La sélection et le modèle de Roy

Sam Gyetvay

ECO8000

September 14, 2025

Sélection

Idée omniprésente en théorie économique : les individus font des choix basés sur un comportement d'optimisation

- Choisir l'emploi qui offre le salaire le plus élevé

Pourquoi les gens font-ils des choix différents ?

- Les avocats sont mieux payés que les économistes
- Pourquoi tous les économistes ne deviennent-ils pas avocats ?
- Avons-nous fait une erreur ?

Les individus sont différents

- Ceux qui choisissent d'être économistes sont bons en maths et aiment coder
- Ceux qui choisissent d'être avocats sont bons en lecture et aiment débattre
- Certains économistes seraient de mauvais avocats, et inversement...

Pourquoi est-ce important ?

Il existe de grandes différences de revenus entre groupes de travailleurs

- Les travailleurs en milieu rural gagnent moins que ceux en ville
- Les travailleurs avec un bac gagnent plus que ceux avec un diplôme de CÉGEP
- Les travailleurs dans des entreprises syndiquées gagnent plus que ceux dans des entreprises non syndiquées

Pourrait-on réduire les inégalités salariales en déplaçant les travailleurs ruraux vers les villes, en obligeant tous les étudiants de CÉGEP à obtenir un baccalauréat, et en transférant tous les travailleurs non syndiqués vers des entreprises syndiquées ?

Si les différences sont causées par la sélection, ces politiques *augmenteraient* les inégalités !

- Les travailleurs ont choisi rural/CÉGEP/non-syndiqué parce qu'ils gagnaient un salaire plus élevé que s'ils étaient urbains/universitaires/syndiqués

SOME THOUGHTS ON THE DISTRIBUTION OF EARNINGS[1]

By A. D. ROY

I

AN attempt has been made elsewhere[2] to show that the output of any individual working by hand is the resultant of a large number of random influences. As a first approximation these influences can be assumed to operate independently, i.e. they are not significantly associated with one another. The rather vague term 'influence' is intended to refer to such factors as health, strength, skill, and so on. The suggestion was made that it is more fruitful to define such factors so that, taken singly, they exercise the same proportionate effect on the output of otherwise similarly situated individuals rather than the same absolute effect. In other words, it is more reasonable to say that a given loss of health will depress a worker's output by, say, 10 per cent., other things being equal, than by, say, 10 units.

Modèle de Roy

Le modèle de Roy nous fournit un cadre qui permet d'analyser la sélection

Il produit des prédictions puissantes reliant la distribution des compétences à la nature de la sélection

- Comment la dispersion des compétences affecte-t-elle la sélection ?
- Comment la corrélation entre différentes compétences affecte-t-elle la sélection ?

Questions clés en économie du travail :

- Les travailleurs qui choisissent d'immigrer sont-ils plus productifs que ceux qui choisissent de rester dans leur pays d'origine ?
- Les femmes qui choisissent de travailler sont-elles plus productives sur le marché du travail que celles qui n'y entrent pas ?

Plan

1. Motivation # 1 : assimilation salariale des immigrants
2. Motivation # 2 : offre de travail féminine et écart salarial entre les sexes
3. Modèle de Roy : chasse et pêche
4. Modèle linéaire normal de Roy
5. Borjas (1987)
6. Mulligan et Rubinstein (2008)
7. Discussion: **Abramitzky, Boustan et Eriksson (2012)**

Motivation #1 : assimilation salariale des immigrants

Les immigrants sont moins bien payés que les natifs comparables à leur arrivée. Leurs salaires rattrapent-ils ceux des natifs à mesure qu'ils accumulent de l'expérience dans le nouveau pays ? Si oui, à quelle vitesse ?

Chiswick (1978) a tenté de répondre à cette question avec des données du recensement américain de 1970.

Le recensement de 1970 est une **coupure transversale unique** (anglais: **single cross-section**). Il enregistre un « instantané » des individus à un moment donné.

Le jeu de données compilé par Chiswick ressemblait à ceci :

ID	Immigrant	AnnéesDepuisImmig	Expérience	Revenu	Éducation
1	Oui	2	5	18,000	<HS
2	Non	-	12	25,000	HS
3	Oui	10	15	32,000	CÉGEP
4	Non	-	20	40,000	Bac
5	Oui	25	30	45,000	Bac

Chiswick (1978)

Chiswick a estimé la régression suivante

$$\log(Revenu_i) = X_i'\theta + \beta Exp_i + \delta Imm_i + \alpha YSM \times Imm_i + \epsilon_i$$

et a obtenu $\hat{\beta} = 0.03$, $\hat{\delta} = -0.2$, et $\hat{\alpha} = 0.01$.

Comment interpréter ces coefficients ?

Chiswick (1978)

Chiswick a estimé la régression suivante

$$\log(Revenu_i) = X_i'\theta + \beta Exp_i + \delta Imm_i + \alpha YSM \times Imm_i + \epsilon_i$$

et a obtenu $\hat{\beta} = 0.03$, $\hat{\delta} = -0.2$, et $\hat{\alpha} = 0.01$.

Comment interpréter ces coefficients ?

- β : effet d'une année d'expérience supplémentaire sur le revenu logarithmique des natifs ($+1$ an $\rightarrow +3\%$ de revenu)

Chiswick (1978)

Chiswick a estimé la régression suivante

$$\log(Revenu_i) = X_i'\theta + \beta Exp_i + \delta Imm_i + \alpha YSM \times Imm_i + \epsilon_i$$

et a obtenu $\hat{\beta} = 0.03$, $\hat{\delta} = -0.2$, et $\hat{\alpha} = 0.01$.

Comment interpréter ces coefficients ?

- β : effet d'une année d'expérience supplémentaire sur le revenu logarithmique des natifs ($+1$ an $\rightarrow +3\%$ de revenu)
- δ : différence de revenu logarithmique entre immigrants et natifs à l'arrivée (20% de moins)

Chiswick (1978)

Chiswick a estimé la régression suivante

$$\log(Revenu_i) = X_i'\theta + \beta Exp_i + \delta Imm_i + \alpha YSM \times Imm_i + \epsilon_i$$

et a obtenu $\hat{\beta} = 0.03$, $\hat{\delta} = -0.2$, et $\hat{\alpha} = 0.01$.

Comment interpréter ces coefficients ?

- β : effet d'une année d'expérience supplémentaire sur le revenu logarithmique des natifs ($+1$ an $\rightarrow +3\%$ de revenu)
- δ : différence de revenu logarithmique entre immigrants et natifs à l'arrivée (20% de moins)
- α : effet d'une année supplémentaire depuis l'immigration sur le revenu logarithmique des immigrants ($+1$ an $\rightarrow +1\%$ de revenu)

Comme chaque année depuis l'immigration apporte aussi une année d'expérience supplémentaire, les revenus des immigrants augmentent 1% plus vite que ceux des natifs. (Question bonus : combien de temps avant qu'ils rattrapent ?)

Borjas c. Chiswick

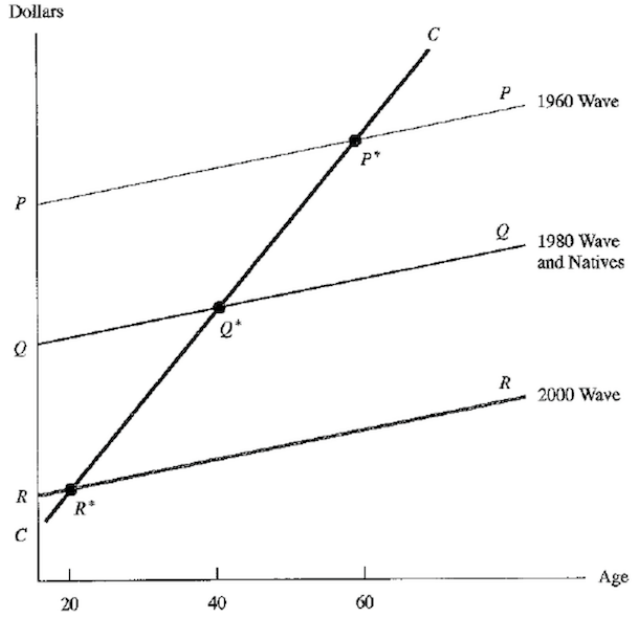
Dans un article de 1985, George Borjas a critiqué la conclusion de Chiswick.

Dans la régression de Chiswick, le coefficient sur les années depuis l'immigration (α) reflète une combinaison de deux effets :

1. Effet réel d'assimilation salariale
2. Différences de productivité fixes entre cohortes

Comme on n'observe qu'un seul « instantané » de chaque cohorte à un moment donné, on ne peut pas distinguer ces deux effets.

Si la productivité des cohortes d'immigrants décline au fil du temps, cela conduirait à une surestimation de l'assimilation.


Dollars
C
P
1960 Wave
P*
P
Q
1980 Wave
and Natives
Q*
Q
R
2000 Wave
R
R*
C
Age
20
40
60

Borjas (1985)

Borjas a combiné les recensements de 1970 et 1980 pour créer un ensemble de données en **coupures transversales répétées** (anglais: **repeat cross-section**).

Cela lui permet d'observer les membres d'une même cohorte d'immigrants à deux périodes différentes. Il peut ainsi estimer des valeurs distinctes de δ pour chaque cohorte :

$$\log(Revenu_{it}) = X_i'\theta_t + \beta Exp_i + \delta_c Imm_i + \alpha YSM \times Imm_i + \epsilon_{it}$$

En effet, Borjas a montré que Chiswick surestimait fortement la croissance des revenus des immigrants, et que les valeurs de δ_c pour les cohortes plus récentes déclinaient.

Dans sa conclusion, Borjas avance des hypothèses sur la raison du déclin des revenus initiaux des cohortes successives d'immigrants :

> *Bien qu'une partie puisse être expliquée par une baisse de la demande de main-d'œuvre immigrée, les résultats sont compatibles avec l'hypothèse d'un* ***déclin séculaire de la qualité des cohortes d'immigrants****.*

...mais qu'est-ce qui a causé ce déclin de la « qualité » (productivité) des immigrants ?

Borjas (1987)

Dans un article subséquent dans l'American Economic Review, Borjas propose une réponse simple et convaincante : la sélection.

Les immigrants ne sont pas un échantillon aléatoire de la population mondiale. Pour immigrer, il faut d'abord choisir de le faire. Pourquoi certains choisissent-ils et pas d'autres ? Comment les candidats à l'immigration décident-ils ?

Borjas :

> *On suppose que les individus comparent leurs revenus potentiels aux États-Unis avec leurs revenus dans leur pays d'origine, et prennent leur décision de migrer en fonction de ces différentiels (nets des coûts de mobilité).*

Cette citation résume l'essence du modèle de Roy.

Comment la sélection peut-elle expliquer l'évolution de la qualité entre cohortes d'immigrants ?

Avant de voir l'argument de Borjas, il nous faut maîtriser le modèle de Roy.

Mais d'abord, une autre motivation...

