Firm wages

Sam Gyetvay

ECO8000

November 10, 2025

Dispersion des salaires par entreprise en théorie

Dans le dernier cours, nous avons étudié des modèles où le même travailleur est payé différemment selon les entreprises

Burdett et Mortensen (1998)

- Tous les travailleurs et toutes les entreprises sont *ex ante* identiques
- En équilibre, chaque entreprise paie un salaire différent

Card, Cardoso, Heining et Kline (2018)

- Les entreprises diffèrent en productivité, les travailleurs ont des préférences idiosyncratiques pour les entreprises
- Les entreprises paient des salaires proportionnels à la productivité avec une décote qui dépend de l'élasticité de l'offre de travail

Pourquoi est-ce important ?

Cela nous aide à distinguer entre différents modèles du marché du travail

- Dans un marché du travail parfaitement compétitif, pas de dispersion des salaires par entreprise
- Comme nous l'avons vu, la plupart des modèles de marchés imparfaitement compétitifs présentent une dispersion des salaires par entreprise

La dispersion des salaires par entreprise nous permet également d'étudier les inégalités

- L'inégalité a-t-elle augmenté en raison de changements dans le capital humain des travailleurs ?
- Ou en raison de changements dans la rémunération des entreprises ?

Important également pour la politique

- Réduire les inégalités en appariant les travailleurs à faible salaire avec des entreprises à salaire élevé ?
- Augmenter l'efficacité en réaffectant les travailleurs vers des entreprises à salaire élevé ?

Slichter (1950) : une enquête sur les salaires de 1940 à Boston

	Average Hourly Earnings in All Plants (cents)	Average Hourly Earnings in Lowest Plant (cents)	Average Hourly Earnings in Highest Plant (cents)	Spread between High and Low Plants (cents)
Common labor	57.9	44.8	74.1	29.3
Janitor	55.3	41.0	70.5	29.5
Watchman	59.6	45.2	74.0	28.8
Producing and processing laborers	64.2	44.8	100.7	55.9
Producing and processing operators	72.0	57.8	88.6	30.8
Receiving and shipping clerks	68.0	50.0	89.6	39.6
Machinists	87.5	70.0	105.0	35.0
Steamfitter	86.4	70.0	105.0	35.0
Electrician	88.0	67.9	105.0	37.1
Carpenter	81.5	65.0	99.5	34.5
Sheet metal workers	85.4	77.8	90.5	12.7
Millwright	86.1	82.5	95.5	13.0
Maintenance helper	67.1	50.7	82.0	31.3
Female producing laborers	45.1	33.8	63.4	29.6
Female producing operators	47.9	37.7	58.3	20.6
Firemen	78.4	63.0	90.8	27.8

Slichter : “ni les taux de salaire ni les gains horaires ne représentent le prix du travail”

Primes salariales par industrie (Krueger et Summers, 1988)

Certaines industries paient-elles des salaires plus élevés ?

Si nous comparons simplement les salaires moyens à travers les industries, nous trouvons de grandes différences. Mais cela pourrait simplement être parce que différents travailleurs travaillent dans différentes industries

- Les secteurs de la technologie, du droit, de la finance paient des salaires élevés
- Mais les travailleurs qui sont très éduqués, travailleurs et intelligents ont tendance à travailler dans ces industries
- Est-ce l'industrie ou les travailleurs ?

Krueger et Summers s'intéressent à estimer ce qu'ils appellent les **primes salariales par industrie** : la hausse de salaire qu'un travailleur donné reçoit lorsqu'il travaille dans cette industrie, toutes les caractéristiques observées et non observées des travailleurs étant constantes

Régression transversale (Krueger et Summers, 1988)

Supposons que nous utilisions une année de données, donc nous observons chaque travailleur une fois (ceci est appelé **données transversales**), et nous estimons la régression suivante

$$\log w_i = \beta_s D_{s(i)} + X_i\delta + \varepsilon_i$$

- $D_{s(i)}$ sont des variables fictives par industrie indiquant dans quelle industrie le travailleur i travaille
- X_i est un vecteur de caractéristiques observées du travailleur (éducation, âge, expérience)
- ε_i sont des déterminants du salaire au niveau du travailleur non observés (capacité, motivation, etc.)

Qu'est-ce que β_s ?

Régression transversale (Krueger et Summers, 1988)

Supposons que nous utilisions une année de données, donc nous observons chaque travailleur une fois (ceci est appelé **données transversales**), et nous estimons la régression suivante

$$\log w_i = \beta_s D_{s(i)} + X_i\delta + \varepsilon_i$$

- $D_{s(i)}$ sont des variables fictives par industrie indiquant dans quelle industrie le travailleur i travaille
- X_i est un vecteur de caractéristiques observées du travailleur (éducation, âge, expérience)
- ε_i sont des déterminants du salaire au niveau du travailleur non observés (capacité, motivation, etc.)

Qu'est-ce que β_s ? Estimations des primes salariales par industrie dans le secteur s

Quand β_s est-il biaisé ?

Régression transversale (Krueger et Summers, 1988)

Supposons que nous utilisions une année de données, donc nous observons chaque travailleur une fois (ceci est appelé **données transversales**), et nous estimons la régression suivante

$$\log w_i = \beta_s D_{s(i)} + X_i\delta + \varepsilon_i$$

- $D_{s(i)}$ sont des variables fictives par industrie indiquant dans quelle industrie le travailleur i travaille
- X_i est un vecteur de caractéristiques observées du travailleur (éducation, âge, expérience)
- ε_i sont des déterminants du salaire au niveau du travailleur non observés (capacité, motivation, etc.)

Qu'est-ce que β_s ? Estimations des primes salariales par industrie dans le secteur s

Quand β_s est-il biaisé ? Lorsque ε_i est corrélé avec $D_{s(i)}$

Régression en panel (Krueger et Summers, 1988)

Supposons maintenant que nous disposons de plusieurs années de données, donc nous pouvons observer chaque travailleur plusieurs fois (ceci est appelé **données en panel**), et nous estimons la régression suivante

$$\log w_{it} = \alpha_i + \beta_s^* D_{s(i,t)} + X_{it}\delta + \varepsilon_{it}$$

- $D_{s(i,t)}$ sont des variables fictives par industrie indiquant dans quelle industrie le travailleur i travaille pendant la période t
- X_{it} est un vecteur de caractéristiques du travailleur variant dans le temps (âge, expérience)
- α_i est un "effet fixe du travailleur" capturant les déterminants du salaire au niveau du travailleur invariants dans le temps
- ε_{it} est

Régression en panel (Krueger et Summers, 1988)

Supposons maintenant que nous disposons de plusieurs années de données, donc nous pouvons observer chaque travailleur plusieurs fois (ceci est appelé **données en panel**), et nous estimons la régression suivante

$$\log w_{it} = \alpha_i + \beta_s^* D_{s(i,t)} + X_{it}\delta + \varepsilon_{it}$$

- $D_{s(i,t)}$ sont des variables fictives par industrie indiquant dans quelle industrie le travailleur i travaille pendant la période t
- X_{it} est un vecteur de caractéristiques du travailleur variant dans le temps (âge, expérience)
- α_i est un "effet fixe du travailleur" capturant les déterminants du salaire au niveau du travailleur invariants dans le temps
- ε_{it} est des déterminants du salaire variant dans le temps non observés (effort, santé, bonus, etc.)

β_s^* sont des estimations des primes salariales par industrie dans le secteur s

Quand β_s^* est-il biaisé ?

Régression en panel (Krueger et Summers, 1988)

Supposons maintenant que nous disposons de plusieurs années de données, donc nous pouvons observer chaque travailleur plusieurs fois (ceci est appelé **données en panel**), et nous estimons la régression suivante

$$\log w_{it} = \alpha_i + \beta_s^* D_{s(i,t)} + X_{it}\delta + \varepsilon_{it}$$

- $D_{s(i,t)}$ sont des variables fictives par industrie indiquant dans quelle industrie le travailleur i travaille pendant la période t
- X_{it} est un vecteur de caractéristiques du travailleur variant dans le temps (âge, expérience)
- α_i est un "effet fixe du travailleur" capturant les déterminants du salaire au niveau du travailleur invariants dans le temps
- ε_{it} est des déterminants du salaire variant dans le temps non observés (effort, santé, bonus, etc.)

β_s^* sont des estimations des primes salariales par industrie dans le secteur s

Quand β_s^* est-il biaisé ? Lorsque ε_{it} est corrélé avec $D_{s(i),t}$.

