Identification and Estimation of Categorical Random Coefficient Models

Zhan Gao [1] M. Hashem Pesaran [1,2]

[1]University of Southern California [2]Trinity College, University of Cambridge;

Categorical Random Coefficient Model

Suppose the single cross-section observations $\{y_i, x_i, \boldsymbol{z}_i\}_{i=1}^n$, follow the model

$$y_i = x_i\beta_i + \boldsymbol{z}_i'\boldsymbol{\gamma} + u_i$$

where $\beta_i \in \{b_1, b_2, \cdots, b_K\}$ follows the categorical distribution,

$$\beta_i = b_k, \text{ w.p. } \pi_k,$$

with $\pi_k \in (0,1)$, $\sum_{k=1}^K \pi_k = 1$, $b_1 < b_2 < \cdots < b_K$.

- $\gamma \in \mathbb{R}^p$ is homogeneous.
- Assume $\beta_i \perp \boldsymbol{w}_i = (x_i, \boldsymbol{z}_i')'$. The idiosyncratic shock $u_i \sim \left(0, \sigma_i^2\right) \perp \boldsymbol{w}_i$. $\beta_i \perp u_i$.
- K is assumed to be **known**.
- Allow for distributions of $\boldsymbol{w}_i$ and u_i are not identical across i while independence across i is maintained.

Goal: Identify and estimate γ and the distributional parameters of β_i,
$\boldsymbol{\theta} = \left(\boldsymbol{\pi}' = (\pi_1, \pi_2, ..., \pi_K), \boldsymbol{b}' = (b_1, b_2, ..., b_K)\right)$.

Identification

Identifying the moments of β_i

Recursively solve the linear systems for $r = 2, 3, \cdots, 2K-1$,

$$\rho_{0,r}\mathrm{E}\left(\beta_i^r\right) + \sigma_r = \rho_{r,0} - \sum_{q=2}^{r-1}\binom{r}{q}\rho_{0,r-q}\sigma_q\mathrm{E}\left(\beta_i^{r-q}\right),$$

$$\rho_{0,2r}\mathrm{E}\left(\beta_i^r\right) + \rho_{0,r}\sigma_r = \rho_{r,r} - \sum_{q=2}^{r-1}\binom{r}{q}\rho_{0,2r-q}\sigma_q\mathrm{E}\left(\beta_i^{r-q}\right),$$

with $\left|n^{-1}\sum_{i=1}^n \mathrm{E}\left(\tilde{y}_i^r x_i^s\right) - \rho_{r,s}\right| = O\left(n^{-1/2}\right)$, $\left|n^{-1}\sum_{i=1}^n \mathrm{E}\left(u_i^r\right) - \sigma_r\right| = O\left(n^{-1/2}\right)$.

Identifying the distribution of β_i

Show the system

$$\mathrm{E}\left(\beta_i^r\right) = \sum_{k=1}^K \pi_k b_k^r,\ \ r = 0, 1, 2, \cdots, 2K-1,$$

has a unique solution $\left(\boldsymbol{\pi}', \boldsymbol{b}'\right)$ based on a linear recurrence structure, under the conditions

$$b_1 < b_2 < \cdots < b_K, \text{ and } \pi_k \in (0,1).$$

Estimation

Two-step estimation procedure

1. $\sqrt{n}$-consistent estimator for $\boldsymbol{\gamma}$, then replace $\boldsymbol{\gamma}$ by $\hat{\boldsymbol{\gamma}}$ in estimation of the distribution of β_i.
   - Let $\boldsymbol{w}_i = (x_i, \boldsymbol{z}_i')'$, $\boldsymbol{\phi} = \left(\mathrm{E}\left(\beta_i\right), \boldsymbol{\gamma}'\right)'$, the least square estimator $\hat{\boldsymbol{\phi}} = \left(n^{-1}\sum_{i=1}^n \boldsymbol{w}_i\boldsymbol{w}_i'\right)^{-1}\left(n^{-1}\sum_{i=1}^n \boldsymbol{w}_i y_i\right)$
2. Generalized method of Moments (GMM) estimator for $\boldsymbol{\theta} = \left(\boldsymbol{\pi}', \boldsymbol{b}'\right)$.
   - The moment conditions,
   $$\mathrm{E}\left(\tilde{y}_i^r x_i^{s_r}\right) = \sum_{q=0}^r \binom{r}{q}\mathrm{E}\left(x_i^{r-q+s_r}\right)\mathrm{E}\left(u_i^q\right)m_{r-q},$$
   $s_r = 0, 1, \cdots, S-r$, where S is a user-specific tuning parameter, chosen such that the highest order moments of x_i included is at most S, where $S > 2K-1$.
   - Sample version, $\hat{g}_n^{(r,s_r)}\left(\boldsymbol{\theta}, \boldsymbol{\sigma}, \hat{\boldsymbol{\gamma}}\right) = \frac{1}{n}\sum_{i=1}^n\left[\sum_{q=0}^r\binom{r}{q}x_i^{r-q+s_r}\sigma_q\left[h\left(\boldsymbol{\theta}\right)\right]_{r-q} - \hat{\tilde{y}}_i^r x_i^{s_r}\right]$, where $\hat{\tilde{y}}_i = y_i - \boldsymbol{z}_i'\hat{\boldsymbol{\gamma}}$.
   - $\left(\hat{\boldsymbol{\theta}}', \hat{\boldsymbol{\sigma}}'\right)' = \arg\min_{\boldsymbol{\theta}\in\Theta, \boldsymbol{\sigma}\in\mathcal{S}} \hat{\boldsymbol{g}}_n\left(\boldsymbol{\theta}, \boldsymbol{\sigma}, \hat{\boldsymbol{\gamma}}\right)' \boldsymbol{A}_n \hat{\boldsymbol{g}}_n\left(\boldsymbol{\theta}, \boldsymbol{\sigma}, \hat{\boldsymbol{\gamma}}\right)$.

