Analyse statistique avancée

OULEDI Victor, TORDJEMAN Robin, TREGOUET Yoann

2023-03-20

M1 APE/RISE

Plan

1) Introduction (en option) et description de la base

2) Statistiques descriptives

3) Spécification des différents modèles

- Régression estimateur EF
- Régression estimateur D1
- Régression between et pooling
- Choix entre EF et EA
- Choix entre D1 et within
- Estimations régresseurs constants avec within

4) Conclusion

Introduction et description de la base

Introduction Optionnelle

La base de données à laquelle nous nous sommes intéressés se nomme "wagepan" et provient du package Wooldridge. Elle contient des informations sur les salaires et les caractéristiques des travailleurs aux États-Unis. Elle a été créée en compilant des données provenant d'enquêtes menées par le Département du Travail des États-Unis sur plusieurs années.La base contient des milliers d'observations sur plusieurs années (545 individus de 1980 à 1987), ce qui permet d'analyser des tendances au fil du temps grâce à des techniques éconoétriques de données de panel.

Cette base de données est utilisée pour étudier les déterminants des salaires aux États-Unis. Les travaux économiques qui ont été réalisés sur cette base avaient pour but de comprendre comment des facteurs tels que l'éducation, l'expérience de travail, le genre, l'origine ethnique et d'autres variables liées à l'emploi influencent les salaires. Ils interrogeaint également des questions telles que la discrimination salariale et les différences de salaires entre les sexes et les différentes origines ethniques.Au total il est possible de s'appuyer sur 44 variables expliquatives pour la détermination de la variable "salaire" à expliquer

Les enjeux économiques abordés par cette base de données sont importants car les salaires sont un élément crucial du niveau de vie des travailleurs et de leur famille. Comprendre les facteurs qui influencent les salaires peut aider à identifier les politiques qui peuvent aider à améliorer les salaires et à réduire les inégalités salariales.

Description des variables

La base de données contient des variables qui ont été collectées à partir d'une enquête menée sur un échantillon de travailleurs américains entre 1985 et 1987. Ces variables comprennent des informations sur les caractéristiques des travailleurs et leurs salaires. On y retrouve : "nr": l'identifiant de chaque individu "year": l'année de l'enquête "age": l'âge du travailleur en années "origin": l'origine ethnique du travailleur (hispanique, noir ou autre) "educ": le niveau d'éducation du travailleur, mesuré en années d'études "geo_zone" : la zone géographique du travailleur : sud, nord est ou nord ouest des États-Unis "union": indique si le travailleur est membre d'un syndicat "wage": le salaire horaire du travailleur en dollars "hours": le nombre d'heures de travail hebdomadaires du travailleur "exper": l'expérience professionnelle du travailleur en années "married": indique si le travailleur est marié "sector": indique le secteur d'activité du travailleur (agriculture, construction, manufacture, transport, trading, entertainment, professional relation services, business, finance, public services ou autre) "poorhlth": état de santé du travailleur "rur": indique si le travaille habite en zone rurale

