The Empirical Content of the Roy Model

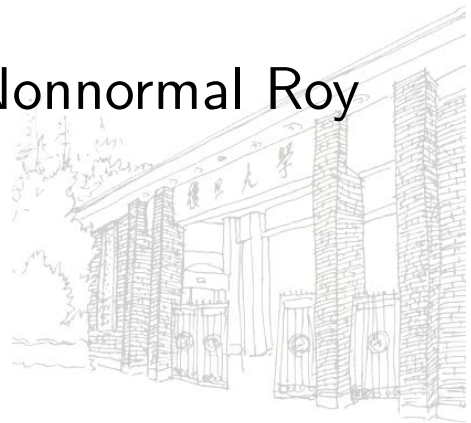
James J. Heckman, Bo E. Honore

Presenter: Qinzhu Sun

December 15, 2020

Overview

1. Setting: Two-skill Roy Model
2. The Roy Model
   - Log Concavity and the Roy Model
   - Consequences of Log Normality
   - Nonrobustness of the Roy Model to Non-Log Concavity
3. Identifiability of the Roy Model and Its Normal Extensions
   - Identifiability of the Log Normal Roy Model from Cross-Section Data
   - Nonidentifiability of a General Nonnormal Roy Model in a Single Cross Section
   - Identification from Multi-market Data in a General Nonnormal Roy Model
4. Appendix



Setting: Two-skill Roy Model

- Income maximizing agents possess two skills S_1 and S_2 with associated positive skill prices π_1 and π_2.

Setting: Two-skill Roy Model

- Income maximizing agents possess two skills S_1 and S_2 with associated positive skill prices π_1 and π_2.
- Population skill distribution: $F(s_1, s_2)$.

Setting: Two-skill Roy Model

- Income maximizing agents possess two skills S_1 and S_2 with associated positive skill prices π_1 and π_2.
- Population skill distribution: $F(s_1, s_2)$.
- Skill i is useful only in sector i.

Setting: Two-skill Roy Model

- Income maximizing agents possess two skills S_1 and S_2 with associated positive skill prices π_1 and π_2.
- Population skill distribution: $F(s_1, s_2)$.
- Skill i is useful only in sector i.
- An agent chooses sector one if $\pi_1 S_1 > \pi_2 S_2$.

Setting: Two-skill Roy Model

- Income maximizing agents possess two skills S_1 and S_2 with associated positive skill prices π_1 and π_2.
- Population skill distribution: $F(s_1, s_2)$.
- Skill i is useful only in sector i.
- An agent chooses sector one if $\pi_1 S_1 > \pi_2 S_2$.
- Assume that (S_1, S_2) has a well-defined density $f(s_1, s_2)$.

Setting: Two-skill Roy Model

- Income maximizing agents possess two skills S_1 and S_2 with associated positive skill prices π_1 and π_2.
- Population skill distribution: $F(s_1, s_2)$.
- Skill i is useful only in sector i.
- An agent chooses sector one if $\pi_1 S_1 > \pi_2 S_2$.
- Assume that (S_1, S_2) has a well-defined density $f(s_1, s_2)$.
- The proportion of the population working in sector one is

$$P_1 = \int_0^{\infty} \int_0^{\pi_1 S_1/\pi_2} f(s_1, s_2)\, ds_2\, ds_1.$$

Setting: Two-skill Roy Model

- The population density of S_1 is

$$f(s_1) = \int_0^\infty f(s_1, s_2)\, ds_2$$

Setting: Two-skill Roy Model

- The population density of S_1 is

$$f(s_1) = \int_0^{\infty} f(s_1, s_2)\, ds_2$$

- The density of skill employed in sector one is

$$g(s_1 | \pi_1 S_1 > \pi_2 S_2) = \frac{1}{P_1} \int_0^{\pi_1 S_1 / \pi_2} f(s_1, s_2)\, ds_2\, ds_1$$

Setting: Two-skill Roy Model

- The population density of S_1 is

$$f(s_1) = \int_0^{\infty} f(s_1, s_2)\, ds_2$$

- The density of skill employed in sector one is

$$g(s_1 | \pi_1 S_1 > \pi_2 S_2) = \frac{1}{P_1} \int_0^{\pi_1 S_1 / \pi_2} f(s_1, s_2)\, ds_2\, ds_1$$

$\Rightarrow$ The distribution of skills observed in sector one differs from the population distribution of skills.

Setting: Two-skill Roy Model

- The density of earnings in the economy at large, $g(w)$, is a weighted average of the densities in each sector

$$g(w) = P_1 g_1(w) + P_2 g_2(w)$$

where the weight applied to the sector i density is the proportion of the population in the sector.

Setting: Two-skill Roy Model

- Define $U_i = lnS_i - \mu_i \quad \Rightarrow \quad lnW_i = ln\pi_i + \mu_i + U_i$
- Define

$$c = ln(\pi_1/\pi_2) + \mu_1 - \mu_2, \quad \sigma^2 = \sigma_{11} + \sigma_{22} - 2\sigma_{12},$$

$$a_1 = \frac{\sigma_{11} - \sigma_{22}}{\sigma^2}, \quad a_2 = a_1 - 1 = \frac{-\sigma_{22} + \sigma_{12}}{\sigma^2},$$

$$D = U_1 - U_2, \quad V = a_1 U_2 - a_2 U_1,$$

$$c_* = c/\sigma, \quad D_* = D/\sigma,$$

$$\rho = corr[D, U_1] = \frac{\sigma_{11} - \sigma_{12}}{\sigma\sqrt{\sigma}_{11}} = a_1\sigma/\sqrt{\sigma}_{11}$$

Setting: Two-skill Roy Model

Then

$$U_i = a_i D + V,$$

$$E[D] = 0, \quad E[V] = 0,$$

$$Var[D] = \sigma^2, \quad Var[V] = \frac{\sigma_{11}\sigma_{22} - \sigma_{12}^2}{\sigma^2} = \sigma_{11}(1 - \rho^2),$$

$$Cov[D, V] = 0.$$

Setting: Two-skill Roy Model

Rewrite the earnings function to be

$$\begin{aligned} ln W_i &= ln\pi_i + \mu_i + a_i D + V \\ &= ln\pi_i + \mu_i + a_i \sigma D_* + V, \end{aligned}$$

Setting: Two-skill Roy Model

Rewrite the earnings function to be

$$\begin{aligned} ln W_i &= ln\pi_i + \mu_i + a_i D + V \\ &= ln\pi_i + \mu_i + a_i \sigma D_* + V, \end{aligned}$$

$$\begin{aligned} \Rightarrow ln W_1 - ln W_2 &= ln\pi_1 + \mu_1 + a_1 D + V - ln\pi_2 - \mu_2 - a_2 D - V \\ &= c + D = \sigma(c_* + D_*) \end{aligned}$$

Setting: Two-skill Roy Model

Derive the sectoral moments of log earnings:

$$E[lnW_1|lnW_1 > lnW_2] = ln\pi_1 + \mu_1 + E[U_1|D > -c]$$
$$= ln\pi_1 + \mu_1 + a_1 E[D|D > -c] + E[V|D > -c]$$

and

$$Var[lnW_1|lnW_1 > lnW_2] = Var[U_1|D > -c]$$
$$= a_1^2 Var[D|D > -c] + Var[V|D > -c]$$
$$+ 2a_1 Cov[D, V|D > -c].$$

Setting: Two-skill Roy Model

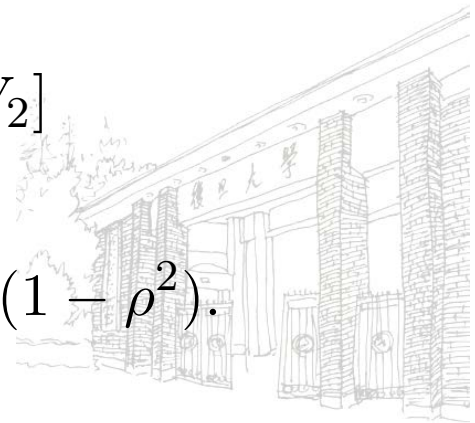
If D and V are *independent*, then $E[V|D > -c] = 0$, $Var[V|D > -c] = Var[V]$ and $Cov[D, V|D > -c] = 0$, so

$$\begin{aligned} &E[ln W_1 | ln W_1 > ln W_2] \\ &= ln\pi_1 + \mu_1 + a_1 E[D|D > -c] \\ &= ln\pi_1 + \mu_1 + a_1 \sigma E[D_*|D_* > -c_*], \end{aligned}$$

$$\begin{aligned} Var[ln W_1 | ln W_1 > ln W_2] =& a_1^2 Var[D|D > -c] + Var[V] \\ =& \sigma_{11}(\rho^2 Var[D_*|D_* > -c_*] + (1 - \rho^2)), \end{aligned}$$

and

$$\begin{aligned} &E[(ln W_1 - E[ln W_1 | ln W_1 > ln W_2])^3 | ln W_1 > ln W_2] \\ &= a_1^3 E[(D - E[D|D > -c])^3 | D > -c] + E[V^3] \\ &= a_1^3 \sigma^3 E[(D_* - E[D_*|D_* > -c_*])^3 | D_* > -c_*] + (1 - \rho^2). \end{aligned}$$



Setting: Two-skill Roy Model

Likewise, the moments of log skills are:

$$E[lnS_1|lnW_1 > lnW_2] = \mu_1 + a_1\sigma E[D_*|D_* > -c_*],$$

$$Var[lnS_1|lnW_1 > lnW_2] = \sigma_{11}(\rho^2 Var[D_*|D_* > -c_*] + (1-\rho^2)),$$

$$\begin{aligned} &E[(lnS_1 - E[lnS_1|lnW_1 > lnW_2])^3|lnW_1 > lnW_2] \\ &= a_1^3\sigma^3 E[(D_* - E[D_*|D_* > -c_*])^3|D_* > -c_*] + E[V^3]. \end{aligned}$$

Log Concavity and the Roy Model

In this subsection we investigate conditions on D that will allow us to characterize conditional moments. One assumption that will allow us to characterize the truncated distribution of D is that D is *log concave*.

