The Empirical Content of the Roy Model

James J. Heckman       Bo Honoré
University of Chicago  Princeton University

I. The Roy Model

This section exposits the two sector log normal Roy model and establishes new properties of it. Incoming maximizing agents possess two skills $S_1 = s_1$ and $S_2 = s_2$ with associated positive skill prices $\pi_1$ and $\pi_2$.

An agent chooses sector one if his earnings are greater there: \textit{i.e.}

\begin{equation}
\pi_1 s_1 > \pi_2 s_2.
\end{equation}
The proportion of the population working in sector one, $P_1$, is the proportion for whom (1) is true:

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

The density of skill employed in sector one differs from the population density of skill. The latter density may be written as

$f_1(s_1) = \int_0^\infty f(s_1, s_2) ds_2.$

The former density is

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

Similarly, the density of skill employed
in sector 2 is:

\[(5) \, g(s_2 | \pi_1 s_1 > \pi_2 s_2) = \frac{1}{P_2} \int_0^{\pi_2 s_2 / \pi_1} f(s_1, s_2) ds_1.\]

We get the density of earnings in sector 1

(using \( w_1 = \pi_1 s_1 \)) as:

\[(6) \, g_1(w_1) = \frac{1}{P_1 \pi_1} \int_0^{w_1 / \pi_2} f(w_1 / \pi_1, s_2) ds_2.\]

The density of earnings in sector two is

\[(7) \, g_2(w_2) = \frac{1}{P_2 \pi_2} \int_0^{w_2 / \pi_1} f(s_1, w_2 / \pi_2) ds_1.\]

The overall density of earnings is then:

\[(8) \, g(w) = P_1 g_1(w) + P_2 g_2(w).\]

2. **Normal Roy Model**

The Normal Roy model assumes that
(lnS₁, lnS₂) are normally distributed with mean (µ₁, µ₂) and covariance matrix Σ.

Define (U₁, U₂) as a mean zero normal vector.

As a consequence of these assumptions and definitions we have”

(9) lnSᵢ = µ₁ + U₁

⇒ lnWᵢ = lnπᵢ + µᵢ + Uᵢ, i = 1, 2

is normally distributed.

Define σ* = [VAR(U₁ − U₂)]^{1/2}

$$cᵢ = (lnπᵢ/πⱼ + µᵢ − µⱼ)/σ^*, \quad i \neq j,$$

and let
\[ P_i = P(\ln W_i > \ln W_j) = 1 - \Phi(-c_i) = \Phi(c_i) \]

\[ i \neq j, \quad i, j = 1, 2. \]

\[ \ln W_i - \ln W_j = \ln(\pi_i/\pi_j) + \mu_i - \mu_j + U_i - U_j, \]

\[ i \neq j. \]

Let \( D_i = U_i - U_j \) and \( c_i^* = \ln(\pi_i/\pi_j) + \mu_i - \mu_j \). Then
Equation 10 is an important relation, describing the expected value of the log wages in sector \( i \), assuming that individuals have self-selected based on the rule described in the beginning of this section (ie choose \( i \) if \( \pi_i S_i > \pi_j S_j \)). To explore this further, we first note some results for the random variable \( D_1 \). We have:
\[ Var(D_i) = E(U_i - U_j)^2 = \sigma_{ii} + \sigma_{jj} - 2\sigma_{ij}, \]

\[ i \neq j, \quad i, j = 1, 2 \]

\[ = (\sigma^*)^2 \]

so that

\[ D_i \sim N(0, (\sigma^*)^2). \]

Conditional expectations of normal variables are linear. Accordingly, we may decompose \( U_i \) into the linear regression

(11)

\[ U_i = a_i D_i + V_i \]

where
\[ a_i = \frac{Cov(D_i, U_i)}{Var(D_i)} = \frac{\sigma_{ii} - \sigma_{ij}}{(\sigma^*)^2} \]

As a consequence of normality \( V_i \) is independent of \( D_i \). Further, \( E(V_i) = 0 \) and \( Var(V_i) = \sigma_{ii}(1 - \rho_i^2) \) where

\[ \rho_i = \frac{Cov(D_i, U_i)}{(Var(D_i)Var(U_i))^{1/2}} = \frac{\sigma_{ii} - \sigma_{ij}}{(\sigma^*)\sigma_{ii}^{1/2}}. \]

Thus

\[ V_i \sim N(0, \sigma_{ii}(1 - \rho_i^2)). \]

Then, as a consequence of (11) and using fact that \( V_i \perp \perp D_i \) we get from (10) that:

8
\begin{align*}
(12) \ E(U_i | U_i - U_j > -c_i^*) = E(U_i | D_i > -c_i^*) \\
& \quad = a_i E(D_i | D_i > -c_i^*) \\
\end{align*}

since \( E(V_i | D_i > -c_i^*) = E(V_i) = 0 \) because \( D_i \) and \( V_i \) are independent. Normalizing by \( \sigma^* \), we define

\[ Z_i = D_i/\sigma^* \text{ and } c_i = c_i^*/\sigma^* \]

where \( Z_i \) is a unit variance mean zero normal variable. Then

\[ E(D_i | D_i > -c_i^*) = \sigma^* E(Z_i | Z_i > -c_i). \]