Statistiques descriptives

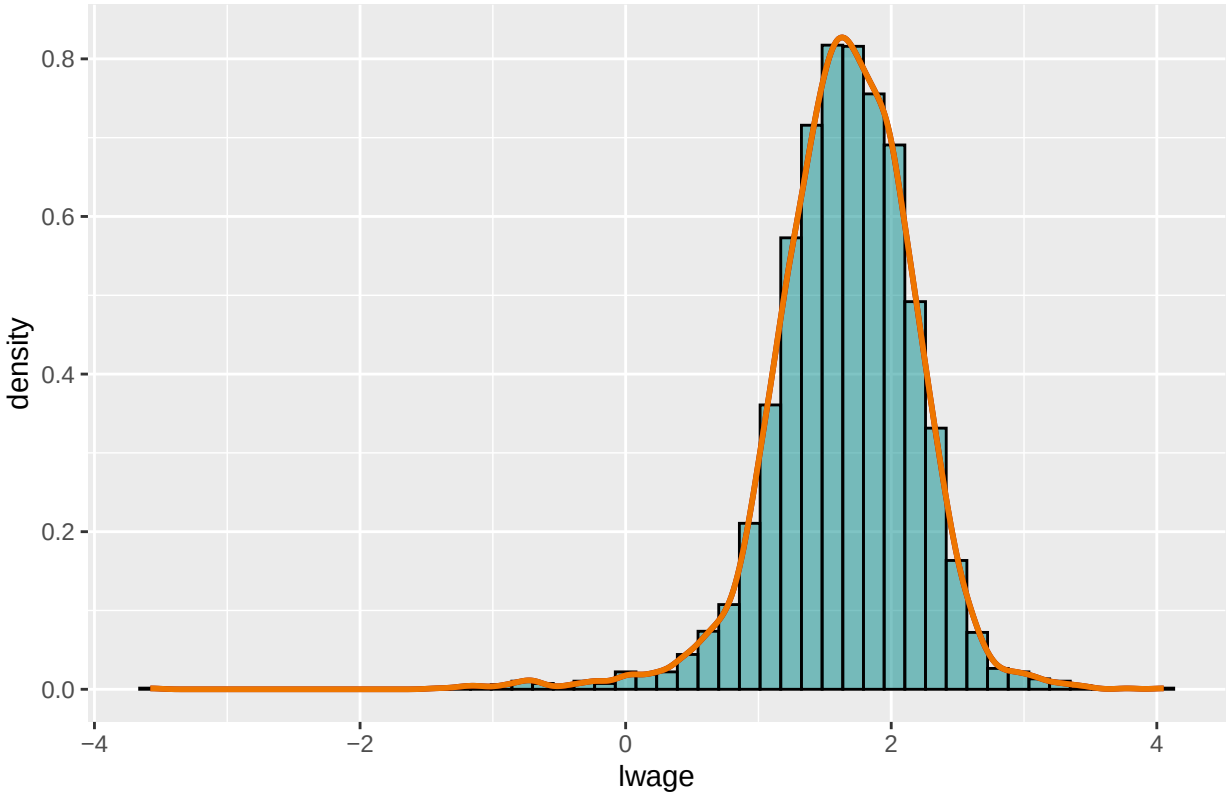
Nous pouvons d'emblée nous intéresser aux statistiques descriptives de nos principales variables numériques avec la fonction summary.

	hours	educ	wage	lwage	exper	expersq
	Min. : 120	Min. : 3.00	Min. : 0.0279	Min. :-3.579	Min. : 0.000	Min. : 0.00
	1st Qu.:2040	1st Qu.:11.00	1st Qu.: 3.8602	1st Qu.: 1.351	1st Qu.: 4.000	1st Qu.: 16.00
	Median :2080	Median :12.00	Median : 5.3182	Median : 1.671	Median : 6.000	Median : 36.00
	Mean :2191	Mean :11.77	Mean : 5.9192	Mean : 1.649	Mean : 6.515	Mean : 50.42
	3rd Qu.:2414	3rd Qu.:12.00	3rd Qu.: 7.3235	3rd Qu.: 1.991	3rd Qu.: 9.000	3rd Qu.: 81.00
	Max. :4992	Max. :16.00	Max. :57.5043	Max. : 4.052	Max. :18.000	Max. :324.00

On se rend compte qu'il y a tout de même de grandes différences entre les individues en termes d'années d'expérience, de salaires, d'heures exercées

Intéressons nous dans un second temps à la distribution des différents salaires :