Motivation #2 : écart salarial entre les sexes

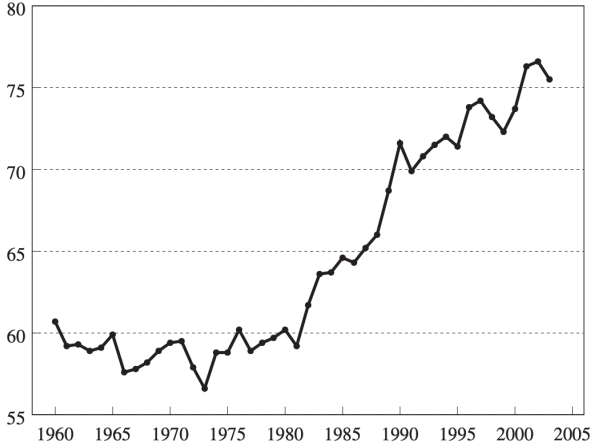


FIGURE 7. WOMEN'S EARNINGS AS A PERCENTAGE OF MEN'S EARNINGS: 1960 TO 2003

Participation au marché du travail selon le sexe (Goldin, 2006)

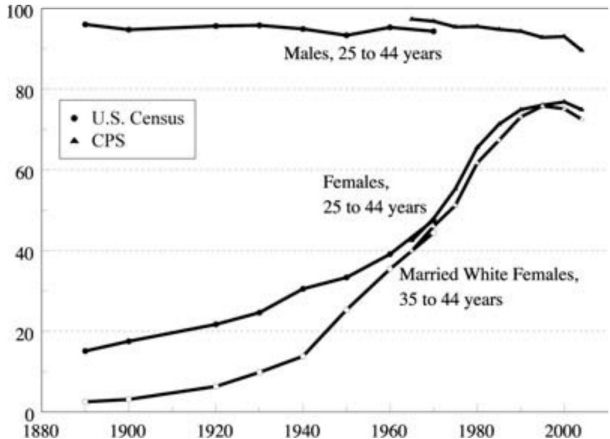


FIGURE 1. LABOR FORCE PARTICIPATION RATES FOR FEMALES AND MALES BY AGE AND MARITAL STATUS: 1890 TO 2004

Sélection et écart salarial entre les sexes

L'écart salarial entre hommes et femmes s'est réduit depuis les années 1960. Cependant, la proportion de femmes dans la population active a aussi beaucoup augmenté pendant cette période.

On n'observe que les revenus des femmes qui travaillent, mais on s'intéresse aussi aux salaires potentiels des femmes qui ne travaillent pas.

Si, dans les années 1960, ce sont surtout les femmes à haut salaire potentiel qui choisissaient de ne pas travailler, alors la baisse de l'écart salarial pourrait être due uniquement à la sélection, et non à une rémunération plus équitable.

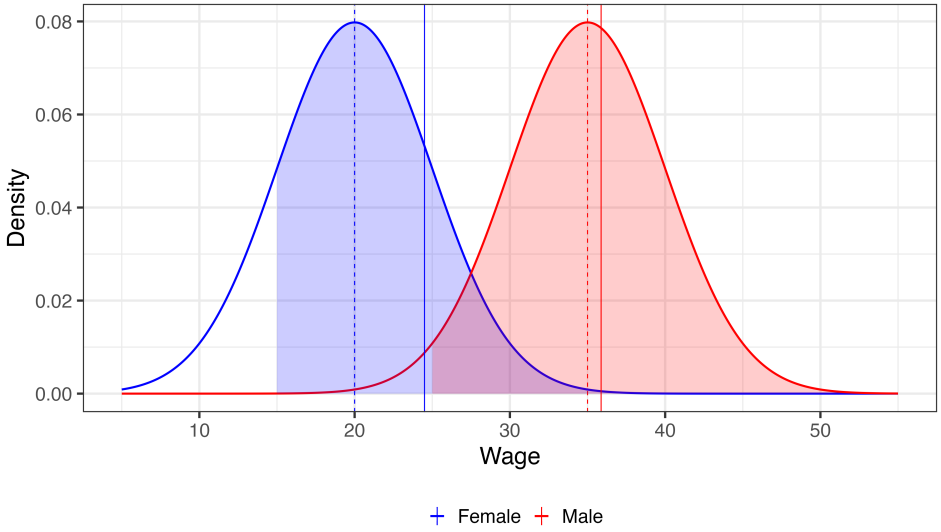
On peut utiliser la régression linéaire pour contrôler certaines différences observables. Mais les ensembles de données n'incluent que quelques variables (âge, éducation), et les compétences non observées expliquent une grande partie des écarts de revenu.

La quasi-totalité des hommes en âge actif travaillent, donc le « biais de sélection » nous préoccupe moins pour eux.

Gronau (1974)

Plus de 95 pour cent des hommes en âge actif (25-55) participent au marché du travail, alors que le taux de participation des femmes, en particulier des femmes mariées, dépasse rarement 55 pour cent. ***La différence de participation peut être expliquée par la distribution plus faible des offres de salaires auxquelles font face les femmes****, et probablement par la valeur plus élevée de leur temps en l'absence d'opportunités de marché. Ces deux facteurs tendent à* ***accroître l'écart entre le salaire minimum acceptable et le salaire moyen offert.***

On a estimé que le salaire horaire moyen des femmes, ajusté pour [race], scolarité, taille de la ville, statut marital, catégorie professionnelle et durée du trajet, représentait en 1959 seulement les deux tiers de celui des hommes hors agriculture. ***Étant donné le « biais de sélection », il semble que ce chiffre sous-estime l'écart salarial « réel » entre hommes et femmes.***


0.08
0.06
0.04
0.02
0.00
Density
10
20
30
40
50
Wage
Female
Male

Coïncidence ?

Axe Y = écart salarial femmes, Axe X = inégalités parmi les hommes

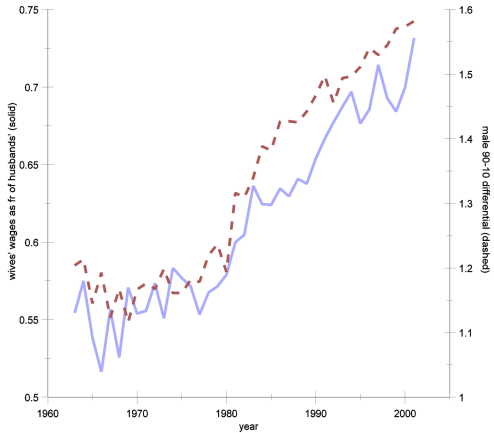


Figure 1 Wage Inequality between and within Genders

… pensez-y :-)

Le monde de Roy

Nous entrons maintenant dans le monde du modèle de Roy.

Dans le monde de Roy, nous supposons :

- Les travailleurs sont très différents de façon inobservable
- Les salaires ne changent pas et chacun obtient son emploi préféré
- Pas de « chance »
- Pas de chômage

Dans le monde de Roy, il n'y a que deux professions : chasseur de lapins et pêcheur.

Mise en place

- R : chasse (aux lapins)
- F : pêche (aux poissons)
- π_R : prix des lapins
- π_F : prix des poissons
- R_i : nombre de lapins attrapés par le travailleur i
- F_i : nombre de poissons attrapés par le travailleur i
- $w_{Ri} = \pi_R R_i$: salaire que i reçoit s'il choisit la chasse
- $w_{Fi} = \pi_F F_i$: salaire que i reçoit s'il choisit la pêche
- $\log w_{Ri} = \log \pi_R + \log R_i$
- $\log w_{Fi} = \log \pi_F + \log F_i$

Pour simplifier, $\log \pi_R = \log \pi_F = 0$

Direction de la sélection

Les chasseurs sont-ils meilleurs à la pêche que les pêcheurs ?

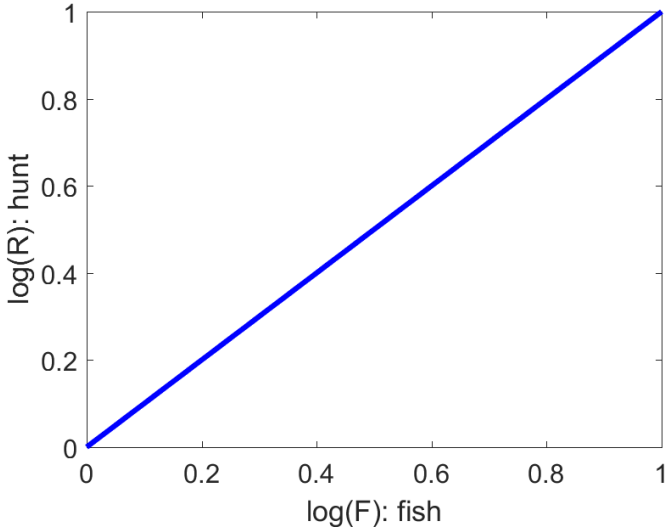
- Si oui, **sélection positive dans la chasse**

Les pêcheurs sont-ils moins bons à la chasse que les chasseurs ?

- Si oui, **sélection négative dans la pêche**

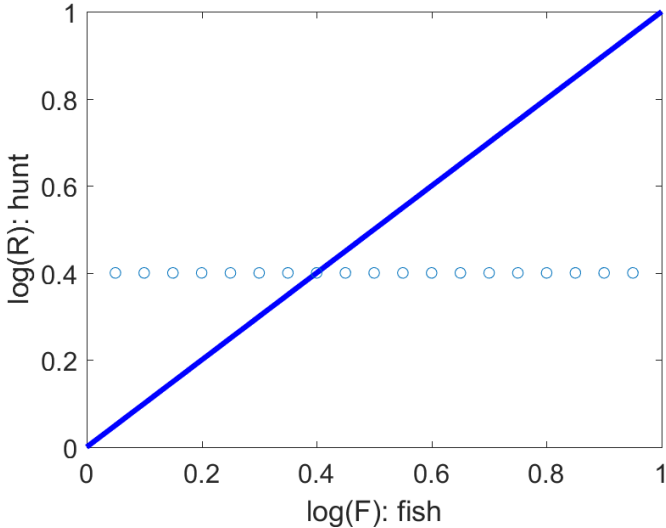
Sinon, pas de sélection dans la chasse/pêche

Qui chasse, qui pêche?