Identification de β_s

β_s est identifié à partir de comparaisons entre travailleurs à un moment donné

Considérons deux travailleurs, i et j ayant le même X_i, qui travaillent dans des industries différentes s et s' respectivement

$$\log w_i = \beta_s + X_i\delta + \varepsilon_i$$
$$\log w_j = \beta_{s'} + X_i\delta + \varepsilon_j$$

Alors si $E[\varepsilon|D, X] = 0$, nous avons

$$\beta_s - \beta_{s'} = E[\log w_i|D_s = 1, X_i = x] - E[\log w_i|D_{s'} = 1, X_i = x]$$

Identification de β_s^*

β_s^* est identifié à partir de comparaisons dans le temps au sein d'un même travailleur

Considérons le travailleur i qui travaille dans l'industrie s pendant la période t et dans l'industrie s' pendant la période $t-1$

$$\begin{aligned}\log w_{it} &= \alpha_i + \beta_s^* + X_{it}\delta + \varepsilon_{it} \\ \log w_{it-1} &= \alpha_i + \beta_{s'}^* + X_{it-1}\delta + \varepsilon_{it-1}\end{aligned}$$

Alors si $E[\varepsilon|D, X] = 0$, nous avons

$$\beta_s^* - \beta_{s'}^* = E[\log w_{it}|D_s = 1, X_i = x] - E[\log w_{it-1}|D_{s'} = 1, X_i = x]$$

TABLE I

ESTIMATED WAGE DIFFERENTIALS FOR ONE-DIGIT INDUSTRIES—MAY CPS[a]
(Standard Errors in Parentheses)

Industry	(1) 1974	(2) 1979	(3) 1984	(4) 1984 Total Compensation
Construction	.195	.126	.108	.091
	(.021)	(.031)	(.034)	(.035)
Manufacturing	.055	.044	.091	.131
	(.020)	(.029)	(.032)	(.032)
Transportation & Public Utilities	.111	.081	.145	.203
	(.021)	(.031)	(.034)	(.034)
Wholesale & Retail Trade	−.128	−.082	−.111	−.136
	(.020)	(.030)	(.033)	(.033)
Finance, Insurance and Real Estate	.047 (.022)	−.010 (.035)	.055 (.034)	.069 (.034)
Services	−.070	−.055	−.078	−.111
	(.021)	(.030)	(.032)	(.032)
Mining	.179	.229	.222	.231
	(.035)	(.058)	(.075)	(.075)
Weighted Adjusted Standard Deviation of Differentials[b]	.097**	.069**	.094**	.126**
Sample Size	29,945	8,978	11,512	11,512

[a] Other explanatory variables are education and its square, 6 age dummies, 8 occupation dummies, 3 region dummies, sex dummy, race dummy, central city dummy, union member dummy, ever married dummy, veteran status, marriage × sex interaction, education × sex interaction, education squared × sex interaction, 6 age × sex interactions, and a constant. Each column was estimated from a separate cross-sectional regression.

[b] Weights are employment shares for each year.

** F test that industry wage differentials jointly equal 0 rejects at the .000001 level.

TABLE IV

THE EFFECTS OF UNMEASURED LABOR QUALITY[a]

Industry	(1) Fixed Effects Unadjusted for Measurement Error	(2) Fixed Effects Adjusted for Measurement Error I[b]	(3) Fixed Effects Adjusted for Measurement Error II[c]	(4) Levels
Construction	.063	.098	.174	.174
	(.033)	(.060)	(.060)	(.024)
Manufacturing	.028	.055	.107	.064
	(.031)	(.058)	(.058)	(.022)
Transportation and Public Utilities	.019	.060	.049	.114
	(.035)	(.059)	(.059)	(.024)
Wholesale and Retail Trade	−.042	−.068	−.125	−.133
	(.031)	(.056)	(.056)	(.023)
Finance, Insurance and Real Estate	.027	.017	.018	.035
	(.036)	(.061)	(.061)	(.025)
Services	−.040	−.088	−.128	−.079
	(.032)	(.056)	(.057)	(.023)
Mining	.067	.122	.142	.156
	(.004)	(.057)	(.058)	(.040)

[a] Data set is three matched May CPS's pooled together: 1974–1975, 1977–1978, and 1979–1980. Sample size is 18,122. Levels are 1974, 1977, and 1979 data pooled. Results of the 1975, 1978, and 1980 sample are qualitatively the same. Controls for fixed effects regressions are change in education and its square, change in occupation, 3 region dummies, change in union membership, experience squared, change in marital status, year dummies, and a constant. Controls for level regressions are the same as Table I plus year dummies.

[b] Adjustment I assumes 3.4 per cent error rate and that misclassifications are proportional to industry size. See Appendix for description.

[c] Adjustment II assumes average error rate is 3.4 per cent and misclassifications are allocated according to employer-employee mismatches. See Appendix for description.

TABLE IX

THE EFFECT OF INDUSTRY WAGE DIFFERENTIALS ON JOB TENURE AND QUITS

Independent Variables	Dependent Variable[a] (1) Tenure	(2) Quit[b]
Industry wage premium	2.198	−.078
	(.676)	(.135)
Union (1 = yes)	3.179	−.164
	(.157)	(.037)
Other variables	Age dummies (6), Age * Sex (6), Education, Education Squared * Sex, Region Dummies (3), Race Dummy, Sex Dummy, Central City Dummy, Firm Size Dummies (4), Plant Size Dummies (4), Marriage Dummy, Marriage * Sex, Veteran Status Dummy	Education, Education Squared, Region Dummies (3), Race Dummy, Sex Dummy, SMSA Dummy, (Age—Education—5) and its square
Sample Size	8,978	633
R^2	.40	.20

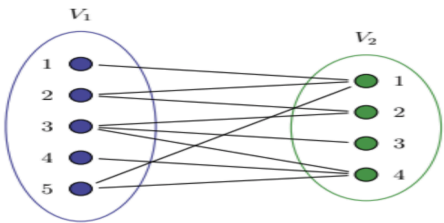
[a] Mean (SD) of Tenure is 5.70 (7.61); Mean (SD) of Quit is .26 (.44).
[b] Quit equation was estimated with a linear probability model.

Données appariées employeur-employé

Les travailleurs et les entreprises sont **appariés** à différents moments

Exemple de structure de données de réseau "bipartite" (Bonhomme, 2020)

- Patients et médecins
- Élèves et enseignants



id	year	firmid	log_dailyw-s
1773242	2001	8401145847	5.186759
1773306	1985	9003001003	5.007141
1773306	1986	9003001003	4.784913
1773306	1987	9004100701	4.612522
1773306	1988	9004474273	4.479647
1773306	1989	9004233946	4.529663
1773306	1990	9002109193	4.600437
1773306	1991	9002109193	4.648328
1773306	1992	9002109193	4.623967
1773306	1993	9002109193	4.610094
1773306	1994	9002109193	4.607891
1773306	1995	9002109193	4.61261
1773306	1996	9002109193	4.590056
1773306	1997	9002109193	4.623967
1773306	1998	9002109193	4.634075
1773306	1999	9002109193	4.456581
1773306	2000	9002109193	4.59681
1773306	2001	9002634691	4.704414
1773323	1985	8400616106	4.801253
1773323	1986	8400616106	4.841681
1773323	1987	8401984114	4.799707
1773323	1988	8401984114	5.126745
1773323	1989	8402812986	4.667272
1773323	1990	8402812986	4.721242
1773323	1991	8402812986	4.758056
1773323	1992	8402812986	4.827647
1773323	1993	8402812986	4.887209

Où obtenir des données appariées employeur-employé

De nombreux pays (la plupart ?) disposent de jeux de données appariées employeur-employé

- Brésil (RAIS), Veneto Italie (VWH) : tout le monde peut télécharger sur un ordinateur portable
- Allemagne (IAB) : postuler via FDZ, envoyer des scripts par courriel pour exécution à distance
- Canada (CEEDD, BEAM) : analyser les données au CRDC de StatCan
- USA (LEHD) : accès via FSRDC au Census. Processus d'approbation étendu
- USA (IRS/Trésor) : Formulaires 1120, 1120S, 1065, W-2
- Scandinavie/N. Europe (Norvège, Danemark, Suède, Finlande) : liens extrêmement riches

Où obtenir des données appariées employeur-employé

De nombreux pays (la plupart ?) disposent de jeux de données appariées employeur-employé

- Brésil (RAIS), Veneto Italie (VWH) : tout le monde peut télécharger sur un ordinateur portable
- Allemagne (IAB) : postuler via FDZ, envoyer des scripts par courriel pour exécution à distance
- Canada (CEEDD, BEAM) : analyser les données au CRDC de StatCan
- USA (LEHD) : accès via FSRDC au Census. Processus d'approbation étendu
- USA (IRS/Trésor) : Formulaires 1120, 1120S, 1065, W-2
- Scandinavie/N. Europe (Norvège, Danemark, Suède, Finlande) : liens extrêmement riches

Ma philosophie : soyez pragmatique, pas patriotique

- Trouvez les données qui correspondent le mieux à votre question de recherche, sous réserve de l'accessibilité

Veneto Worker Histories Database (VWH)

Nous utiliserons ce base de données dans le second ensemble de problèmes

Vous pouvez lire à propos du base de données ici :

`https://www.jarellanobover.com/veneto-workers-histories-dataset-additional-resources`

Vous devriez demander l'accès au VWH ici

`https://www.frdb.org/en/dati/dati-inps-carriere-lavorative-in-veneto/`

Ce base de données pourrait être intéressant pour un projet de thèse de maitrise ou de doctorat avec une substance méthodologique

Un terrain de jeu intéressant pour apprendre l'analyse des données appariées employeur-employé

Additional resources for users of the Veneto Workers History dataset (VWH)

Here, I provide some additional information on the VWH dataset. My hope is that these prove to be useful tips for a first-time user of the data. Some of the information on this page is already in the *readme* file of the replication package in *Differences in On-the-Job Learning across Firms* (Arellano-Bover and Saltiel, 2026), but some other things mentioned here go beyond the strictly necessary to replicate the paper. Click here for the actual replication files for *Differences in On-the-Job Learning across Firms*.

Background and access

The VWH was originally developed by the Economics Department in Università Ca' Foscari Venezia under the supervision of Giuseppe Tattara. A copy of the dataset can easily be obtained through the Fondazione Rodolfo Debenedetti (fRDB), following the instructions here. The dataset description provided by fRDB is **here**.

Here is a non-comprehensive list of published papers using these data: Card et al., 2014; Battisti, 2017; Bartolucci et al., 2018; Serafinelli 2019; Kline et al., 2020; Hong and Lattanzio, 2025. Their replication files can be useful resources too.

Files

The VWH is composed of three data files, which can be linked to each other:

- Person-level dataset (*anagr.dta*)
- Firm-level dataset (*azien.dta*)
- Contract-level dataset (*contr.dta*)

Workers are uniquely identified by the variable *cod_pgr*. Firms are uniquely identified by the variable *matr_az*. These variables can be used to link the person-level dataset *anagr.dta* and the firm-level dataset *azien.dta* to the contract-level dataset *contr.dta*.