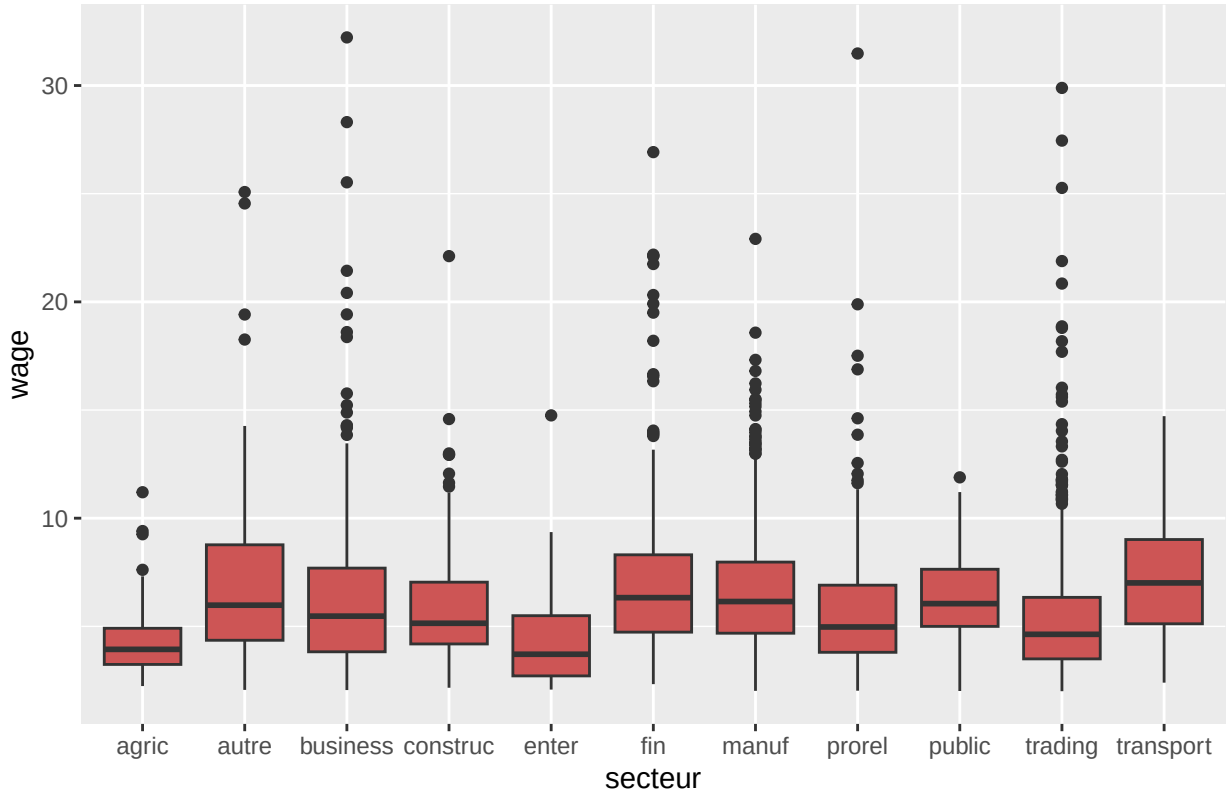
Graph 1 : Histogramme de la répartition des salaires



On peut déjà observer que ces derniers suivent plus ou moins une loi normale. Le travail économétrique sur les salaires sera d'autant plus adéquat avec des méthodes classiques.

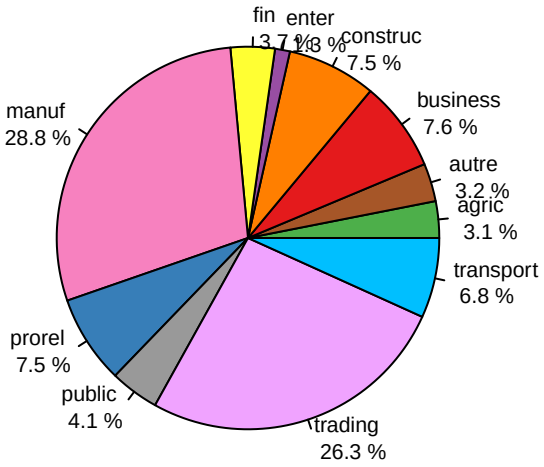
On créé ensuite des variables catégorielles à partir de nos variables muettes. On pourra ainsi réaliser des statistiques descriptives de ces dernières variables.