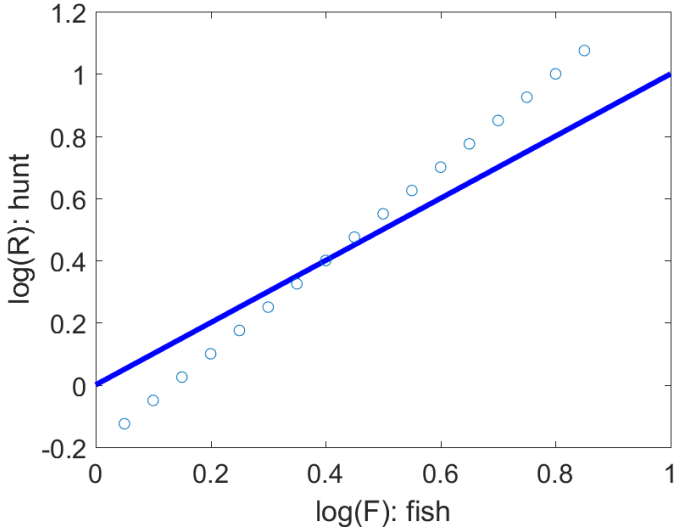
Cas 1 : seulement une variance dans la capacité de pêche

Selection: +F, 0R



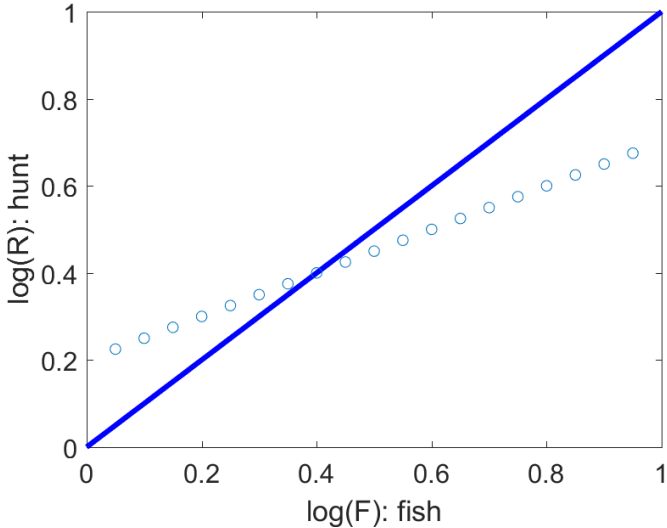
Cas 2a : $\log R_i = \alpha_0 + \alpha_1 \log F_i$, $\alpha_1 > 1$

Sélection: -F, +R



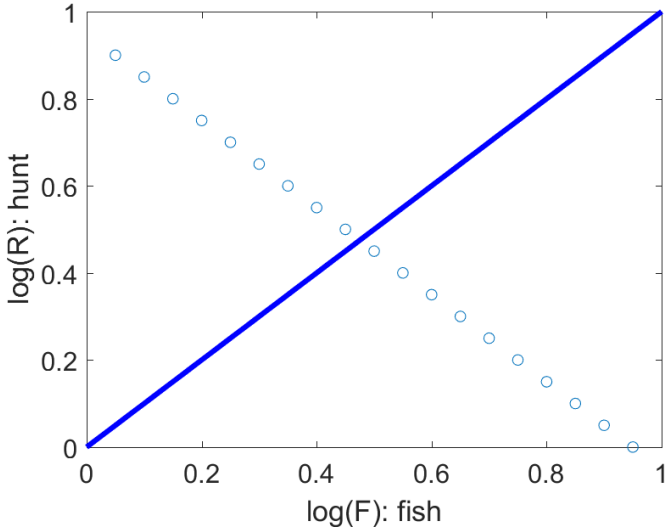
Cas 2b : $\log R_i = \alpha_0 + \alpha_1 \log F_i$, $\alpha_1 < 1$

Sélection: +F, -R



Cas 3 : corrélation négative

Sélection: +F, +R



Modèle de Roy linéaire normal

$$\begin{bmatrix} \log F_i \\ \log R_i \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_F \\ \mu_R \end{bmatrix}, \begin{bmatrix} \sigma_{FF}^2 & \sigma_{RF} \\ \sigma_{RF} & \sigma_{RR}^2 \end{bmatrix} \right)$$

- μ_F : capacité moyenne de pêche
- μ_R : capacité moyenne de chasse
- σ_{FF} : écart-type de la capacité de pêche
- σ_{RR} : écart-type de la capacité de chasse
- $\frac{\sigma_{RF}}{\sqrt{\sigma_{FF}\sigma_{RR}}}$: corrélation entre capacité de chasse et de pêche

Pour simplifier l'algèbre, nous supposerons $\log(\pi_R) = \log(\pi_F) = 0$

Rappel sur la loi normale

Entièrement décrite par ses deux premiers moments (moyenne et variance)

$X \sim N(\mu, \sigma^2)$ signifie « X suit une loi normale de moyenne μ et de variance σ^2 »

Densité $f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(X-\mu)^2}{2\sigma^2}\right)$

$\mu = E[X] = \int_{\infty}^{\infty} xf(x)dx$, $\sigma^2 = Var[X] = E[(X-\mu)^2]$

Si $X \sim N(\mu, \sigma^2)$, alors $aX + b \sim N(a\mu + b, b^2\sigma^2)$.

Si $X_1 \sim N(\mu_1, \sigma_1^2)$ et $X_2 \sim N(\mu_2, \sigma_2^2)$ sont indépendants, alors $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

La loi normale standard $Z \sim N(0, 1)$ a un symbole spécial pour sa densité $\phi(\cdot)$ et sa fonction de répartition (anglais: CDF) $\Phi(\cdot)$

Rapport de Mills inverse (anglais: Inverse Mills Ratio)

Si $X \sim N(\mu, \sigma^2)$, alors

$$\begin{aligned}\mathbb{E}[X \mid X > a] &= \mu + \sigma \frac{\phi\left(\frac{a-\mu}{\sigma}\right)}{1 - \Phi\left(\frac{a-\mu}{\sigma}\right)} \\ &= \mu + \sigma\lambda\left(\frac{a-\mu}{\sigma}\right)\end{aligned}$$

$\lambda(\cdot) = \frac{\phi(\cdot)}{1-\Phi(\cdot)}$ est appelé le **rapport de Mills inverse**

Rappel sur la loi normale bivariée

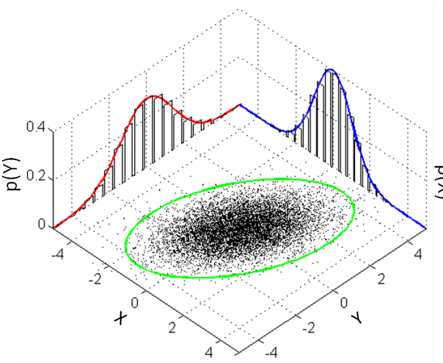
$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right)$$

en mots : « X_1 et X_2 suivent une loi normale conjointe »

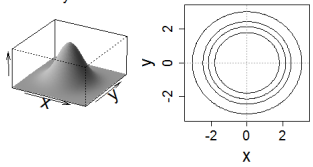
- $X_1 \sim N(\mu_1, \sigma_1^2)$
- $X_2 \sim N(\mu_2, \sigma_2^2)$
- $\text{Cov}(X_1, X_2) = \sigma_{12}$

$X_1 + X_2 \sim N(\mu_1 + \mu_2,\ \sigma_1^2 + \sigma_2^2 + 2\sigma_{12})$

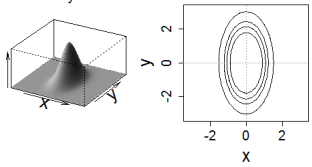
Si X_1 et X_2 sont conjointement normaux, toute combinaison linéaire $Y_1 = aX_1 + bX_2$ et $Y_2 = cX_1 + dX_2$ est aussi conjointe normale.



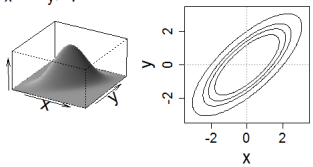
$\sigma_x = \sigma_y,\ \rho = 0$



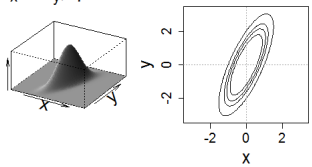
$2\sigma_x = \sigma_y,\ \rho = 0$



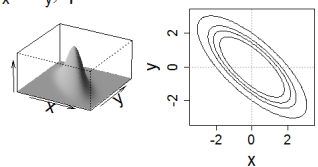
$\sigma_x = \sigma_y,\ \rho = 0.75$



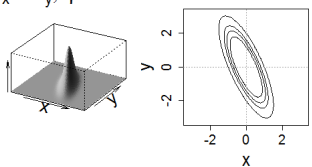
$2\sigma_x = \sigma_y,\ \rho = 0.75$



$\sigma_x = \sigma_y,\ \rho = -0.75$



$2\sigma_x = \sigma_y,\ \rho = -0.75$



Espérance conditionnelle de normales conjointes

Si $\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix}\right)$, alors :

$$\mathbb{E}[Y \mid X] = \mu_Y + \frac{\sigma_{XY}}{\sigma_X^2}(X - \mu_X).$$

Preuve : On peut écrire $Y = \alpha_0 + \alpha_1 X + U$ où U est normal, de moyenne nulle et indépendant de X.

$$\begin{aligned} \mathrm{Cov}(Y, X) &= \mathrm{Cov}(\alpha_0 + \alpha_1 X + U, X) \\ \sigma_{XY} &= \alpha_1 \,\mathrm{Var}(X) \\ \alpha_1 &= \sigma_{XY}/\sigma_X^2 \end{aligned}$$

$$\begin{aligned} \mathbb{E}[Y] &= \mathbb{E}[\alpha_0 + \alpha_1 X + U] \\ &= \alpha_0 + \alpha_1 \mu_X \\ \alpha_0 &= \mu_Y - \alpha_1 \mu_X \end{aligned}$$

En combinant les deux,

$$\mathbb{E}[Y \mid X] = \mu_Y - \alpha_1 \mu_X + \alpha_1 X = \mu_Y + \alpha_1 (X - \mu_X).$$

□

Retour au modèle de Roy linéaire normal

Le travailleur i choisit de chasser si $\log(R_i) > \log(F_i)$.

$$E[\log R_i | \log R_i > \log F_i]$$

Qu'est-ce que c'est ?

Retour au modèle de Roy linéaire normal

Le travailleur i choisit de chasser si $\log(R_i) > \log(F_i)$.

$$E[\log R_i | \log R_i > \log F_i]$$

Qu'est-ce que c'est ? $\rightarrow$ **Capacité moyenne de chasse des travailleurs qui choisissent de chasser**

Retour au modèle de Roy linéaire normal

Le travailleur i choisit de chasser si $\log(R_i) > \log(F_i)$.

$$E[\log R_i | \log R_i > \log F_i]$$

Qu'est-ce que c'est ? $\rightarrow$ **Capacité moyenne de chasse des travailleurs qui choisissent de chasser**

$$E[\log R_i]$$

Et ça ?