Collecting results, we get:
(13) $E(\ell \ln W_i | \ell \ln W_i > \ell \ln W_j)$

$$= \ell n \pi_i + \mu_i + a_i E(D_i | D_i > -c_i^*)$$

$$= \ell n \pi_i + \mu_i + \left( \frac{\sigma_{ii} - \sigma_{ij}}{(\sigma^*)^2} \right) (\sigma^*) E(Z_i | Z_i > -c_i)$$

$$= \ell n \pi_i + \mu_i + \left( \frac{\sigma_{ii} - \sigma_{ij}}{\sigma^*} \right) E(Z_i | Z_i > -c_i)$$
\[(14) \ Var(U_i \mid U_i - U_j > -c_i^*) \]

\[= Var(a_i D_i + V_i \mid U_i - U_j > -c_i^*) \]

\[= Var(a_i D_i \mid U_i - U_j > c_i^*) + Var(V_i) \]

\[= a_i^2 Var(D_i \mid D_i > -c_i^*) + Var(V_i) \]

\[= (a_i \sigma^*)^2 Var(Z_i \mid Z_i > -c_i) + Var(V_i). \]

Using the fact that \((a_i \sigma^*)^2 = \sigma_{ii} \rho_i^2\), we conclude that

\[(15) \ Var(\ln W_i \mid \ln W_i > \ln W_j) \]

\[= Var(U_i \mid U_i - U_j > -c_j^*) \]

\[= \sigma_{ii} \left\{ \rho_i^2 Var(Z_i \mid Z_i > c_i) + (1 - \rho_i^2) \right\}. \]
Now we will list some useful facts for the expectation and variance of the standard normal conditional on it being greater than a given real number \( \{Z_i/Z_i > c\} \).

**Some Facts**

(P-1) \[
E(Z \mid Z > c) = \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{c^2}{2} \right) \Phi(-c) \equiv \lambda(c);
\]
\( \lambda(c) \geq 0 \) and \( \lambda(c) \geq c \).

(P-2) \[
Var(Z \mid Z > c) = 1 + \lambda(c)c - (\lambda(c))^2
\]

(P-3) \[
\lim_{c \to \infty} \lambda(c) = \infty
\]

(P-4) \[
\lim_{c \to -\infty} \lambda(c) = 0
\]

(P-5) \[
\frac{\partial \lambda(c)}{\partial c} > 0 \quad -\infty < c < \infty
\]

(P-6) \[
\lim_{c \to -\infty} \frac{\partial \lambda(c)}{\partial c} = 0
\]

(P-7) \[
\lim_{c \to \infty} \frac{\partial \lambda(c)}{\partial c} = 1
\]
(P-8) \[ 0 < \frac{\partial \lambda(c)}{\partial c} = \lambda'(c) = \lambda(c)(\lambda(c) - c) < 1, \quad -\infty < c < \infty \]

(P-9) \[ \frac{\partial^2 \lambda(c)}{\partial c^2} > 0, \quad c < \infty \]

(P-10) \[ \frac{\partial \text{Var}(Z \mid Z > c)}{\partial c} < 0, \quad c < \infty \]

(P-11) \[ \lim_{c \to \infty} \text{Var}(Z \mid Z > c) = 0 \]

(P-12) \[ \lim_{c \to -\infty} \text{Var}(Z \mid Z > c) = 1 \]

(P-13) \[ E([Z - \lambda(c)]^3 \mid Z > c) > 0, \quad c < \infty \]

and \[ \lim_{c \to \infty} E[[Z - \lambda(c)]^3 \mid Z > c] = 0 \]

so \( Z \) is positively skewed if \( c < \infty \)

mean of \( Z \geq \) mode of \( Z \).
Continuing from Equation 14 and utilizing the facts above, we can write:

\[(16) \quad E(lnW_i \mid lnW_i > lnW_j) = ln\pi_i + \mu_i + \left(\frac{\sigma_{ii} - \sigma_{ij}}{\sigma^*}\right) \lambda(-c_i), \quad i \neq j.\]

Since \(W_i = \pi_i S_i\) it also follows that

\[E(lnS_i \mid lnW_i) > lnW_j = \mu_i + \left(\frac{\sigma_{ii} - \sigma_{ij}}{\sigma^*}\right) \lambda(-c_i) \quad i \neq j.\]

Now \(0 \leq Correl(lnS_1, lnS_2) = \frac{\sigma_{12}^2}{\sigma_{11}\sigma_{22}} \leq 1\) so
\[
0 \leq \left( \frac{\sigma_{12}}{\sigma_{11}} \right) \left( \frac{\sigma_{12}}{\sigma_{22}} \right) \leq 1.
\]

Both terms in parentheses cannot exceed one.

Thus if \( \sigma_{11} < \sigma_{12} \) so \( 1 < \frac{\sigma_{12}}{\sigma_{11}} \) then \( \frac{\sigma_{12}}{\sigma_{22}} < 1. \)

The effect of an increase in \( \pi_i \) on mean skill in sector \( i \) is given by

\[
(17) \quad \frac{\partial E(\ln S_i | \ln W_i > \ln W_j)}{\partial \ln \pi_i} = \frac{(\sigma_{ii} - \sigma_{ij})}{(\sigma^*)^2} \lambda'(-c_i)
\]

\[
= \frac{(\sigma_{ii} - \sigma_{ij})}{(\sigma^*)^2} [\lambda(-c_i)(\lambda(-c_i) + c_i)].
\]

Is it possible for an increase in \( \pi_i \) to reduce the average wage paid in both sectors?