Sampling

VWH covers the employment histories of individuals who ever work in the Italian region of Veneto between 1975-2001.

- If a worker leaves Veneto and is employed elsewhere in Italy, information on that non-Veneto contract and non-Veneto employer is also recorded in VWH. As such, there is no sample attrition due to inter-regional migration. However, this implies that the sample is representative of individuals who ever work in Veneto, and the firm-level data is representative of Veneto firms, even if it includes information on some non-Veneto firms.
- Note that the fRDB VWH documentation file mentions that the data only covers the provinces of Vicenza and Treviso. However, the data clearly encompasses the entire Veneto region. Kline et al. (2020) reach a similar conclusion.
- There are some missing data issues before the mid 1980s. In *Differences in On-the-Job Learning across Firms,* we carry out our analysis 1984-2001. Kline et al. (2020) do 1985-2001 instead.

Firms' tax number

The dataset *azien.dta* includes firm's national tax number (*codice fiscale)*, which implies that firms in the VWH are not anonymized. This is recorded in the variable *cod_fis*. Card et al., 2014 use this information to match VWH to firms' balance sheet records, provided by the AIDA database (Bureau van Djik).

The firm identifiers *matr_az and cod_fis* may in principle not coincide, as the first identifies the contribution ID of the firm, while the second identifies the tax number. However, in VWH, 99.96% of *matr_az* observations have a 1-to-1 mapping to *cod_fis*, and only a few hundred *matr_az* have a many-to-1 mapping to *cod_fis*.

Sectors

The dataset *azien.dta* includes the variable *ateco81* reflecting a firm's 3-digit sector, following the ATECO classification from 1981. **This PDF file** lists the descriptions of each of the 3-digit 1981 ATECO codes.

Geography

Firms: The dataset *azien.dta* includes firms' full addresses:

- Address (*indirizzo*)
- Municipality name (*comune*)
- Postal code (*cap*)
- Province (*prov*)
- Municipality statistical code (*cod_com*)

The statistical code *cod_com* can be linked to other municipality-level datasets. Here is a spreadsheet with Italian municipalities and their statistical code.

Workers: The dataset *anagr.dta* includes geographical information on workers' place of birth:

- Municipality of birth name (*com_n*)
- Province of birth (*prov_n*)

Note, however, that *anagr.dta* does not include municipality code, so any linkage to external data would have to be done using municipality names.

Also note that *anagr.dta* includes workers' municipality and province of residence (*com_r* and *prov_r*). However, it is unclear which is the time period in which an individual's residence is recorded.

Occupations

The dataset *contr.dta* includes information on the type of position each employment contract is assigned to. This information is recorded in the variable *qualif*. This variable does not capture occupation in the sense of ISCO or SOC, but in essence classifies workers into managers (*dirigente*), white-collar (*impiegato*), blue-collar (*operaio*), and apprentices (*apprendista*):

- The variable *qualif* takes on many different values, which are listed and named in **this spreadsheet.**
- In *Differences in On-the-Job Learning across Firms*, we use the names attached to qualif values to map qualif into managers, white-collar, blue-collar, and apprentices. **This is the crosswalk**.

Econometrica, Vol. 67, No. 2 (March, 1999), 251–333

HIGH WAGE WORKERS AND HIGH WAGE FIRMS

BY JOHN M. ABOWD, FRANCIS KRAMARZ, AND DAVID N. MARGOLIS[1]

We study a longitudinal sample of over one million French workers from more than five hundred thousand employing firms. We decompose real total annual compensation per worker into components related to observable employee characteristics, personal heterogeneity, firm heterogeneity, and residual variation. Except for the residual, all components may be correlated in an arbitrary fashion. At the level of the individual, we find that person effects, especially those not related to observables like education, are a very important source of wage variation in France. Firm effects, while important, are not as important as person effects. At the level of firms, we find that enterprises that hire high-wage workers are more productive but not more profitable. They are also more capital and high-skilled employee intensive. Enterprises that pay higher wages, controlling for person effects, are more productive and more profitable. They are also more capital intensive but are not more high-skilled labor intensive. We find that person effects explain about 90% of inter-industry wage differentials and about 75% of the firm-size wage effect while firm effects explain relatively little of either differential.

KEYWORDS: Wage determination, person effects, firm effects, inter-industry wage differentials, heterogeneity.

Abowd, Kramarz, Margolis (1999)

AKM fait référence à la spécification suivante de l'**équation des salaires** :

$$y_{it} = \alpha_i + \psi_{J(i,t)} + W'_{it}\beta + \epsilon_{it}$$

- y_{it} : salaire logarithmique du travailleur i en année t
- α_i : « effets individuels »
- $\psi_{J(i,t)}$: « effets d'entreprise »
- W_{it} : âge, expérience

α_i représente le capital humain « portable » que les travailleurs apportent avec eux d'une entreprise à l'autre

ψ_j sont des primes salariales communes de l'entreprise payées à tous les travailleurs

Ce que vous gagnez dépend de qui vous êtes **plus** où vous travaillez

La croissance salariale provient de la montée de l'« échelle professionnelle » vers des entreprises à ψ_j plus élevés

Décomposition de la variance AKM

La principale raison de s'intéresser à AKM est la décomposition de variance suivante

$$\mathbb{V}_n(y_{it} - W_{it}'\beta) = \underbrace{\mathbb{V}_n(\alpha_i)}_{\text{variance de l'effet individuel}} + \underbrace{\mathbb{V}_n(\psi_j)}_{\text{variance de l'effet entreprise}} + \underbrace{2\mathbb{C}_n(\psi_j, \alpha_i)}_{\text{appariement}} + \underbrace{\mathbb{V}_n(\epsilon_{it})}_{\text{bruit}}$$

Décomposition de la variance AKM

La principale raison de s'intéresser à AKM est la décomposition de variance suivante

$$\mathbb{V}_n(y_{it} - W_{it}'\beta) = \underbrace{\mathbb{V}_n(\alpha_i)}_{\text{variance de l'effet individuel}} + \underbrace{\mathbb{V}_n(\psi_j)}_{\text{variance de l'effet entreprise}} + \underbrace{2\mathbb{C}_n(\psi_j, \alpha_i)}_{\text{appariement}} + \underbrace{\mathbb{V}_n(\epsilon_{it})}_{\text{bruit}}$$

$\mathbb{V}_n(\psi_j)$ mesure le rôle de la politique salariale des entreprises dans l'inégalité des salaires

Appariement $\mathbb{C}_n(\psi_j, \alpha_i)$: les travailleurs à haut salaire travaillent dans des entreprises à haut salaire

Décomposition de la variance AKM

La principale raison de s'intéresser à AKM est la décomposition de variance suivante

$$\mathbb{V}_n(y_{it} - W_{it}'\beta) = \underbrace{\mathbb{V}_n(\alpha_i)}_{\text{variance de l'effet individuel}} + \underbrace{\mathbb{V}_n(\psi_j)}_{\text{variance de l'effet entreprise}} + \underbrace{2\mathbb{C}_n(\psi_j, \alpha_i)}_{\text{appariement}} + \underbrace{\mathbb{V}_n(\epsilon_{it})}_{\text{bruit}}$$

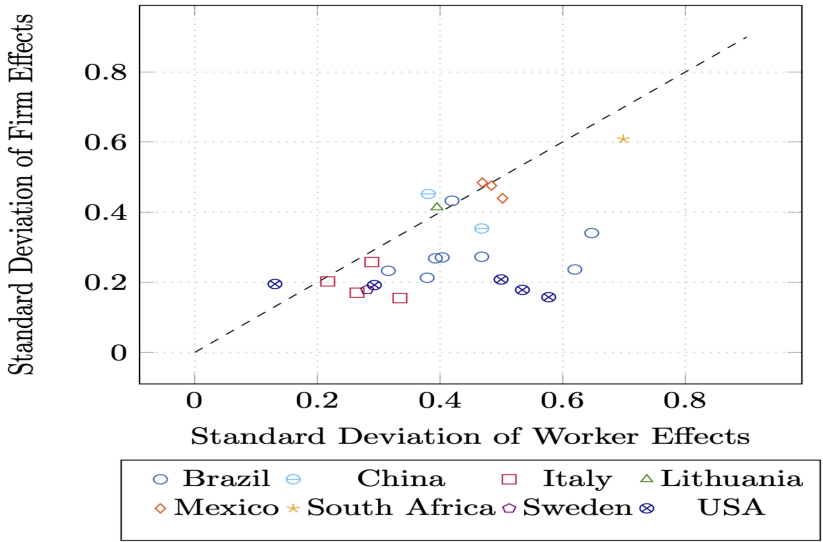
$\mathbb{V}_n(\psi_j)$ mesure le rôle de la politique salariale des entreprises dans l'inégalité des salaires

Appariement $\mathbb{C}_n(\psi_j, \alpha_i)$: les travailleurs à haut salaire travaillent dans des entreprises à haut salaire

L'estimation des composantes de variance était économétriquement difficile en raison du biais de mobilité limitée des estimateurs standards « plug-in ». Problème maintenant résolu

- Kline, Saggio, Solvsten (2021) : correction de biais par ajustement croisé
- Bonhomme, Holzheu, Lamadon, Manresa, Mogstad, Setzler (2023) : effets aléatoires corrélés

Kline (2025)


Standard Deviation of Firm Effects
0
0.2
0.4
0.6
0.8
0
0.2
0.4
0.6
0.8
Standard Deviation of Worker Effects
Brazil
China
Italy
Lithuania
Mexico
South Africa
Sweden
USA