Log Concavity and the Roy Model

In this subsection we investigate conditions on D that will allow us to characterize conditional moments. One assumption that will allow us to characterize the truncated distribution of D is that D is *log concave*.

Definition 1: A *log concave random variable* X is one for which the density f satisties the condition that

$$f(\lambda x_1 + (1-\lambda)x_2) \geq [f(x_1)]^{\lambda}[f(x_2)]^{1-\lambda},$$

$0 \leq \lambda \leq 1$, for x_1,x_2 in the support of X.

Log Concavity and the Roy Model

Proposition 1: If D is a *log concave random variable*, then

$$0 \leq \frac{\partial E[D|D > d]}{\partial d} \leq 1$$

$$0 \leq \frac{\partial E[D|D \leq d]}{\partial d} \leq 1$$

and

$$\frac{\partial Var[D|D > d]}{\partial d} \leq 0$$

$$\frac{\partial Var[D|D \leq d]}{\partial d} \geq 0.$$

Corollary 1: If D is *log concave*, then $Var[D|D \leq d] \leq \sigma^2$.

Log Concavity and the Roy Model

The effect of an increase of π_i on the mean log skill and earnings in sector one:

$$\frac{\partial E[ln W_1 | ln W_1 > ln W_2]}{\partial ln \pi_i} = \begin{cases} 1 - a_1 \frac{\partial E[D|D > d]}{\partial d} \mid_{d=-c} & if \quad i = 1, \\ a_1 \frac{\partial E[D|D > d]}{\partial d} \mid_{d=-c} & if \quad i = 2, \end{cases}$$

$$\frac{\partial E[ln S_1 | ln W_1 > ln W_2]}{\partial ln \pi_i} = \begin{cases} -a_1 \frac{\partial E[D|D > d]}{\partial d} \mid_{d=-c} & if \quad i = 1, \\ a_1 \frac{\partial E[D|D > d]}{\partial d} \mid_{d=-c} & if \quad i = 2. \end{cases}$$

Proof: Recall that

$$c = ln(\pi_1/\pi_2) + \mu_1 - \mu_2,$$

$$E[ln W_1 | ln W_1 > ln W_2] = ln \pi_1 + \mu_1 + a_1 E[D|D > -c],$$

$$E[ln S_1 | ln W_1 > ln W_2] = \mu_1 + a_1 E[D|D > -c].$$

$Q.E.D.$

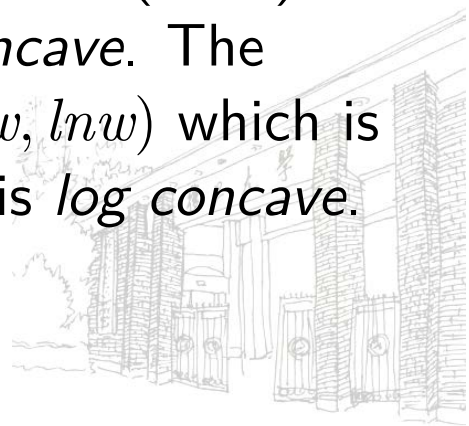
Log Concavity and the Roy Model

Theorem 1: If lnS_1, lnS_2 are joint *log concave* random variables with *log concave* densities, the aggregate log income distribution

$$G(lnw) = P_1 G_1(lnw) + P_2 G_2(lnw)$$

is log concave.

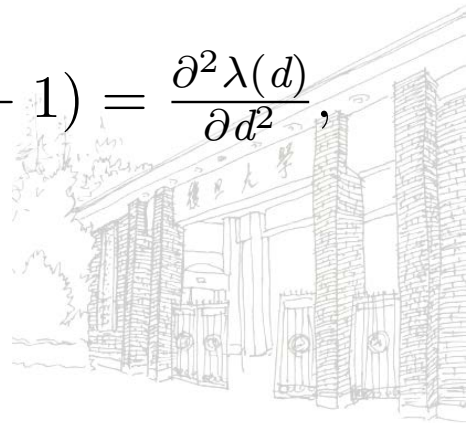
Proof: If (lnS_1, lnS_2) is *log concave* with density $f(lns_1, lns_2)$, then so is $ln W_1, ln W_2$ because translations of *log concave* random variables are *log concave* (Prekopa (1973, Theorem 7)). By the Brascamp-Lieb (1975) theorem, the distribution function $F(lnw_1, lnw_2)$ is *log concave*. The observed wage is $ln W = max(ln W_1, ln W_2)$ with cdf $F(lnw, lnw)$ which is obviously *log concave* if the distribution of $(ln W_1, ln W_2)$ is *log concave*. $Q.E.D.$



Consequences of Log Normality

Let Z be a standard *normal* random variable and let $\lambda(d) = E[Z|Z > d]$; then for $d \in (-\infty, \infty)$, we prove the following results:

- $\lambda(d) = \frac{\frac{1}{\sqrt{2\pi}} exp\{-d^2/2\}}{\Phi(-d)} > max(0, d)$,
- $0 < \frac{\partial \lambda(d)}{\partial d} = \lambda'(d) = \lambda(d)(\lambda(d) - d) < 1$,
- $\frac{\partial^2 \lambda(d)}{\partial d^2} > 0$,
- $0 < Var[Z|Z > d] = 1 + \lambda(d)d - \lambda^2(d) < 1$,
- $\frac{\partial Var[Z|Z>d]}{\partial d} < 0$,
- $E[(Z - \lambda(d))^3|Z > d] = \lambda(d)(2\lambda^2(d) - 3d\lambda(d) + d^2 - 1) = \frac{\partial^2 \lambda(d)}{\partial d^2}$,
- $E[Z|Z > d] \geq mode[Z|Z > d]$.



Consequences of Log Normality

Furthermore,

$$lim_{d\to-\infty}\lambda(d) = 0, \quad lim_{d\to\infty}\lambda(d) = \infty,$$

$$lim_{d\to-\infty}\frac{\partial\lambda(d)}{\partial d} = 0, \quad lim_{d\to\infty}\frac{\partial\lambda(d)}{\partial d} = 1,$$

$$lim_{d\to-\infty} Var[Z|Z > d] = 1, \quad lim_{d\to\infty} Var[Z|Z > d] = 0.$$

Consequences of Log Normality

Theorem 2: For a *log normal* Roy economy, any random assignment of persons to sectors with the same proportion of persons in each sector as in the Roy economy has higher variance of log earnings provided the proportions lie strictly in the unit interval. This is true whether or not skill prices in the two economies are the same.

Proof:

记P_1为 Roy 经济中处于部门一的总体比例，那么$P_2 = 1 - P_1$即为处于部门二的总体比例。在 Roy 经济中的总体收入方差为

$$V = P_1 Var[lnW_1 | lnW_1 > lmW_2] + P_2 Var[lnW_2 | lnW_1 \leq lnW_2]$$
$$+P_1 P_2 (E[lnW_1 | lnW_1 > lnW_2] - E[lnW_2 | lnW_1 \leq lnW_2])^2.$$

假设在随机指定的经济中的技能工资分别为$\tilde{\pi}_1$和$\tilde{\pi}_2$，在 Roy 经济下的技能工资为π_1和π_2。则在随机指定的经济中对数收入方差为

$$\tilde{V} = P_1 \sigma_{11} + P_2 \sigma_{22} + P_1 P_2 (ln\tilde{\pi}_1 + \mu_1 - ln\tilde{\pi}_2 - \mu_2)^2.$$

我们需要证明的是$\tilde{V} > V$。定义$c_* = (ln\pi_1 + \mu_1 - ln\pi_2 - \mu_2)/\sigma$，$\tilde{c_*} = (ln\tilde{\pi}_1 + \mu_1 - ln\tilde{\pi}_2 - \mu_2)/\sigma$，$\rho_1 = a_1\sigma/\sqrt{\sigma_{11}}$，$\rho_2 = a_2\sigma/\sqrt{\sigma_{22}}$。

在 Roy 经济中，$P_1 = \Phi(c_*)$，$P_2 = \Phi(-c_*)$[4]。利用公式(10)、公式(11)、公式(R-1)、公式(R-4)，代入得：

$$V = \sigma_{11}\Phi(c_*)\left[1 + \rho_1^2\left(-c_*\lambda(-c_*) - \lambda^2(-c_*)\right)\right]$$
$$+\sigma_{22}\Phi(-c_*)\left[1 + \rho_2^2\left(c_*\lambda(c_*) - \lambda^2(c_*)\right)\right]$$
$$+\sigma^2\Phi(c_*)\Phi(-c_*)\left(c_* + \frac{\rho_1\sqrt{\sigma_{11}}}{\sigma}\lambda(-c_*) - \frac{\rho_2\sqrt{\sigma_{22}}}{\sigma}\lambda(c_*)\right)^2$$

且有

$$\tilde{V} = \Phi(c_*)\sigma_{11} + \Phi(-c_*)\sigma_{22} + \sigma^2\Phi(c_*)\Phi(-c_*)\tilde{c}_*^2.$$

Consequences of Log Normality

Proof:(Cont.)

