET and THRETS, respectively. Similarly, Table 3 presents the relative MSEs of the GMM estimates of the slope coefficient of the endogenous covariate. Figures 3-5 present the corresponding kernel density estimates.

In the interests of robustness, we also investigated what happened when we varied the degree of the correlation between the instrumental variables $z$ and the exogenous slope variables $x_2$. As in the case of Heckman's estimator, THRET-C2SLS and THRETS-C2SLS become more efficient as $\omega$ decreases and the degree of multicollinearity between $\pi'z$ and $x$ is small. Furthermore, our findings are also robust to regime dependent heteroskedasticity. Due to space limitations these experiments are not reported in detail but they are all available from the authors on request.

# 5 Empirical Example

In this section, we revisit the institutions versus geography debate using our THRET methods, as discussed in the Introduction. The data we use comes primarily from Easterly and Levine (2003). As mentioned above, the dependent variable is the log of GDP per capita in 1995. We include a variable that measures trade openness and a variable that measures ethnic diversity. We also include a proxy for institutional quality, the average (over 1985-95) expropriation risk variable, from the International Country Risk Guide (ICRG). Finally, we augment the Easterly-Levine dataset with Sachs' preferred malaria variable (MALFAL94p) from the Harvard Center for International Development (CID). Following Acemoglu, Johnson, and Robinson (2001) we instrument institutional quality (which is assumed to be endogenous) using the log of European settler mortality.

We contrast results where the model is assumed to be linear against those where the model is a THRETS model with institutional quality as the (endogenous) threshold variable. Table 4 presents the results. Our objective in these exercises is not to embark on a thorough re-examination of this important debate, but rather to highlight how taking Sachs' methodological critique above (see, Introduction) seriously can lead to new and important insights. In all cases, we find that our THRETS-C2SLS results deliver more nuance interpretations of established findings.

For example, column 1 of Table 4 shows the linear 2SLS results for a regression of per capita GDP on institutional quality and malaria. These results for the linear model appear to support Sachs' finding that "malaria transmission, which is strongly affected by ecological conditions, directly affects the level of per capita income after controlling for the quality of institutions [Sachs (2003), Abstract]". However, our THRETS-C2SLS results (see, column 2 of Table 4) suggest that this finding for malaria is only true when the quality of institutions is above a threshold level. Below that threshold level, neither institutions nor disease ecology appears to have any effect on a country's economic performance. This result is maintained even if we include Easterly and Levine's ethnic diversity variable as another growth determinant (see column 4 of Table 4). In columns 5 and 6 of Table 4 where we also include the trade openness variable as a growth determinant, we find that Rodrik, Subramanian, and Trebbi (2004) may have under-sold the importance of macroeconomic policies that promote a more open economy when they claim that "once institutions are controlled for, trade is almost always insignificant [Rodrik, et al, Abstract]". Their claims certainly appear to be true when we assume a linear model (see, column 5 of Table 4). However, our THRETS-C2SLS results suggest that for low-quality institutions countries, trade openness may be one of the only factors that has a significant positive impact on economic performance. For high-quality institutions countries, on the other hand, good institutions promote economic performance while higher levels of ethnic diversity detract from it.[2]

We also carried out a series of robustness checks (unreported) where we included other macroeconomic policy variables that are commonly employed in the literature, such as inflation and real exchange rate overvaluation, as additional growth determinants. We also included other fundamental determinants such as religious affiliation shares for Catholics, Muslims, and Other Religions on the righthand-side. We carried out exercises that included various combinations of these regressors along with those described above. In most cases, we found that the results for the THRETS-C2SLS model differed substantially from those for the linear model. Overall, we conclude that there is much evidence to suggest that there exists substantial heterogeneity in the growth experiences of countries, and that studies that seek to promote mono-causal explanations for the variation in long-run economic performance across countries are potentially misleading.

## 6 Conclusion

In this paper we propose an extension of Hansen (2000) and Caner and Hansen (2004) that deals with the endogeneity of the threshold variable. We developed a concentrated two stage least squares

[2] As in Hansen (2000) we compute a heteroskedasticity corrected asymptotic confidence interval for threshold estimate using a quadratic polynomial. One difference is that the nuisance parameters in the conditional variance is estimated via a polynomial regression in $\widehat{q}$ and $\widehat{q^2}$ instead of $q$ and $q^2$. $\widehat{q}$ and $\widehat{q^2}$ are the fitted values from LS regressions of $q$ and $q^2$ on the set of instruments $z$. Simulated coverage probabilities of a nominal coverage of 90% interval provides support to our proposal. Due to space limitations these experiments are not reported but they are all available from the authors on request.

estimator that deals with the problem of endogeneity in the threshold variable by generating a correction term based on the inverse Mills ratios to produce consistent estimates for the threshold parameter and the slope coefficients. By means of an extensive simulation study we examine the performance of our estimator when compared with its competitors. Our proposed estimator performs well for a variety of sample sizes and parameter combinations. We illustrate the usefulness of the proposed estimator by means of an empirical example from economic growth.

## References


[1] Acemoglu, D. Johnson, S. and J. A. Robinson (2001), "The Colonial Origins of Comparative Development: An Empirical Investigation," *American Economic Review,* 91, p. 1369-1401.

[2] Caner, M. and B. Hansen. (2004), "Instrumental Variable Estimation of a Threshold Model," *Econometric Theory*, 20, p. 813-843.

[3] Chan, K, . S. (1993), "Consistency and Limiting Distribution of the Least Squares Estimator of a Threshold Autoregressive Model," *The Annals of Statistics*, 21, p. 520-533.

[4] Dhrymes, P. J. (2004), "Limited Information Estimation of Simultaneous Equations with Indicator Endogenous Variables", Columbia University, mimeo.

[5] Durlauf, S., A. Kourtellos, and A. Minkin, (2001), "The Local Solow Growth Model," *European Economic Review*, 45, 4-6, p. 928-40.

[6] Durlauf, S. N., and P. A. Johnson (1995), "Multiple Regimes and Cross Country Behavior", *Journal of Applied Econometrics*, 10(4), p. 365-384.