Retour au modèle de Roy linéaire normal

Le travailleur i choisit de chasser si $\log(R_i) > \log(F_i)$.

$$E[\log R_i | \log R_i > \log F_i]$$

Qu'est-ce que c'est ? $\rightarrow$ **Capacité moyenne de chasse des travailleurs qui choisissent de chasser**

$$E[\log R_i]$$

Et ça ? $\rightarrow$ **Capacité moyenne de chasse de tous les travailleurs** (y compris ceux qui ne choisissent pas de chasser)

Retour au modèle de Roy linéaire normal

Le travailleur i choisit de chasser si $\log(R_i) > \log(F_i)$.

$$E[\log R_i | \log R_i > \log F_i]$$

Qu'est-ce que c'est ? $\rightarrow$ **Capacité moyenne de chasse des travailleurs qui choisissent de chasser**

$$E[\log R_i]$$

Et ça ? $\rightarrow$ **Capacité moyenne de chasse de tous les travailleurs** (y compris ceux qui ne choisissent pas de chasser)

Si $E[\log R_i | \log R_i > \log F_i] > E[\log R_i]$, alors il y a

(a) positive (b) négative (c) aucune

sélection dans la chasse.

Retour au modèle de Roy linéaire normal

Le travailleur i choisit de chasser si $\log(R_i) > \log(F_i)$.

$$E[\log R_i | \log R_i > \log F_i]$$

Qu'est-ce que c'est ? $\rightarrow$ **Capacité moyenne de chasse des travailleurs qui choisissent de chasser**

$$E[\log R_i]$$

Et ça ? $\rightarrow$ **Capacité moyenne de chasse de tous les travailleurs** (y compris ceux qui ne choisissent pas de chasser)

Si $E[\log R_i | \log R_i > \log F_i] > E[\log R_i]$, alors il y a

(a) positive ~~(b) négative~~ ~~(c) aucune~~

sélection dans la chasse.

Combien de travailleurs choisissent la chasse ?

La différence entre la capacité de chasse et de pêche du travailleur i est

$$\log R_i - \log F_i \sim N(\mu_R - \mu_F, \sigma^2).$$

Alors la part des travailleurs qui chassent est :

$$\begin{aligned} P(\text{travailleur } i \text{ chasse}) &= P(\log R_i - \log F_i > 0) \\ &= P\left(\frac{\log R_i - \log F_i}{\sigma} > 0\right) \\ &= P\left(\underbrace{\frac{\log R_i - \log F_i - (\mu_R - \mu_F)}{\sigma}}_{\sim N(0,1)} > \frac{\mu_F - \mu_R}{\sigma}\right) \\ &= \Phi\left(\frac{\mu_R - \mu_F}{\sigma}\right) \end{aligned}$$

La part des travailleurs qui chassent augmente avec μ_R et diminue avec μ_F

Dérivation du salaire des chasseurs

Notre objectif est de dériver $E[\log R_i | \log R_i - \log F_i > 0]$ comme fonction des paramètres.

Nous procédons en deux étapes.

Étape 1 : Inverse Mills Ratio

$$E[\log R_i - \log F_i \mid \log R_i - \log F_i > 0] = \mu_R - \mu_F + \sigma\lambda\left(\frac{\mu_F - \mu_R}{\sigma}\right)$$

Étape 2 : l'espérance conditionnelle

$$E[\log R_i \mid \log R_i - \log F_i] = \mu_R + \frac{\sigma_R^2 - \sigma_{RF}}{\sigma^2}\left[(\log R_i - \log F_i) - (\mu_R - \mu_F)\right]$$

appliquer maintenant l'Étape 1 :

$$E[\log R_i \mid \log R_i - \log F_i > 0] = \mu_R + \frac{\sigma_R^2 - \sigma_{RF}}{\sigma^2}\lambda\left(\frac{\mu_F - \mu_R}{\sigma}\right)$$

Sélection

Réarranger

$$E[\log R_i \mid \log R_i - \log F_i > 0] - E[\log R_i] = \underbrace{\frac{\sigma_F \sigma_R}{\sigma}}_{\geq 0} \cdot \left(\frac{\sigma_R}{\sigma_F} - \rho \right) \cdot \underbrace{\lambda\left(\frac{\mu_F - \mu_R}{\sigma} \right)}_{\geq 0}$$

Le premier et le troisième terme du produit sont non négatifs, donc la sélection est déterminée par le deuxième terme.

$-1 < \rho < 1$ (pourquoi ?), donc si $\sigma_R > \sigma_F$, alors il doit y avoir une sélection positive dans la chasse !

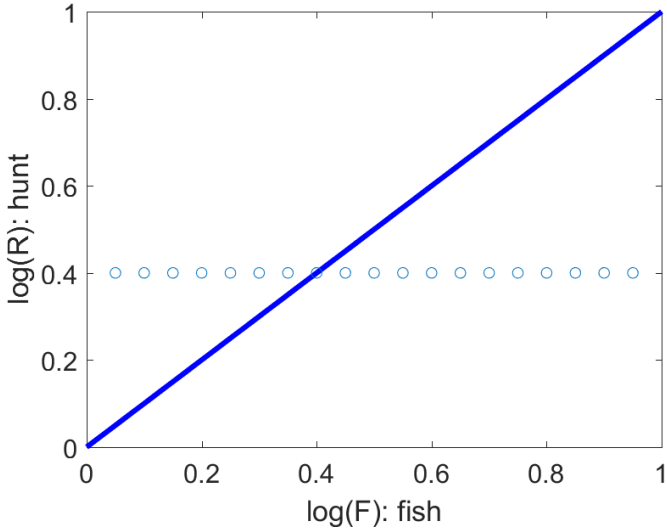
Si $\sigma_R < \sigma_F$ et ρ est légèrement positif, il peut encore y avoir une sélection positive dans la chasse

Si $\sigma_R < \sigma_F$ et ρ est fortement positif, il y a une sélection négative dans la chasse

Si $\rho < 0$, il y a une sélection positive à la fois dans la chasse et la pêche !

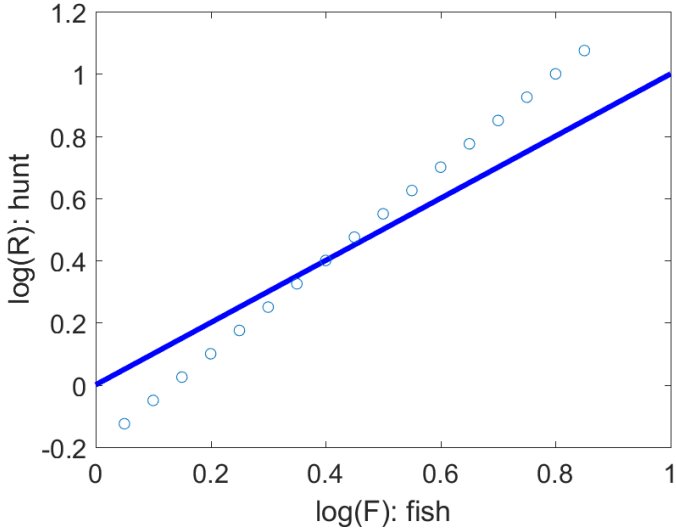
Cas 1 : seulement une variance dans la capacité de pêche

$\sigma_R = 0$, $\sigma_F > 0$, $\rho = 0 \implies$ pas de sélection dans la chasse, sélection positive dans la pêche



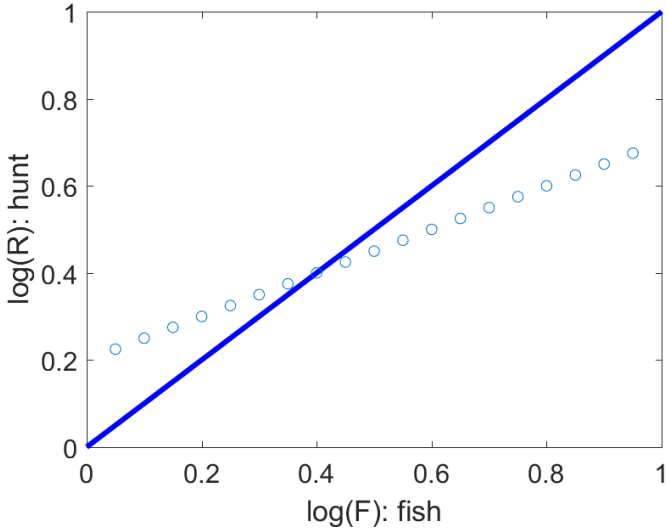
Cas 2a : $\log R_i = \alpha_0 + \alpha_1 \log F_i$, $\alpha_1 > 1$

$\sigma_R > \sigma_F$, $\rho \approx 1 \implies$ sélection positive dans la chasse, sélection négative dans la pêche



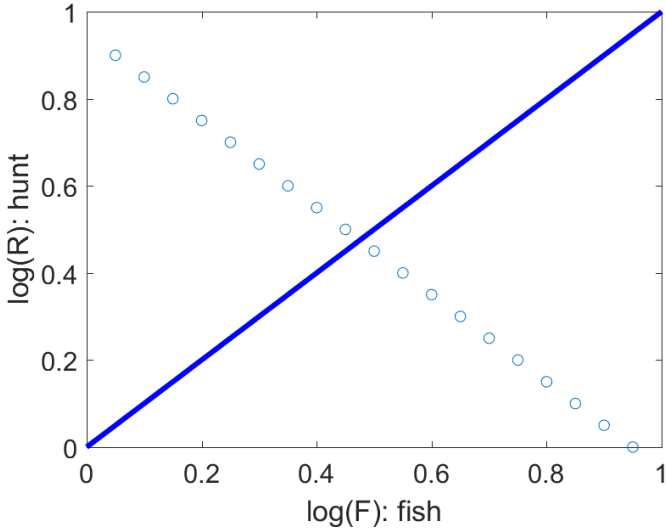
Cas 2b : $\log R_i = \alpha_0 + \alpha_1 \log F_i$, $\alpha_1 < 1$

$\sigma_R < \sigma_F$, $\rho \approx 1 \implies$ sélection positive dans la pêche, sélection négative dans la chasse



Cas 3 : corrélation négative

$\sigma_R = \sigma_F$, $\rho = -1 \implies$ sélection positive dans la pêche et la chasse



La semaine prochaine

Comment estimer empiriquement le modèle de Roy à partir de données

- Méthode en deux étapes (two-step) de Heckman (« Heckit »)
- Identification non paramétrique (« identification at infinity »)

Deux applications classiques :

1. Borjas (1987) : la sélection explique-t-elle l'assimilation salariale des immigrants ?
2. Mulligan et Rubinstein (2008) : la sélection explique-t-elle le déclin de l'écart salarial entre hommes et femmes ?

Discussion en classe : **Abramitzky, Boustan et Eriksson (2012)**

- Lisez l'article pour le cours de la semaine prochaine, venez préparés à discuter

Estimation d'un modèle de Roy

Pour chaque travailleur i, nous observons son occupation $J_i \in \{H, F\}$ et salaire w_{Ji}

- Nous n'observons que le salaire de l'occupation qu'il a choisie
- Nous n'observons jamais le salaire de l'autre occupation qu'il n'a pas choisie

Il serait souhaitable de pouvoir observer (J_i, w_{Hi}, w_{Fi}) pour chaque travailleur i. Mais nous n'observons que $(H_i, w_{Hi}, \cdot)$ pour les travailleurs qui choisissent de chasser, et $(F_i, \cdot, w_{Fi})$ pour ceux qui choisissent de pêcher.