我们先证明$\tilde{V} - \sigma^2\Phi(c_*)\Phi(-c_*)\tilde{c}_*^2 \geq V$，即证

(C-1)

$$\sigma_{11}\Phi(c_*)\left[\rho_1^2\left(-c_*\lambda(-c_*) - \lambda^2(c_*)\right)\right] + \sigma_{22}\Phi(-c_*)\left[\rho_2^2\left(c_*\lambda(c_*) - \lambda^2(c_*)\right)\right]$$

$$+\sigma^2\Phi(c_*)\Phi(-c_*)\left(c_* + \frac{\rho_1\sqrt{\sigma_{11}}}{\sigma}\lambda(-c_*) - \frac{\rho_2\sqrt{\sigma_{22}}}{\sigma}\lambda(c_*)\right)^2 \leq 0.$$

注意到$a_1 = \rho_1\sqrt{\sigma_{11}}/\sigma$且$1 - a_1 = \rho_2\sqrt{\sigma_{22}}/\sigma$，则(C-1)的 LHS 可转化为

$$\sigma^2\Big(a_1^2\Phi(c_*)\left(-c_*\lambda(-c_*) - \lambda^2(c_*)\right) + (1-a_1)^2\Phi(-c_*)\left(c_*\lambda(c_*) - \lambda^2(c_*)\right)$$

$$+ \Phi(c_*)\Phi(-c_*)\left[a_1\left(c_* + \lambda(-c_*)\right) + (1-a_1)\left(c_* - \lambda(c_*)\right)\right]^2\Big)$$

上式可写为

$$\sigma^2(a_1^2\eta_1 + (1-a_1)^2\eta_2 + 2a_1(1-a_1)\eta_3)$$

其中

$$\eta_1 = \Phi(c_*)\left(-c_*\lambda(-c_*) - \lambda^2(-c_*)\right) + \Phi(c_*)\Phi(-c_*)\left(c_* + \lambda(-c_*)\right)^2$$

$$= \Phi(c_*)\left(c_* + \lambda(-c_*)\right)\left(-\lambda(-c_*) + \Phi(-c_*)\left(c_* + \lambda(-c_*)\right)\right),$$

Consequences of Log Normality

Proof:(Cont.)

$$\eta_2 = \Phi(-c_*)\big(c_*\lambda(c_*) - \lambda^2(c_*)\big) + \Phi(-c_*)\Phi(c_*)\big(c_* - \lambda(c_*)\big)^2$$

$$= \Phi(-c_*)\big(c_* - \lambda(c_*)\big)\Big(\lambda(c_*) + \Phi(c_*)\big(c_* - \lambda(c_*)\big)\Big),$$

$$\eta_3 = \Phi(c_*)\Phi(-c_*)\big(c_* - \lambda(c_*)\big)\big(c_* + \lambda(-c_*)\big).$$

$$\eta_1 - \eta_3 = \Phi(c_*)\big(c_* + \lambda(c_*)\big)$$

$$\cdot\Big(-\lambda(-c_*) + \Phi(-c_*)\big(c_* + \lambda(-c_*)\big) - \Phi(-c_*)\big(c_* - \lambda(c_*)\big)\Big)$$

$$= \Phi(c_*)\big(c_* + \lambda(-c_*)\big)\Big(-\lambda(-c_*) + \Phi(-c_*)\big(\lambda(c_*) + \lambda(-c_*)\big)\Big)$$

$$= \Phi(c_*)\big(c_* + \lambda(-c_*)\big)\big(-\lambda(-c_*) + \lambda(-c_*)\big)$$

$$= 0$$

其中利用了$\Phi(-c_*)\big(\lambda(c_*) + \lambda(-c_*)\big) = \lambda(-c_*)$的结论。因此有$\eta_1 = \eta_3$。类似的，有

Consequences of Log Normality

Proof:(Cont.)

$$\begin{aligned}\eta_2 - \eta_3 &= \Phi(-c_*)\big(c_* - \lambda(c_*)\big) \\ &\quad \cdot \Big(\lambda(c_*) + \Phi(c_*)\big(c_* - \lambda(c_*)\big) - \Phi(c_*)\big(c_* + \lambda(-c_*)\big)\Big) \\ &= \Phi(-c_*)\big(c_* - \lambda(c_*)\big)\Big(\lambda(c_*) - \Phi(c_*)\big(\lambda(c_*) + \lambda(-c_*)\big)\Big) \\ &= \Phi(-c_*)\big(c_* - \lambda(c_*)\big)\big(\lambda(c_*) - \lambda(c_*)\big) \\ &= 0\end{aligned}$$

其中利用了$\Phi(c_*)\big(\lambda(-c_*) + \lambda(c_*)\big) = \lambda(c_*)$的结论。因此，$\eta_1 = \eta_2 = \eta_3$。

因此可将(C-1)的左侧改写为

$$\sigma^2\eta_3\big(a_1^2 + (1-a_1)^2 + 2a_1(1-a_1)\big) = \sigma^2\eta_3$$

由于$\lambda(c_*) - c_* > 0$，$\lambda(-c_*) + c_* > 0$，因此

$$\eta_3 = \Phi(c_*)\Phi(-c_*)\big(c_* - \lambda(c_*)\big)\big(c_* + \lambda(-c_*)\big) < 0.$$

Consequences of Log Normality

Proof:(Cont.)

因此可将(C-1)的左侧改写为

$$\sigma^2\eta_3\left(a_1^2+(1-a_1)^2+2a_1(1-a_1)\right)=\sigma^2\eta_3$$

由于$\lambda(c_*)-c_*>0$，$\lambda(-c_*)+c_*>0$，因此

$$\eta_3=\Phi(c_*)\Phi(-c_*)\left(c_*-\lambda(c_*)\right)\left(c_*+\lambda(-c_*)\right)<0.$$

因此(C-1)成立，即$\tilde{V}-\sigma^2\Phi(c_*)\Phi(-c_*)\tilde{c}_*^2\geq V$，故$\tilde{V}\geq V$成立，即随机指派的经济比 Roy 经济的对数收入方差（收入不平等）更大。

Q.E.D

Consequences of Log Normality

Theorem 3: In a *log normal* skill Roy economy, aggregate log earnings distributions are right skewed as long as some positive fraction of the population works in each sector.

Note that

- Sectoral distribution skewness:

$$E[(lnW_1 - E[lnW_1|lnW_1 > lnW_2])^3|lnW_1 > lnW_2]$$
$$= a_1^3 E[(D - E[D|D > -c])^3|D > -c] + E[V^3]$$
$$= a_1^3\sigma^3 E[(D_* - E[D_*|D_* > -c_*])^3|D_* > -c_*] + (1-\rho^2),$$

- Aggregate distribution skewness:

$$E[(\max\{lnW_1, lnW_2\} - E[\max\{lnW_1, lnW_2\}])^3].$$

Nonrobustness of the Roy Model to Non-Log Concavity

It is natural to ask whether the results obtained for *log concave* models can be generalized to all distributions of (U_1, U_2). As might be expected, the answer is "no". For *log convex* random variables, inequalities in Proposition 1 and the Corollary of the Proposition are reversed.

Nonrobustness of the Roy Model to Non-Log Concavity

It is natural to ask whether the results obtained for *log concave* models can be generalized to all distributions of (U_1, U_2). As might be expected, the answer is "no". For *log convex* random variables, inequalities in Proposition 1 and the Corollary of the Proposition are reversed.

Definition 2: A *log convex* random variable is one for which the density of f satisties the condition that

$$f(\lambda x_1 + (1-\lambda)x_2) \leq [f(x_1)]^{\lambda}[f(x_2)]^{1-\lambda},$$

$0 \leq \lambda \leq 1$, for x_1,x_2 in the support of $\mathbf{X}$.

Nonrobustness of the Roy Model to Non-Log Concavity

Proposition 2:
If D is a *log convex* random variable and $D \geq 0$, with support in $[0, \infty)$,

$$\frac{\partial E[D|D > d]}{\partial d} \geq 1$$

and

$$\frac{\partial Var[D|D > d]}{\partial d} \geq 0.$$

Nonrobustness of the Roy Model to Non-Log Concavity

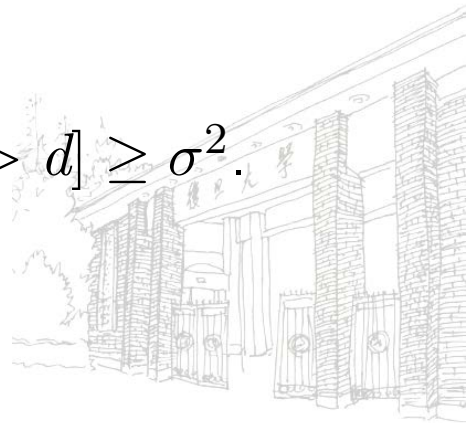
Proposition 2:
If D is a *log convex* random variable and $D \geq 0$, with support in $[0, \infty)$,

$$\frac{\partial E[D|D > d]}{\partial d} \geq 1$$

and

$$\frac{\partial Var[D|D > d]}{\partial d} \geq 0.$$

Corollary 2:
If D is *log convex* with support in $[0, \infty)$, then $Var[D|D > d] \geq \sigma^2$.



Comparision

- log concave: very good!

$$0 \leq \frac{\partial E[D|D > d]}{\partial d} \leq 1$$

$$0 \leq \frac{\partial E[D|D \leq d]}{\partial d} \leq 1$$

$$\frac{\partial Var[D|D > d]}{\partial d} \leq 0$$

$$\frac{\partial Var[D|D \leq d]}{\partial d} \geq 0,$$

- log convex: nonrobust :(

$$\frac{\partial E[D|D > d]}{\partial d} \geq 1$$

$$\frac{\partial Var[D|D > d]}{\partial d} \geq 0.$$

Identifiability of the Roy Model and Its Normal Extensions

Three steps:

- Identification of *log normal Roy model* from cross-section data;
- Nonidentifiability of a general *nonnormal Roy model* in a single cross section;
- Identification from multi-market data in a general *nonnormal Roy model*
  - Pooled cross-section
  - Panel data

Identifiability of the Log Normal Roy Model from Cross-Section Data

We establish that

- [Theorem 4] the log normal skills Roy model is identified using a single cross-section of data on earnings and sectoral choices of agents. Thus it is possible to identify μ and $\sum$ of Section 1;
- [Theorem 5] it is possible to identify μ and $\sum$ except for their subscripts, from the knowledge of aggregate earnings distribution;
- [Theorem 6] it is possible to recover the distribution of skills when only the proportion working in each sector and the sectoral earnings distribution in one sector are observed, as occurs in the housewife case where nonmarket output is not observed.