[7] Easterly, W. and R. Levine (2003), "Tropics, Germs, and Crops: How Endowments Influence Economic Development," *Journal of Monetary Economics*, 50(1), p. 3-39.

[8] Gonzalo, J. and M. Wolf (2005), "Subsampling Inference in Threshold Autoregressive Models," *Journal of Econometrics*, 127(2), p. 201-224.

[9] Hansen, B. E. (2000), "Sample Splitting and Threshold Estimation," *Econometrica*, 68(3), p. 575-604.

[10] Heckman, J. (1979), "Sample Selection Bias as a Specification Error," *Econometrica*, 47(1), p. 153-161.

[11] Mamuneas, T., A. Savvides, and T. Stengos, (2006), "Economic Development and Return to Human Capital: A Smooth Coefficient Semiparametric Approach," *Journal of Applied Econometrics*, 21(1), p. 111-132.

[12] Rodrik, D., A. Subramanian, and F. Trebbi (2004), "Institutions Rule: The Primacy of Institutions over Geography and Integration in Economic Development," *Journal of Economic Growth*, 9(2), p.131-165.

[13] Sachs, J. (2003), "Institutions Don't Rule: Direct Effects of Geography on Per Capita Income," National Bureau of Economic Research Working Paper No. 9490.

[14] Seo, M. H. and O. Linton (2007), "A Smoothed Least Squares Estimator For Threshold Regression Models," *Journal of Econometrics*, 141(2), p. 704-735.

[15] Tan, C. M. (2005), "No One True Path: Uncovering the Interplay between Geography, Institutions, and Fractionalization in Economic Development", Tufts University, Working Paper No. 2005-12.

# Appendix

## A Consistency of $\widehat{\gamma}_{C2SLS}$. Proof of Proposition 1

Let us define the $n \times 1$ vector $Y$, the $n \times p$ matrix $X$ and the $n \times l$ matrix $Z$ by stacking $y_i$, $x_i$ and, $z_i$, respectively. We also define $X_\gamma$ to be the $n \times l$ matrix with typical $i-th$ row $x_i(\gamma) = x_i d_i(\gamma)$, where as before $d(\gamma) = I(q_i \leq \gamma)$ and similarly $\Lambda_\gamma$ to be the matrix with typical element $d(\gamma)\lambda_{1i}$. Let us also define the $n \times 1$ vector $\Lambda(\gamma) = d(\gamma)\lambda_1(\gamma) + (1 - d(\gamma))\lambda_2(\gamma)$.

At $\gamma_0$, $\Lambda(\gamma_0) = \Lambda(0)$, $X_{\gamma_0} = X_0$, $\Lambda_{\gamma_0} = \Lambda_0$. In the spirit of Hansen (2000), we define $\widetilde{X}_\gamma = (X_\gamma, \Lambda_\gamma)$, $\widetilde{X}(\gamma) = (X, \Lambda(\gamma))$, $\widetilde{X}^*_\gamma = (\widetilde{X}_\gamma, \widetilde{X}(\gamma) - \widetilde{X}_\gamma))$ and using similar regularity conditions we assume that

$$\begin{gathered}
\tfrac{1}{n}\widetilde{X}'_\gamma \widetilde{X}_\gamma \xrightarrow{p} M(\gamma) \\
\tfrac{1}{n}\widetilde{X}'_{\gamma_0} \widetilde{X}_{\gamma_0} \xrightarrow{p} M(\gamma_0) \\
\tfrac{1}{n}\widetilde{X}'_{\gamma_0} \widetilde{X}_\gamma \xrightarrow{p} M(\gamma_0) \\
\tfrac{1}{n}(\widetilde{X}(\gamma_0) - \widetilde{X}_{\gamma_0})'\widetilde{X}_\gamma \xrightarrow{p} 0
\end{gathered}$$

The last condition guarantees that asymptotically the matrix of cross products between $\widetilde{X}^*_{\gamma_0}$ and $\widetilde{X}^*_\gamma$ for $\gamma \geq \gamma_0$ is diagonal.

We also have that

$$p\lim_{n\to\infty} \frac{1}{n} \begin{pmatrix} \widetilde{X}'_0 \widetilde{X}_0 & 0 \\ \left(\widetilde{X}(0) - \widetilde{X}_0\right)' \widetilde{X}_\gamma & \left(\widetilde{X}(\gamma) - \widetilde{X}_\gamma\right)' \left(\widetilde{X}(\gamma) - \widetilde{X}_\gamma\right) \end{pmatrix} =$$

$$\begin{pmatrix} M(\gamma_0) & 0 \\ 0 & p\lim_{n\to\infty} \frac{1}{n} \left(\widetilde{X}(\gamma) - \widetilde{X}_\gamma\right)' \left(\widetilde{X}(\gamma) - \widetilde{X}_\gamma\right) \end{pmatrix} = M(\gamma_0, \gamma)$$

We then define the projection matrix spanned by the columns of $\widetilde{X}^*_\gamma$.

$$\widetilde{P}^*_\gamma = \widetilde{X}^*_\gamma \left(\widetilde{X}^{*\prime}_\gamma \widetilde{X}^*_\gamma\right)^{-1} \widetilde{X}^{*\prime}_\gamma \tag{A.1}$$

Let us rewrite the model as

$$Y = X\theta + X_0\delta + \rho\Lambda(0) + \phi\Lambda_0 + \varepsilon \tag{A.2}$$

or

$$Y = \widetilde{X}(0)\alpha + \widetilde{X}_0\psi + \varepsilon \tag{A.3}$$

So we have

$$S_n(\gamma) - \varepsilon'\varepsilon = Y'\left(I - \widetilde{P}_\gamma^*\right) Y - \varepsilon'\varepsilon \tag{A.4}$$