Yes, this can occur if

\[
\left( \frac{\sigma_{ii} - \sigma_{ij}}{\sigma^*} \right) \lambda'(-c_i) > 1
\]

15
(*)

since
\[
\frac{\partial E(\ln W_i | \ln W_i > \ln W_j)}{\partial \ln \pi_i} = 1 - \frac{(\sigma_{ii} - \sigma_{ij})}{(\sigma^*)^2} \lambda'(-c_i).
\]

Now for (*) to occur, a necessary condition
(since \(\lambda'(c) \in [0, 1] \forall c\), by fact (P-8) is that

\[
\frac{\sigma_{ii} - \sigma_{ij}}{(\sigma_{ii} - 2\sigma_{ij} + \sigma_{jj})} > 1.
\]

But we can pick this and make \(\lambda\) big enough
(alter means), so that it is possible to have
an increase in wages in one sector leading
to a reduction in average wages paid in both
sectors.
Now we look at some results for the variance in sector wages. From Equation 15 and the facts for $\lambda(c)$, we get:

$$(18) \quad \text{Var}(\ln W_i | \ln W_i > \ln W_j) = \sigma_{ii} \left\{ \rho_i^2 \left[ 1 - c_i \lambda(-c_i) \right] - \lambda^2(-c_i) \right\}$$

$$+ \left( 1 - \rho_i^2 \right).$$
Theorem 1: For any random assignment of persons to sectors, a Roy economy with the same proportion of persons in each sector as the random assignment economy has lower variance in log earnings provided some individuals are assigned to each sector.

We also look at one result for the skewness of the sector wages. We have expression for skewness of wage in sector $i$ as:

$$E \left\{ \left[ \ln W_i - E(\ln W_i \mid \ln W_i > \ln W_j) \right]^3 \mid \ln W_i > \ln W_j \right\}$$

$$= E \left\{ (a_i)^3[D_i - E[D_i \mid D_i > c_i^*]]^3 \mid D_i > -c_i^* \right\}$$

$$= (\sigma_{ii})^{3/2}(\rho_i)^3 E \left\{ [Z - \lambda(-c_i)]^3 \mid Z > -c_i \right\}. $$
Theorem 2: In a Roy economy, log earnings distributions are right skewed as long as some positive fraction of the population works in each sector. □
**Theorem 3:** The right tail of the Roy density of earnings $f(w)$ within sectors or in the overall economy are thinner than Pareto tails from density $g(w)$ in the sense that

$$\lim_{w \to \infty} \frac{f(w)}{g(w)} \to 0.$$  

We state without proof, some other results for the normal Roy Model (refer Heckman and Honore (1990) for proofs).

(i) Self-selection raises the mean of employed log skill $i$ above the population mean $\mu_i$ if $\sigma_{ii} > \sigma_{ij}$.

(ii) Self-selection reduces the mean of employed log skill $i$ below the population mean $\mu_i$ if $\sigma_{ii} < \sigma_{ij}$ (the “unusual case”).

(iii) The “unusual” case can arise in at most one sector.
(iv) In response to a 10% increase in \( \pi_i \), mean log wages in \( i \) rise by more than 10%, 10% or less than 10% depending on whether or not \( \sigma_{ii} \leq \sigma_{ij} \). If \( \sigma_{ii} > \sigma_{ij} \), a 10% increase in \( \pi_i \) may result in a decrease in sector \( i \) earnings.

(v) In response to an increase in \( \pi_i \), mean log wages in \( j \) rise unless the unusual case occurs in sector \( j \).

(vi) Self-selection reduces sectoral and aggregate earnings inequality (measured by the variance of log earnings) compared to a random assignment economy.

(vii) As \( \pi_1 \) increases, the variance of log earnings in sector \( i \) increases and the variance of log earnings in sector \( j \) decreases.

(viii) Self-selection produces right skewness in sector \( i \) log earnings distributions if \( \sigma_{ii} > \sigma_{ij}, \ i \neq j \). If \( \sigma_{ii} < \sigma_{ij} \) (the “unusual” case), sector \( i \) earnings
distributions are left skewed. Self-selection produces right skewness in the aggregate log earnings distribution with the exception of cases where all agents work in one sector.
3. Robustness of The Roy Model to Non-Normality

Here we drop the assumption that \((lnS_1, lnS_2)\) are normally distributed. We present some results and proofs in the non-normal case.

**Theorem 4.** For any skill distribution with income maximizing agents, selection from the bottom in earnings or skill distributions (the “unusual” case) can occur in at most one sector, i.e. if

(i) \(E(lnW_i \mid lnW_i > lnW_j) < E(lnW_i)\)

then

(ii) \(E(lnW_j \mid lnW_i \leq lnW_j) > E(lnW_j), \quad i \neq j.\)
It is also true that if

(iii) \( E(\ln S_i \mid \ln W_i \leq \ln W_j) > E(\ln S_i) \) then
(iv) \( E(\ln S_j \mid \ln W_j \leq \ln W_i) > E(\ln S_j), \quad i \neq j. \)

**Proof:** We shall prove by contradiction.

Assume that, contrary to assertion (i) and (ii):

(a) \( E(\ln W_i \mid \ln W_i > \ln W_j) \leq E(\ln W_i) \) and
(b) \( E(\ln W_j \mid \ln W_i \leq \ln W_j) \leq E(\ln W_j), \quad i \neq j \)

where the inequality is strict in at least one of the two cases. With no loss of generality we may assume that

\( E(\ln W_i) \geq E(\ln W_j), \quad i \neq j. \) Now, since:
\[ E(\ln W_i - \ln W_j \mid \ln W_i - \ln W_j \leq 0) \leq 0 \]

we may use (b) to prove that
\[ E(\ln W_i \mid \ln W_i < \ln W_j) \leq \]
\[ E(\ln W_j \mid \ln W_i < \ln W_j) \leq \]
\[ E(\ln W_j) \leq E(\ln W_i), \]
\[ i \neq j. \]

But then
\[ E(\ln W_i) = E(\ln W_i \mid \ln W_i > \ln W_j)P_i \]
\[ +E(\ln W_i \mid \ln W_j > \ln W_i)(1 - P_i) \]
\[ < E(\ln W_i), i \neq j, \]

since both expectations are less than \( E(\ln W_i) \).