Identifiability of the Log Normal Roy Model from Cross-Section Data

Theorem 4: Under the conditions postulated for the *log normal* Roy model, μ and $\sum$ can be identified from data on wages paid in each sector and sectoral choices.

Identifiability of the Log Normal Roy Model from Cross-Section Data

Theorem 4: Under the conditions postulated for the *log normal* Roy model, μ and $\sum$ can be identified from data on wages paid in each sector and sectoral choices.

Proof:

注：笔者仿照参考文献 Basu and Ghosh(1978)给出证明，参考文献基于$Z = \min(X_1, X_2)$的体系建模，而本文是基于$Z' = \max(lnW_1, lnW_2)$，为此进行转化：

$$Z' = \max(lnS_1, lnS_2) \Leftrightarrow -Z' = \min(-lnS_1, -lnS_2).$$

可见两者具共通性，以下只考虑对$Z = \min(X_1, X_2)$的识别，以与 Basu and Ghosh(1978)保持一致。

Identifiability of the Log Normal Roy Model from Cross-Section Data

Proof:(Cont.)

记$(X_1, X_2) \sim BVN(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho_{12})$，$Z = \min(X_1, X_2)$，定义当$Z = X_i$ $(i = 1,2)$时，有$I = i$。此为不可观测的分布。

类似的，记$(X_3, X_4) \sim BVN(\mu_3, \mu_4, \sigma_3, \sigma_4, \rho_{34})$，$Z' = \min(X_3, X_4)$，同样定义当$Z' = X_{i'}$ $(i' = 3,4)$时，有$I' = i'$。

定义

$$\alpha_1 = 1 - \frac{\rho_{12}\sigma_1}{\sigma_2}, \qquad \alpha_2 = 1 - \frac{\rho_{12}\sigma_2}{\sigma_1},$$

$$\alpha_3 = 1 - \frac{\rho_{34}\sigma_3}{\sigma_4}, \qquad \alpha_4 = 1 - \frac{\rho_{34}\sigma_4}{\sigma_3}.$$

定义

$$\mu_1' = \begin{cases} \dfrac{\mu_1 - \dfrac{\rho_{12}\sigma_1\mu_2}{\sigma_2}}{\alpha_1} & (\alpha_1 \neq 0), \\ (\mu_1 - \mu_2) & (\alpha_1 = 0), \end{cases} \qquad \sigma_1' = \begin{cases} \dfrac{\sigma_1(1-\rho_{12}^2)^{\frac{1}{2}}}{|\alpha_1|} & (\alpha_1 \neq 0), \\ \sigma_1(1-\rho_{12}^2)^{\frac{1}{2}} & (\alpha_1 = 0). \end{cases}$$

Identifiability of the Log Normal Roy Model from Cross-Section Data

Proof:(Cont.)

即需证：如果(Z, I)与(Z', I')有相同的分布，则

$$(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho_{12}) = (\mu_3, \mu_4, \sigma_3, \sigma_4, \rho_{34}).$$

记$f_i(\cdot)$表示给定$Z = X_i$时的条件概率密度，因此有

(4.1)

$$p_1 f_1(t) = \phi_{11}(t)\left[1 - \Phi\left(\frac{t - m_2(t)}{\sigma_2(1-\rho_{12}^2)^{\frac{1}{2}}}\right)\right],$$

其中$p_i = \Pr(I = i)$，$\phi_{ii}(\cdot) \sim N(\mu_i, \sigma_i^2)$，$\Phi(\cdot)$为标准正态分布的累积概率函数，且

$$m_i(t) = E[X_i | X_{1-i} = t] = \mu_i + \rho_{12}\left(\frac{\sigma_i}{\sigma_{1-i}}\right)(t - \mu_{1-i}).$$

注意到

$$\frac{t - m_1(t)}{\sigma_1(1-\rho_{12}^2)^{\frac{1}{2}}} = \frac{\left(1 - \frac{\rho_{12}\sigma_1}{\sigma_2}\right)t - \left(\mu_1 - \rho_{12}\left(\frac{\sigma_1}{\sigma_2}\right)\mu_2\right)}{\sigma_1(1-\rho_{12}^2)^{\frac{1}{2}}}.$$

Identifiability of the Log Normal Roy Model from Cross-Section Data

Proof:(Cont.)

如果$\alpha_1 > 0$，我们有

$$p_2 f_2(t) = \phi_{22}(t)[1 - \Phi_{1'1'}(t)],$$

此处$\Phi_{1'1'}(\cdot) \sim N(\mu_1', \sigma_1'^2)$，$\mu_1'$，$\sigma_1'$定义如前。

如果$\alpha_1 < 0$，我们有

$$p_2 f_2(t) = \phi_{22}(t)\Phi_{1'1'}(t).$$

如果$\alpha_1 = 0$，我们有

$$\begin{aligned} p_2 f_2(t) &= \phi_{22}(t)[1 - \Phi_{1'1'}(0)] \\ &= \phi_{22}(t)p_2. \end{aligned}$$

Identifiability of the Log Normal Roy Model from Cross-Section Data

Proof:(Cont.)

接下来证明定理。由于$\left(\frac{\rho_{12}\sigma_2}{\sigma_1}\right)\left(\frac{\rho_{12}\sigma_1}{\sigma_2}\right)=\rho_{12}^2<1$[1]，故$\alpha_1\left(=1-\frac{\rho_{12}\sigma_1}{\sigma_2}\right)$与$\alpha_2\left(=1-\frac{\rho_{12}\sigma_2}{\sigma_1}\right)$中至少有一个为正；类似的，$\alpha_3$与$\alpha_4$中至少有一个为正。

情形一[2]：$\alpha_i>0\ (i=1,2,3,4)$

可以计算得[3]

$$f_1(t)=\begin{cases}p_1^{-1}\phi_{11}(t)\left(1-\dfrac{\Phi(t-\mu^*)}{\sigma^*}\right) & if\ \rho_{12}\sigma_2\neq\sigma_1,\\ \phi_{11}(t) & otherwise.\end{cases}$$

其中

$$\mu^*=\frac{1}{1-\frac{\rho\sigma_2}{\sigma_1}}\mu_2+\left(1-\frac{1}{1-\frac{\rho\sigma_2}{\sigma_1}}\right)\mu_1,$$

$$\sigma^*=\frac{\sigma_2\sqrt{1-\rho^2}}{1-\rho\sigma_2/\sigma_1}.$$

$f_2(t)$可类似定义（只需交换下标的 1 与 2）.

Identifiability of the Log Normal Roy Model from Cross-Section Data

Proof:(Cont.)

令$t \to -\infty$，有

$$\lim_{t\to-\infty} p_1 f_1(t)/\phi_{11}(t) = \lim_{t\to-\infty} p_1 f_1(t)/\phi_{33}(t) = 1,$$

所以

$$\lim_{t\to-\infty} \phi_{11}(t)/\phi_{33}(t) = 1.$$

同样地，

$$\lim_{t\to-\infty} \phi_{22}(t)/\phi_{44}(t) = 1.$$

因此有

$$(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho_{12}) = (\mu_3, \mu_4, \sigma_3, \sigma_4, \rho_{34}).$$

Identifiability of the Log Normal Roy Model from Cross-Section Data

Proof:(Cont.)

情形二：(α_1, α_2)与(α_3, α_4)各有一参数为正

不妨假设$\alpha_1 > 0$，$\alpha_2 < 0$。那么有$\alpha_3 > 0$，$\alpha_4 < 0$或者$\alpha_3 < 0$，$\alpha_4 > 0$。首先假设前者成立，即$\alpha_3 > 0$，$\alpha_4 < 0$。由于(X_1, X_2)与(X_3, X_4)的最小值分布相同，则根据(4.1)式有

$$p_1 f_1(t) = \phi_{11}(t)\Phi_{2'2'}(t) = \phi_{33}(t)\Phi_{4'4'}(t)$$
$$\Rightarrow \Phi_{2'2'}(t) = \{\phi_{11}(t)\}^{-1}\phi_{33}(t)\Phi_{4'4'}(t) \quad \forall t.$$

令$t \to \infty$，有

$$\phi_{11}(t) = \phi_{33}(t)$$
$$\Rightarrow \mu_1 = \mu_3, \sigma_1 = \sigma_3.$$

与情形一类似，同样可以推得$\mu_2 = \mu_4, \sigma_2 = \sigma_4$。

由$p_1 = \Pr(X_1 < X_2) = \Pr(X_3 < X_4)$，可得

$$\Phi\left(\frac{\mu_2 - \mu_1}{\sigma_1^2 + \sigma_2^2 - 2\rho_{12}\sigma_1\sigma_2}\right) = \Phi\left(\frac{\mu_4 - \mu_3}{\sigma_3^2 + \sigma_4^2 - 2\rho_{34}\sigma_3\sigma_4}\right)$$
$$\Rightarrow \rho_{12} = \rho_{34}.$$

类似的，如果假设$\alpha_3 < 0$，$\alpha_4 > 0$，那么

$$p_1 f_1(t) = \phi_{11}(t)\Phi_{2'2'}(t) = \phi_{33}(t)[1 - \Phi_{4'4'}(t)]$$
$$\Rightarrow [\phi_{33}(t)]^{-1}\phi_{11}(t)\Phi_{2'2'}(t) = 1 - \Phi_{4'4'}(t) \quad \forall t.$$

令$t \to -\infty$，上式右手侧$\to 1$，但是左手侧并不趋近于1，矛盾。因此，$\alpha_3 < 0$，$\alpha_4 > 0$的假设不成立。

Identifiability of the Log Normal Roy Model from Cross-Section Data

Proof:(Cont.)