Graph 2 : Boxplot des salaires en fonction du secteur

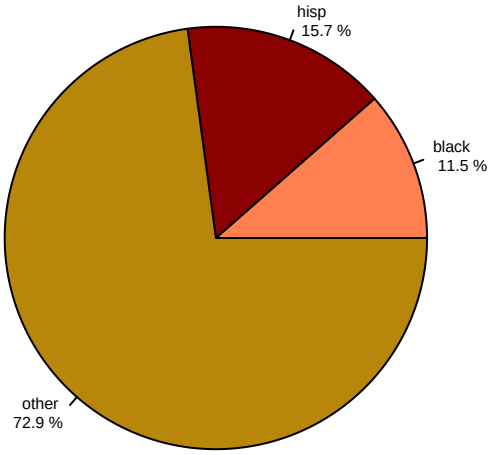


On observe à travers ce booxplot que la plupart des salaires sont situés sous la barre des 10 dollars de l'heure. Il y a néanmoins beaucoup d'outliers vers qui représentent des salaires extrèmement élevés allant jsuqu'à presque 60 dollars de l'heure, cela pour plusieurs secteurs.

Graph 3 : répartition des emplois



Graph 4 : origines ethniques

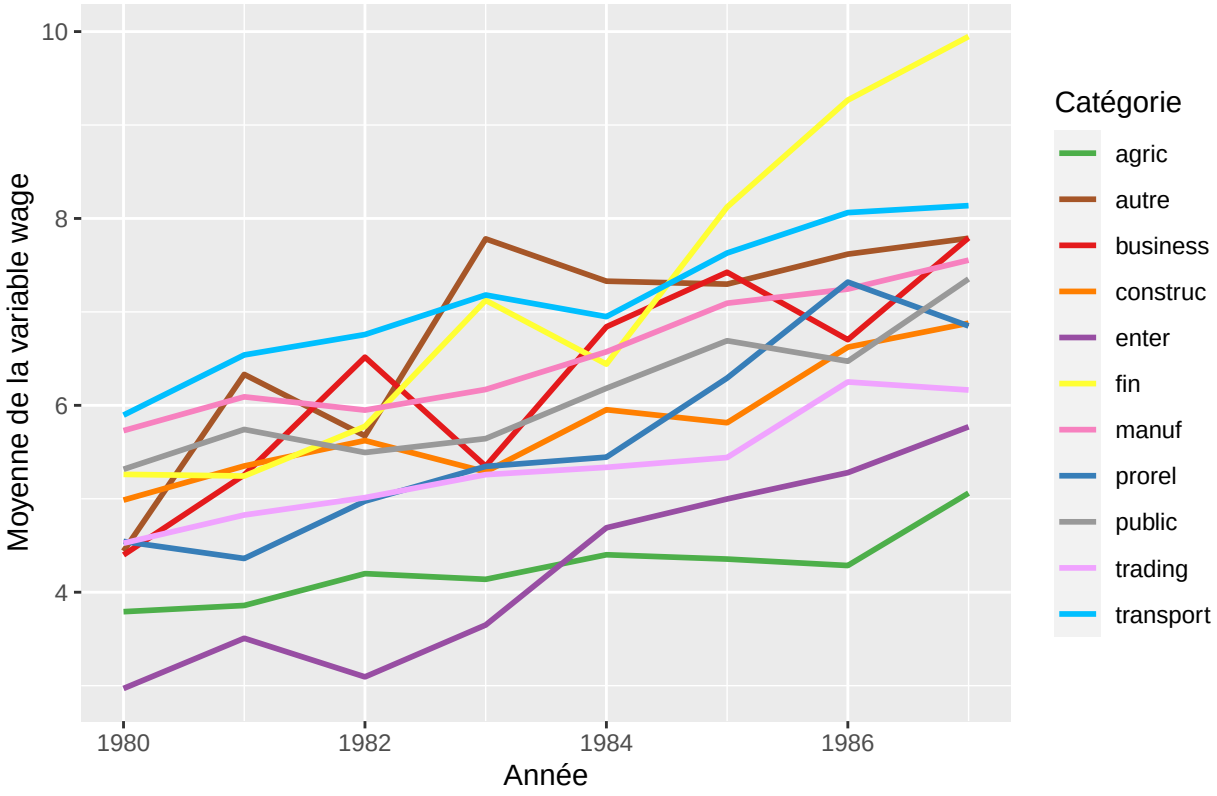


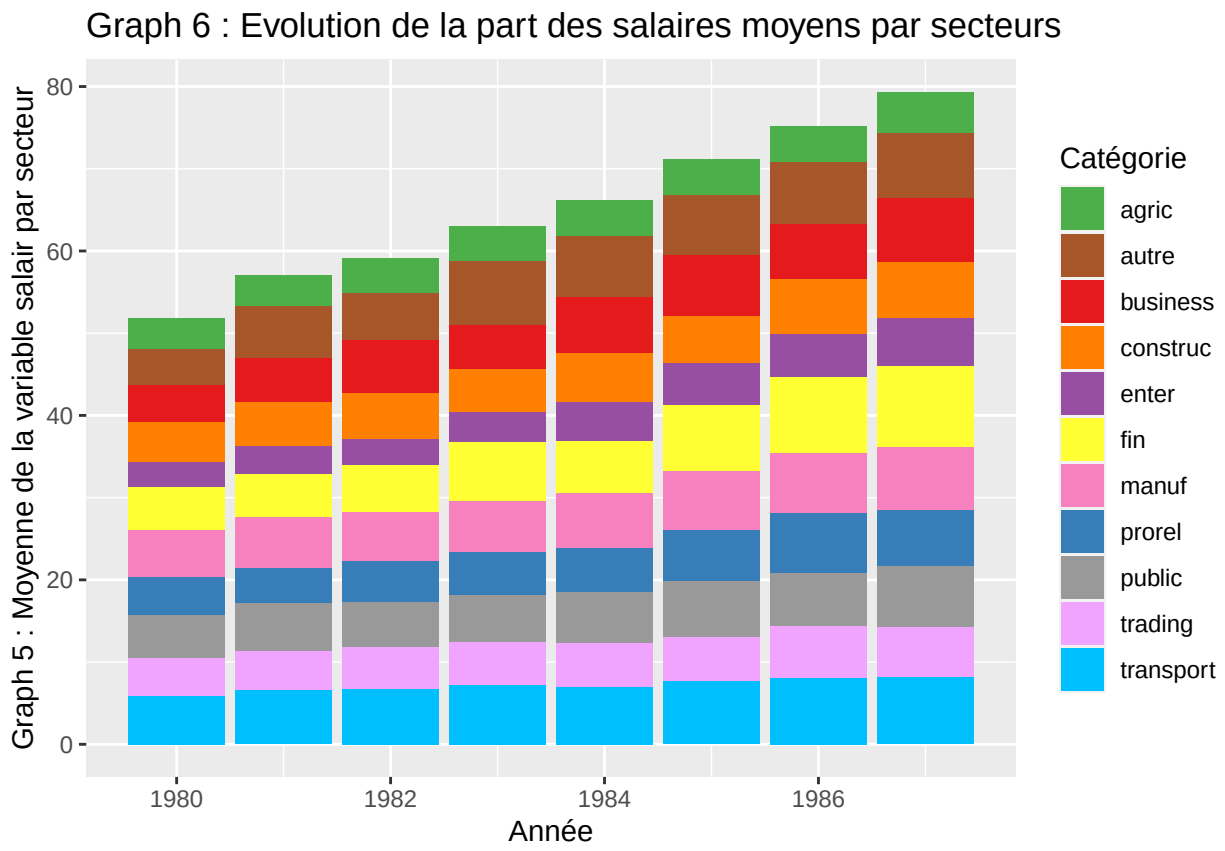
Les pie-charts ci-dessus nous montrent la répartiion des emplois et des origines ethniques. On remarque que beaucoup des salariés observé travaillent dans les secteurs “trading” et “manufacture” et très peu dans les secteurs “entertainment” et “agriculture”