Bonhomme, Holzheu, et al. (2023)

	sample information						$Var(\psi)$				$2 \times Cov(\alpha, \psi)$			
	T	set	firms	workers	movers	var(y)	btw firm	FE	FE-HO	FE-HE	CRE	FE	FE-HO	FE-HE	CRE
US	6	connected	2,568	55,464	14,888	0.414	39.6%	12.2%	5.5%	-	6.2%	1.1%	13.5%	-	15.0%
	6	leave-out	1,689	52,484	13,968	0.416	38.8%	9.5%	5.5%	5.8%	5.9%	5.9%	13.0%	12.5%	14.6%
US	3	connected	1,241	36,826	4,252	0.436	38.2%	16.3%	4.1%	-	5.3%	-12.0%	11.7%	-	12.5%
	3	leave-out	670	33,031	3,645	0.440	37.6%	10.4%	4.3%	4.5%	5.0%	-0.8%	11.0%	10.6%	12.1%
Austria	6	connected	206	3,396	1,123	0.187	45.5%	18.7%	15.3%	-	11.7%	4.7%	10.5%	-	19.6%
	6	leave-out	140	3,240	1,055	0.182	43.7%	15.5%	12.7%	12.9%	11.1%	8.7%	13.5%	13.0%	18.9%
Austria	3	connected	117	2,845	387	0.183	43.7%	19.7%	12.1%	-	10.1%	-5.3%	9.3%	-	17.5%
	3	leave-out	68	2,604	336	0.178	41.8%	15.0%	10.7%	13.9%	9.2%	1.5%	9.7%	3.2%	16.2%
Italy	6	connected	92	1,111	379	0.167	46.1%	23.1%	17.5%	-	12.7%	-1.3%	8.7%	-	20.0%
	6	leave-out	61	1,034	346	0.168	44.8%	19.3%	15.8%	15.7%	12.3%	4.7%	11.1%	11.2%	19.3%
Italy	3	connected	54	864	148	0.176	44.9%	24.1%	15.7%	-	11.0%	-8.4%	7.7%	-	17.7%
	3	leave-out	30	755	121	0.181	43.5%	18.5%	14.6%	10.9%	10.2%	1.3%	8.8%	16.1%	17.2%
Norway	6	connected	114	1,286	556	0.239	47.2%	24.4%	13.9%	-	11.8%	-7.7%	11.3%	-	16.8%
	6	leave-out	78	1,199	519	0.236	45.8%	19.2%	12.5%	12.3%	11.0%	0.8%	12.6%	12.8%	16.3%
Norway	3	connected	63	986	203	0.229	44.5%	37.8%	14.9%	-	11.5%	-41.3%	2.5%	-	12.2%
	3	leave-out	37	856	175	0.227	42.6%	24.2%	12.1%	10.2%	10.3%	-16.7%	6.3%	10.1%	11.3%
Sweden	6	connected	63	1,921	608	0.164	31.6%	14.6%	8.2%	-	5.0%	-8.1%	3.9%	-	10.3%
	6	leave-out	52	1,850	596	0.164	30.9%	11.6%	7.8%	7.1%	4.7%	-3.2%	3.7%	5.0%	10.0%
Sweden	3	connected	42	1,497	237	0.161	31.3%	23.6%	11.6%	-	4.6%	-28.5%	-5.4%	-	9.0%
	3	leave-out	29	1,377	221	0.161	30.2%	15.5%	8.9%	7.4%	4.3%	-14.1%	-1.3%	1.5%	8.1%

Identification AKM

AKM est juste un modèle de régression linéaire de haute dimension. Réécrivons AKM en notation matricielle :

$$y_{it} = \alpha_i + \psi_{J(i,t)} + W_{it}\gamma + \epsilon_{it}$$

$$\underbrace{y}_{n\times 1} = \underbrace{D}_{n\times N}\alpha + \underbrace{F}_{n\times J}\psi + \underbrace{W}_{n\times P}\gamma + \underbrace{\epsilon}_{n\times 1}$$

$$y = X\beta + \epsilon$$

β est identifié si (1) $X'X$ est inversible + (2) $E[\epsilon|X] = 0$. Concentrons-nous sur (1) pour l'instant

X contient des effets fixes pour des dizaines de millions de travailleurs et des millions d'entreprises. Quand est-ce que $X'X$ est non singulier ? À quoi ressemble la multicollinéarité ?

Identification AKM

AKM est juste un modèle de régression linéaire de haute dimension. Réécrivons AKM en notation matricielle :

$$y_{it} = \alpha_i + \psi_{J(i,t)} + W_{it}\gamma + \epsilon_{it}$$

$$\underbrace{y}_{n\times 1} = \underbrace{D}_{n\times N}\alpha + \underbrace{F}_{n\times J}\psi + \underbrace{W}_{n\times P}\gamma + \underbrace{\epsilon}_{n\times 1}$$

$$y = X\beta + \epsilon$$

β est identifié si (1) $X'X$ est inversible + (2) $E[\epsilon|X] = 0$. Concentrons-nous sur (1) pour l'instant

X contient des effets fixes pour des dizaines de millions de travailleurs et des millions d'entreprises. Quand est-ce que $X'X$ est non singulier ? À quoi ressemble la multicollinéarité ?

Si les travailleurs ne changeaient jamais d'entreprise, α_i et ψ_j seraient-ils identifiés ?

Identification AKM

AKM est juste un modèle de régression linéaire de haute dimension. Réécrivons AKM en notation matricielle :

$$y_{it} = \alpha_i + \psi_{J(i,t)} + W_{it}\gamma + \epsilon_{it}$$

$$\underbrace{y}_{n\times 1} = \underbrace{D}_{n\times N}\alpha + \underbrace{F}_{n\times J}\psi + \underbrace{W}_{n\times P}\gamma + \underbrace{\epsilon}_{n\times 1}$$

$$y = X\beta + \epsilon$$

β est identifié si (1) $X'X$ est inversible + (2) $E[\epsilon|X] = 0$. Concentrons-nous sur (1) pour l'instant

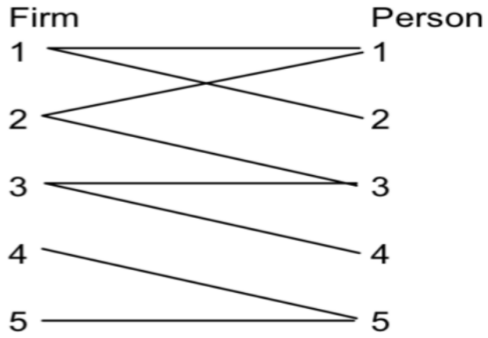
X contient des effets fixes pour des dizaines de millions de travailleurs et des millions d'entreprises. Quand est-ce que $X'X$ est non singulier ? À quoi ressemble la multicollinéarité ?

Si les travailleurs ne changeaient jamais d'entreprise, α_i et ψ_j seraient-ils identifiés ? **Non !**

ψ_j sont identifiés parmi des ensembles d'entreprises connectées par des mouvements, à la fois directement et indirectement (c.-à-d. à travers des « chemins » de mouvements par différents travailleurs)

Identification AKM Partie I : Inversibilité

Firm	Person	Period
1	1	1
1	2	1
2	1	2
2	3	1
3	3	2
3	4	1
4	5	1
5	5	2

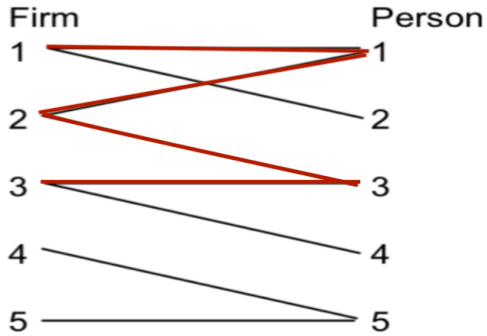


$\psi_3 - \psi_1$ est-il identifié ?

$$
\begin{aligned}
& E[y_{3,2} - y_{3,1}] + E[y_{1,2} - y_{1,1}] \\
&= \alpha_3 + \psi_3 - \alpha_3 - \psi_2 + \alpha_1 + \psi_2 - \alpha_1 - \psi_1 \\
&= \psi_3 - \psi_1
\end{aligned}
$$

Identification AKM Partie I : Inversibilité

Firm	Person	Period
1	1	1
1	2	1
2	1	2
2	3	1
3	3	2
3	4	1
4	5	1
5	5	2

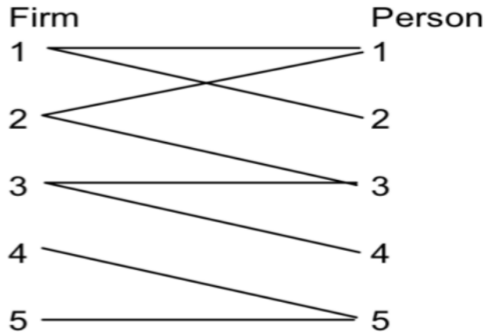


$\psi_3 - \psi_1$ est-il identifié ?

$$
\begin{aligned}
&E[y_{3,2} - y_{3,1}] + E[y_{1,2} - y_{1,1}] \\
&\quad = \alpha_3 + \psi_3 - \alpha_3 - \psi_2 + \alpha_1 + \psi_2 - \alpha_1 - \psi_1 \\
&\quad = \psi_3 - \psi_1
\end{aligned}
$$

Identification AKM Partie I : Inversibilité

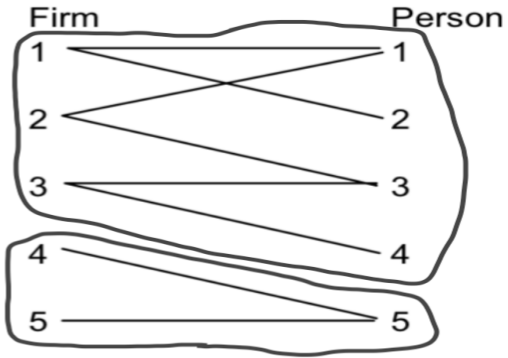
Firm	Person	Period
1	1	1
1	2	1
2	1	2
2	3	1
3	3	2
3	4	1
4	5	1
5	5	2



Est-ce que $\psi_5 - \psi_1$ est identifié ?