情形三：$(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$中有一参数为负，不妨设$\alpha_1 > 0, \alpha_2 > 0, \alpha_3 > 0, \alpha_4 < 0$。同样可以得到与上述情形相同的矛盾。

其他情形：接下来讨论α_i中某一（几）个值为 0 的情形，证明省略。

$Q.E.D.$

Identifiability of the Log Normal Roy Model from Cross-Section Data

Theorem 5: Under the conditions postulated for the *log normal* Roy model, μ and $\sum$ can be identified, except for their subscripts, from knowledge of the aggregate earnings distribution.

$\Rightarrow \sigma_{12}$ is uniquely identified.

Identifiability of the Log Normal Roy Model from Cross-Section Data

Proof:

识别策略为 MLE，以下证明给出$Z = \min(X_1, X_2)$对应的概率密度函数推导过程：

$$\Pr(Z \geq t) = \Pr(t \leq X_1 \leq X_2) + \Pr(t \leq X_2 \leq X_1),$$

$$\Rightarrow f(t) = -\frac{dPr(Z \geq t)}{dt} = -\left[\frac{d\Pr(t \leq X_1 \leq X_2)}{dt} + \frac{d\Pr(t \leq X_2 \leq X_1)}{dt}\right]$$

其中，

$$\Pr(t \leq X_1 \leq X_2)$$

$$= \int_t^\infty \int_x^\infty \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right)\right] dy\, dx$$

$$= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \int_t^\infty \exp\left[-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right] \int_x^\infty \exp\left[-\frac{1}{2(1-\rho^2)}\left[\frac{y-\mu_2}{\sigma_2} - \frac{\rho(x-\mu_1)}{\sigma_1}\right]^2\right] dydx,$$

Identifiability of the Log Normal Roy Model from Cross-Section Data

Proof:(Cont.)

因此，

$$-\frac{d\Pr(t \le X_1 \le X_2)}{dt}$$

$$= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{(t-\mu_1)^2}{2\sigma_1^2}\right] \int_t^{\infty} \exp\left[-\frac{1}{2(1-\rho^2)}\left[\frac{y-\mu_2}{\sigma_2} - \frac{\rho(x-\mu_1)}{\sigma_1}\right]^2\right] dy$$

$$= \phi_{11}(t)\left\{1 - \Phi\left(\frac{\sigma_1 - \rho\sigma_2}{\sigma_1\sigma_2}t + \frac{\rho\sigma_2\mu_1 - \sigma_1\mu_2}{\sigma_1\sigma_2}\right)\right\},$$

其中，$\phi_{ii}(\cdot) \sim N(\mu_i, \sigma_i^2)$，$\Phi(\cdot)$为标准正态分布的累积概率函数。

类似的，有

$$-\frac{d\Pr(t \le X_2 \le X_1)}{dt} = \phi_{22}(t)\left\{1 - \Phi\left(\frac{\sigma_2 - \rho\sigma_1}{\sigma_1\sigma_2}t + \frac{\rho\sigma_1\mu_2 - \sigma_2\mu_1}{\sigma_1\sigma_2}\right)\right\}.$$

由此，可以得到

$$f(t) = \phi_{11}(t)\left\{1 - \Phi\left(\frac{\sigma_1 - \rho\sigma_2}{\sigma_1\sigma_2}t + \frac{\rho\sigma_2\mu_1 - \sigma_1\mu_2}{\sigma_1\sigma_2}\right)\right\}$$
$$+ \phi_{22}(t)\left\{1 - \Phi\left(\frac{\sigma_2 - \rho\sigma_1}{\sigma_1\sigma_2}t + \frac{\rho\sigma_1\mu_2 - \sigma_2\mu_1}{\sigma_1\sigma_2}\right)\right\}.$$

Identifiability of the Log Normal Roy Model from Cross-Section Data

Proof:(Cont.)

从而，似然函数式为

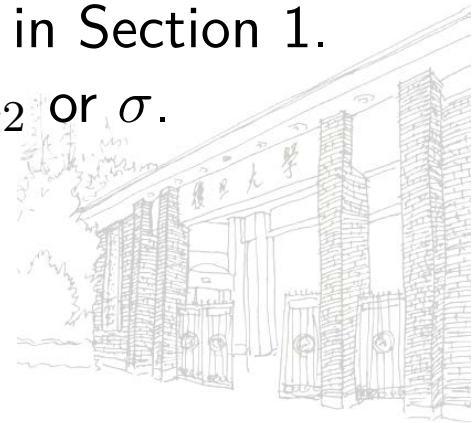
$$L = \prod_i f(z_i).$$

可利用 MLE 方法识别出均值μ与协方差矩阵$\sum$，但无法识别其归属的部门。 $Q.E.D.$

Identifiability of the Log Normal Roy Model from Cross-Section Data

In the analysis of female labor supply, it is often assumed that one of the sectors is home production, where only the proportion working in each sector and the sectoral earnings distribution in one sector are observed.$\Rightarrow$

Theorem 6: If only the earnings density in sector one, $f(lnw_1|lnW_1 > lnW_2)$ and $Pr(lnW_1 > lnW_2)$ are known in the *log normal* Roy model, it is possible to identify μ_1 and σ_{11}. It is also possible to identify $(\mu_1 - \mu_2)/\sigma$ and ρ, where σ and ρ are as defined in Section 1. Without further restrictions it is not possible to identify μ_2 or σ.



Identifiability of the Log Normal Roy Model from Cross-Section Data

Proof:

直觉上看，利用(7-a)与(7-b)可知，$f(lnw_1|lnW_1 > lnW_2)$与$\Pr(lnW_1 > lnW_2)$依赖于$\sigma_{11}(1-\rho^2)$，μ_1，$c_* = (\mu_1 - \mu_2)/\sigma$，$\sqrt{\sigma_{11}} = a_1\sigma$，因此最多只可能识别出$\mu_1, \sigma_{11}, c_*$。事实上，这三个参数确实可以识别出来，证明如下：

利用

$$\Pr(lnW_1 > lnW_2) = \Phi(c_*),$$

可识别出c_*。从而可以计算出$k_1 = E[Z|Z > -c_*], k_2 = Var[Z|Z > -c_*], k_3 = E[(Z - E[Z|Z > -c_*])^3|Z > -c_*]$。由(10)-(12)，给定$lnW_1 > lnW_2$条件下的$lnW_1$的前三阶矩的公式如下：

$$E[lnW_1|lnW_1 > lnW_2] = \mu_1 + \rho\sqrt{\sigma_{11}}k_1 \tag{C-2}$$

$$Var[lnW_1|lnW_1 > lnW_2] = \sigma_{11}\left(\rho^2 k_2 + (1-\rho^2)\right) = \sigma_{11} + (k_2 - 1)\rho^2\sigma_{11} \tag{C-3}$$

$$E[(lnW_1 - E[lnW_1|lnW_1 > lnW_2])^3|lnW_1 > lnW_2] = \rho^3\sigma_{11}^{\frac{3}{2}}k_3 \tag{C-4}$$

由(C-4)可识别$\rho\sigma_{11}^{\frac{1}{2}}$；接下来由(C-2)可识别$\mu_1$；再由(C-3)识别$\sigma_{11}$，从而识别出$\rho$。

Q.E.D.

The Nonidentifiability of a General Nonnormal Roy Model in a Single Cross Section

The results above are not robust when examined in the context of a general *nonnormal* model of skill distribution.

The Nonidentifiability of a General Nonnormal Roy Model in a Single Cross Section

The results above are not robust when examined in the context of a general *nonnormal* model of skill distribution.

- Any cross-section wage distribution can be rationalized by a model with two independent skills ([Theorem 7]) or two highly correlated skills ([Theorem 8]).

Nonidentifiability of a General Nonnormal Roy Model in a Single Cross Section

Theorem 7: It is possible to rationalize sectoral wage data in a single cross-section by a two-skill model with independence. More precisely, it is possible to rationalize data on $f(s_1|S_1 > S_2)$, $f(s_2|S_2 > S_1)$, and $P(S_1 > S_2)$ by an independent skill model $f(s_1, s_2) = f_1(s_1)f_2(s_2)$.

Nonidentifiability of a General Nonnormal Roy Model in a Single Cross Section

Proof:

注：笔者仿照参考文献Tsiatis(1975)进行改写[1]，自行给出证明，或有不严谨之处。

记

$$Q_i(t) = \Pr\left(S_i \le t \bigcap_{j \ne i} S_i \ge S_j\right)$$

$$H^{(2)}(t_1, t_2) = \Pr(S_1 \le t_1, S_2 \le t_2)$$

则

$$Q_1(t+h) - Q_1(t) = \Pr(S_1 \le t+h, S_1 \ge S_2)$$
$$= \Pr(t < S_1 \le t+h, S_1 \ge S_2)$$

$Q_1(t+h) - Q_1(t)$的下界为

$$\Pr(t < S_1 \le t+h, t \ge S_2)$$
$$= H^{(2)}(t+h, t) - H^{(2)}(t, t),$$

上界为

$$\Pr(t < S_1 \le t+h, t+h \ge S_2)$$
$$= H^{(2)}(t+h, t+h) - H^{(2)}(t, t+h).$$

Nonidentifiability of a General Nonnormal Roy Model in a Single Cross Section

Proof:(Cont.)

故有

$$H^{(2)}(t+h,t)-H^{(2)}(t,t)\leq Q_1(t+h)-Q_1(t)\leq H^{(2)}(t+h,t+h)-H^{(2)}(t,t+h)$$

对上述不等式除以h并取极限$h\to 0$，利用极限的夹逼定理可知

$$Q_1'(t)=\left.\frac{\partial H^{(2)}}{\partial t_1}\right|_{t_1=t_2=t}\triangleq H_1^{(2)}(t),$$

从而有

$$Q_i(t)=\int_0^t H_i^{(2)}(x)\,dx.$$

如果不考虑独立性，则$Q_i(t)$可能对应多个$H^{(2)}(t)$的边缘分布。以下利用技能的独立性证明解的唯一性：

由独立性知

$$H^{(2)}(t_1,t_2)=H_1(t_1)H_2(t_2),$$

Nonidentifiability of a General Nonnormal Roy Model in a Single Cross Section

Proof:(Cont.)