Then as in Hansen (2000) for $\psi_n = Cn^{-\mu}$ with $C \neq 0$ and $0 < \mu < \frac{1}{2}$

$$n^{-1+2\mu}\left(S_n(\gamma) - \varepsilon'\varepsilon\right)$$

$$= n^{-1+2\mu}\left[\left(\widetilde{X}(0)\alpha_n + \widetilde{X}_0\psi_n + \varepsilon\right)'\left(I - \widetilde{P}_\gamma^*\right)\left(\widetilde{X}(0)\alpha_n + \widetilde{X}_0\psi_n + \varepsilon\right) - \varepsilon'\varepsilon\right]$$

$$= n^{-1}\left[C_1'\widetilde{X}'(0)\left(I - \widetilde{P}_\gamma^*\right)\widetilde{X}(0)C_1 + C_2'\widetilde{X}'(0)\left(I - \widetilde{P}_\gamma^*\right)\widetilde{X}_0C_2 + C_3'\widetilde{X}_0'\left(I - \widetilde{P}_\gamma^*\right)\widetilde{X}_0C_3\right] + o_p(1)$$

$$= C_1'\left(\frac{\widetilde{X}'(0)\widetilde{X}(0)}{n}\right)C_1 - C_1'\left(\frac{\widetilde{X}'(0)\widetilde{X}_\gamma^*}{n}\right)\left(\frac{\widetilde{X}_\gamma^{*\prime}\widetilde{X}_\gamma^*}{n}\right)^{-1}\left(\frac{\widetilde{X}_\gamma^{*\prime}\widetilde{X}(0)}{n}\right)C_1 +$$

$$C_2'\left(\frac{\widetilde{X}'(0)\widetilde{X}_0}{n}\right)C_2 - C_2'\left(\frac{\widetilde{X}'(0)\widetilde{X}_\gamma^*}{n}\right)\left(\frac{\widetilde{X}_\gamma^{*\prime}\widetilde{X}_\gamma^*}{n}\right)^{-1}\left(\frac{\widetilde{X}_\gamma^{*\prime}\widetilde{X}_0}{n}\right)C_2 +$$

$$C_3'\left(\frac{\widetilde{X}_0'\widetilde{X}_0}{n}\right)C_3 - C_3'\left(\frac{\widetilde{X}_0'\widetilde{X}_\gamma^*}{n}\right)\left(\frac{\widetilde{X}_\gamma^{*\prime}\widetilde{X}_\gamma^*}{n}\right)^{-1}\left(\frac{\widetilde{X}_\gamma^{*\prime}\widetilde{X}_0}{n}\right)C_3 + o_p(1)$$

$$= C'\left[\left(\frac{\widetilde{X}_0^{*\prime}\widetilde{X}_0^*}{n}\right) - \left(\frac{\widetilde{X}_0^{*\prime}\widetilde{X}_\gamma^*}{n}\right)\left(\frac{\widetilde{X}_\gamma^{*\prime}\widetilde{X}_\gamma^*}{n}\right)^{-1}\left(\frac{\widetilde{X}_\gamma^{*\prime}\widetilde{X}_0^*}{n}\right)\right]C + o_p(1)$$

That is,

$$\begin{aligned} & n^{-1+2\mu}\left(S_n(\gamma) - \varepsilon'\varepsilon\right) \\ = \quad & C'\left[\left(\frac{\widetilde{X}_0^{*\prime}\widetilde{X}_0^*}{n}\right) - \left(\frac{\widetilde{X}_0^{*\prime}\widetilde{X}_\gamma^*}{n}\right)\left(\frac{\widetilde{X}_\gamma^{*\prime}\widetilde{X}_\gamma^*}{n}\right)^{-1}\left(\frac{\widetilde{X}_\gamma^{*\prime}\widetilde{X}_0^*}{n}\right)\right]C + o_p(1) \end{aligned} \tag{A.5}$$

It can be shown that

$$n^{-1+2\mu}\left(S_n(\gamma) - \varepsilon'\varepsilon\right) \xrightarrow{p} C'\left[M(\gamma_0) - M(\gamma_0, \gamma)M(\gamma)^{-1}M(\gamma_0, \gamma)'\right]C \tag{A.6}$$

Let

$$b_1(\gamma) = C' \left[ M(\gamma_0) - M(\gamma_0, \gamma) M(\gamma)^{-1} M(\gamma_0, \gamma)' \right] C$$

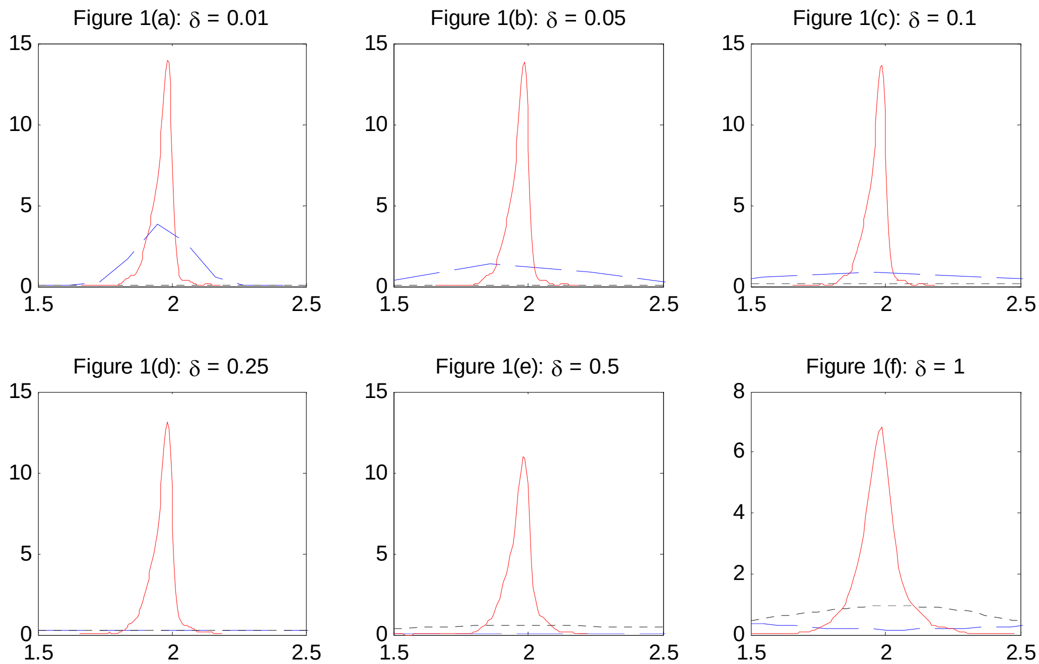
Following similar arguments as in Hansen (2000, Lemma A.5) it can be shown that $\frac{d}{d\gamma} b_1(\gamma_0) > 0$ and $b_1(\gamma)$ is continuous and weakly increasing so that $b_1(\gamma)$ is uniquely minimized at $\gamma_0$ over $[\gamma_0, \overline{\gamma}]$. A similar argument can be made for $\gamma \in [\underline{\gamma}, \gamma_0]$, so that $b_2(\gamma)$ which is suitably defined is uniquely minimized at $\gamma_0$.