Les méthodes économétriques pour traiter ce problème ont été développées par **Jim Heckman** à la fin des années 1970, ce qui lui a valu le Prix Nobel en 2000

L'idée clé de Heckman a été de montrer que la sélection n'est pas seulement un problème de **données manquantes** (**missing data**), mais qu'elle peut être utilement reformulée comme un problème de **variables omises** (**omitted variable bias**)

Heckman (1977) : mise en place

Chaque femme i a un **salaire potentiel de marché** w_i qu'elle gagnerait si elle travaillait sur le marché du travail et un **salaire potentiel hors marché** qu'elle gagnerait par la "production domestique" r_i. Les femmes choisissent le secteur (marché vs. hors marché) qui paie le salaire le plus élevé

- Dans les données, nous observons si chaque femme travaille ou non, et pour celles qui travaillent, nous observons w_i. Nous n'observons pas r_i
- Nous observons également certaines covariables X_i et Z_i

Supposons que les salaires dans le secteur de marché et hors marché soient donnés par

$$\log w_i = \beta_1 X_i + \gamma_1 Z_i + u_i$$
$$\log r_i = \beta_2 X_i + \gamma_2 Z_i^r + v_i$$

$$\begin{bmatrix} u_i \\ v_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{bmatrix} \right)$$

- X_i affecte à la fois les salaires de marché et hors marché, Z_i est une variable qui affecte les salaires de marché mais pas les salaires hors marché, et Z_i^r est une variable qui affecte les salaires hors marché mais pas les salaires de marché

Offre de travail (labor supply) et équation de salaire (wage equation)

Une femme travaille ($D_i = 1$) si son salaire de marché dépasse son salaire hors marché :

$$D_i = 1 \iff \log w_i \geq \log r_i \iff (\beta_1 - \beta_2)X_i + \gamma_1 Z_i - \gamma_2 Z_i^r \geq v_i - u_i$$

$$D_i = \Phi\left(\frac{(\beta_1 - \beta_2)X_i + \gamma_1 Z_i - \gamma_2 Z_i^r}{\sigma}\right)$$

Offre de travail (labor supply) et équation de salaire (wage equation)

Une femme travaille ($D_i = 1$) si son salaire de marché dépasse son salaire hors marché :

$$D_i = 1 \iff \log w_i \geq \log r_i \iff (\beta_1 - \beta_2)X_i + \gamma_1 Z_i - \gamma_2 Z_i^r \geq v_i - u_i$$

$$D_i = \Phi\left(\frac{(\beta_1 - \beta_2)X_i + \gamma_1 Z_i - \gamma_2 Z_i^r}{\sigma}\right)$$

Salaires des femmes qui travaillent :

$$\mathbb{E}[\log w_i \mid D_i = 1, X_i, Z_i] = \beta_1 X_i + \gamma_1 Z_i + \mathbb{E}[u_i \mid D_i = 1, X_i, Z_i]$$

où

$$\mathbb{E}[u_i \mid D_i = 1, X_i, Z_i] = \underbrace{\frac{\sigma_u \sigma_v}{\sigma}\left(\frac{\sigma_u}{\sigma_v} - \rho\right)}_{=\tau} \cdot \lambda\left(\frac{(\beta_1 - \beta_2)X_i + \gamma_1 Z_i - \gamma_2 Z_i^r}{\sigma}\right)$$

donc

$$\mathbb{E}[\log w_i \mid D_i = 1, X_i, Z_i] = \beta_1 X_i + \gamma_1 Z_i + \tau\lambda\left(\frac{(\beta_1 - \beta_2)X_i + \gamma_1 Z_i - \gamma_2 Z_i^r}{\sigma}\right)$$

Heckman two-step

Pour estimer Roy, nous allons confronter ces deux équations aux données :

$$D_i = \Phi\left(\frac{(\beta_1 - \beta_2)X_i + \gamma_1 Z_i - \gamma_2 Z_i^r}{\sigma}\right) \tag{1}$$

$$\log w_i = \beta_1 X_i + \gamma_1 Z_i + \tau\lambda\left(\frac{(\beta_1 - \beta_2)X_i - \gamma Z_i}{\sigma}\right) \tag{2}$$

Nous procédons par une procédure en deux étapes (**two-step**)

1. Estimer (1) avec un **probit**
2. Utiliser les estimations de l'étape 1 pour former le terme $\lambda()$ et estimer (2) par MCO

Cette procédure est connue sous le nom de **Heckman two-step** ou **Heckit**

Rappel : probit

Le probit est un modèle pour variables dépendantes binaires $D_i \in \{0, 1\}$ où l'on modélise D_i comme provenant d'une variable latente D_i^*

$$D_i^* = X_i\beta + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, 1) \qquad\qquad D_i = \begin{cases} 1 & \text{si } D_i^* > 0, \\ 0 & \text{sinon.} \end{cases}$$

En supposant que ϵ_i est normal, nous pouvons modéliser D_i

$$\begin{aligned} \Pr(D_i = 1 \mid X_i) &= \Pr(\varepsilon_i > -X_i\beta) \\ &= \Phi(X_i\beta) \end{aligned}$$

Cette équation est non linéaire en β, donc nous l'estimons par **maximum likelihood**, en choisissant $\hat{\beta}$ pour maximiser la probabilité d'observer l'échantillon $\{D_i, X_i\}$

$$L(\beta) = \prod_i \Phi(X_i\beta)^{D_i} \left[1 - \Phi(X_i\beta)\right]^{1-D_i}$$

Identification des paramètres du modèle : Heckit

Quels sont les paramètres du modèle ?

$$\log w_i = \beta_1 X_i + \gamma_1 Z_i + u_i$$
$$\log r_i = \beta_2 X_i + \gamma_2 Z_i^r + v_i$$

$$\begin{bmatrix} u_i \\ v_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{bmatrix} \right)$$

Identification des paramètres du modèle : Heckit

Quels sont les paramètres du modèle ?

- β_1, β_2, γ_1, γ_2, σ_u, σ_v, σ_{uv}

Identification des paramètres du modèle : Heckit

Quels sont les paramètres du modèle ?

- β_1, β_2, γ_1, γ_2, σ_u, σ_v, σ_{uv}

Quels paramètres sont identifiés par Heckit ?

Étape 1 : dans le probit, nous identifions

- $\frac{(\beta_1 - \beta_2)}{\sigma}$: coefficient sur X_i
- $\frac{\gamma_1}{\sigma}$: coefficient sur Z_i
- $-\frac{\gamma_2}{\sigma}$: coefficient sur Z_i^r

Identification des paramètres du modèle : Heckit

Quels sont les paramètres du modèle ?

- β_1, β_2, γ_1, γ_2, σ_u, σ_v, σ_{uv}

Quels paramètres sont identifiés par Heckit ?

Étape 1 : dans le probit, nous identifions

- $\frac{(\beta_1-\beta_2)}{\sigma}$: coefficient sur X_i
- $\frac{\gamma_1}{\sigma}$: coefficient sur Z_i
- $-\frac{\gamma_2}{\sigma}$: coefficient sur Z_i^r

Étape 2 : dans la régression, nous identifions

- β_1 : coefficient sur X_i
- γ_1 : coefficient sur Z_i^r
- $\tau = \frac{\sigma_u \sigma_v}{\sigma}\left(\frac{\sigma_u}{\sigma_v} - \rho\right)$: coefficient sur le rapport de Mills inverse $\lambda(\cdot)$

Identification des paramètres du modèle

Jusqu'à présent, nous avons identifié $\frac{(\beta_1-\beta_2)}{\sigma}$, $\frac{\gamma_1}{\sigma}$ $-\frac{\gamma_2}{\sigma}$, β_1, γ_1, et τ

Notre objectif est d'identifier β_1, β_2, γ_1, γ_2, σ_u, σ_v, σ_{uv}

Nous pouvons utiliser les paramètres déjà identifiés pour en déduire d'autres

Identification des paramètres du modèle

Jusqu'à présent, nous avons identifié $\frac{(\beta_1-\beta_2)}{\sigma}$, $\frac{\gamma_1}{\sigma}$ $-\frac{\gamma_2}{\sigma}$, β_1, γ_1, et τ

Notre objectif est d'identifier β_1, β_2, γ_1, γ_2, σ_u, σ_v, σ_{uv}

Nous pouvons utiliser les paramètres déjà identifiés pour en déduire d'autres

- Nous pouvons utiliser γ_1/σ et γ_1 pour résoudre σ
- Ensuite, utiliser γ_2/σ et σ pour résoudre γ_2
- Ensuite, utiliser σ, β_1 et $\frac{\beta_1-\beta_2}{\sigma}$ pour résoudre β_2

Nous avons donc maintenant identifié β_1, β_2, γ_1, γ_2, et σ. Rappelons que

$$\sigma^2 = Var(v_i - u_i) = \sigma_u^2 + \sigma_v^2 - 2\sigma_{uv}$$

σ_u, σ_v et σ_{uv} (ou ρ) peuvent être identifiés à l'aide de cette équation et de moments associés des données (détails dans vos devoirs)

τ identifie la direction de la sélection

Rappelons que dans la deuxième étape de la procédure Heckman two-step, nous avons montré que le coefficient sur le terme du rapport de Mills inverse $\lambda(\cdot)$ était $\tau = \frac{\sigma_u \sigma_v}{\sigma}\left(\frac{\sigma_u}{\sigma_v} - \rho\right)$

Le signe de ce terme reflète la **direction de la sélection** (par ex. sélection positive, sélection négative)

Pourquoi ?

τ identifie la direction de la sélection

Rappelons que dans la deuxième étape de la procédure Heckman two-step, nous avons montré que le coefficient sur le terme du rapport de Mills inverse $\lambda(\cdot)$ était $\tau = \frac{\sigma_u \sigma_v}{\sigma}\left(\frac{\sigma_u}{\sigma_v} - \rho\right)$

Le signe de ce terme reflète la **direction de la sélection** (par ex. sélection positive, sélection négative)

Pourquoi ?

$$\mathbb{E}[\log w_i | X_i, Z_i] = \beta_1 X_i + \gamma_1 Z_i$$
$$\mathbb{E}[\log w_i \mid D_i = 1, X_i, Z_i] = \beta_1 X_i + \gamma_1 Z_i + \tau\lambda(\cdot)$$
$$\mathbb{E}[\log w_i \mid D_i = 1, X_i, Z_i] - \mathbb{E}[\log w_i | X_i, Z_i] = \tau\lambda(\cdot)$$

Comme $\lambda(\cdot) \geq 0$, si $\tau > 0$, **sélection positive**. Si $\tau < 0$, **sélection négative**. Si $\tau = 0$, **pas de sélection**

Restrictions d'exclusion

Nous avons supposé que nous pouvions trouver des variables Z_i et Z_i^r qui affectent uniquement w_i et r_i respectivement, mais pas l'autre

Dans la réalité, elles sont très difficiles (impossibles ?) à trouver. Pouvez-vous penser à des exemples de Z_i ?