从而可转化为

$$Q_j'(t) = -r_j(t)H_1(t)H_2(t),$$

其中$r_j(t) = -\frac{d}{dt}\log H_j(t)$，积分得$H_j(t) = \exp\{-\int_t^\infty r_j(x)\,dx\}$。

记$r(x) = \sum_j r_j(x)$，从而有$Q_j'(t) = -r_j(t)\exp\{-\int_t^\infty r(x)\,dx\}$。

假设存在$H_j^*(t)$ $(j = 1,2)$对应相同的$Q_i(t)$ $(i = 1,2)$。则

$$r_j^*(t) = -\frac{d}{dt}\log H_j^*(t),$$

有

$$Q_j'(t) = -r_j^*(t)\exp\left\{-\int_t^\infty r^*(x)\,dx\right\},$$

Nonidentifiability of a General Nonnormal Roy Model in a Single Cross Section

Proof:(Cont.)

累加得

$$\sum_{j=1}^{k} Q_j'(t) = -r^*(t)\exp\left\{-\int_t^{\infty} r^*(x)\,dx\right\}$$

$$= \frac{d}{dt}\exp\left\{-\int_t^{\infty} r^*(x)\,dx\right\},$$

其中$r^*(x) = \sum_j r_j^*(x)$，第二处等号来自于求导的逆运算。积分得

$$\sum_{j=1}^{k} Q_j(t) = \exp\left\{-\int_t^{\infty} r^*(x)\,dx\right\}.$$

代回前式$Q_j'(t) = -r_j^*(t)\exp\{-\int_t^{\infty} r^*(x)\,dx\}$，从而得到

$$r_j^*(t) = -\frac{Q_j'(t)}{\sum_{j=1}^{k} Q_j(t)},$$

代回$H_j(t) = \exp\{-\int_t^{\infty} r_j(x)\,dx\}$得到

$$H_j^*(t) = \exp\left\{\int_t^{\infty} \frac{Q_j'(x)}{\sum_{j=1}^{k} Q_j(x)}\,dx\right\}.$$

给定$Q_j(\cdot)$时所得到的$H_j^*(\cdot)$是唯一的。

$Q.E.D$

Nonidentifiability of a General Nonnormal Roy Model in a Single Cross Section

Theorem 8: For any $c < 1$, it is always possible to rationalize sectoral wage data in a single cross-section by a two-skill model with correlation greater than c, provided that $Var[max\{S_1, S_2\}]$ exists.

Proof:

Let Y be the observed wage and let R be an indicator giving the sector associated with that wage. Theorem 7 informs us that any distribution of (Y, R) can be explained by a Roy model with independent skills. On the other hand, imagine that for each (Y, R) we define

$$(S_1, S_2) = \begin{cases} (Y, Y - \varepsilon), & if \quad R = 1, \\ (Y - \varepsilon, Y), & if \quad R = 2. \end{cases}$$

The correlation (provided that it exixts) between S_1 and S_2 constructed in this manner will depebd on ε, but it can be made arbitrarily close to 1 by making ε close to 0. The theorem naturally follows. $Q.E.D.$

Identification from Multi-market Data in a General Nonnormal Roy Model

- These negative conclusions can be reversed with access to data on earnings distributions from markets with *the same distributions of skills but different relative skill prices*.

- With sufficient price variation, it is possible to identify the population skill distribution
  - [Theorem 9] *even if an agent's sectoral choice is unknown;*
  - [Theorem 10] *even if earnings in one sector are not observed.*
- [Theorem 11] Access to panel data greatly reduces the required amount of price variation.
- [Theorem 12] Access to regressors that affect the location of the log-skill distribution substitutes for price variation and secures identification in a single cross-section.

Identification from Multi-market Data in a General Nonnormal Roy Model

Theorem 9: Let S_1 and S_2 be positive random variables with distribution function $F(s_1, s_2)$. If we only observe $Z = max\{S_1, \pi_2 S_2\}$ and π_2 takes all possible values in the interval $(0, \infty)$, then F is identifiable.

Proof:

By assumption we know $Pr(max\{S_1, \pi_2 S_2\} \leq x)$ for all x and π_2, but

$$\begin{aligned}\{max\{S_1, \pi_2 S_2\} \leq x\} &= \{S_1 \leq x, \pi_2 S_2 \leq S_1\} \cup \{\pi_2 S_2 \leq x, S_1 \leq \pi_2 S_2\} \\ &= \{S_1 \leq x, S_2 \leq x/\pi_2\},\end{aligned}$$

so for any $s_1, s_2 > 0$ we have (setting $x = s_1$ and $\pi_2 = s_1/s_2$),

$$\begin{aligned}F(s_1, s_2) &= Pr(S_1 \leq s_1, S_2 \leq s_2) = Pr(S_1 \leq s_1, S_2 \leq s_1/\pi_2) \\ &= Pr(max\{S_1, \pi_2 S_2\} \leq x),\end{aligned}$$

which completes the proof. $Q.E.D.$

Identification from Multi-market Data in a General Nonnormal Roy Model

Theorem 10: If we observe the distribution of Z given by

$$Z = \begin{cases} \pi_2 S_2 & if \quad S_1 < \pi_2 S_2, \\ 0 & if \quad S_1 \geq \pi_2 S_2. \end{cases}$$

and π_2 traverses the interval $(0, \infty)$, then $F(S_1, S_2)$ is identified from multimarket data on aggregate earnings.

Proof:

Let s_1, s_2 be given. We will then show that we can find $F(s_1, s_2)$.
Let $\varepsilon > 0$ be given. For given π_2, we know the probability of events of the type $\{S_1 < \pi_2 S_2 \leq x\}$ for all $x \in (0, \infty)$. This means that we know the probability of the event given by *OAB* in Figure 1. By the same argument, we also know the probability of the event given by the set *OCD*. We therefore know the probability of the difference *DCAB*.

Identification from Multi-market Data in a General Nonnormal Roy Model

Proof:(Cont.) By exactly the same reasoning we know the probability of the event *FGED*, and therefore of the event *FGECAB*. If we continue this process, we will converge to a number μ satisfying

$$F(s_1, s_2) \leq \mu \leq F(s_1 + \varepsilon, s_2).$$

Do this for each ε, and the limit as $\varepsilon \to 0$, and we obtain $F(s_1, s_2)$. $Q.E.D.$