Table 1: Tableau de moyennes salariales par secteur et année

secteur	1980	1981	1982	1983	1984	1985	1986	1987
agric	3.791914	3.858426	4.199009	4.138336	4.401166	4.354754	4.285881	5.060549
autre	4.439785	6.331036	5.674023	7.781662	7.328790	7.297205	7.618613	7.787921
business	4.398994	5.255770	6.515661	5.348814	6.840407	7.424383	6.702432	7.792633
construc	4.986037	5.351048	5.623197	5.288719	5.953379	5.813976	6.622987	6.879296
enter	2.969751	3.507446	3.094072	3.649327	4.689979	4.998087	5.279654	5.770435
fin	5.261630	5.243538	5.774821	7.123903	6.440673	8.122180	9.266870	9.946993
manuf	5.729268	6.091097	5.949119	6.169803	6.572875	7.093066	7.244447	7.553584
prorel	4.543777	4.361239	4.974791	5.346625	5.445473	6.291960	7.318981	6.849633
public	5.316101	5.742141	5.495308	5.643847	6.182695	6.691598	6.472101	7.352381
trading	4.524922	4.826589	5.012427	5.258308	5.336553	5.441666	6.250365	6.164187
transport	5.895251	6.539591	6.757708	7.180986	6.949007	7.630110	8.062734	8.137270

Graph 5 : Evolution du salaire moyen par catégorie de 1980 à 1987





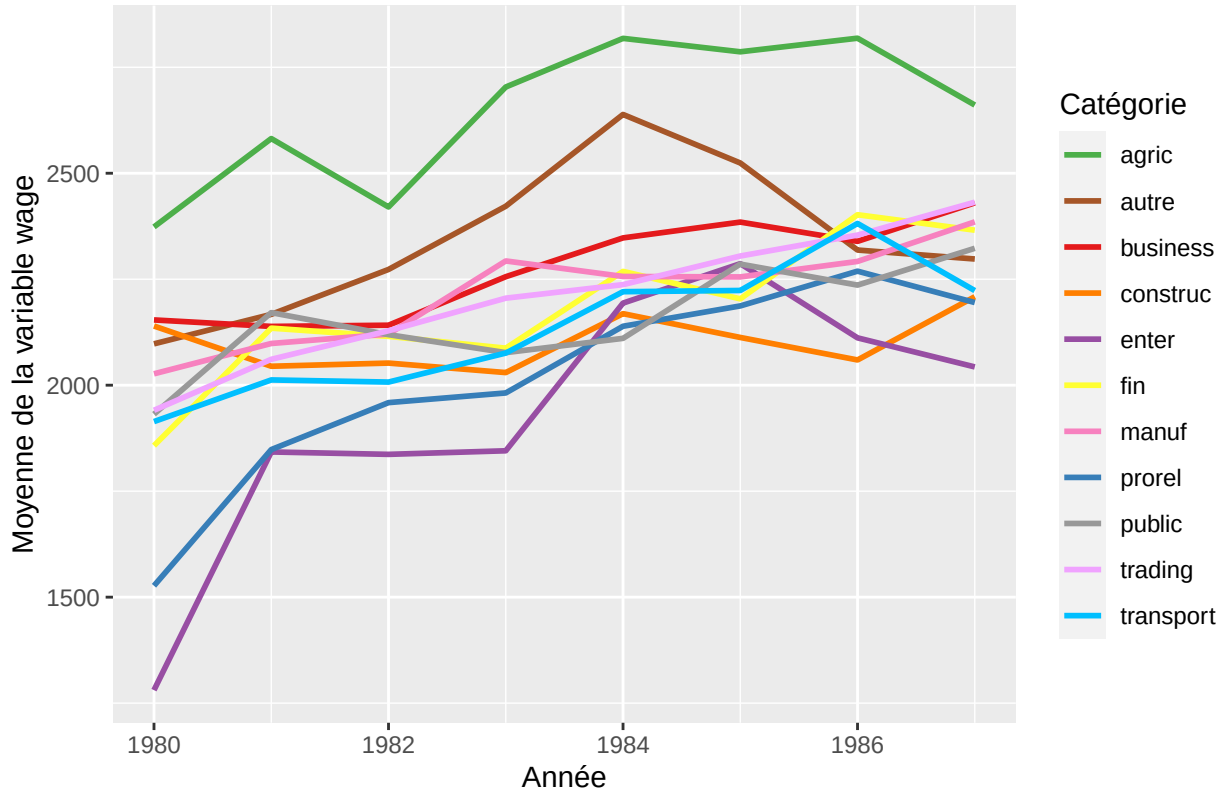
Graphiquement on peut dans un premier temps voir que les salaires moyens de chaques secteurs ont tous augmenté. Cependant, le graphique en barre nous montre que le poids de chaque secteur dans le revenu total n'est pas resté constant.Certains secteur ayant vu leur masse salariale au sein de la somme des salaires augmenter au détriment d'autres secteurs.