But this is a contradiction. This proves (i) and (ii). To prove (iii) and (iv) note that
\[
E(\ln S_i \mid \ln W_i > \ln W_j)
\]
\[
= (\ln W_i - \ln \pi_i \mid \ln W_i > \ln W_j)
\]
\[
= -\ln \pi_i + E[\ln W_i \mid \ln W_i > \ln W_j]
\]
and \( E[\ln W_i - \ln \pi_i] = E[\ln S_i] \). Thus
\[
E[\ln S_i \mid \ln W_i > \ln W_j] =
\]
\[
E[\ln W_i - \ln \pi_i \mid \ln W_i > \ln W_j]
\]
\[
\ge E[\ln S_i] = E[\ln W_i - \ln \pi_i], \quad i \neq j,
\]
\[
\iff E[\ln W_i \mid \ln W_i > \ln W_j] \ge E[\ln W_i].
\]
Thus (i) and (ii) imply (iii) and (iv). \(\square\)

Other conclusions of the Roy model are
not robust when skills are not log normal. For example, the variance of log observed skill may exceed the population variance of log skill. Thus selection may increase sectoral inequality compared to the random sample case.

**Example:** Consider the following distribution of skills.

<table>
<thead>
<tr>
<th>$(\ln S_1, \ln S_2)$</th>
<th>$P(\ln S_1, \ln S_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(1, \frac{1}{2})$</td>
<td>$1/3$</td>
</tr>
<tr>
<td>$(3, \frac{3}{2})$</td>
<td>$1/3$</td>
</tr>
<tr>
<td>$(5, \frac{5}{2})$</td>
<td>$1/3$</td>
</tr>
</tbody>
</table>

Set $\pi_1 = \pi_2 = 1$. Then $\text{Var}(\ln S_1) = 2$ and $2/3$
Selection may also increase economy-wide inequality rather than decrease it as in the Roy model.

In the same vein, we previously established that as \( \pi_1 \) increases holding \( \pi_2 \) fixed, the variance of employed skills in one cannot decrease, and the variance of log skills employed in sector two cannot increase. But this is not true in general.

**Example:** Let \( \ln \pi_1 = \ln \pi_2 = 0 \)
initially. Then raise $\pi_1$ so $\ln \pi_1 = 1$. Let the
distribution be

\[
\begin{array}{ccc}
\text{(ln } S_1, \text{ln } S_2) & P(\text{ln } S_1, \text{ln } S_2) & \text{Sector Selected At Old Prices} \\
\left(1, \frac{1}{3}\right) & \frac{1}{5} & 1 \\
\left(5, \frac{1}{2}\right) & \frac{1}{5} & 1 \\
\left(1, \frac{3}{2}\right) & \frac{1}{5} & 2 \\
\left(1, \frac{5}{2}\right) & \frac{1}{5} & 2 \\
\left(3, \frac{3}{2}\right) & \frac{1}{5} & 2 \\
\end{array}
\]
At the old prices,

\[ Var(\ln W_1 \mid \ln W_1 > \ln W_2) = 4 \]

\[ Var(\ln W_2 \mid \ln W_2 \geq \ln W_1) = \frac{24}{27} \]

At the new prices,

\[ Var(\ln W_1 \mid \ln W_1 > \ln W_2) = 2 \quad \text{and} \quad \frac{2}{3} \]

\[ Var(\ln W_2 \mid \ln W_2 \geq \ln W_1) = 1. \]

Thus the variance of employed skills in sector one decreases and the variance of employed skills in sector two increases, both contrary to the result in the normal Roy model.

\[ \square \]
3.1 Log Concave Version of Roy Model

To define the class of models that preserves many of the features of the Roy model assume that $\ln S_i$ can be decomposed into two independent components, $D$ and $V_i$ so that

\begin{equation}
\ln S_i = a_i D + V_i \quad i = 1, 2
\end{equation}

where $E(V_i) = 0$ and $a_i$ is as defined in (11) and

\[ D = (\ln S_1 - \ln S_2). \]

We assume that $D$ is log concave.
**Definition:** A log concave random variable $D$ with density $f(d)$ is one for which

$\ln f(d)$ is a concave function of $d$.

From (19) and the definition of $D$ it follows that

$$(\ln S_1 - \ln S_2) = (a_1 - a_2)(\ln S_1 - \ln S_2) + V_1 - V_2$$

so that $V_1 = V_2$ with probability one and

$\alpha_1 = \alpha_2 + 1$. Thus the proposed model writes that each $\ln S_i$ can be decomposed into two components:

$$\ln S_1 = a_1(\ln S_1 - \ln S_2) + V$$
\ln S_2 = \alpha_2 (\ln S_1 - \ln S_2) + V
**Fact 1.** If $D$ is a log-concave random variable

(21) \[ 0 \leq \frac{\partial E(D \mid D > c)}{\partial c} \leq 1. \]
Fact 2. If $D$ is a log concave random variable then

$$\frac{\partial \text{Var}(D \mid D > c)}{\partial c} \leq 0.$$  \(\square\)
We also have the following relations:

(22) \( E(\ln S_1 \mid \ln W_1 > \ln W_2) = \alpha_1 E(D \mid D > -\ln(\frac{\pi_1}{\pi_2})) \)