Identification from Multi-market Data in a General Nonnormal Roy Model

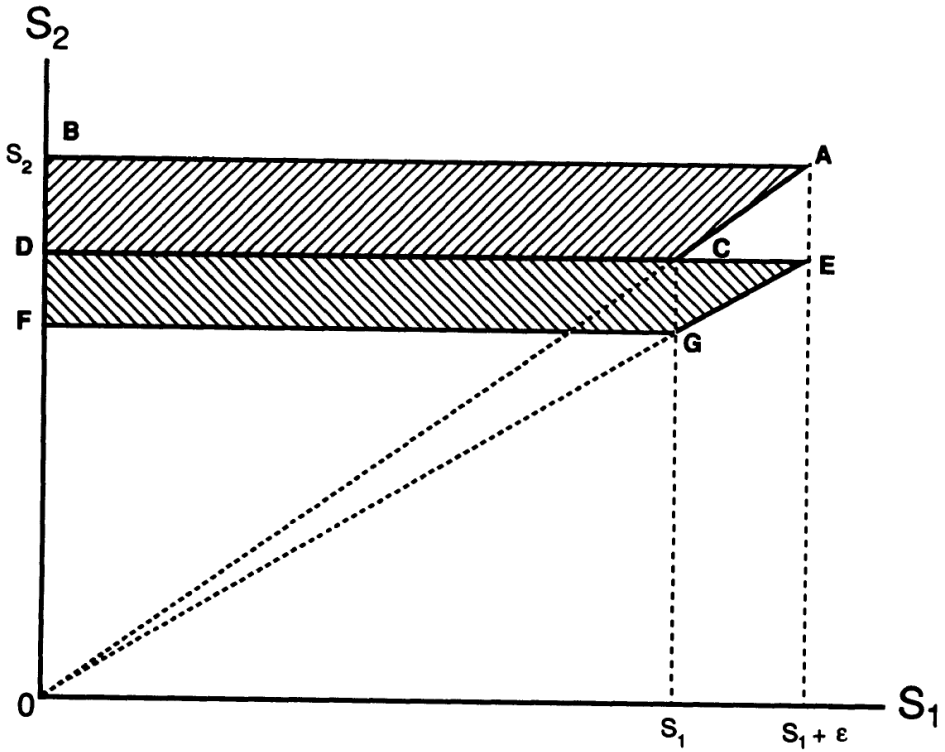


FIGURE 1

Identification from Multi-market Data in a General Nonnormal Roy Model

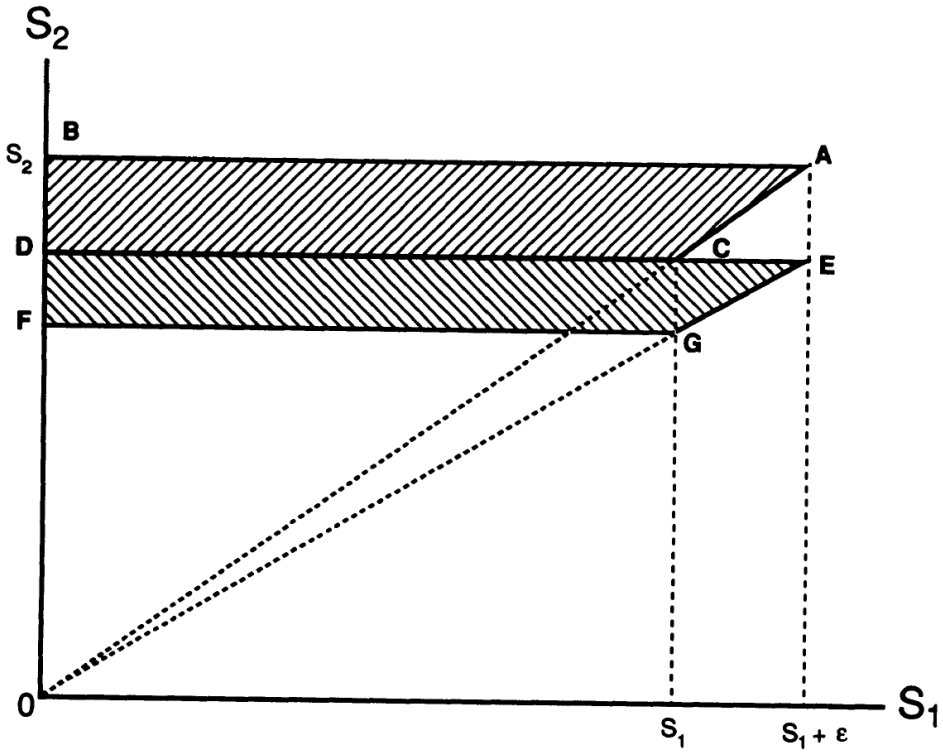


FIGURE 1

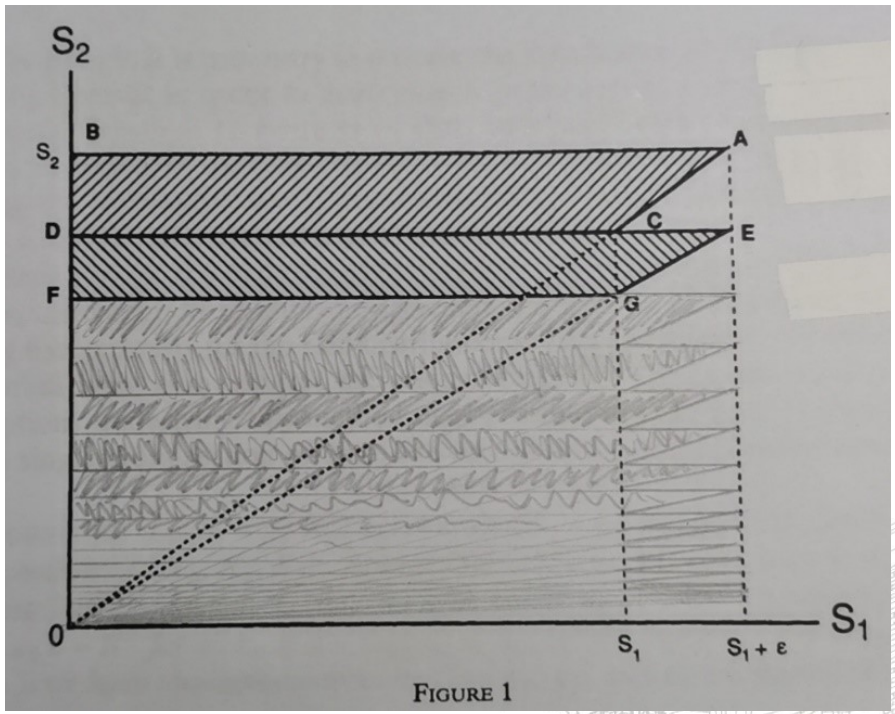


FIGURE 1

Identification from Multi-market Data in a General Nonnormal Roy Model

Theorem 11: Suppose that we have panel data on aggregate earnings of individuals and that individual skills do not change over time. If we observe $(Z, Z') = (max\{S_1, \pi_2 S_2\}, max\{S_1, \pi_2' S_2\})$ for $\pi_2 < \pi_2'$, then we can identify $F(s_1, s_2)$ over the region $\pi_2 s_2 \leq s_1 \leq \pi_2' s_2$.

Proof: By hypothesis we know $Pr(Z' \leq z', Z \leq z)$ for all z', z. Let $s_1, s_2 > 0$ be given such that $\pi_2 s_2 \leq s_1 \leq \pi_2' s_2$. Now over this region

$$\begin{aligned} F(s_1, s_2) &= Pr(S_1 \leq s_1, S_2 \leq s_2) \\ &= Pr(S_1 \leq s_1, S_1 \leq \pi_2' s_2, S_2 \leq s_2, S_2 \leq s_1/\pi_2) \\ &= Pr(S_1 \leq s_2, \pi_2 S_2 \leq s_1, S_1 \leq \pi_2' s_2, \pi_2' S_2 \leq \pi_2' s_2) \\ &= Pr(max\{S_1, \pi_2 S_2\} \leq s_1, max\{S_1, \pi_2' S_2\} \leq \pi_2' s_2) \end{aligned}$$

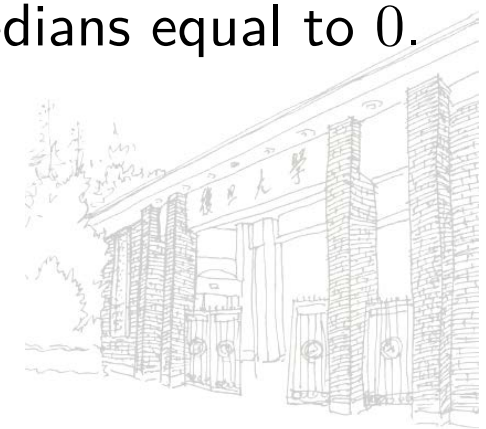
which, by hypothesis, is known. Hence $F(s_1, s_2)$ is identified for all $\pi_2 s_2 \leq s_1 \leq \pi_2' s_2$. *Q.E.D.*

Identification from Multi-market Data in a General Nonnormal Roy Model

Theorem 12: Let $S_1 = g_1(X_1, X_0) + \varepsilon_1$ and $S_2 = g_2(X_2, X_0) + \varepsilon_2$ where $(\varepsilon_1, \varepsilon_2)$ is independent of (X_0, X_1, X_2). Assume that

- (a) $(\varepsilon_1, \varepsilon_2)$ is continuously distributed with distribution function G and support equal to R^2;
- (b) $support(g_1(X_1, x_0), g_2(X_2, x_0)) = R^2$ for all x_0 in the support of X_0;
- (c) marginal distributions of ε_1 and ε_2 both have medians equal to 0.

Then g_1, g_2, and G are identified.



Identification from Multi-market Data in a General Nonnormal Roy Model

Proof: By assumption we know

$$(A) \qquad Pr(S_1, S_2) = Pr(g_1(x_1, x_0) + \varepsilon_1 > g_2(x_2, x_0) + \varepsilon_2),$$

$$(B) \qquad Pr(S_1 \leq y, S_1 > S_2) = Pr(g_1(x_1, x_0) + \varepsilon_1 \leq y, \; g_1(x_1, x_0) + \varepsilon_1 > g_2(x_2, x_0) + \varepsilon_2),$$

$$(C) \qquad Pr(S_2 \leq y, S_2 \geq S_1) = Pr(g_2(x_2, x_0) + \varepsilon_2 \leq y), \; g_2(x_2, x_0) + \varepsilon_2 > g_1(x_1, x_0) + \varepsilon_1),$$

for all (x_0, x_1, x_2) in the support of (X_0, X_1, X_2) and for all y.
Fix x_0. Let $\bar{x_1}$ and $\bar{x_2}$ be in the support of X_1 and X_2, respectively. From (A), we can then find

$$\begin{aligned}\{(x_1, x_2) : & Pr(g_1(x_1, x_0) + \varepsilon_1 > g_2(x_2, x_0) + \varepsilon_2)\} \\ &= Pr(g_1(\bar{x_1}, x_0) + \varepsilon_1 > g_2(\bar{x_2}, x_0) + \varepsilon_2) \\ &= \{(x_1, x_2) : g_1(x_1, x_0) + l = g_2(x_2, x_0)\}\end{aligned}$$

for some unknown constant l.

Identification from Multi-market Data in a General Nonnormal Roy Model

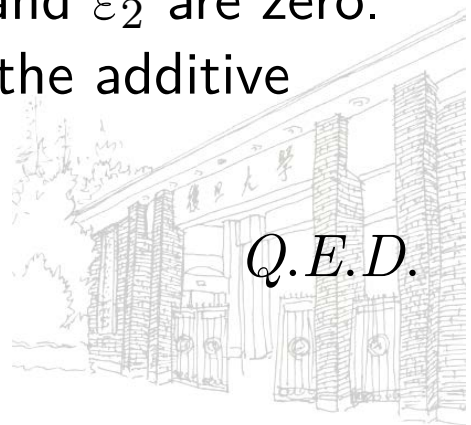
Proof:(Cont.)

For any point in that set, we can use (B) to find

$$Pr(g_1(x_1, x_0) + \varepsilon_1 \leq y, \varepsilon_1 > \varepsilon_2 + l)$$

for all y. This identifies $g_1(\cdot, x_0)$ except for an additive constant. In a similar way, $g_2(\cdot, x_0)$ is identified (except for an additive constant). G is then identified by Theorem 9, except for the location. The location of G is determined by exploiting the fact that the medians of ε_1 and ε_2 are zero. Having determined the location of G, we can determined the additive constants in $g_1(\cdot, x_0)$ and $g_2(\cdot, x_0)$.