So uniformly over all values of $\gamma$,

$$n^{-1+2\mu} \left( S_n^*(\gamma) - \varepsilon'\varepsilon \right) \xrightarrow{p} b_1(\gamma) 1_{\{\gamma > \gamma_0\}} + b_2(\gamma) 1_{\{\gamma \leq \gamma_0\}} \tag{A.7}$$

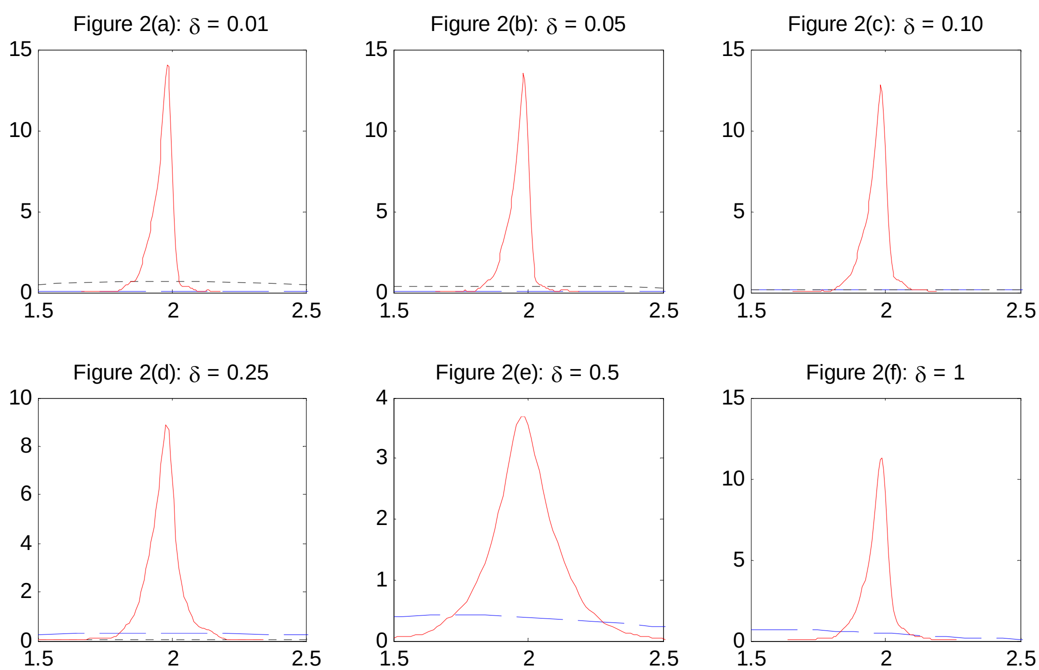
Since $\widehat{\gamma}_{C2SLS} = \arg\min \left( S_n^*(\gamma) - \varepsilon'\varepsilon \right)$, we get that $\widehat{\gamma}_{C2SLS} \xrightarrow{p} \gamma_0$.

**Figures 1(a) – (f) : MC Kernel Densities of the Threshold Estimate (endogeneity in the threshold alone)**



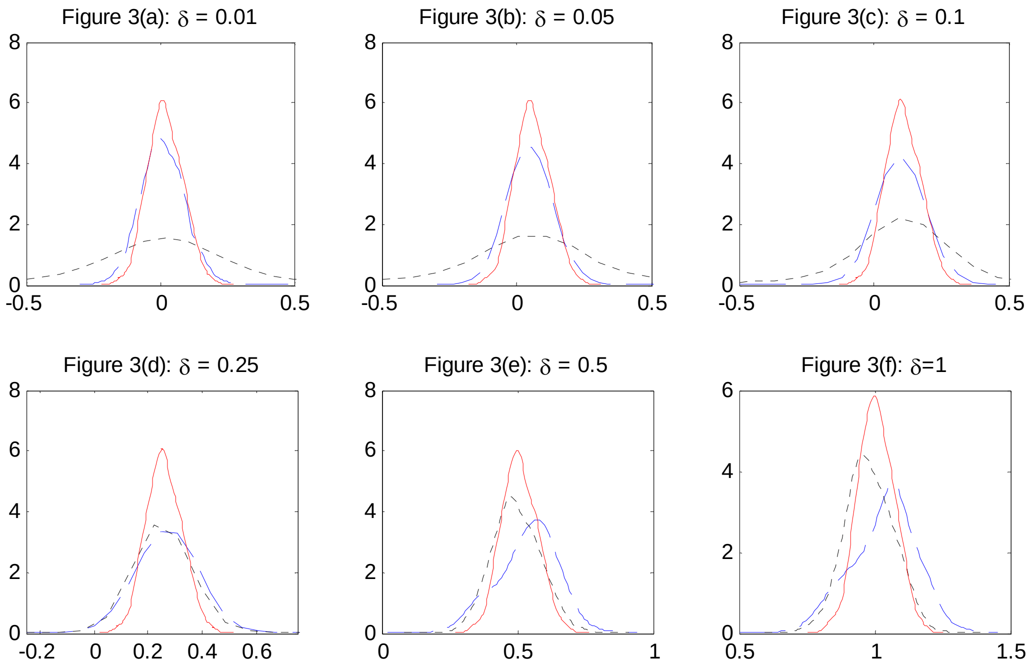

Note: "The solid line represents THRET-C2SLS, the dashed line represents TR-CLS, and the dotted line represents NAÏVE-CLS."