\( E(\ln S_2 \mid \ln W_2 \geq \ln W_1) = a_2 E(D \mid D \geq -\ln(\frac{\pi_2}{\pi_1})) \)

and

(23) \( Var(\ln W_1 \mid \ln W_1 > \ln W_2) = \)

\( (a_1)^2 Var(D \mid D > -\ln(\frac{\pi_1}{\pi_2})) + Var(V) \)

\( Var(\ln W_2 \mid \ln W_2 > \ln W_1) = b^2 \)

\( Var(D \mid D \geq -\ln(\frac{\pi_2}{\pi_1})) + Var(V). \)
Fact 1': If $D$ is a log convex random variable

$$\frac{\partial E(D \mid D > c)}{\partial c} \geq 1.$$ 

See Appendix B of Heckman and Honoré for the proof. $\square$
Fact 2': If $D$ is a log convex random variable

$$\frac{\partial \text{Var}(D \mid D > c)}{\partial c} \geq 0.$$
4. The Identifiability of the Roy Model and Its Normal Extensions 4.1 Identifiability of The Log Normal Roy Model From Cross-Section Data

In the models considered in the previous sections, let:

\[ \mu_i = Z\beta, \quad i = 1, 2 \]

where \( \beta \) is a vector of parameters, and provided that sectoral choices are known, it is possible to identify \( \sum \beta_1 \) and \( \beta_2 \) from a single cross section using standard sample selection methods.
Theorem 5. Under the conditions postulated for the normal Roy model, \( \mu, \sum \) can be identified from data on wages paid in each sector and sectoral choices. More precisely, from knowledge of the sectoral earnings densities

\[ f(\ln w_i | \ln W_i > \ln W_j), i \neq j, i = j = 1, 2 \]

it is possible to identify \( \mu, \sum \).

Theorem 6. Under the conditions postulated for the normal Roy model, \( \mu, \sum \)
can be identified, except for their subscripts,
from knowledge of the aggregate earnings
density $g(lnw)$.
Theorem 7. If only $f(\ln w_1 \mid \ln W_1 > \ln W_2)$ and $P(\ln W_1 > \ln W_2)$ are known, in the normal Roy model it is possible to identify $\mu_1, \sigma_{11}, \sigma_{22}$ and either $\mu_2$ or $\sigma_{12}$ or some combination of the two parameters.

Proof: See Appendix B.
4.2. The Nonidentifiability of a General Non-Normal Roy Model in A Single Cross Section

**Theorem 8.** It is always possible to rationalize sectoral wage data in a single cross section by a two skill model. More precisely it is possible to rationalize data on

\[ f(s_1 \mid S_2 > S_2) = f_1(s_1)f_2(s_2). \]

**Proof:**

Define

\[ Q'_1(s) = \int_0^s f(s, s_2)ds_2 \]
and \[ Q'_2(s) = \int_0^s f(s_1, s) ds_1. \]

By hypothesis, \( Q'_{1}(s) \) and \( Q'_{2}(s) \) are observed.

Assume that skills are independent so
\[ \begin{align*}
Q'_{1}(s) &= f_{1}(s) F_{2}(s) \\
Q'_{2}(s) &= F_{1}(s) f_{2}(s).
\end{align*} \]

Define:
\[ \tilde{Q}(t) = \int_0^s [Q'_{1}(x) + Q'_{2}(x)] dx = F_{1}(s) F_{2}(s). \]

Now:
\[ - \int_s^\infty \frac{Q'_i(x)}{\tilde{Q}(x)} dx = - \int_s^\infty \frac{f_i(x)}{F_i(x)} dx = \ln F_i(s), \]
\( i = 1, 2 \)
because \( \lim_{s \to \infty} F_i(s) = 1 \). Thus:

\[
F_1(s) = \exp \left\{ - \int_s^\infty \frac{Q'_1(x)}{\tilde{Q}(x)} \, dx \right\}
\]

and by the same reasoning

\[
F_2(s) = \exp \left\{ - \int_s^\infty \frac{Q'_2(x)}{\tilde{Q}(x)} \, dx \right\}.
\]

Thus \( F_1 \) and \( F_2 \) are uniquely implied by the data and the independence assumption. But \( F_1, F_2 \) and the assumption of independence can always explain the data. Observe that \( Q'_1 \) and \( Q'_2 \) summarize the data.

Then

\[
\Pr(S_1 \leq t \mid S_1 > S_2) \Pr(S_1 > S_2)
\]
\[
= \int_0^t \int_0^{s_1} f(s_1, s_2) ds_2 ds_1 \\
= \int_0^t f_1(s_1) F_2(s_1) ds_1 \\
= \int_0^t \exp \left( - \int_{s_1}^\infty \frac{Q'_1(x)}{\tilde{Q}(x)} dx \right) \frac{Q'_1(s_1)}{\tilde{Q}(s_1)} \\
\exp \left( - \int_{s_1}^\infty \frac{Q'_2(x)}{\tilde{Q}(x)} dx \right) ds_1 \\
= \int_0^t \exp \left( - \int_{s_1}^\infty \frac{d\ln \tilde{Q}(x)}{dx} dx \right) \frac{Q'_1(s_1)}{\tilde{Q}(s_1)} ds_1 \\
= \int_0^t Q'_1(s_1) ds_1 = Q_1(t)
\]

so we can always reproduce \( Q_1(t) \). By a

similar argument \( Q_2(t) \) can be reproduced.