Identification AKM Partie I : Inversibilité

Firm	Person	Period
1	1	1
1	2	1
2	1	2
2	3	1
3	3	2
3	4	1
4	5	1
5	5	2



Est-ce que $\psi_5 - \psi_1$ est identifié ? **Non !** Ils sont dans deux composantes connexes séparées.

Identification AKM Partie II : Exogénéité

$$y = D\alpha + F\psi + W\gamma + \epsilon = X\beta + \epsilon$$

Autre condition pour l'identification de β est l'exogénéité $E[X'\epsilon]$

Identification AKM Partie II : Exogénéité

$$y = D\alpha + F\psi + W\gamma + \epsilon = X\beta + \epsilon$$

Autre condition pour l'identification de β est l'exogénéité $E[X'\epsilon]$

Focus sur $E[F'\epsilon] = 0$, appelé « mobilité exogène ». Ça paraît fou... mais l'est-ce vraiment ?

Identification AKM Partie II : Exogénéité

$$y = D\alpha + F\psi + W\gamma + \epsilon = X\beta + \epsilon$$

Autre condition pour l'identification de β est l'exogénéité $E[X'\epsilon]$

Focus sur $E[F'\epsilon] = 0$, appelé « mobilité exogène ». Ça paraît fou... mais l'est-ce vraiment ?

$$E[F'\epsilon] = 0 \iff P(J(i,t) = j|\epsilon_{it}) = P(J(i,t) = j) = G(\alpha_i, \psi_1, ..., \psi_J)$$

- Les travailleurs *peuvent* se trier selon α_i et ψ_j
- OK si le tri est motivé par des avantages non salariaux spécifiques à l'entreprise

Identification AKM Partie II : Exogénéité

$$y = D\alpha + F\psi + W\gamma + \epsilon = X\beta + \epsilon$$

Autre condition pour l'identification de β est l'exogénéité $E[X'\epsilon]$
Focus sur $E[F'\epsilon] = 0$, appelé « mobilité exogène ». Ça paraît fou... mais l'est-ce vraiment ?

$$E[F'\epsilon] = 0 \iff P(J(i,t) = j|\epsilon_{it}) = P(J(i,t) = j) = G(\alpha_i, \psi_1, ..., \psi_J)$$

- Les travailleurs *peuvent* se trier selon α_i et ψ_j
- OK si le tri est motivé par des avantages non salariaux spécifiques à l'entreprise

$$\epsilon_{it} = \underbrace{m_{j(i,t),i}}_{\text{Effet de Correspondance}} + \underbrace{v_{it}}_{\text{Composante de Racine Unité}} + \underbrace{e_{it}}_{\text{Erreur de Mesure}}$$

- Les mouvements entre entreprises ne sont pas motivés par des effets de correspondance ou par une « dérive » dans le salaire attendu
- Ces éléments ont des prédictions empiriques testables : symétrie et absence de tendances

Symétrie

Sous AKM, le **gain** de salaire qu'un travailleur obtient en passant de l'entreprise 1 à l'entreprise 2 est le même que la **perte** de salaire subie par les travailleurs passant de l'entreprise 2 à 1 :

$$E[y_{it} - y_{i,t-1} | j(i,t) = 2, j(i,t-1) = 1] = \psi_2 - \psi_1$$
$$E[y_{it} - y_{i,t-1} | j(i,t) = 1, j(i,t-1) = 2] = \psi_1 - \psi_2$$

Les mouvements ascendants et descendants dans l'échelle professionnelle sont **symétriques** :
$\psi_2 - \psi_1 + \psi_1 - \psi_2 = 0$

Symétrie

Sous AKM, le **gain** de salaire qu'un travailleur obtient en passant de l'entreprise 1 à l'entreprise 2 est le même que la **perte** de salaire subie par les travailleurs passant de l'entreprise 2 à 1 :

$$E[y_{it} - y_{i,t-1} | j(i,t) = 2, j(i,t-1) = 1] = \psi_2 - \psi_1$$
$$E[y_{it} - y_{i,t-1} | j(i,t) = 1, j(i,t-1) = 2] = \psi_1 - \psi_2$$

Les mouvements ascendants et descendants dans l'échelle professionnelle sont **symétriques** : $\psi_2 - \psi_1 + \psi_1 - \psi_2 = 0$

> *« Faire une promenade » le long du graphe ne devrait avoir aucun effet sur les salaires tant qu'on termine au même endroit où la promenade a commencé. - Kline (2025)*

Symétrie

Sous AKM, le **gain** de salaire qu'un travailleur obtient en passant de l'entreprise 1 à l'entreprise 2 est le même que la **perte** de salaire subie par les travailleurs passant de l'entreprise 2 à 1 :

$$E[y_{it} - y_{i,t-1} | j(i,t) = 2, j(i,t-1) = 1] = \psi_2 - \psi_1$$
$$E[y_{it} - y_{i,t-1} | j(i,t) = 1, j(i,t-1) = 2] = \psi_1 - \psi_2$$

Les mouvements ascendants et descendants dans l'échelle professionnelle sont **symétriques** : $\psi_2 - \psi_1 + \psi_1 - \psi_2 = 0$

> *« Faire une promenade » le long du graphe ne devrait avoir aucun effet sur les salaires tant qu'on termine au même endroit où la promenade a commencé. - Kline (2025)*

Si la mobilité est motivée par la sélection sur les effets de correspondance, alors le gain (perte) de salaire pour un travailleur passant de l'entreprise 1 à l'entreprise 2 (et vice versa) est

$$\psi_2 - \psi_1 + E[m_{i2} - m_{i1} | j(i,t) = 2, j(i,t-1) = 1]$$
$$\psi_1 - \psi_2 + E[m_{i1} - m_{i2} | j(i,t) = 1, j(i,t-1) = 2]$$

S'il y a une sélection positive de Roy sur les effets de correspondance, alors les deux termes de correspondance seront positifs

Absence de tendances préalables

La dérive peut survenir dans les modèles d'apprentissage (Gibbons, Katz, Lemieux, Parent, 2005) ou dans les modèles où les entreprises font des contre-offres (Postel-Vinay et Robin, 2002)

Supposons que le travailleur i se révèle plus productif que prévu lorsqu'il est employé par l'entreprise j

- Expérience d'une tendance à la hausse de la croissance salariale lorsqu'employé chez j
- Plus susceptible de se déplacer dans des entreprises qui emploient des travailleurs hautement qualifiés

Implique généralement des tendances dans les salaires avant et après le déplacement

Absence de tendances préalables

La dérive peut survenir dans les modèles d'apprentissage (Gibbons, Katz, Lemieux, Parent, 2005) ou dans les modèles où les entreprises font des contre-offres (Postel-Vinay et Robin, 2002)

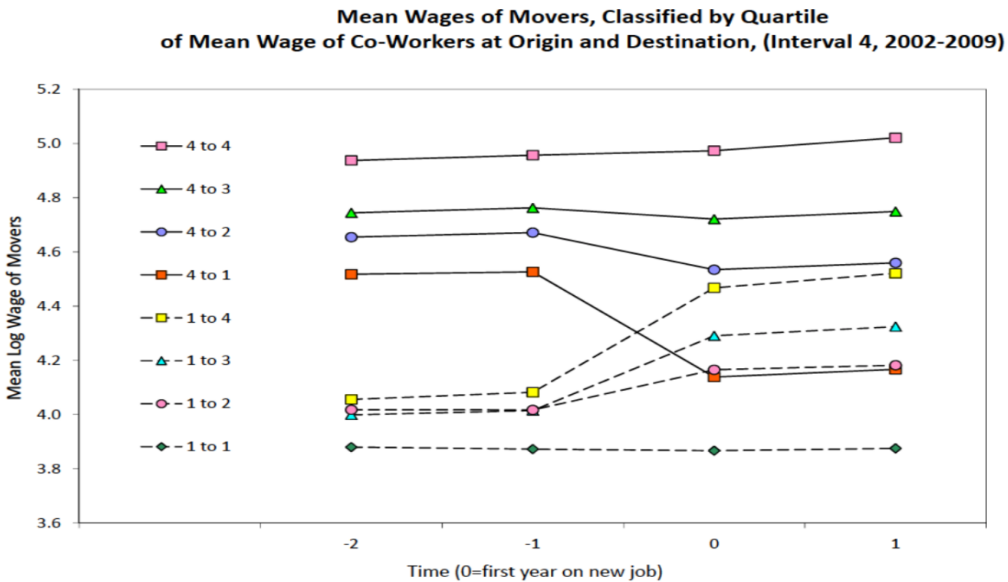
Supposons que le travailleur i se révèle plus productif que prévu lorsqu'il est employé par l'entreprise j

- Expérience d'une tendance à la hausse de la croissance salariale lorsqu'employé chez j
- Plus susceptible de se déplacer dans des entreprises qui emploient des travailleurs hautement qualifiés

Implique généralement des tendances dans les salaires avant et après le déplacement

La symétrie et l'absence de tendances préalables ont reçu une confirmation empirique surprenante dans l'une des figures les plus importantes de l'économie du travail : l'étude d'événement de Card Heining Kline (2013)