Restrictions d'exclusion

Nous avons supposé que nous pouvions trouver des variables Z_i et Z_i^r qui affectent uniquement w_i et r_i respectivement, mais pas l'autre

Dans la réalité, elles sont très difficiles (impossibles ?) à trouver. Pouvez-vous penser à des exemples de Z_i ? Z_i^r ?

Restrictions d'exclusion

Nous avons supposé que nous pouvions trouver des variables Z_i et Z_i^r qui affectent uniquement w_i et r_i respectivement, mais pas l'autre

Dans la réalité, elles sont très difficiles (impossibles ?) à trouver. Pouvez-vous penser à des exemples de Z_i ? Z_i^r ?

Ceci est appelé une **restriction d'exclusion**

- Lié aux **variables instrumentales**
- Z_i^r affecte seulement le salaire via son effet sur l'offre de travail
- Heckit est un cas particulier d'une généralisation non linéaire des VI appelée **fonction de contrôle**

Quels paramètres reposent sur l'exclusion pour leur identification ?

- Nous n'avons pas eu besoin d'exclusion pour identifier β_1 et γ_1
- Mais nous l'avons utilisée pour identifier β_2 et γ_2

Sujet avancé : identification non paramétrique

Pourquoi avons-nous supposé que u et v sont joint normal ?

Sujet avancé : identification non paramétrique

Pourquoi avons-nous supposé que u et v sont joint normal ?

Existe-t-il quelque chose dans la théorie économique ou dans le monde réel qui nous dit que les log-salaires dans les secteurs de marché et hors marché sont normaux ?

Sujet avancé : identification non paramétrique

Pourquoi avons-nous supposé que u et v sont joint normal ?

Existe-t-il quelque chose dans la théorie économique ou dans le monde réel qui nous dit que les log-salaires dans les secteurs de marché et hors marché sont normaux ?

- **Non.** Nous avons fait cette hypothèse uniquement par commodité. Nous avons supposé la normalité car cela simplifie la résolution du modèle

Et si les log-salaires ne sont pas normaux ? Nos résultats seront-ils erronés ?

Sujet avancé : identification non paramétrique

Pourquoi avons-nous supposé que u et v sont joint normal ?

Existe-t-il quelque chose dans la théorie économique ou dans le monde réel qui nous dit que les log-salaires dans les secteurs de marché et hors marché sont normaux ?

- **Non.** Nous avons fait cette hypothèse uniquement par commodité. Nous avons supposé la normalité car cela simplifie la résolution du modèle

Et si les log-salaires ne sont pas normaux ? Nos résultats seront-ils erronés ?

- En général, oui
- Le biais pourrait ne pas être trop grave si la réalité est « proche » de la normalité, mais il est difficile de le dire *a priori*
- La littérature en économétrie sur la **misspecification** est très complexe

Et si nous n'assumons pas la normalité ? Pouvons-nous toujours identifier les paramètres ?

- Oui.

Sujet avancé : identification non paramétrique

Pourquoi avons-nous supposé que u et v sont joint normal ?

Existe-t-il quelque chose dans la théorie économique ou dans le monde réel qui nous dit que les log-salaires dans les secteurs de marché et hors marché sont normaux ?

- **Non.** Nous avons fait cette hypothèse uniquement par commodité. Nous avons supposé la normalité car cela simplifie la résolution du modèle

Et si les log-salaires ne sont pas normaux ? Nos résultats seront-ils erronés ?

- En général, oui
- Le biais pourrait ne pas être trop grave si la réalité est « proche » de la normalité, mais il est difficile de le dire *a priori*
- La littérature en économétrie sur la **misspecification** est très complexe

Et si nous n'assumons pas la normalité ? Pouvons-nous toujours identifier les paramètres ?

- Oui. **Mais**, il faut faire d'autres hypothèses
- Cela s'appelle l'identification **non paramétrique**, par contraste avec **paramétriques**

Identification non paramétrique du modèle de Roy

Je donne une brève discussion informelle. Pour un excellent exposé complet, voir French et Taber (2011), chapitre du *Handbook of Labor Economics*.

Considérons une version non linéaire du modèle de Roy sans normalité

$$\log w_i = g(X_i, Z_i) + u_i$$
$$r_i = g_r(X_i, Z_i^r) + v_i$$

Sous quelles conditions pouvons-nous identifier $g(\cdot)$, $g_r(\cdot)$, et la distribution de u_i et v_i ?

Hypothèse clé : **grand support** ou **identification à l'infini**

- Pour toute valeur de $g(x, z)$, $g_r(X, Z)$ varie sur toute la droite réelle (et vice versa)

En mots: nous pouvons trouver un groupe de travailleuses qui reçoivent un salaire hors marché très faible ($r_i \to -\infty$) de sorte qu'environ 100% d'entre elles travaillent. Dans ce cas, il n'y a **pas de sélection**

Applications du modèle de Roy

Cela conclut notre analyse théorique du modèle de Roy

Maintenant nous passons à deux applications, en revenant aux motivations avec lesquelles nous avons commencé le cours

1. **Borjas (1987, AER)**: *Self-selection and the earnings of immigrants*
2. **Mulligan and Rubinstein (2008, QJE)**: *Selection, investment and women's relative wages over time*

Ensuite, nous conclurons avec notre discussion en classe de **Abramitzky, Boustan et Eriksson (2012, AER)**

Modèle de Borjas (1987)

$$\ln w_0 = \mu_0 + \epsilon_0 \qquad \text{revenus dans le pays d'origine}$$

$$\ln w_1 = \mu_1 + \epsilon_1 \qquad \text{revenus aux États-Unis}$$

$$I = \ln w_1 - \ln w_0 - C \qquad \text{si } I > 0, \text{ immigrer aux États-Unis}$$

$$E[\ln w_0 | I > 0] = \mu_0 + \frac{\sigma_0 \sigma_1}{\sigma}\left(\rho - \frac{\sigma_0}{\sigma_1}\right)\lambda \qquad \text{revenus des immigrants dans le pays d'origine}$$

$$E[\ln w_1 | I > 0] = \mu_1 + \frac{\sigma_0 \sigma_1}{\sigma}\left(\frac{\sigma_0}{\sigma_1} - \rho\right)\lambda \qquad \text{revenus des immigrants aux États-Unis}$$

Questions

- Les immigrants sont-ils positivement sélectionnés ?
- Si les immigrants sont positivement sélectionnés, seront-ils plus riches que l'américain moyen ?
- Qu'est-ce qui explique l'évolution de la sélection des immigrants au fil du temps ?

Prédictions du modèle de Borjas (1987)

Cas 1 : *sélection positive*

- les immigrants sont positivement sélectionnés dans leur pays d'origine
- les immigrants gagnent plus que l'américain moyen
- se produit si $\rho > \min(\sigma_0/\sigma_1, \sigma_1/\sigma_0)$ et $\sigma_1 > \sigma_0$

Cas 2 : *sélection négative*

- les immigrants sont négativement sélectionnés dans leur pays d'origine
- les immigrants gagnent moins que l'américain moyen
- se produit si $\rho > \min(\sigma_0/\sigma_1, \sigma_1/\sigma_0)$ et $\sigma_1 < \sigma_0$

Cas 3 : *tri des réfugiés*

- les immigrants sont négativement sélectionnés dans leur pays d'origine
- les immigrants gagnent plus que l'américain moyen
- se produit si $\rho < \min(\sigma_0/\sigma_1, \sigma_1/\sigma_0)$

Stratégie empirique de Borjas (1987)

Le modèle de Roy prédit que les changements dans l'inégalité dans le pays d'origine mèneront à des différences de sélection

Borjas estime une régression avec les recensements de 1970 et 1980 où la variable dépendante est une mesure des écarts de revenus entre immigrants et natifs aux États-Unis, et les variables explicatives sont des chocs spécifiques aux pays qui peuvent accroître l'inégalité

TABLE 4—DEFINITION OF COUNTRY-SPECIFIC VARIABLES

Variable	Definition and Source	Mean	Mini-mum	Maxi-mum	U.S. Value
Politically Competitive System	= 1 if the country had a competitive party system during the entire 1950–73 period; 0 otherwise. *Source*: Cross-National Time-Series Archive (CNTSA)	.41	–	–	1
Recent Loss of Freedom	= 1 if the country had a competitive party system at the beginning of the period but had a non-competitive party system at the end of the period; 0 otherwise. *Source*: CNTSA.	.20	–	–	0
Number of Assassinations	Number of politically motivated murders or attempted murders of high government officials or politicians in 1950–73. *Source*: CNTSA.	3.27	0	22	12
Income Inequality	Ratio of household income of the top 10 percent of the households to the income of the bottom 20 percent of the households. *Source*: World Bank (various issues) and United Nations (1977).	7.50	1.42	30.0	5.91
Distance from U.S.	Number of air miles (in thousands) between the country's capital and the nearest U.S. gateway (Los Angeles, Miami, or New York). *Source*: Airline offices contacted by author.	3.37	.18	7.49	–
English Proficiency	Fraction of 1975–80 cohort of immigrants who speak English well or very well. *Source*: 5/100 A Sample of the 1980 U.S. Census.	.74	.24	1.00	–
Age at Migration	Mean age at migration. *Source*: 5/100 A Sample of the 1980 U.S. Census.	24.56	12.40	32.40	–
ln (per capital *GNP*)	(ln) 1980 per capita *GNP* in dollars. *Source*: U.S. Arms Control and Disarmament Agency (1984).	8.17	5.42	9.62	9.39
Rate of Change in Per Capita *GNP*	Annual rate of change in per capita *GNP* between 1963 and 1980, defined by: $\ln(GNP_{1980}/GNP_{1963})/17$. *Source*: U.S. Arms Control and Disarmament Agency (1975, 1984).	.03	.004	.07	.02
Rate of Change in Central Government Expenditures	Annual Change in the Percentage of GNP that is accounted for by central government expenditures, defined by $(GOVT_{1980} - GOVT_{1950})/30$, where $GOVT_t$ is the percent of GNP attributable to central government expenditures in year t. *Source*: CNTSA and U.S. Arms Control and Disarmament Agency (1984).	.41	−1.69	2.08	.26
Change in Quota	Change in fraction of population eligible for migration to the U.S., defined by (20000/1979 population) ÷ ($QUOTA$/1950 population), where 20,000 is the maximum number of visas	38.90	.28	149.67	–

TABLE 5—DETERMINANTS OF THE ENTRY WAGE DIFFERENTIAL BETWEEN THE 1979 IMMIGRANT COHORT AND NATIVES[a]

Country of Origin Characteristics	Regression			
	1	2	3	4
Intercept	−.2214	.1838	−.9934	−.9469
	(−3.88)	(1.06)	(−3.41)	(−3.30)
Politically Competitive System	.2743	.1306	.1101	.1264
	(4.49)	(2.01)	(2.16)	(2.39)
Recent Loss of Freedom	−.0010	−.0511	−.0062	.0136
	(−.01)	(−.75)	(−.12)	(.25)
Number of Assassinations	−.0072	−.0028	.0021	.0044
	(−1.20)	(−.54)	(.51)	(.92)
Income Inequality	−.0084	−.0038	.0039	.0046
	(−1.78)	(−.89)	(1.02)	(1.13)
Distance from U.S.	–	−.0114	−.0031	.0018
		(−.89)	(−.31)	(.09)
English Proficiency	–	.2596	.1980	.2030
		(2.20)	(2.12)	(2.21)
Mean Age at Migration	–	−.0217	−.0149	−.0119
		(−3.55)	(−2.99)	(2.28)
ln (per capita *GNP*)	–	–	.1164	.1015
			(4.57)	(3.77)
Country in Asia or Africa	–	–	–	−.1145
				(−1.58)
Country in North or South America	–	–	–	−.0640
				(−.73)
R^2	.504	.681	.808	.826

[a] The *t*-ratios are presented in parentheses.