**Figures 2(a) – (f) : MC Kernel Densities of the Threshold Estimate (endogeneity in both the threshold and slope)**



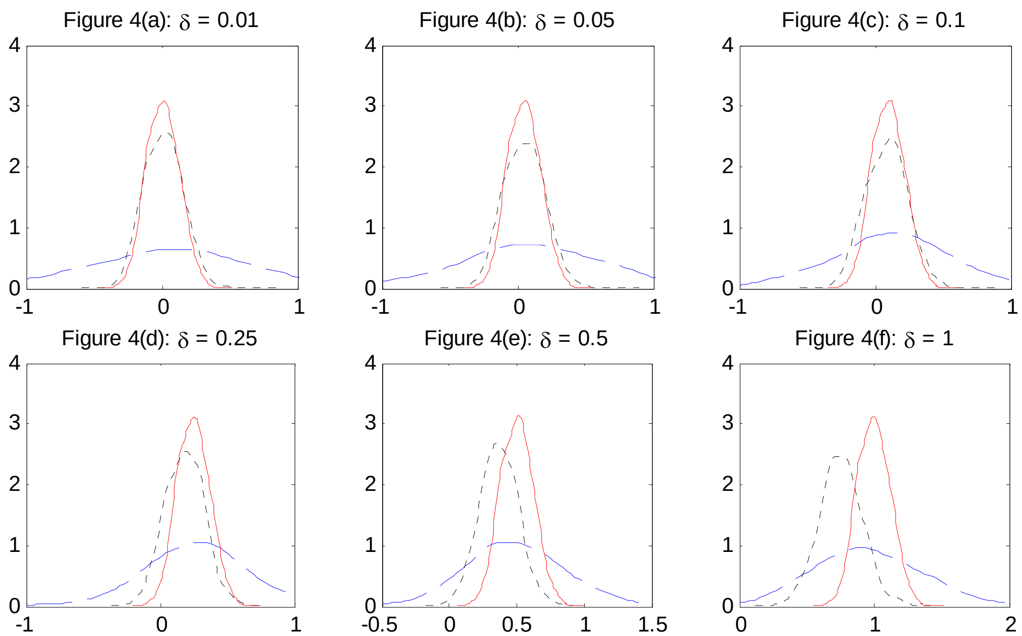

Note: "The solid line represents THRETS-C2SLS, the dashed line represents IVTR-C2LS, and the dotted line represents NAÏVE-C2LS."

**Figures 3(a) – (f) : MC Kernel Densities of the Slope Coefficient of the Exogenous Covariate (endogeneity in the threshold alone)**



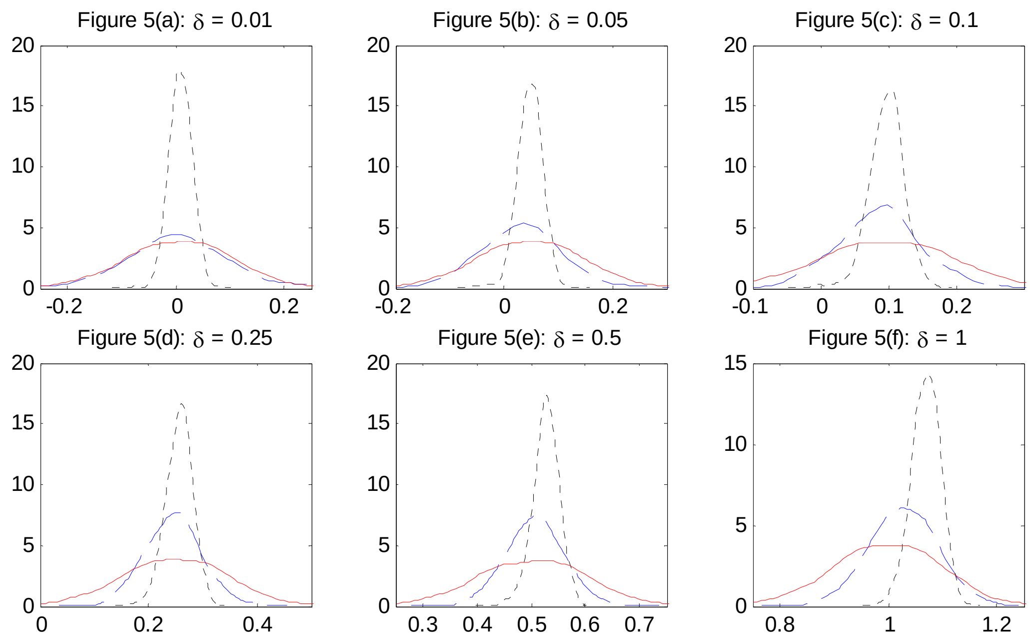

Note: "The solid line represents THRET-LS, the dashed line represents TR-LS, and the dotted line represents NAÏVE-LS."

**Figures 4(a) – (f) : MC Kernel Densities of the Slope Coefficient of the Exogenous Covariate (endogeneity in both the threshold and slope)**



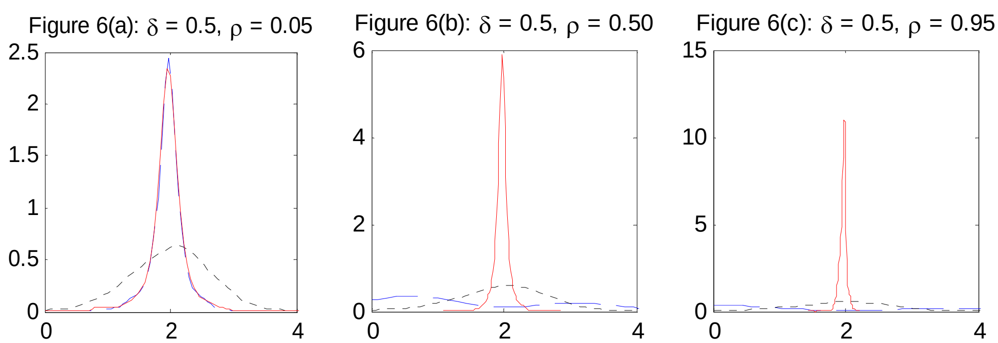

Note: "The solid line represents THRET-GMM, the dashed line represents IVTR-GMM, and the dotted line represents NAÏVE-GMM."