Étude d'événement CHK

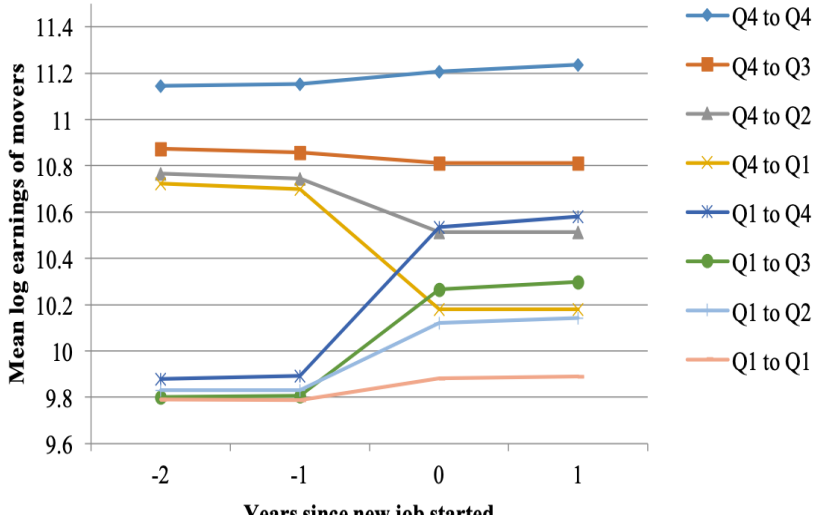


Étude d'événement CHK : version tableau

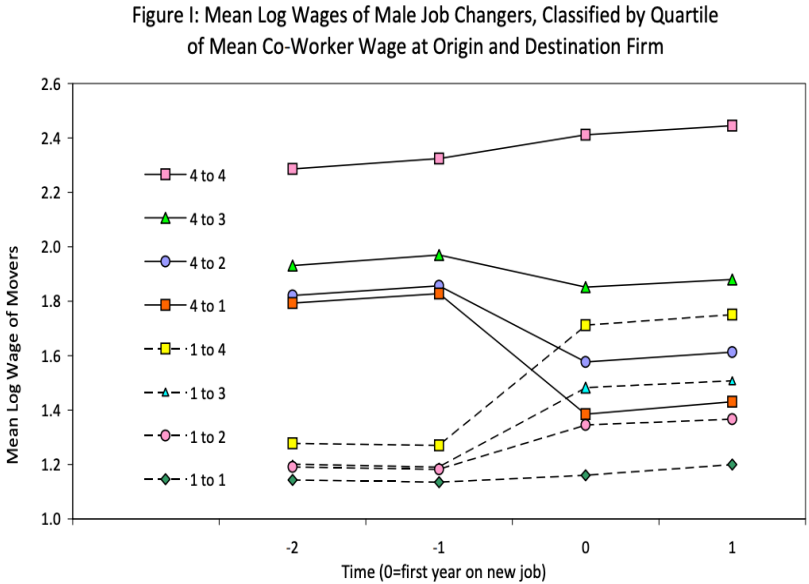
Appendix Table 3: Mean Log Wages Before and After Job Change, for Movers with Two or More Years of Wage Data Before and After Job Change

Origin/destination quartile*	Number of Changes: (1)	Mean Log Wages of Movers: 2 years before (2)	1 year before (3)	1 year after (4)	2 years after (5)	4 Year Change: Raw (6)	4 Year Change: Adjusted** (7)
Interval 1: 1985-1991							
1 to 1	333,648	4.003	4.025	4.085	4.113	0.110	0.000
1 to 2	206,251	4.063	4.085	4.207	4.248	0.185	0.075
1 to 3	136,119	4.064	4.087	4.271	4.323	0.260	0.150
1 to 4	82,193	4.102	4.132	4.380	4.444	0.342	0.232
2 to 1	125,376	4.160	4.178	4.144	4.175	0.015	-0.072
2 to 2	204,787	4.229	4.251	4.286	4.316	0.087	0.000
2 to 3	158,360	4.258	4.278	4.359	4.395	0.137	0.051
2 to 4	86,038	4.298	4.324	4.474	4.529	0.231	0.144
3 to 1	59,334	4.245	4.261	4.163	4.194	-0.051	-0.153
3 to 2	91,474	4.315	4.337	4.333	4.371	0.056	-0.046
3 to 3	173,160	4.384	4.409	4.452	4.486	0.102	0.000
3 to 4	136,569	4.460	4.487	4.594	4.635	0.175	0.073
4 to 1	30,110	4.373	4.396	4.252	4.284	-0.089	-0.220
4 to 2	41,079	4.459	4.488	4.447	4.487	0.028	-0.103
4 to 3	91,177	4.552	4.584	4.596	4.633	0.080	-0.051
4 to 4	290,921	4.678	4.710	4.777	4.809	0.131	0.000
Interval 4: 2002-2009							
1 to 1	541,307	3.880	3.873	3.867	3.875	-0.005	0.000
1 to 2	197,982	4.018	4.017	4.165	4.182	0.164	0.054
1 to 3	88,768	3.999	4.015	4.291	4.325	0.325	0.215
1 to 4	49,167	4.056	4.083	4.468	4.521	0.465	0.355
2 to 1	208,184	4.202	4.190	4.021	4.030	-0.171	-0.258
2 to 2	333,219	4.315	4.309	4.305	4.309	-0.007	0.000
2 to 3	137,528	4.381	4.377	4.437	4.456	0.075	-0.012

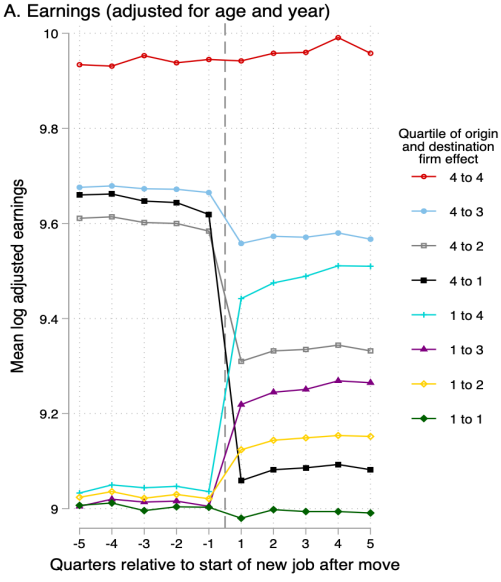
Étude d'événement CHK dans l'État de Washington (Lachowska, Mas, Woodbury, 2018)



Étude d'événement CHK au Portugal (Card, Cardoso, Kline, 2016)



Étude d'événement CHK dans LEHD (Card, Rothstein, Yi, 2023)



Carte thermique 3D CHK

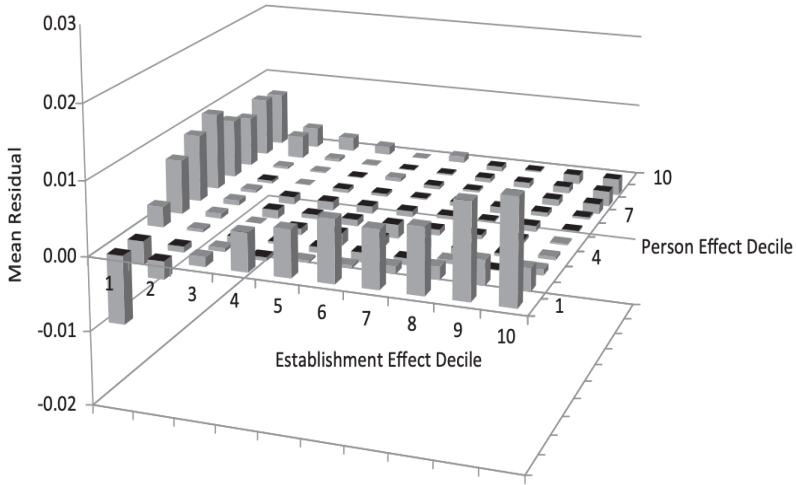


FIGURE VI

Mean Residuals by Person/Establishment Deciles, 2002–2009

Lectures complémentaires sur AKM

Utilisation de AKM pour décomposer les écarts de salaires moyens entre groupes (AKM-Oaxaca-Blinder)

- Écart salarial de genre (Card, Cardoso, Kline, 2016), écart salarial entre noirs et blancs (Gerard, Lagos, Severnini, Card, 2021), écart salarial entre natifs et immigrants (Dostie, Li, Card, Parent, 2020)
- Relaxation de la séparabilité additive

Échelle salariale double (AKM-Postel-Vinay-Robin)

- AKM avec des salaires différents pour l'entreprise d'origine vs. entreprise de destination
- Tests des « modèles d'enchères séquentielles » (Postel-Vinay Robin, 2002)
- Di Addario, Kline, Saggio, Solvsten (2023) : « Ce n'est pas d'où vous venez, c'est où vous êtes »

Divers

- Primes salariales par industrie/ville : (Card, Rothstein, Yi, 2021, 2023)
- AKM variant dans le temps (Lachowska, Mas, Woodbury, 2023)
- Différentiels compensatoires (Sorkin, 2018)

FIRMING UP INEQUALITY*

JAE SONG
DAVID J. PRICE
FATIH GUVENEN
NICHOLAS BLOOM
TILL VON WACHTER

We use a massive, matched employer-employee database for the United States to analyze the contribution of firms to the rise in earnings inequality from 1978 to 2013. We find that one-third of the rise in the variance of (log) earnings occurred within firms, whereas two-thirds of the rise occurred due to a rise in the dispersion of average earnings between firms. However, this rising between-firm variance is not accounted for by the firms themselves but by a widening gap between firms in the composition of their workers. This compositional change can be split into two roughly equal parts: high-wage workers became increasingly likely to work in high-wage firms (i.e., sorting increased), and high-wage workers became increasingly likely to work with each other (i.e., segregation rose). In contrast, we do not find a rise in the variance of firm-specific pay once we control for the worker composition in firms. Finally, we find that two-thirds of the rise in the within-firm variance of earnings occurred within mega (10,000+ employee) firms, which saw a particularly large increase in the variance of earnings compared with smaller firms. *JEL* Codes: E23, J21, J31.