Theorem 1: Consistency and Asymptotic Normality of $\hat{\phi}$ and $\hat{\theta}$

Under regularity conditions,

1. *$\hat{\phi}$ is a consistent estimator for ϕ, and $\sqrt{n}\left(\hat{\phi} - \phi\right) \to_d N\left(\mathbf{0}, \mathbf{Q}_{ww}^{-1}\mathbf{V}_{w\xi}\mathbf{Q}_{ww}^{-1}\right)$.*
2. *Let $\eta = (\theta', \sigma')'$ and $\eta_0 = (\theta_0', \sigma_0')'$, as $n \to \infty$, $\hat{\eta} \to_p \eta_0$.*
3. *Let $\sqrt{n}\left(\hat{\gamma} - \gamma\right) \to_d \zeta_\gamma \sim N(0, V_\gamma)$, as $n \to \infty$,*
$$\sqrt{n}\left(\hat{\eta} - \eta_0\right) \to_d \left(\mathbf{G}_0'\mathbf{A}\mathbf{G}_0\right)^{-1}\mathbf{G}_0'\mathbf{A}\left(\zeta + \mathbf{G}_{0,\gamma}\zeta_\gamma\right).$$

Multiple Regressors with Random Coefficients

The model

$$y_i = \mathbf{x}_i'\beta_i + \mathbf{z}_i'\gamma + u_i,$$

where the $p \times 1$ vector of random coefficients, $\beta_i \in \mathbb{R}^p$ follows the multivariate distribution

$$\Pr\left(\beta_{i1} = b_{1k_1}, \beta_{i2} = b_{2k_2}, \cdots, \beta_{ip} = b_{pk_p}\right) = \pi_{k_1,k_2,\cdots,k_p},$$

with $k_j \in \{1, 2, \cdots, K\}$, $b_{j1} < b_{j2} < \cdots < b_{jK}$, and $\sum_{k_1,k_2,\cdots,k_p\in\{1,2,\cdots,K\}}\pi_{k_1,k_2,\cdots,k_p} = 1$.

- Identification and estimation of the marginal distributions of $\boldsymbol{\beta}_i$ follow as corollaries.
- The joint distribution can be identified in special cases, $p = 2$ and $K = 2$ for example.