Thus without further restrictions it is possible

to rationalize any sectoral wage distributions

by an independent skill distribution.\( \square \)
**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 $\pi_2$ 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 $\pi_2$, but

$$\{\max \{S_1, \pi_2 S_2\} \leq x\} = \{S_1 \leq \}$$

50
\[ x, \pi_2 S_2 \leq S_1 \} \cup \]

\[ \{ \pi_2 S_2 \leq x, S_1 \leq \pi_2 S_2 \} = \{ S_1 \leq x, \]

\[ S_2 \leq x/\pi_2 \}, \]

so for any \( s_1, s_2 > 0 \) we have

(-setting \( x = s_1 \) and \( \pi_2 = s_1/s_2 \),)

\[ F(s_1, s_2) = \Pr(S_1 \leq s_1, S_2 \leq s_2) \]

\[ = \Pr(S_1 \leq s_1, S_2 \leq s_1/\pi_2) \]

\[ = \Pr(\max\{S_1, \pi_2 S_2\} \leq x), \]

which completes the proof. Q.E.D.

51
**Theorem 10:** If we observe the distribution of $Z$ given by

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

and $\pi_2$ 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 $\epsilon > 0$ be given. For given $\pi_2$, we know the probability of events of the type

$$\{S_1 < \pi_2 S_2 \leq x\} \text{ for all } x \in (0, \infty).$$

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.
1. 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 \( \mu \) satisfying
\[ F(s_1, s_2) \leq \mu \leq F(s_1 + \epsilon, s_2). \]

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

This theorem can be proved for a general multiple sector economy in which the wage is not observed in one sector (available on request from the authors).

Thus far we have assumed access only to repeated cross-section data. Access to panel data on earnings that follows the same persons over time greatly facilitates identification if
individual skills do not change over time.

**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 $\Pr(Z' \leq 56\ldots$
\( z', Z \leq z \) 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
\[ F(s_1, s_2) = \Pr(S_1 \leq s_1, S_2 \leq s_2) \]

\[ = \Pr(S_1 \leq s_1, S_1 \leq \pi'_2 s_2, S_2 \leq s_2, S_2 \leq s_1/\pi_2) \]

\[ = \Pr(S_1 \leq s_1, \pi'_2 S_2 \leq s_1, S_1 \leq \pi'_2 s_2, \pi'_2 S_2 \leq \pi'_2 s_2) \]

\[ = \Pr(\max\{S_1, \pi_2 S_2\} \leq s_1, \max\{S_1, \pi'_2 S_2\} \leq \pi'_2 s_2) \]

which by hypothesis, is known. Hence

\[ F(s_1, s_2) \text{ is identified for all } \pi_2 s_2 \leq s_1 \leq \pi'_2 s_2. \quad \text{Q.E.D.} \]
Theorem 12: Let \( S_1 = g_1(X_1, X_0) + \epsilon_1 \)

and

\[ S_2 = g_2(X_2, X_0) + \epsilon_2 \]

where \( \epsilon_1, \epsilon_2 \) is independent of \( (X_0, X_1, X_2) \). Assume that

(a) \( \epsilon_1, \epsilon_2 \) is continuously distributed with distribution function \( G \) and support equal to \( \mathbb{R}^2 \); (b) support \( (g_1(X_1, x_0), g_2(X_2, x_0)) = \mathbb{R}^2 \)

for all \( x_0 \) in the support of \( X_0 \); (c) the marginal distributions of \( \epsilon_1 \) and \( \epsilon_2 \) both have medians equal to 0. Then \( g_1, g_2, \) and \( G \) are identified.
Proof: By assumption we know

(A) \( \Pr(S_1 > S_2) = \)

\( \Pr(g_1(x_1, x_0) + \epsilon_1 > g_2(x_2, x_0) + \epsilon_2), \)

(B) \( \Pr(S_1 \leq y, S_1 > S_2) = \Pr(g_1(x_1, x_0) + \epsilon_1 \leq y, g_1(x_1, x_0) + \epsilon_1 > g_2(x_2, x_0) + \epsilon_2), \)

(C) \( \Pr(S_2 \leq y, S_2 \geq S_1) = \Pr(g_2(x_2, x_0) + \epsilon_2 \leq y, g_2(x_2, x_0) + \epsilon_2 > g_1(x_1, x_0) + \epsilon_1), \)

for all \((x_0, x_1, x_2)\) in the support of \((X_0, X_1, X_2)\)

and for all \(y\).

60
Fix $x_0$. Let $\tilde{x}_1$ and $\tilde{x}_2$ be in the support of $X_1$ and $X_2$, respectively. From (A), we can then find
\[(x_1, x_2) : \Pr(g_1(x_1, x_0) + \epsilon_1 > g_2(x_2, x_0) + \epsilon_2)\]
\[= \Pr(g_1(\hat{x}_1, x_0) + \epsilon_1 > g_2(\tilde{x}_2, x_0) + \epsilon_2)\]
\[= \{(x_1, x_2) : g_1(x_1, x_0) + l = g_2(x_2, x_0)\}\]
for some unknown constant \(l\),

For any point in that set we can use (B) to find
\[Pr(g_1(x_1, x_0) + \epsilon_1 \leq y, \epsilon_1 > \epsilon_2 + l)\]
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 $\epsilon_1$ and $\epsilon_2$ are zero. Having determined the location of $G$, we can determine the additive constants in $g_1(\cdot, x_0)$ and $g_2(\cdot, x_0)$.