II. Data

II.A. The Master Earnings File

The main source of data used in this article is the Master Earnings File (MEF), which is a confidential database compiled and maintained by the U.S. Social Security Administration (SSA). The MEF contains earnings records for every individual who has ever been issued a U.S. Social Security number. In addition to basic demographic information (sex, race, date of birth, etc.), the MEF contains annual labor earnings information from 1978 to (as of this writing) 2013. Earnings records are derived from Box 1 of Form W-2, which is sent directly by employers to the SSA. These earnings data are uncapped and include wages and salaries, bonuses, tips, exercised stock options, the dollar value of vested restricted stock units, and other sources of income deemed as remuneration for labor services by the Internal Revenue Service.[3] Because of potential measurement issues prior to 1981 (see Guvenen, Kaplan, and Song (2014), Kopczuk, Saez, and Song 2010), we start most of our analysis in 1981, although results back to 1978 look similar. All earnings are converted to 2013 real values using the personal consumption expenditures (PCE) deflator.

Because earnings data are based on the W-2 form, the data set includes one record for each individual, for each firm they

worked in, for each year. Crucially for our purposes, the MEF also contains a unique employer identification number (EIN) for each W-2 earnings record. Because the MEF covers the entire U.S. population and has EIN records for each job of each worker, we can use worker-side information to construct firm-level variables. In particular, we assign all workers who received wage earnings from the same EIN in a given year to that firm. Workers who hold multiple jobs in the same year are linked to the firm providing their largest source of earnings for the year. Many workers have multiple W-2s, but few have multiple W-2s consistently: in 2013, 30.5% of workers had multiple W-2s, but only 4.3% had multiple W-2s every year from 2009 to 2013. The resulting matched employer-employee data set contains information for each firm on total employment, wage bill, and earnings distribution, as well as the firm's gender, age, and job tenure composition.

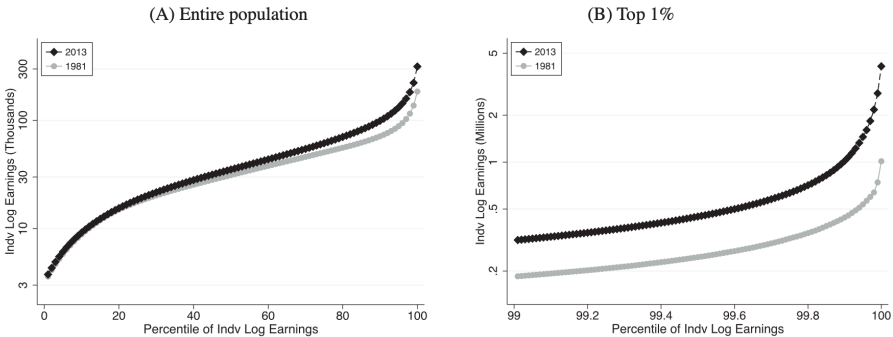


FIGURE I

Cumulative Distributions of Annual Earnings in the SSA Data

II.B. What Is a Firm?

Throughout the article, we use EINs as the boundary of a firm. The EIN is the level at which companies file their tax returns with the IRS, so it reflects a distinct corporate unit for tax (and therefore accounting) purposes. Government agencies, such as the Bureau of Labor Statistics, commonly use EINs to define firms.[6] They are often used in research on firms based on administrative data.

An EIN is not always the same, however, as the ultimate parent firm. Typically, this is because large firms file taxes at a slightly lower level than the ultimate parent firm.[7] Although it is

unclear what level of aggregation is appropriate to define a "firm," we follow much of the existing literature and view the EIN as a sensible concept reflecting a unit of tax and financial accounting. An EIN is a concept distinct from an "establishment," which typically represents a single geographic production location and is another commonly used unit of analysis to study the behavior of "firms" (this is the definition used by Barth et al. 2016, who study inequality using U.S. Census data). Around 30 million U.S. establishments in the Longitudinal Business Database in 2012 are owned by around 6 million EIN firms, so an establishment is a more disaggregated concept. As Online Appendix Figure A.4 shows, 84% of the increase in cross-establishment inequality can be accounted for by firms, so firms are an appropriate unit of analysis.

II.D. Baseline Sample

For our descriptive analysis in Section III, we restrict our baseline sample to individuals aged 20 to 60 who were employed, where "employed" is defined as earning at least that year's minimum wage for one quarter full-time (so for 2013, 13 weeks for 40 hours at $7.25 per hour, or $3,770). These restrictions reduce the effect on our results of individuals who are not strongly attached to the labor market. We also restrict to firms (and workers in firms) with 20+ employees to help ensure that within-firm statistics are meaningful. We exclude firms (and workers in firms) in the government or educational sectors because organizations in those sectors are schools and government agencies. This yields a sample of, on average, 72.6 million workers and 477,000 firms a year, rising from 55.5 million and 371,000 in 1981 to 85.2 million and 517,000 in 2013, respectively. None of our results are sensitive to these assumptions. Although there is some variation, the results look similar using all ages, all firm sizes, all industries, and minimum earnings thresholds up to full-time (2,080 hours) at minimum wage. Some statistics describing the sample are shown in Table I. More details about the data procedures are discussed in Online Appendix B.

TABLE I

PERCENTILES OF VARIOUS STATISTICS FROM THE DATA

Year	Group	Statistic	10%ile	25%ile	50%ile	75%ile	90%tile
1981	Firm	Earnings (unwgt)	12.6	16.6	23.8	32.5	41.9
1981	Firm	Earnings (wgted)	15.2	21.5	30.6	43.2	52.1
1981	Firm	Employees	22	26	38	73	169
1981	Individual	Earnings	9.46	18.2	31.9	51.7	73.8
1981	Individual	Earnings/firm avg	0.43	0.724	1.05	1.45	2.06
1981	Individual	Employees	42	127	1,153	12,418	62,718
2013	Firm	Earnings (unwgt)	13.8	19.3	30.5	43.8	61.4
2013	Firm	Earnings (wgted)	15.3	21.4	35.8	52.1	73.6
2013	Firm	Employees	22	26	39	79	189
2013	Individual	Earnings	9.82	19.2	36	63.2	104
2013	Individual	Earnings/firm avg	0.421	0.681	1.03	1.5	2.22
2013	Individual	Employees	45	157	1,381	14,197	78,757

Notes. Values indicate various percentiles for the data for individuals or firms. All dollar values are in thousands of 2013 dollars, adjusted for inflation using the PCE deflator. Only firms and individuals in firms with at least 20 employees are included. Firm statistics are based on mean earnings at firms and are either unweighted or weighted by number of employees, as indicated. Only employed individuals aged 20 to 60 are included in all statistics, where "employed" is defined as earning the equivalent of minimum wage for 40 hours per week in 13 weeks. Individuals and firms in public administration or educational services are not included.

III.A. A Simple Variance Decomposition

We decompose the overall (cross-sectional) variance of log earnings into within- and between-firm components. In particular, let $y_t^{i,j}$ be the log earnings of worker i employed by firm j in period t.[10] This can be broken down into two components:

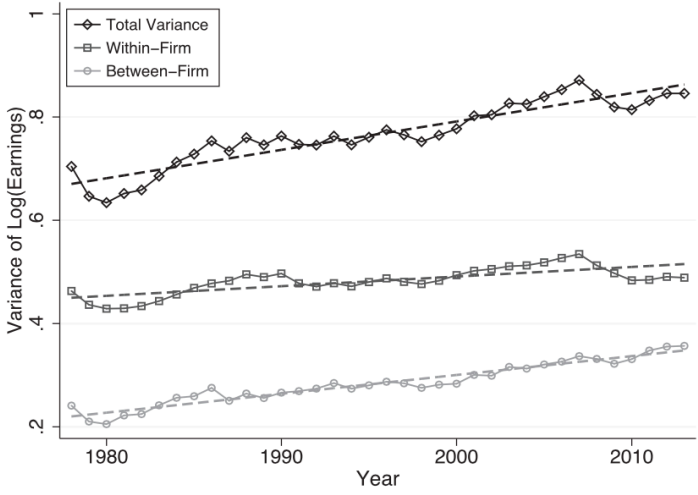
$$y_t^{i,j} \equiv \bar{y}_t^j + \left[y_t^{i,j} - \bar{y}_t^j \right], \tag{1}$$

where $\bar{y}_t^j$ is the firm average earnings for firm j. Some simple algebra shows that the overall variance can be decomposed into two terms:

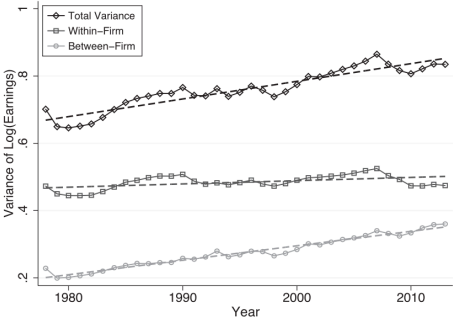
$$\text{var}(y_t^{i,j}) = \underbrace{\text{var}_j(\bar{y}_t^j)}_{\text{Between-firm dispersion}} + \underbrace{\sum_j \omega_j \times \text{var}_i(y_t^{i,j} | i \in j)}_{\text{Within-firm dispersion}}. \tag{2}$$

The first term is the between-firm variance of firm average earnings, and the second term is the employment-weighted mean of within-firm dispersion in employee earnings, where ω_j denotes the employment share of firm j in the sample.