**Figures 5(a) – (f) : MC Kernel Densities of the Slope Coefficient of the Endogenous Covariate**



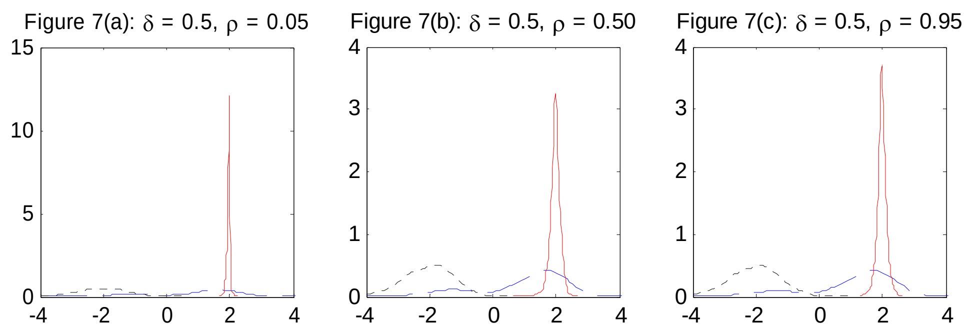

Note: “The solid line represents THRET-GMM, the dashed line represents IVTR-GMM, and the dotted line represents NAÏVE-GMM.”

**Figures 6(a) – (c) : MC Kernel Densities of the Threshold Estimate for various degrees of endogeneity (endogeneity in the threshold alone)**




Note: “The solid line represents THRET-C2SLS, the dashed line represents TR-CLS, and the dotted line represents NAÏVE-CLS.”

**Figures 7(a) – (c) : MC Kernel Densities of the Threshold Estimate for various degrees of endogeneity (endogeneity in both the threshold and slope)**




Note: “The solid line represents THRETS-C2SLS, the dashed line represents IVTR-C2LS, and the dotted line represents NAÏVE-C2LS.”

**Table 1(a): Relative Efficiency of Threshold Estimator $\hat{\gamma}$ (Endogeneity in the threshold alone)**

| Quantiles | $MSE_{TR} / MSE_{THRET}$ | | | $MSE_{NAIVE} / MSE_{THRET}$ | | |
|---|---|---|---|---|---|---|
| | 0.05 | 0.50 | 0.95 | 0.05 | 0.50 | 0.95 |
| δ = 0.01 | | | | | | |
| n = 100 | 2.565 | 4.134 | 239.8 | 506.7 | 614.7 | 275.4 |
| n = 200 | 2.184 | 2.095 | 727.7 | 2474 | 2471 | 2189.8 |
| n = 500 | 1.298 | 1.372 | 1.544 | 7432 | 28140 | 23941 |
| δ = 0.05 | | | | | | |
| n = 100 | 3.948 | 6.270 | 244.2 | 536.2 | 649.7 | 267.1 |
| n = 200 | 2.940 | 2.683 | 1096.3 | 2946 | 2285 | 2139 |
| n = 500 | 1.193 | 1.411 | 1.809 | 5312 | 14255 | 22876 |
| δ = 0.10 | | | | | | |
| n = 100 | 4.475 | 23.33 | 228.5 | 512.4 | 483.5 | 247.2 |
| n = 200 | 3.027 | 4.646 | 1289 | 2384 | 1610 | 2072 |
| n = 500 | 1.326 | 1.630 | 198.8 | 1767 | 4114 | 19016 |
| δ = 0.25 | | | | | | |
| n = 100 | 17.71 | 230.2 | 121.1 | 220.8 | 137.9 | 114.5 |
| n = 200 | 19.35 | 708.5 | 1371 | 742.0 | 292.3 | 1191 |
| n = 500 | 5.367 | 676.1 | 2118 | 726.7 | 724.0 | 948.1 |
| δ = 0.50 | | | | | | |
| n = 100 | 570.7 | 109.0 | 34.63 | 39.79 | 22.33 | 13.64 |
| n = 200 | 4113 | 432.3 | 137.2 | 118.8 | 68.49 | 43.99 |
| n = 500 | 67296 | 2964 | 574.8 | 152.6 | 242.9 | 138.1 |
| δ = 1.0 | | | | | | |
| n = 100 | 182.3 | 21.57 | 5.248 | 13.64 | 5.197 | 2.585 |
| n = 200 | 750.7 | 61.32 | 16.00 | 23.86 | 10.74 | 5.701 |
| n = 500 | 4288 | 398.4 | 79.38 | 38.88 | 40.27 | 17.07 |

**Table 1(b): Relative Efficiency of Threshold Estimator of $\hat{\gamma}$ (Endogeneity in both the threshold and the slope)**