Since x_0 was arbitrary, this completes the proof. $Q.E.D.$



Summary

- This paper derive a general class of nonnormal models for which the main conclusion of the Roy model remains valid. It assumes that skills can be decomposed into two components: a log concave random variable and an independent, additive component that can be freely specified. *Selection depends on the log concave component but not on the other component.*

Summary

- Identifiability is necessary for consistent estimation of the Roy model. This paper only investigate the necessary first step.
- *Under Roy's normality assumptions*, it is possible to recover underlying skill distributions from a single cross-section of earnings.
- The strong identification results *vanish in a general nonnormal model*, where any single cross-section distribution of wages can be rationalized by a model with independent, positively correlated or negatively correlated skills.
- Access to data from *markets with different relative skill prices* facilitates identification; panel data further helps.
- With available independent regressors and relative assumptions, *cross-section variation in regressors can substitute for multimarket variation in skill prices*.

The End

Appendix

Proposition 1: If D is a *log concave random variable*, then

$$0 \leq \frac{\partial E[D|D > d]}{\partial d} \leq 1$$

$$0 \leq \frac{\partial E[D|D \leq d]}{\partial d} \leq 1$$

and

$$\frac{\partial Var[D|D > d]}{\partial d} \leq 0$$

$$\frac{\partial Var[D|D \leq d]}{\partial d} \geq 0$$

Proof of Proposition 1:

定义函数$L(x,y)$是TP_2的：当$L(x,y) \geq 0 \; \forall x,y$，且$\forall x_1 < x_2, y_1 < y_2$时，有

$$\begin{vmatrix} L(x_1,y_1) & L(x_1,y_2) \\ L(x_2,y_1) & L(x_2,y_2) \end{vmatrix} \geq 0.$$

根据 Polya & Szego (1972)定理：如果$K(x,q)$与$L(q,y)$是TP_2的，那么由复合公式（composition formula），

$$M(x,y) = \int_{q \in \underline{Q}} K(x,q)L(q,y)\, dq$$

也是TP_2的。

注意到

$$R(x,q) \overset{\text{def}}{=} \begin{cases} 1, & x \geq q, \\ 0, & x < q, \end{cases}$$

与

$$R^*(x,q) \overset{\text{def}}{=} \begin{cases} 1, & x \leq q, \\ 0, & x > q, \end{cases}$$

都是TP_2的。

Proof of Proposition 1:(Cont.)

记J表示实值函数，定义$L(q,y)=J(q-y)$。“$L(q,y)$是TP_2的”等价于“J是对数凹的”。类似的，“$L(q,y)=J(q+y)$是TP_2的（$q\geq 0, y\geq 0$）”等价于“J是对数凸的”。如果$J(z)$是对数凹（凸）、正、二次可微的，那么“$J(q-y)$（$J(q+y)$）是TP_2的”等价于

$$\frac{J''}{J}-\left(\frac{J'}{J}\right)^2 \lesseqgtr 0 \quad \begin{matrix} J(q-y)TP_2, \\ J(q+y)TP_2, \end{matrix}$$

也即（若$J'\neq 0$）

$$\frac{J''J}{(J')^2} \lesseqgtr 1 \quad \begin{matrix} J(q-y)TP_2, \\ J(q+y)TP_2. \end{matrix}$$

令$K(x,q)=R(x,q)$或$K(x,q)=R^*(x,q)$，可以发现：对于$x=a$为常数，复合公式（composition formula）意味着：

当$J(q\pm y)$是TP_2时，

(B-1)

$$M(a,y)=\int_{q\in \underline{Y}} R(a,q)J(q\pm y)\,dq=\int_{\{q|q\leq a,\, q\in \underline{Y}\}} J(q\pm y)\,dq,$$

(B-2)

$$M^*(a,y)=\int_{q\in \underline{Y}} R^*(a,q)J(q\pm y)\,dq=\int_{\{q|q\geq a,\, q\in \underline{Y}\}} J(q\pm y)\,dq$$

都是TP_2的。（Y是J的定义域）

Proof of Proposition 1:(Cont.)

如果随机变量Z在支集(m,n)内的概率密度是对数凹的，那么：如果m或n之一是有限数，就可以通过定义$f(z)=0\ z\in(-\infty,m]$与$f(z)=0\ z\in[n,\infty)$将Z延展至$(-\infty,\infty)$的范围。如此重新定义的概率密度仍然是对数凹的。

但是，对于对数凸概率密度的支集的延展无法得到新的对数凸的函数，无法在支集为$(-\infty,\infty)$时定义对数凸的随机变量。以下证明假设Z的支集为$[0,\infty)$。

Proof of Proposition 1:(Cont.)

接下来分别讨论对数凹与对数凸两种情形：

1. 如果Z是对数凹的，其密度为J，改写(B-1)与(B-2)如下：

(B-1)'

$$M(a,y) = N(a-y) = \int_{-\infty}^{a-y} J(t)\,dt,$$

(B-2)'

$$M^*(a,y) = N^*(a-y) = \int_{a-y}^{\infty} J(t)\,dt.$$

利用前述的 Polya-Szego 定理，$\Pr(Z \le a)$与$\Pr(Z \ge a)$都是a的对数凹函数。类似的，有

$$\int_{-\infty}^{b} \Pr(Z \le a)\,da$$

与

$$\int_{b}^{\infty} \Pr(Z \ge a)\,da$$

对于b都是对数凹的（只要积分存在有限）。对于这些积分作为被积函数的继续积分可类似推理。

Proof of Proposition 1:(Cont.)

2. 如果J是支集$Z \geq 0$的对数凸随机变量，有

$$M^*(a,y) = N^*(a+y) = \int_{a+y}^{\infty} J(t)\,dt,$$

利用 Polya-Szego 定理可知$N^*(a+y)$是TP_2的，因此也是a的对数凸函数。因此，$\Pr(Z \geq a)$是a的对数凸函数。

为推导本文结论，定义

$$S_{j+1}(a) = \int_a^{\infty} S_j(z)\,dz, \qquad \text{其中} S_0 = S(a) \stackrel{\text{def}}{=} Pr(Z > a),$$

$$F_{j+1}(a) = \int_{-\infty}^{a} F_j(z)\,dz, \qquad \text{其中} F_0 = F(a) \stackrel{\text{def}}{=} Pr(Z \leq a).$$

那么，对于$E[|Z|] < \infty$，有

(B-3)

$$E[Z|Z > a] = a + \frac{S_1(a)}{S_0(a)},$$

Proof of Proposition 1:(Cont.)

证[1]:

$$
\begin{aligned}
E[Z|Z>a] &= a + E[Z-a|Z-a>0] \\
&= a + \frac{E\left[\int_0^{Z-a} 1\,dt\right]}{\Pr(Z-a>0)} \\
&= a + \frac{\int_0^{\infty} E[1\{Z-a\geq t\}]\,dt}{\Pr(Z-a>0)} \\
&= a + \frac{\int_a^{\infty} E[1\{Z\geq t\}]\,dt}{\Pr(Z-a>0)} \\
&= a + \frac{\int_a^{\infty} \Pr(Z\geq t)\,dt}{\Pr(Z>a)} \\
&= a + \frac{S_1(a)}{S_0(a)}.
\end{aligned}
$$

Proof of Proposition 1:(Cont.)

(B-4)

$$E[Z|Z \leq a] = a - \frac{F_1(a)}{F_0(a)}.$$

证：

$$E[Z|Z \leq a] = a + E[Z - a|Z - a \leq 0]$$

$$= a + \frac{E\left[\int_{z-a}^{0} 1\, dt\right]}{\Pr(Z - a \leq 0)}$$

$$= a - \frac{E\left[\int_{0}^{a-z} a\, dt\right]}{\Pr(Z - a \leq 0)}$$

$$= a - \frac{\int_{0}^{\infty} E[1\{a - Z \geq t\}]\, dt}{\Pr(Z - a \leq 0)}$$

$$= a - \frac{\int_{0}^{\infty} \Pr(a - Z \geq t)\, dt}{\Pr(Z - a \leq 0)}$$

$$= a - \frac{\int_{-\infty}^{0} \Pr(Z \leq a + t)\, dt}{\Pr(Z - a \leq 0)}$$

$$= a - \frac{\int_{-\infty}^{a} \Pr(Z \leq t)\, dt}{\Pr(Z - a \leq 0)}$$

$$= a - \frac{F_1(a)}{F_0(a)}.$$

Proof of Proposition 1:(Cont.)

对于$E[Z^2] < \infty$，有

(B-5)

$$Var[Z|Z > a] = \frac{2S_2(a)}{S_0(a)} - \left(\frac{S_1(a)}{S_0(a)}\right)^2,$$

(B-6)

$$Var[Z|Z \leq a] = \frac{2F_2(a)}{F_0(a)} - \left(\frac{F_1(a)}{F_0(a)}\right)^2.$$

Proof of Proposition 1:(Cont.)

可以发现，如果Z的概率密度是对数凹（凸）的，那么(B-3)-B(6)是a的可微函数。求导得

(B-7)

$$\frac{\partial E[Z|Z>a]}{\partial a}=\frac{S_1(a)S_1''(a)}{\left(S_1'(a)\right)^2}\begin{cases}\leq 1 & \left(Z\text{对数凹}\right),\\ \geq 1 & \left(Z\text{对数凸}\right),\end{cases}$$

(B-8)

$$\frac{\partial E[Z|Z\leq a]}{\partial a}=\frac{F_1(a)F_1''(a)}{\left(F_1'(a)\right)^2}\leq 1 \quad \left(Z\text{对数凹}\right),$$

(B-9)

$$\frac{\partial Var[Z|Z>a]}{\partial a}=-\frac{\left(2S_0'(a)\right)}{S_0^2(a)}\left(S_2(a)-\frac{S_1^2(a)}{S_0(a)}\right)\begin{cases}\leq 0 & \left(Z\text{对数凹}\right),\\ \geq 0 & \left(Z\text{对数凸}\right),\end{cases}$$

(B-10)

$$\frac{\partial Var[Z|Z\leq a]}{\partial a}=\frac{\left(2F_0'(a)\right)}{F_0^2(a)}\left(-F_2(a)+\frac{F_1^2(a)}{F_0(a)}\right)\geq 0 \quad \left(Z\text{对数凹}\right).$$