Since $x_0$ was arbitrary, this completes the proof. □
Appendix A
Proofs of Results (R-1) to (R-10)

The moment generating function for a truncated normal distribution with truncation point \(d\) is

\[
mgf(\beta) = e^{\beta/2} \frac{\int_{d-\beta}^{\infty} \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1}{2} u^2 \right) du}{\int_{d}^{\infty} \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1}{2} u^2 \right) du}.
\]

The equality in (R-1) follows from

\[
\lambda(d) = E[Z \mid Z > d] = \frac{\partial mgf}{\partial \beta} \bigg|_{\beta=0}.
\]

The inequality is obvious. By direct calculation,

\[
\lambda'(d) = \lambda(d)(\lambda(d) - d).
\]

Now note that
\[ E[Z^2 | Z > d] = \frac{\partial^2 mgf}{\partial \beta^2} |_{\beta=0} = 1 + \lambda(d)d \]

\[ E[Z^3 | Z > d] = \frac{\partial^3 mgf}{\partial \beta^2} |_{\beta=0} = \lambda(d)(2 + d^2). \]

Therefore

\[ \text{Var}[Z | Z > d] = 1 + \lambda^2(d) = 1 - \frac{\partial \lambda(d)}{\partial d}. \]

As \( \text{Var}[Z | Z > d] > 0 \) and \( \lambda(d)(\lambda(d) - d) > 0 \) by (R-1), this proves (R-2) and (R-4).

To prove (R-3) notice that \( \text{Var}[Z | Z > d] = 1 - \frac{\partial \lambda(d)}{\partial d} \), and therefore

\[ \frac{\partial^2 \lambda(d)}{\partial d^2} = -\frac{\partial \text{Var}[Z | Z > d]}{\partial d} > 0 \]

where the inequality follows from Proposition 65.
1.
(R-5) follows from Proposition 1, whereas (R-6) follows by direct calculation from the expression for

\[ E[(Z - \lambda(d))^3 | Z > d]. \] (R-7) is trivial.

(R-8) is obvious. The first part of (R-9) follows directly from \( \ell' \) Hospital’s rule. (R-2) and (R-3) imply that \( \frac{\partial \lambda(d)}{\partial d} \) is increasing and bounded by 1. Therefore \( \lim_{d \to \infty} \frac{\partial \lambda(d)}{\partial d} \) exists and does not exceed 1. If

\[ \lim_{d \to \infty} \frac{\partial \lambda(d)}{\partial d} < 1 \]

then \( \lambda(d) \) would eventually be less than \( d \),

67
contradicting (R-1). (Details of this proof are available on request from the authors.) This proves the second part of (R-9). (R-9) and (R-4) imply (R-10).
Appendix B: Total Positivity, Log Convexity and Log Concavity

We use results from the theory of total positivity (Karlin, 1968). A function \( L(x, y) \) of two real variables is TP\(_2\) if \( L(x, y) \geq 0 \) for all \( x, y \) and if for all \( x_1 < x_2, y_1 < y_2 \)

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

By a theorem of Polya and Szego (1972, Vol. 1, part II, problem 68) as used by Karlin (1968, p. 17), if \( K(x, q) \) is TP\(_2\) and \( L(q, y) \) is TP\(_2\) then the composition formula states that
\[ M(x, y) = \int_{q \in Q} K(x, q) L(q, y) dq \]

\((Q\) is an interval that does not depend on \(x\) and \(y\)) is \(TP_2\). Note that

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

and

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

are both \(TP_2\).

Let \(J\) be a real valued function and define \(L(q, y) = J(q - y)\). \(L(q, y)\) \(TP_2\) is then equivalent to \(J\) log concave. Likewise

\[ L(q, y) = J(q + y) TP_2, \quad q \geq 0, \quad y \geq 0 \]
is equivalent to $J$ log convex. If $J(z)$

is log concave (convex), positive and twice differentiable then $J(q - y) \, TP_2(J(q + y) TP_2$

is equivalent to

$$\frac{J''}{J} - \left( \frac{J'}{J} \right)^2 \leq 0$$

$$(J(q - y) TP_2)$$

$$(J(q + y) TP_2)$$

or for $J' \neq 0$

$$\frac{J'' J}{(J')^2} \leq 1$$

$$(J(q - y) TP_2)$$

$$(J(q + y) TP_2)$$

Setting $K(x, q) = R(x, q)$ or $K(x, q) = R^*(x, q)$, we obtain that for $x = a$, a constant,

the composition formula implies that

71
(B-1) \( M(a, y) = \int_{q \in Y} R(a, q) J(q \pm y) dq = \int_{\{q | q \leq a, q \in Y\}} J(q \pm y) dq \)

(B-2) \( M^*(a, y) = \int_{q \in Y} R^*(a, q) J(q \pm y) dq = \int_{\{q | q \leq a, q \in Y\}} J(q \pm y) dq \)

are TP_2 if \( J(q \pm y) \) is TP_2 (\( Y \) is the range of definition of \( J \)).

If \( Z \) has a log concave density with support in \((m,n)\), then if either \( m \) or \( n \) is finite we can extend \( Z \) to \((-\infty, \infty)\) by defining
\( f(z) = 0 \) for \( z \in (-\infty, m] \) and \( f(z) = 0 \) for \( z \in [n, \infty) \). The redefined density is log concave. The same extension of the support for a log convex model does not produce a log convex function. It is not possible to have a log convex random variable defined with support on the interval \((-\infty, \infty)\). We follow convention and assume that the support of \( Z \) is \([0, \infty)\).

For the case of log concave random variable \( Z \) with density \( J \) we may use the
extension argument to write (B-1) and (B-2) as

\[ (B-1)' \quad M(a, y) = N(a - y) = \int_{-\infty}^{a-y} J(t)dt \]

\[ (B-2)' \quad M^*(a, y) = N^*(a - y) = \int_{a-y}^{\infty} J(t)dt. \]

Thus by the Polya-Szego theorem \( \Pr(Z \leq a) \) and \( \Pr(Z \geq a) \) are log concave functions of \( a \). By the same reasoning

\[ \int_{-\infty}^{b} \Pr(Z \leq a)da \quad \text{and} \quad \int_{b}^{\infty} \Pr(Z \geq a)da \]

are log concave in \( b \) provided the integrals...
exist. The argument may be repeated for integrals of these integrals and so forth.