| Quantiles | $MSE_{IVTR} / MSE_{THRETS}$ | | | $MSE_{NAIVE} / MSE_{THRETS}$ | | |
|---|---|---|---|---|---|---|
| | 0.05 | 0.50 | 0.95 | 0.05 | 0.50 | 0.95 |
| δ = 0.01 | | | | | | |
| n = 100 | 15.26 | 53.15 | 176.4 | 547.2 | 559.3 | 190.2 |
| n = 200 | 17.97 | 34.63 | 891.1 | 1607 | 2905 | 1325 |
| n = 500 | 6.346 | 9.382 | 1115 | 11380 | 23037 | 10522 |
| δ = 0.05 | | | | | | |
| n = 100 | 21.34 | 175.6 | 134.1 | 644.5 | 499.1 | 142.5 |
| n = 200 | 38.18 | 128.6 | 947.8 | 1859 | 2509 | 1208 |
| n = 500 | 9.463 | 22.74 | 4553 | 6246 | 9967 | 8900 |
| δ = 0.10 | | | | | | |
| n = 100 | 60.27 | 542.13 | 55.11 | 642.9 | 413.2 | 52.01 |
| n = 200 | 159.4 | 2987 | 875.9 | 1256 | 1169 | 903.8 |
| n = 500 | 36.60 | 14798 | 5003 | 2942 | 3266 | 4860 |
| δ = 0.25 | | | | | | |
| n = 100 | 3259 | 494.7 | 42.04 | 154.8 | 121.8 | 31.96 |
| n = 200 | 94242 | 2699 | 248.5 | 350.5 | 233.2 | 167.3 |
| n = 500 | 904607 | 16650 | 2045 | 697.6 | 622.7 | 1068 |
| δ = 0.50 | | | | | | |
| n = 100 | 7756 | 276.9 | 39.92 | 21.28 | 43.63 | 28.26 |
| n = 200 | 67537 | 626.0 | 121.3 | 71.90 | 37.84 | 73.39 |
| n = 500 | 232434 | 2798 | 328.5 | 54.19 | 76.45 | 164.3 |
| δ = 1.0 | | | | | | |
| n = 100 | 49970 | 636.9 | 108.1 | 56.79 | 88.51 | 84.17 |
| n = 200 | 247379 | 2345 | 281.7 | 131.5 | 111.5 | 176.2 |
| n = 500 | 1188079 | 13853 | 1071 | 325.5 | 364.9 | 531.2 |

**Table 2(a): Relative Efficiency of the LS estimates of the Slope Coefficient of Exogenous Covariate $\hat{\delta}_2$ (Endogeneity in the threshold alone)**

| Quantiles | $MSE_{TR} / MSE_{THRET}$ | | | $MSE_{NAIVE} / MSE_{THRET}$ | | |
|---|---|---|---|---|---|---|
| | 0.05 | 0.50 | 0.95 | 0.05 | 0.50 | 0.95 |
| δ = 0.01 | | | | | | |
| n = 100 | 1.713 | 2.182 | 5.736 | 1.301 | 4.998 | 16.06 |
| n = 200 | 1.416 | 1.363 | 2.559 | 4.301 | 4.676 | 33.18 |
| n = 500 | 2.165 | 1.432 | 1.531 | 13.05 | 7.151 | 75.22 |
| δ = 0.05 | | | | | | |
| n = 100 | 1.780 | 2.455 | 7.457 | 2.105 | 4.960 | 16.60 |
| n = 200 | 1.556 | 1.481 | 4.298 | 3.766 | 4.701 | 31.43 |
| n = 500 | 2.226 | 1.415 | 1.544 | 10.908 | 6.194 | 59.80 |
| δ = 0.10 | | | | | | |
| n = 100 | 2.123 | 2.691 | 10.36 | 1.614 | 4.236 | 14.30 |
| n = 200 | 1.834 | 1.743 | 8.277 | 4.002 | 3.791 | 29.69 |
| n = 500 | 2.331 | 1.545 | 1.771 | 5.682 | 3.764 | 25.32 |
| δ = 0.25 | | | | | | |
| n = 100 | 1.644 | 3.483 | 9.848 | 2.744 | 2.613 | 10.82 |
| n = 200 | 3.244 | 2.619 | 7.732 | 2.313 | 2.074 | 5.593 |
| n = 500 | 3.963 | 2.532 | 2.867 | 2.524 | 1.812 | 2.015 |
| δ = 0.50 | | | | | | |
| n = 100 | 1.724 | 2.548 | 2.761 | 2.018 | 1.601 | 1.481 |
| n = 200 | 2.841 | 2.71 | 2.333 | 1.488 | 1.403 | 1.765 |
| n = 500 | 7.758 | 3.702 | 2.789 | 4.257 | 1.726 | 1.733 |
| δ = 1.0 | | | | | | |
| n = 100 | 1.200 | 1.649 | 1.423 | 0.712 | 1.244 | 1.204 |
| n = 200 | 1.434 | 2.168 | 2.066 | 1.530 | 1.509 | 1.513 |
| n = 500 | 12.10 | 3.981 | 3.237 | 5.066 | 1.920 | 1.772 |

**Table 2(b): Relative Efficiency of the GMM Estimates of the Slope Coefficient of Exogenous Covariate $\hat{\delta}_3$ (Endogeneity in both the threshold and the slope)**

| Quantiles | $MSE_{IVTR} / MSE_{THRETS}$ | | | $MSE_{NAIVE} / MSE_{THRETS}$ | | |
|---|---|---|---|---|---|---|
| | 0.05 | 0.50 | 0.95 | 0.05 | 0.50 | 0.95 |
| δ = 0.01 | | | | | | |
| n = 100 | 2.114 | 1.758 | 2.023 | 49.84 | 21.87 | 25.30 |
| n = 200 | 1.739 | 1.459 | 1.510 | 27.32 | 27.32 | 28.93 |
| n = 500 | 1.896 | 1.463 | 1.327 | 18.94 | 22.75 | 30.15 |
| δ = 0.05 | | | | | | |
| n = 100 | 2.555 | 1.782 | 2.258 | 25.98 | 20.81 | 26.86 |
| n = 200 | 1.676 | 1.477 | 1.759 | 18.88 | 23.78 | 29.95 |
| n = 500 | 2.187 | 1.594 | 1.529 | 28.21 | 17.91 | 22.35 |
| δ = 0.10 | | | | | | |
| n = 100 | 2.276 | 1.653 | 2.112 | 25.19 | 17.72 | 20.09 |
| n = 200 | 3.079 | 1.564 | 2.029 | 17.23 | 18.91 | 19.72 |
| n = 500 | 2.008 | 1.608 | 1.590 | 11.58 | 11.40 | 14.64 |
| δ = 0.25 | | | | | | |
| n = 100 | 2.077 | 1.603 | 1.958 | 12.53 | 12.16 | 14.25 |
| n = 200 | 1.927 | 1.712 | 1.734 | 14.48 | 12.57 | 11.42 |
| n = 500 | 2.400 | 1.607 | 1.770 | 9.173 | 8.474 | 10.49 |
| δ = 0.50 | | | | | | |
| n = 100 | 2.294 | 1.508 | 1.988 | 17.518 | 10.52 | 11.74 |
| n = 200 | 0.794 | 2.093 | 1.811 | 9.817 | 11.56 | 11.71 |
| n = 500 | 3.031 | 2.743 | 2.405 | 10.20 | 7.856 | 10.05 |
| δ = 1.0 | | | | | | |
| n = 100 | 2.649 | 2.857 | 2.585 | 20.35 | 17.57 | 16.11 |
| n = 200 | 4.032 | 4.132 | 3.338 | 7.165 | 12.57 | 15.39 |
| n = 500 | 19.07 | 8.875 | 4.877 | 9.201 | 11.09 | 12.86 |