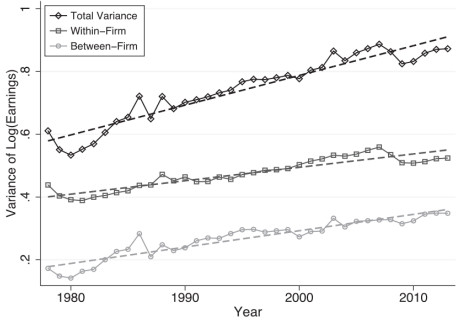
(A) Overall decomposition



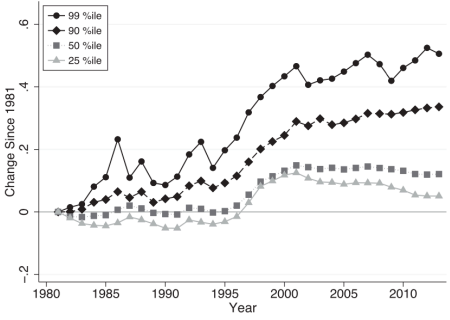
(B) Workers at firms with 20 to 10,000 employees



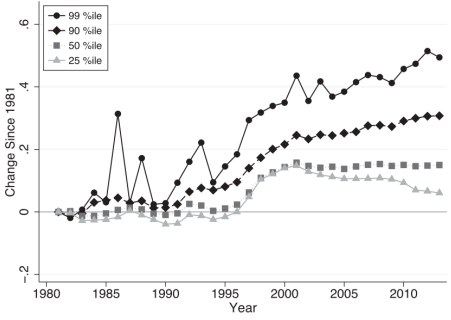
(C) Workers at mega firms (10,000+ employees)



(A) Individuals



(B) Their firms



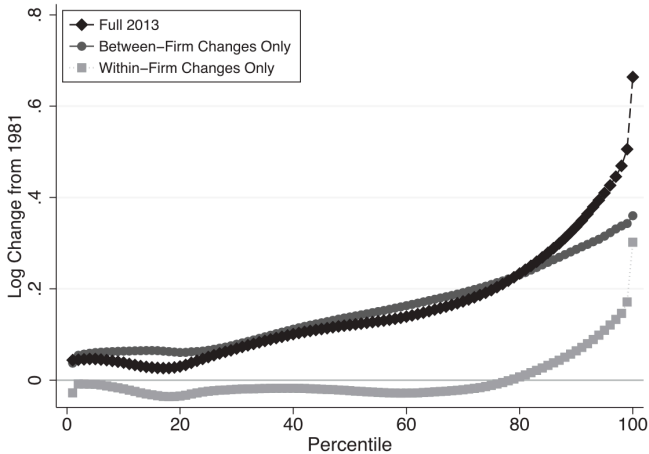


FIGURE IV

Counterfactual Rise in Inequality with Between- or Within-Firm Changes Only

IV.A. Econometric Model of Worker and Firm Effects

To analyze the worker and firm movements in earnings, we follow the CHK implementation of the model introduced by AKM and solved by Abowd, Creecy, and Kramarz (2002).[14] We divide our time period into five seven-year periods, as discussed further below, and estimate a separate model for each period p. The regression model we estimate in each period is

$$y_t^{i,j} = \theta^{i,p} + X_t^i \beta^p + \psi^{j,p} + \epsilon_t^{i,j}, \tag{3}$$

where $\theta^{i,p}$ captures earnings related to fixed worker characteristics (such as returns to formal schooling or to innate ability), β^p captures the effect of time-varying worker characteristics (in our case, a polynomial in age and year fixed effects), and $\psi^{j,p}$ captures persistent earnings differences related to firm j (such as sharing of rents or compensating differentials). The residual, $\epsilon_t^{i,j}$, captures purely transitory earnings fluctuations. In addition, the residual will also contain any worker-firm specific (match) components in earnings, which we denote by $m^{i,j}$.

The estimates of the parameters of the econometric model in equation (3) can be used to further decompose the within- and between-firm components of the variance. Ignoring time-varying worker characteristics $X_t^i \beta^p$ for now and variation across periods (dropping superscript p), the standard approach to decompose the variance into components related to worker effects and firm effects used in AKM, CHK, and related work is

$$\text{(4)} \qquad \text{var}(y_t^{i,j}) = \text{var}(\theta^i) + \text{var}(\epsilon_t^{i,j}) + \text{var}(\psi^j) + 2\text{cov}(\theta^i, \psi^j),$$

where the moments in the last two components are weighted by the number of worker-years in the respective time interval.

$$\mathbf{var}(y_t^{i,j}) = \underbrace{\mathbf{var}(\theta^i - \overline{\theta}^j) + \mathbf{var}(\epsilon_t^{i,j})}_{\text{Within-firm component}}$$

$$+ \underbrace{\mathbf{var}(\psi^j) + 2\mathbf{cov}(\overline{\theta}^j, \psi^j) + \mathbf{var}(\overline{\theta}^j)}_{\text{Between-firm component}}, \tag{5}$$

TABLE III

BASIC DECOMPOSITION OF THE RISE IN INEQUALITY OF ANNUAL EARNINGS

		Interval 1 (1980–1986)		Interval 2 (1987–1993)		Interval 3 (1994–2000)		Interval 4 (2001–2007)		Interval 5 (2007–2013)		Change from 1 to 5	
		Comp. (1)	Share (2)	Comp. (3)	Share (4)	Comp. (5)	Share (6)	Comp. (7)	Share (8)	Comp. (9)	Share (10)	Comp. (11)	Share (12)
Total variance	Var(y)	0.708	—	0.776	—	0.828	—	0.884	—	0.924	—	0.216	—
Components of variance	Var(WFE)	0.330	46.6	0.375	48.3	0.422	51.0	0.452	51.2	0.476	51.5	0.146	67.6
	Var(FFE)	0.084	11.9	0.075	9.7	0.067	8.1	0.075	8.5	0.081	8.7	−0.003	−1.6
	Var(Xb)	0.055	7.8	0.065	8.4	0.079	9.5	0.061	6.9	0.059	6.4	0.004	1.8
	Var(ϵ)	0.154	21.7	0.148	19.1	0.146	17.6	0.149	16.8	0.136	14.7	−0.018	−8.2
	2*Cov(WFE, FFE)	0.033	4.7	0.057	7.3	0.076	9.2	0.094	10.6	0.108	11.7	0.075	34.8
	2*Cov(WFE, Xb)	0.028	3.9	0.029	3.7	0.013	1.6	0.028	3.1	0.036	3.9	0.009	4.1
	2*Cov(FFE, Xb)	0.022	3.1	0.025	3.3	0.023	2.7	0.024	2.7	0.027	2.9	0.005	2.2
Sum of firm components	Cov(y, FFE)	0.112	15.8	0.116	14.9	0.117	14.1	0.134	15.1	0.148	16.0	0.037	16.9
Counterfactuals	1. No rise in corr(WFE, FFE)	0.708		0.750	96.7	0.784	94.6	0.826	93.4	0.854	92.4	0.146	67.5
	2. No fall in var(FFE)	0.708		0.788	101.4	0.854	103.1	0.898	101.6	0.929	100.6	0.221	102.4
	3. Both 1 and 2	0.708		0.763	98.3	0.807	97.4	0.838	94.8	0.859	92.9	0.150	69.7

Notes. Var(y): variance of annual earnings, Var(WFE): variance of worker fixed effects, Var(FFE): variance of firm fixed effects, Var(Xb): variance of covariates, Var(ϵ): variance of residual.

Sum of firm-related components is equal to var(FFE) + Cov(WFE, FFE) + Cov(FFE, Xb). Only men are included in these statistics. Only firms and individuals in firms with at least 20 employees are included. Only employed individuals aged 20 to 60 are included in all statistics, where "employed" is defined as earning the equivalent of 2013 minimum wage, adjusted for inflation with the PCE, for 40 hours per week in 13 weeks. Individuals and firms in public administration or educational services are not included.

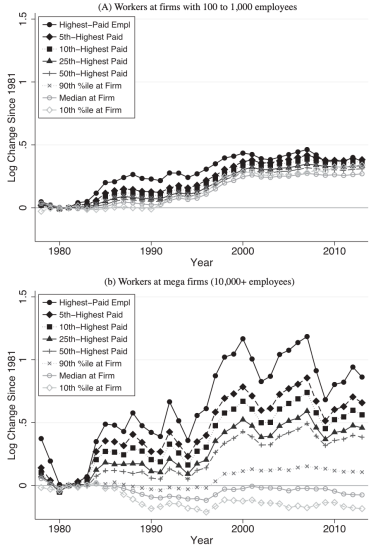


FIGURE VI

Change in Within-Firm Distribution of Earnings: Small versus Mega Firms

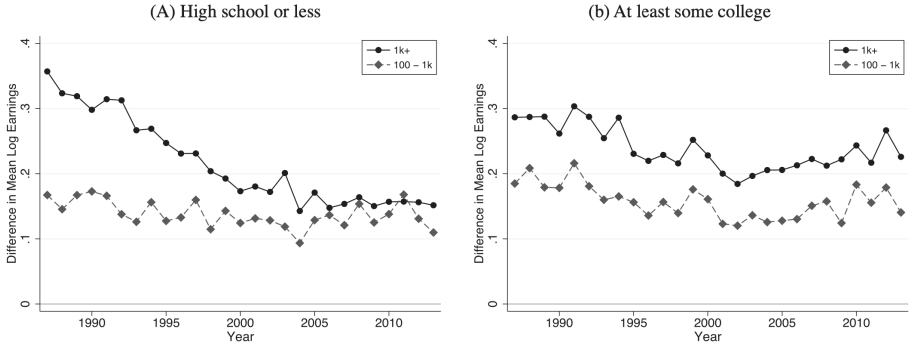


FIGURE VII

The Earnings Premium between Larger Firms (100+ Employees) and Smaller Firms (Less than 100 Employees), by Education