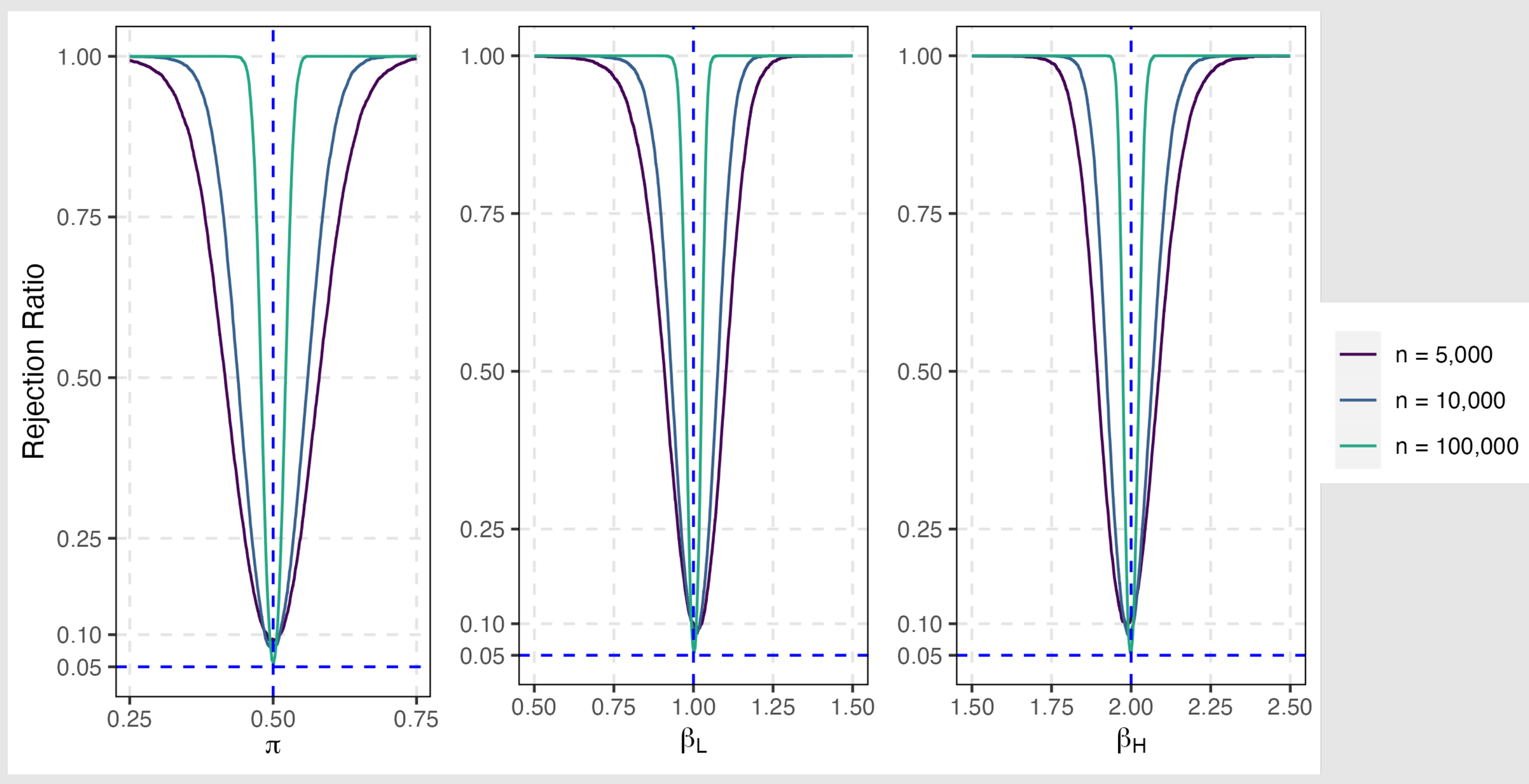
Next Steps

- Determine K. The model is useful when K is not very large.
- Extension to panel and spatial setups.

Monte Carlo Experiments

Data generating process: $y_i = \alpha + x_i\beta_i + z_{i1}\gamma_1 + z_{i2}\gamma_2 + u_i$, for $i = 1, 2, ..., n$. generate $x_i = (\tilde{x}_{1i} - 2)/2$ where $\tilde{x}_{1i} \sim \text{IID}\chi^2(2)$ for $i = 1, 2, \cdots, \lfloor n/2 \rfloor$, and $x_i = (\tilde{x}_{2i} - 2)/4$ where $\tilde{x}_{2i} \sim \text{IID}\chi^2(4)$, for $i = \lfloor n/2 \rfloor + 1, \cdots, n$. The additional regressors, z_{ij}, for $j = 1, 2$ with homogeneous slopes are generated as $z_{i1} = x_i + v_{i1}$ and $z_{i2} = z_{i1} + v_{i2}$, with $v_{ij} \sim \text{IID } N(0,1)$, for $j = 1, 2$. The error term u_i is generated as $u_i = \sigma_i\varepsilon_i$, where σ_i^2 are generated as $0.5(1 + \text{IID}\chi^2(1))$, and $\varepsilon_i \sim \text{IID}N(0,1)$.