TABLE 6—DETERMINANTS OF THE RATE OF ASSIMILATION[a]

Country of Origin Characteristics	Regression			
	1	2	3	4
Intercept	.0076	−.0240	−.0237	−.0280
	(2.96)	(−3.88)	(−1.50)	(−2.32)
Politically Competitive System	−.0029	−.0068	−.0068	−.0091
	(−1.06)	(−2.66)	(−2.60)	(−4.28)
Recent Loss of Freedom	.0063	.0029	.0030	.0021
	(1.81)	(1.21)	(1.15)	(1.06)
Number of Assassinations	.0008	.0006	.0006	.0008
	(2.68)	(2.36)	(2.14)	(3.07)
Income Inequality	−.0001	−.00002	−.00002	.0002
	(−.50)	(−.11)	(−.10)	(.90)
Distance from U.S.	–	.0003	.0003	−.0027
		(.74)	(.70)	(−2.89)
English Proficiency	–	.0138	.0138	.0122
		(3.27)	(3.20)	(3.70)
Mean Age at Migration	–	.0009	.0009	.0009
		(4.28)	(3.95)	(4.72)
ln (per capita *GNP*)	–	–	−.00002	.0021
			(−.01)	(1.83)
Country in Asia or Africa	–	–	–	.0151
				(5.11)
Country in North or South America	–	–	–	−.0080
				(−2.08)
R^2	.302	.704	.704	.842

[a] The *t*-ratios are presented in parentheses.

Discussion de Borjas (1987)

À mon avis, cet article présente des preuves très faibles (pratiquement inexistantes) de son mécanisme central selon lequel l'inégalité dans le pays d'origine et aux États-Unis détermine le sélection

L'article date des années 1980 et était innovant pour l'époque. Mais l'analyse empirique est vraiment mauvaise selon les standards d'aujourd'hui. N'espérez pas écrire un article comme ça pour votre thèse

À mon sens, la nouveauté et la contribution de l'article étaient l'application du modèle de Roy à l'immigration et les dérivations mathématiques

Si personne ne l'a fait, il *pourrait* être intéressant d'explorer cela avec des données mises à jour. Je connais un article qui fait un peu ça : Lull (2018, JHR). Mais c'est un article sur les effets salariaux de l'immigration, pas sur la sélection. Il y en a peut-être d'autres

Le problème est que trop de choses influencent la composition des immigrants en dehors de la sélection. L'immigration est très complexe et bureaucratique, avec de multiples voies, des quotas, des loteries de visas, l'immigration familiale. Aujourd'hui, l'immigration illégale est très importante

Mulligan et Rubinstein (2008)

La participation des femmes au marché du travail a fortement augmenté entre les années 1960 et 2000. Sur la même période, l'écart salarial entre les sexes a diminué...

Mulligan et Rubinstein (2008)

La participation des femmes au marché du travail a fortement augmenté entre les années 1960 et 2000. Sur la même période, l'écart salarial entre les sexes a diminué... **et** l'inégalité entre les hommes a augmenté

- suggère que la dispersion des rendements de la compétence a augmenté (le σ dans Roy)

Le modèle de Roy prédit que les changements d'inégalité mèneront à des changements de sélection

- faible inégalité dans le secteur marchand $\rightarrow$ sélection négative
- forte inégalité dans le secteur marchand $\rightarrow$ sélection positive
- donc, une augmentation de l'inégalité mènera à un passage de la sélection négative à la sélection positive

La baisse observée de l'écart salarial entre les sexes est-elle simplement le résultat d'un changement de sélection ?

- teste cette théorie en estimant un modèle de Roy
- les résultats suggèrent que la réponse est **oui** !

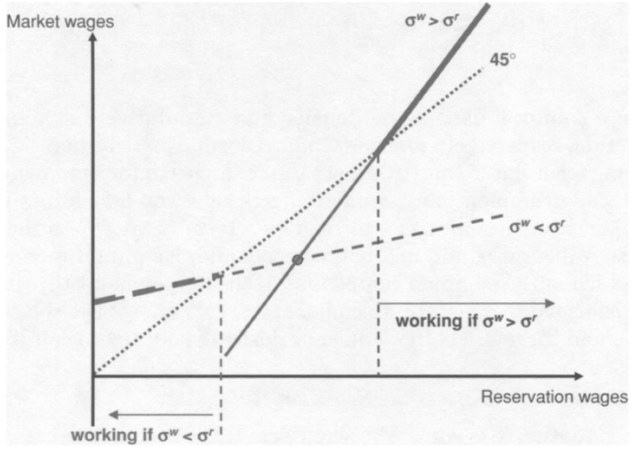


FIGURE II

GHR Model: Inequality Has Composition Effects on Measured Wages

The figure illustrates a comparative static for the Gronau-Heckman-Roy model with respect to σ^w; σ^w (σ^r) denotes the standard deviation of market (reservation) wages. The 45-degree line partitions market workers from nonworkers. Two model parameterizations are shown: σ^w greater (solid line) and σ^r greater (dashed line). The thick portions of the line are above the 45-degree line and indicate workers.

III.B. Estimates Using Heckman's Two-Step Estimator

In the Heckman two-step model, demographic characteristics are assumed to linearly affect μ_t^w and μ_t^r, but not affect ρ, σ_t^w, or σ_t^r. In particular, $\mathbf{X}$ is a row vector of demographic characteristics affecting market wages (and polynomials thereof), and $\mathbf{Z}$ is the row vector $\mathbf{X}$ plus a vector of additional demographic characteristics affecting only reservation wages. In addition, we assume that selection bias is zero for men. For the purposes of estimation, these assumptions imply that inequality (5) becomes a probit equation (13) for the female employment rate $P_t(\mathbf{Z})$ by demographic group and year and a log market wage equation (14) for employed persons:[11]

$$P_t(\mathbf{Z}) \equiv \text{Prob}(L = 1 \mid \mathbf{Z}, g = 1) = \Phi(\mathbf{Z}\,\delta_t) \tag{13}$$

$$w_{it} = \mathbf{X}_{it}\beta_t + g_i\gamma_t + g_i\theta_t\lambda(\mathbf{Z}_{it}\delta_t) + u_{it} \tag{14}$$

The vector $\mathbf{X}$ includes educational attainment dummies, marital status, a potential work experience quartic interacted with education dummies, and region. The vector $\mathbf{Z}$ has the same elements, plus the number of children aged 0–6 interacted with marital status; β and δ are coefficient vectors. The error term u_{it} is the unobserved component of wages $\sigma_t^w\varepsilon_{it}^w$ from equation (1) minus the inverse Mills ratio term $\theta_t\lambda$.[12]

TABLE I

CORRECTING THE GENDER WAGE GAP USING THE HECKMAN TWO-STEP ESTIMATOR

Period	Method		Bias
	OLS	Two-Step	
	Panel A: Variable Weights		
1975–1979	−0.414	−0.337	−0.077
	(0.003)	(0.014)	(0.015)
1995–1999	−0.254	−0.339	0.085
	(0.003)	(0.014)	(0.015)
Change	0.160	−0.002	0.162
	(0.005)	(0.020)	(0.021)
	Panel B: Fixed Weights		
1975–1979	−0.404	−0.330	−0.075
	(0.003)	(0.014)	(0.014)
1995–1999	−0.264	−0.353	0.089
	(0.004)	(0.015)	(0.016)
Change	0.140	−0.024	0.164
	(0.005)	(0.021)	(0.021)

Notes. Each table entry summarizes regression results (reported in full in Appendix II). The entries are female minus male log wages, which differ from each other in terms of (a) rows, i.e., time period used for estimation (1975–1979 vs. 1995–1999); (b) columns, i.e., whether the regression includes the inverse Mills ratio (OLS does not include it, two-step does); and (c) panels, i.e., the weighting used to average the regression results across demographic groups (variable vs. fixed weights). The "Bias" column is the difference between the OLS and two-step columns. The "change" row is the difference between the 1995–1999 and 1975–1979 rows. Weights are factions of working women in each demographic group and are time-specific (variable) or pool both time periods (fixed).

The regressions control for demographics interacted with gender and use our CPS wage sample of white persons aged 25–54, trimming outliers and adjusting topcodes as described in Appendix I.

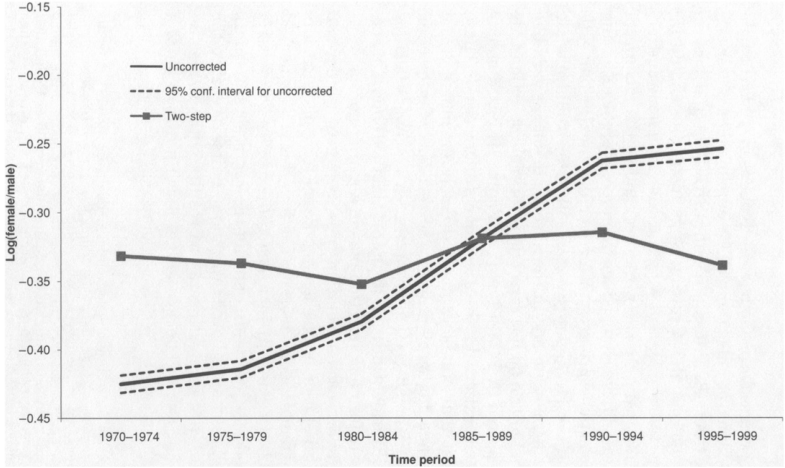


FIGURE III

Correcting the Gender Wage Gap: The Heckman Two-step Estimator

The figure graphs two time series of women's log wages relative to men's (unmarked and square-marked), plus a 95% confidence interval for one of them (dashed). Both relative wage series are net of measured demographic characteristics and averaged across demographic groups using time-specific female workforce weights. Only the two-step series (square-marked) is net of the inverse Mills ratio. The calculations use our CPS sample of white persons aged 25–54, trimming outliers and adjusting topcodes as described in Appendix I.