For the case of $J$ log convex with the support of $Z \geq 0$, we may write

$$\int_{a+y}^{\infty} \int J(t) \, dt$$

and use the Polya-Szego theorem to prove that $N^*(a + y)$ is TP$_2$ and hence log convex in $a$. Thus $\Pr(Z \geq a)$ is log convex in $a$. The analogous expression corresponding to (B-1)'
\[ M(a, y) = \int_{0}^{a} J(q + y) dq = \int_{y}^{a+y} J(t) dt \]

is \( TP_2 \). However, this does not imply that \( \Pr(Z \leq a) \) is log convex in \( a \). Thus we cannot assert that \( \Pr(Z \leq a) \) is log convex in \( a \). This is so because \( M(a, y) \) is a function of both \( a \) and \( y \) and not just \( a + y \).

For example, if the density of \( Z \) is Pareto with

\[ f(z) = (\gamma - 1)z^{-\gamma}, \quad z \geq 1 \quad \gamma > 1, \]

\[ \Pr(Z > a > 1) = a^{1-\gamma} \] which is log convex.
in \( a \). But \( \Pr(Z \leq a) = 1 - a^{1-\gamma} \) is log convex in \( a \) only if \( a < (\gamma)^{-1/(\gamma-1)} \) which is not always true (e.g. \( \gamma = 2, a = 3 \)).

To derive the results stated in the text, it is useful to define

\[
S_j + 1(a) = \int_a^\infty S_j(z) \, dz
\]

where

\[
S_0 = S(a) \text{def} P(z > a)
\]

and

\[
F_j + 1(a) = \int_{-\infty}^a F_j(z) \, dz
\]

where \( F_0 = F(a) \text{def} P(z \leq a) \).

Then for \( E \, | \, Z \, | < 0 \), using integration by parts,

\[
E(Z \, | \, Z > a) = a + \frac{S_1(a)}{S_0(a)}
\]
(B-4) \[ E(Z \mid Z \leq a) = a - \frac{F_1(a)}{F_0(a)}. \]

For \( E(Z^2) < \infty \), integration by parts produces

(B-5) \[ Var(Z \mid Z > a) = \frac{2S_2(a)}{S_0(a)} - \left(\frac{S_1(a)}{S_0(a)}\right)^2. \]

(B-6) \[ Var(Z \mid Z \leq a) = \frac{2F_2(a)}{F_0(a)} - \left(\frac{F_1(a)}{F_0(a)}\right)^2. \]

Observe that if the density of \( Z \) is log convex (concave), then (B-3) - (B-6) are differentiable with respect to \( a \) almost everywhere with respect to \( Z \). Taking
derivatives,

\[
(B-7) \quad \frac{\partial \text{Var}(Z \mid Z > a)}{\partial a} = \frac{S_1(a)S_1''(a)}{(S_1'(a))^2} \leq 1 \quad (Z \text{ log concave})
\]

\[
\geq 1 \quad (Z \text{ log convex})
\]

\[
(B-8) \quad \frac{\partial \text{Var}(Z \mid Z > a)}{\partial a} = \frac{\partial \text{Var}(Z \mid Z > a)}{\partial a} \leq 1 \quad (Z \text{ log concave})
\]
\[
\text{(B-9)} \quad \frac{\partial \text{Var}(Z \mid Z > a)}{\partial a} = -\frac{(2S'_0(a))}{S^2_0(a)} \left( S_2(a) - \frac{S^2_1(a)}{S_0(a)} \right) \leq 0
\]

\[ (Z \log \text{concave}) \]

\[ \geq 0 \quad (Z \log \text{concave}) \]

\[
\text{(B-10)} \quad \frac{\partial \text{Var}(Z \mid Z \leq a)}{\partial a} = \frac{2F'_0(a)}{F^2_0(a)} \left( -F_2(a) + \frac{F^2_1(a)}{F_0(a)} \right)
\]

\[ \geq 0 \quad (Z \log \text{concave}). \]

Note that the hypothesis that the density of \( Z \) is log concave (convex) is \textit{stronger} than is required to produce the orderings in (B-7) - (B-10). If \( S_1(a) \) is log concave (log convex) the slope of the truncated mean (B-3) is less
than or equal to (greater than or equal to) one. 

$S_1(a)$ can be log concave or convex without the density of $Z$ being log concave or convex. Similarly the log concavity of $F_1(a)$ is all that is required in (B-8). By parallel logic, propositions (B-9) and (B-10) hinge on the log convexity (concavity) of $S_2(a)$ and log concavity of $F_2(a)$. The inequalities are strict if convexity (concavity) is strict. Note that if $Z$ does not have a log concave density, then $S_1(a)$ may be log concave while $F_1(a)$ need
not be and vice versa.

Note that if $Z$ were defined to be log convex only on the support $[0,m]$, $m < ∞$, (B-7) and (B-9) or (20) and (21) in the text need not be true. For example if $Z$ is $U(0, 1)$, it is log convex (but not strictly) on the support $[0,1]$ but for $0 \leq a \leq 1$

$$\frac{∂E(Z \mid Z > a)}{∂a} = \frac{1}{2} < 1$$

and

$$\frac{∂Var(Z \mid Z > a)}{∂a} = \frac{a - 1}{6} \leq 0.$$  

Partial versions of the results given here for the log concave case are presented by