	$\pi = 0.5$			$\beta_L = 1$			$\beta_H = 2$		
n	Bias	RMSE	Size	Bias	RMSE	Size	Bias	RMSE	Size
500	-0.0234	0.2384	0.3678	-0.0882	0.6297	0.3217	-0.0216	0.5816	0.2186
1,000	-0.0185	0.1769	0.2981	-0.0362	0.4285	0.2767	-0.0198	0.3216	0.2083
2,000	-0.0069	0.1233	0.2376	-0.0151	0.2274	0.2370	-0.0123	0.1574	0.1828
5,000	-0.0029	0.0677	0.1586	-0.0020	0.0988	0.1504	-0.0060	0.0735	0.1434
10,000	-0.0010	0.0414	0.1112	0.0008	0.0535	0.1060	-0.0032	0.0463	0.1050
100,000	0.0001	0.0114	0.0610	0.0006	0.0135	0.0666	-0.0003	0.0135	0.0620

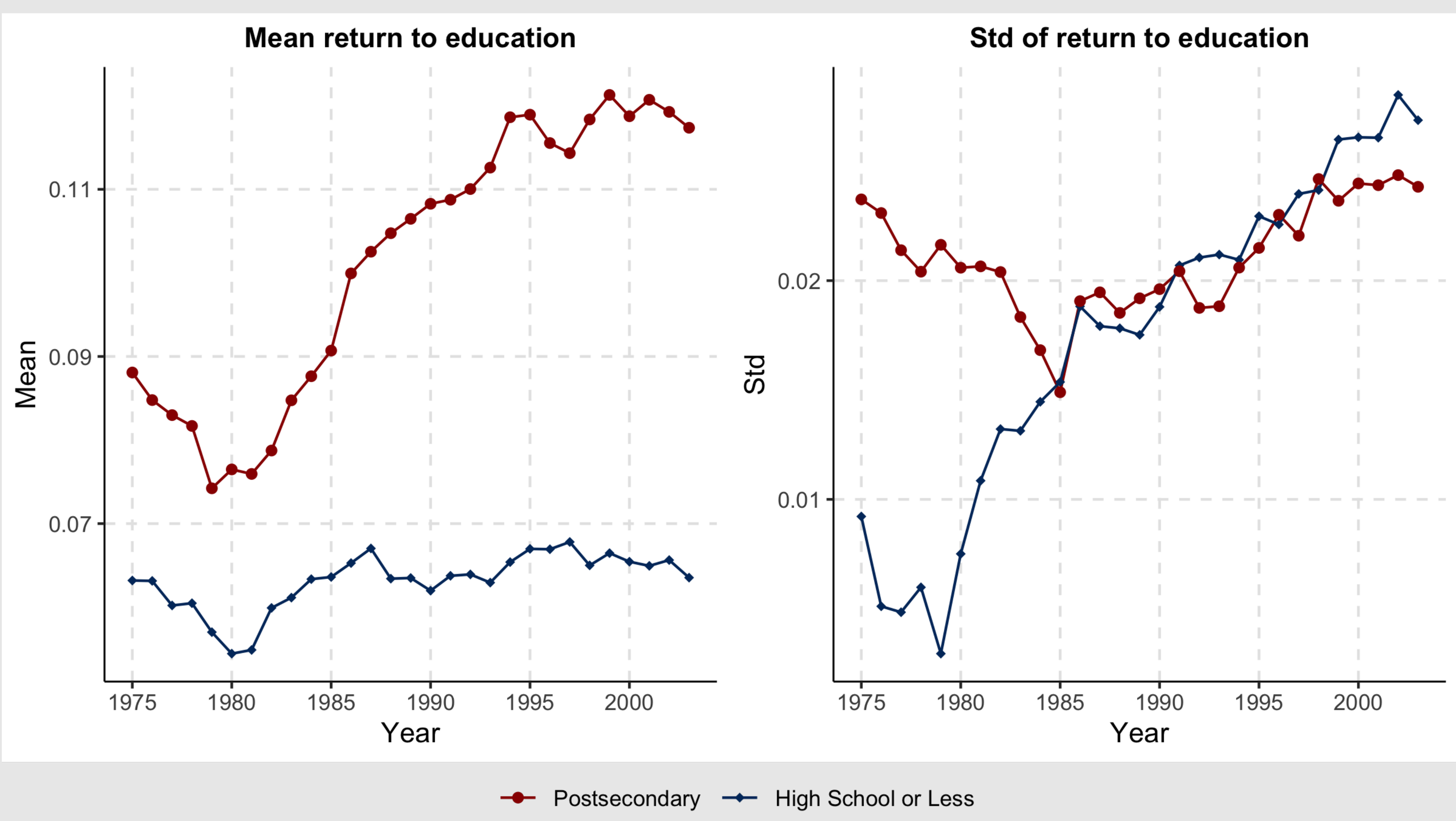


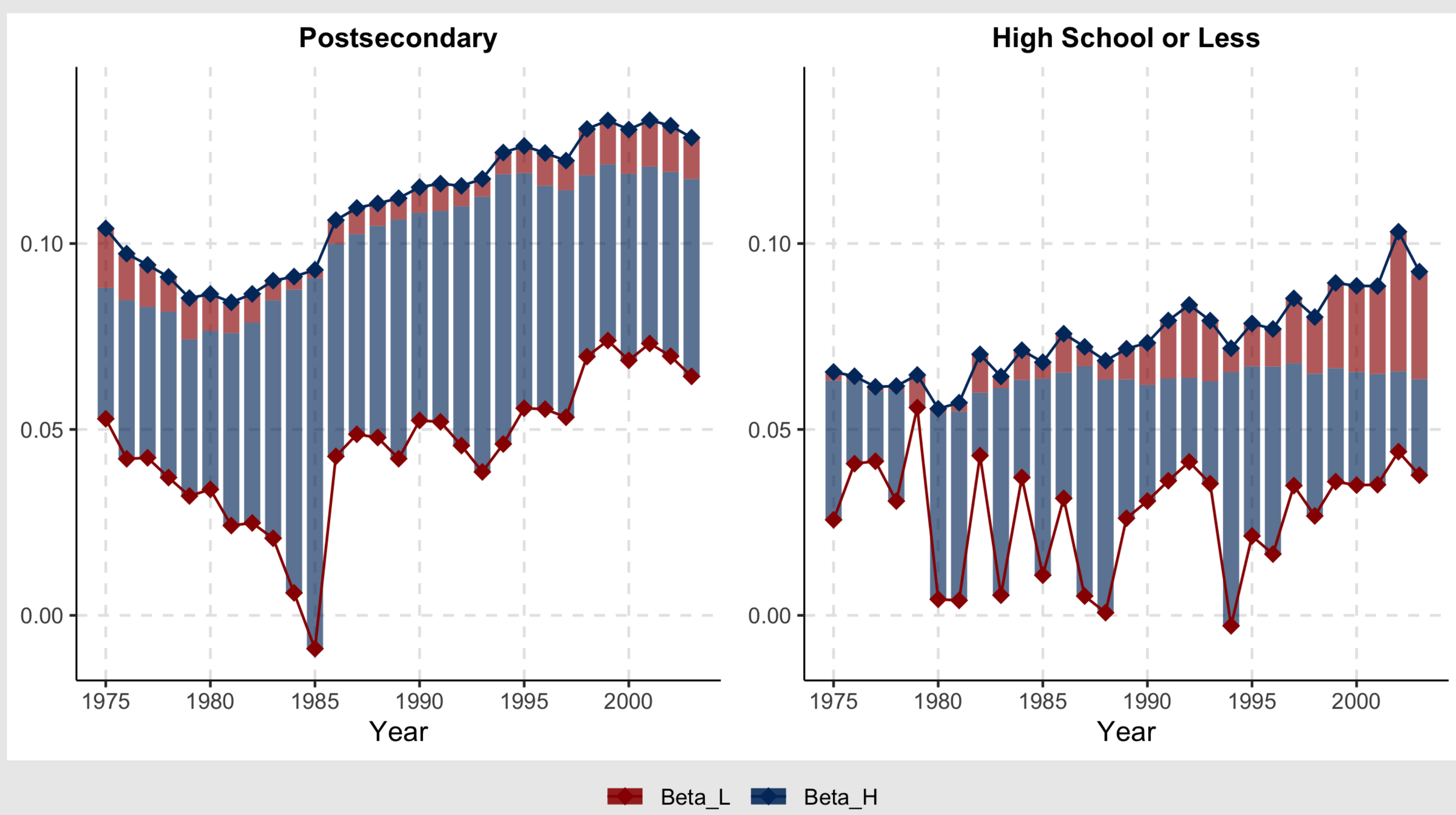
Heterogeneous Return to Education: An Empirical Application

- Estimate the Mincerian equation with repeated cross-sectional data by education groups,

$$\log \text{wage}_{it} = \alpha_t + \beta_{it}\text{edu}_{it} + \rho_{1t}\text{exper}_{it} + \rho_{2t}\text{exper}_{it}^2 + \tilde{\boldsymbol{z}}_{it}'\tilde{\boldsymbol{\gamma}}_t + u_{it}, \text{ where } \beta_{it} = \begin{cases} b_{tL} & \text{w.p. } \pi_t, \\ b_{tH} & \text{w.p. } 1 - \pi_t. \end{cases}$$

- **Data**: The May and Outgoing Rotation Group (ORG) supplements of the Current Population Survey (CPS) data, 1973 - 2003.
- Between group heterogeneity ↑ due to the postsecondary education group; Within group heterogeneity ↑ in high school or less education group.





Correspondence: zhangao@usc.edu