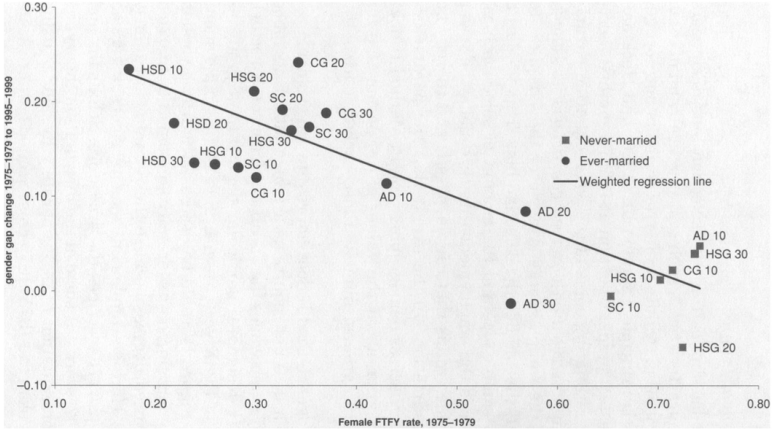


FIGURE IV

Measured Wage Growth Declines with Labor Supply

The scatter diagram shows the gender-gap change 1975–1979 to 1995–1999 against the FTFY employment rate 1975–1979 for the 21 demographic groups with at least 40 observations of female FTFY workers per year in the 1970s. The demographic groups are the cross-product of marital status (never-married vs. ever-married), schooling (high school dropout, HSD; high school grad, HSG; some college, SC; college grad, CG; and advanced degree, AD), and potential experience (10 denotes 5–14 years, 20 denotes 15–24 years, and 30 denotes 25–34 years). The calculations use our CPS sample of white persons aged 25–54, trimming outliers and adjusting topcodes as described in Appendix I.

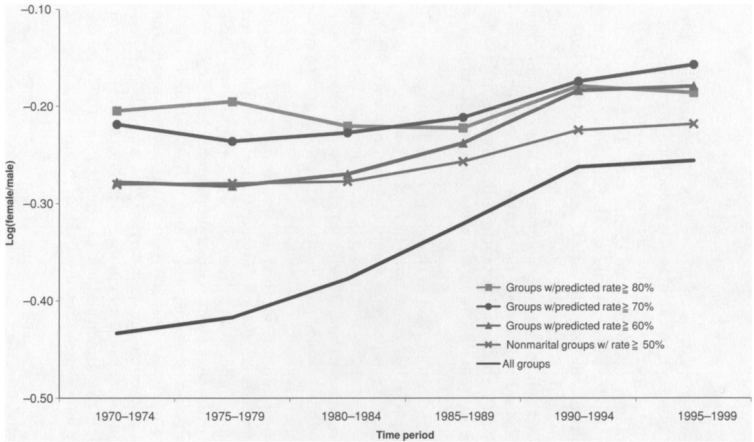


FIGURE V

Gender Wage Gaps Among Strongly Attached Groups, Various Thresholds

The figure graphs five times series of women's log wages relative to men's, net of demographic characteristics. The series differ according to the demographic groups (defined according to gender, schooling, marital status, and potential experience, except for the x-marked series that does not use marital status) included in the estimation. The unmarked series includes all demographic groups. For the other series, demographic groups are selected based on their FTFY employment rate for the years 1975–1979. The calculations use our CPS sample of white persons aged 25–54, trimming outliers and adjusting topcodes as described in Appendix I.

Discussion de Mulligan et Rubinstein (2008)

Hypothèse très provocante. Roy fait une prédiction très forte qui semble confirmée par les tendances agrégées et par les méthodes standard du modèle de Roy

Les preuves sont-elles convaincantes ?

Discussion de Mulligan et Rubinstein (2008)

Hypothèse très provocante. Roy fait une prédiction très forte qui semble confirmée par les tendances agrégées et par les méthodes standard du modèle de Roy

Les preuves sont-elles convaincantes ?

Les restrictions d'exclusion dans Heckit ne sont pas crédibles

L'identification à l'infini fournit un bon complément, mais il n'est pas impossible d'imaginer d'autres explications à ces tendances

Je pense que ce serait une question intéressante à revisiter avec des données administratives contenant un long panel. Le défi est de trouver des données qui remontent assez loin et des variables qui satisfont la restriction d'exclusion

- Le Longitudinal Administrative Databank (LAD) canadien remonte à 1982
- La base de données allemande Integrated Employment Biography (IEB) remonte à 1975
- Le recensement historique complet des États-Unis avec noms complets (permettant l'appariement) sort tous les 10 ans

Abramitzky, Boustan et Eriksson (2012, AER)

Quelle est la principale question de recherche ?

Comment répondent-ils à leur question ?

Les données ne contiennent pas d'information sur les salaires. Comment contournent-ils ce problème ?

Comment relient-ils les individus ? Expliquer l'appariement

Quelles sont les principales conclusions sur la sélection des immigrants ?

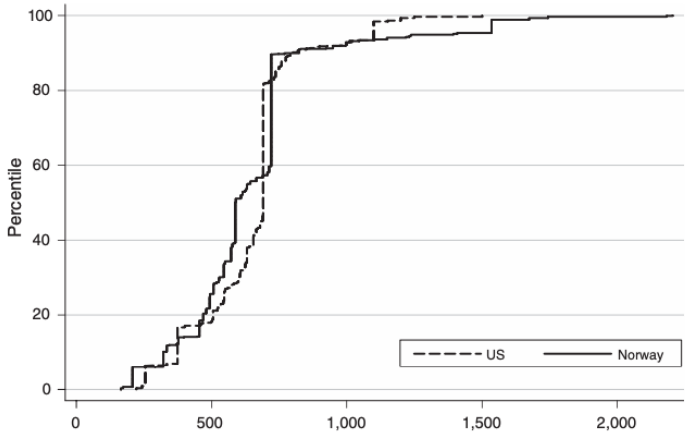
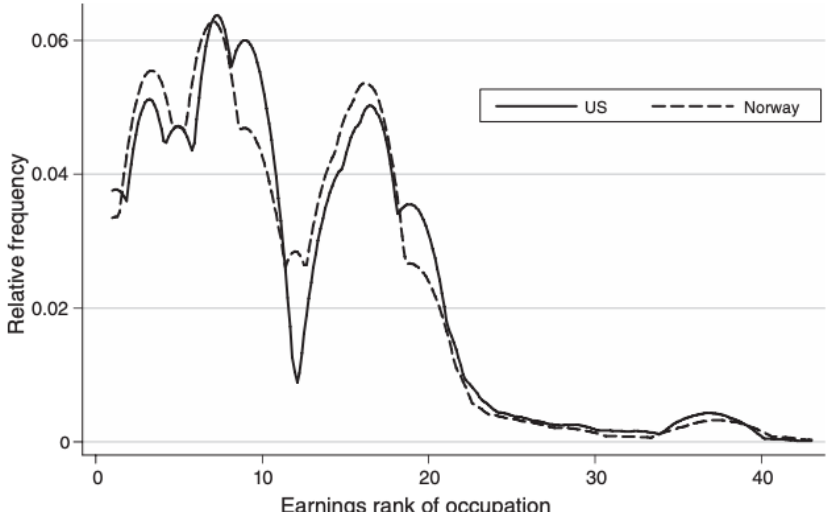
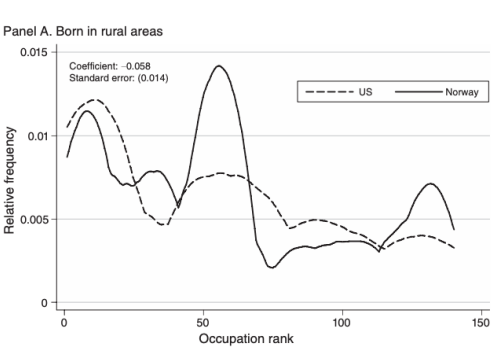


FIGURE 1. CUMULATIVE INCOME DISTRIBUTION FUNCTIONS IN THE UNITED STATES AND NORWAY IN 1900

Notes: US and Norwegian distributions contain all men aged 38 to 50 in the respective censuses of 1900. The *x*-axis is scaled in 1900 US dollars. Individuals are assigned the mean earnings for their occupation and are arrayed from lowest- to highest-paid occupations. The Norwegian distribution is rescaled to have the same mean as the US distribution (the actual Norwegian and US means are US$(1900)350 and US$(1900)643, respectively).


0.06
0.04
0.02
0
Relative frequency
US
Norway
0
10
20
30
40
Earnings rank of occupation



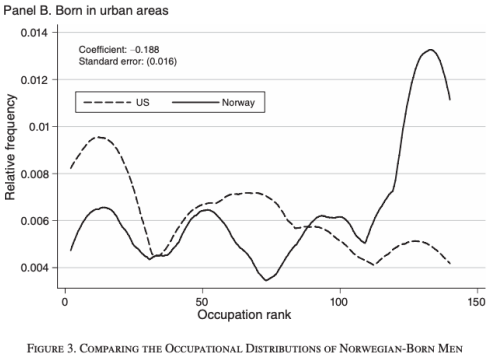


FIGURE 3. COMPARING THE OCCUPATIONAL DISTRIBUTIONS OF NORWEGIAN-BORN MEN IN THE UNITED STATES AND NORWAY IN 1900

TABLE 3—OLS AND WITHIN-HOUSEHOLD ESTIMATES OF THE RETURN TO MIGRATION. HOUSEHOLDS WITH TWO OR MORE MEMBERS IN THE MATCHED SAMPLE

Dependent variable = ln(earnings); Coefficient on = 1 if migrant	Full sample, 1865	Rural, 1865	Urban, 1865
Panel A. Unweighted			
OLS	0.545	0.607	0.384
	(0.027)	(0.034)	(0.044)
Within household	0.511	0.508	0.508
	(0.035)	(0.045)	(0.057)
Chi-squared	1.49	7.47	8.31
p-value	0.2218	0.0063	0.0039
N	2,655	1,823	832
Number of migrant-stayer pairs	326	167	159
Panel B. Weighted			
OLS	0.586	0.609	0.443
	(0.029)	(0.033)	(0.067)
Within household	0.542	0.529	0.561
	(0.039)	(0.042)	(0.049)
Chi-squared	2.13	4.60	5.65
p-value	0.1441	0.0320	0.0175
N	2,241	1,666	306
Number of migrant-stayer pairs	269	140	129

Notes: Each cell contains coefficient estimates and standard errors from regressions of ln(earnings) on a dummy variable equal to one for individuals living in the United States in 1900. Regressions also include controls for age and age squared. In each panel, the first row conducts an OLS regression for the restricted sample of households that have at least two matched members in the dataset and the second row adds household fixed effects. Panel B contains results from regressions weighted to reflect the urban status (full sample only), asset holdings, and occupational distribution of fathers in the full population. We conduct chi-squared tests of the null hypothesis that the OLS and within-household coefficients are equal.