**Table 3: Relative Efficiency of the GMM Estimates of the Slope Coefficient of Endogenous Covariate $\hat{\delta}_2$**

| Quantiles | $MSE_{IVTR} / MSE_{THRETS}$ | | | $MSE_{NAIVE} / MSE_{THRETS}$ | | |
|---|---|---|---|---|---|---|
| | 0.05 | 0.50 | 0.95 | 0.05 | 0.50 | 0.95 |
| δ = 0.01 | | | | | | |
| n = 100 | 440.8 | 337.9 | 179.4 | 4924 | 1671.2 | 3784 |
| n = 200 | 3508 | 703.2 | 269.2 | 10816 | 2004 | 4654 |
| n = 500 | 43847 | 1965 | 443.6 | 17849 | 2123 | 5150 |
| δ = 0.05 | | | | | | |
| n = 100 | 326.1 | 312.8 | 175.8 | 4559 | 1567 | 3027 |
| n = 200 | 1009 | 612.0 | 270.6 | 6998 | 1710 | 3629 |
| n = 500 | 2935 | 1845 | 450.9 | 4844 | 1288 | 2365 |
| δ = 0.10 | | | | | | |
| n = 100 | 296.3 | 236.1 | 171.8 | 3430 | 1327 | 2369 |
| n = 200 | 315.9 | 296.8 | 266.4 | 4913 | 1112 | 2402 |
| n = 500 | 305.2 | 397.7 | 435.8 | 1267 | 650.3 | 623.5 |
| δ = 0.25 | | | | | | |
| n = 100 | 157.4 | 114.7 | 126.7 | 517.3 | 650.3 | 710.7 |
| n = 200 | 126.6 | 126.1 | 129.0 | 362.1 | 402.0 | 414.8 |
| n = 500 | 124.8 | 133.2 | 104.8 | 265.0 | 221.8 | 305.2 |
| δ = 0.50 | | | | | | |
| n = 100 | 268.1 | 151.9 | 99.06 | 402.9 | 389.4 | 484.2 |
| n = 200 | 306.8 | 235.9 | 132.0 | 201.1 | 265.6 | 446.0 |
| n = 500 | 1441 | 495.5 | 208.9 | 450.3 | 324.7 | 399.3 |
| δ = 1.0 | | | | | | |
| n = 100 | 755.1 | 473.1 | 215.6 | 258.8 | 569.1 | 809.8 |
| n = 200 | 3581 | 1039 | 362.1 | 669.3 | 616.6 | 780.2 |
| n = 500 | 101211 | 2608 | 657.4 | 14470 | 1863 | 833.2 |

**Table 4[♦]: Regressions of log GDP per capita in 1995**

| | Linear Regression (2SLS) | THRETS (GMM) | | Linear Regression (2SLS) | THRETS (GMM) | | Linear Regression (2SLS) | THRETS (GMM) | |
|---|---|---|---|---|---|---|---|---|---|
| | | *Avg. Expr.Risk* ≤ 0.515 | *Avg. Expr.Risk* > 0.515 | | *Avg. Expr.Risk* ≤ 0.515 | *Avg. Expr.Risk* > 0.515 | | *Avg. Expr.Risk* ≤ 0.547 | *Avg. Expr.Risk* > 0.547 |
| | | 90% CI = [0.483, 0.769] | | | 90% CI = [0.483, 0.652] | | | 90% CI = [0.500, 0.720] | |
| **Dependent Variable: log GDP per capita (PPP basis) in 1995** | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| **Average Expropriation Risk 1985-95** | 7.697*** (2.197) | 5.516 (6.311) | 6.204*** (1.553) | 7.656*** (2.103) | 7.938 (7.113) | 7.558*** (2.232) | 6.684** (2.841) | -6.998 (14.909) | 8.570*** (2.838) |
| **MALFAL94P** | -1.277*** (0.365) | -0.726 (0.604) | -1.604*** (0.410) | -0.878* (0.460) | -0.943 (0.791) | -1.056** (0.450) | -0.876** (0.422) | 0.204 (1.133) | -0.759 (0.485) |
| **Ethnic Diversity** | - | - | - | -0.872* (0.448) | 0.176 (0.512) | -0.965** (0.443) | -0.831* (0.428) | -0.198 (0.383) | -1.156** (0.529) |
| **Openness** | - | - | - | - | - | - | 0.504 (0.619) | 2.158** (0.991) | -0.401 (0.966) |
| **No. of observations** | 60 | 14 | 46 | 60 | 14 | 46 | 60 | 17 | 43 |

[♦] All the regressions include a constant. Robust standard errors are in parentheses. "***" denotes significance at 1%, "**" at 5%, and "*" at 10%. The Average Expropriation Risk variable defers from Acemoglu, Johnson, and Robinson (2001) only in that it has been rescaled to take values from 0 to 1, with a higher score indicating higher less risk of expropriation. The lowest score for expropriation risk was 0.355 (Haiti) and the highest 1 (United States). We follow Acemoglu, Johnson, and Robinson (2001) and instrument for average expropriation risk using log settler mortality.