Table 2: Tableau de moyennes d'heures éffectuées par secteur et année

secteur	1980	1981	1982	1983	1984	1985	1986	1987
agric	2373.364	2581.800	2420.444	2703.474	2818.188	2786.500	2818.583	2660.917
autre	2097.571	2167.125	2273.062	2422.167	2638.652	2524.158	2318.867	2297.714
business	2153.943	2138.658	2141.487	2255.925	2347.459	2384.743	2339.370	2430.137
construc	2139.600	2044.682	2052.195	2029.812	2168.750	2112.600	2059.548	2208.682
enter	1280.600	1842.500	1836.833	1845.286	2193.636	2286.667	2111.714	2043.111
fin	1857.333	2134.571	2115.235	2087.368	2268.476	2203.625	2402.115	2365.708
manuf	2026.870	2098.419	2120.605	2293.108	2256.489	2255.415	2291.980	2385.329
prorel	1526.868	1848.085	1958.902	1981.700	2139.065	2187.000	2269.026	2195.029
public	1931.667	2171.200	2119.222	2076.810	2110.480	2285.696	2236.088	2322.912
trading	1940.820	2061.026	2128.395	2205.282	2237.230	2304.766	2354.158	2431.697
transport	1914.133	2012.432	2007.419	2075.833	2220.600	2223.474	2381.474	2223.205

Le graphique ci-dessous nous montre l'évolution du temps de travail. On remarque que le temps de travail a augmenté dans tous des secteurs. Les travailleurs du secteur "agriculture" travaillent plus que les autres. Cependant, leur temps de travail a légèrement diminué à partir de 1985.

Graph 7 :Evolution du temps de travail par catégorie de 1980 à 1987



Modèle théorique $lwage = \beta 1_{education} + \beta 2_{origine} + \beta 3_{syndicalisation} + \beta 4_{expérience} + \beta 5_{secteur}$

Régression within avec correction de white

Dans un premier temps, nous utilisons un modèle à effet fixe avec une estimatation within. Spécifier un tel modèle implique d'émettre l'hypothèse que les effets individuels inobservés sont corrélés avec les variables explicatives.Pour notre thématique on pourrait ainsi supposer que des effets individuels tels que la motivation peut par exemple conditionner le fait de poursuivre des études longues et ainsi infuencer les salaires plus tard. De même les carectéristiques génétiques individuelles peuvent être corrélées avec notre variable poorhealth qui traduit l'état de santé et peut potentiellement avoir un impact sur la détermination du salaire. Dans notre modèle, l'utilisation de l'estimateur within permet de soustraire les moyennes individuelles de chaque unité de la variable d'intérêt, et ainsi contrôler les effets fixes non observables et se concentrer sur les différences pour les régresseurs variables au fil du temps. Ainsi pour notre première régression nous ne pouvons pas estimer l'effet de régresseurs constant dans le temps, tels que l'origine ethnique ou le niveau d'étude. Cela voudrait dire que les résultats obtenus pour les effets des régresseurs variables dans le temps seront plus consitants que ceux pouvant être obtenus pour une spécification de modèle à effets aléatoires.

term	estimate	std.error	statistic	p.value	coefficients
exper	0.06	0.00	21.67	0.00	0.06***
secteurautre	0.04	0.05	0.80	0.42	0.04
secteurbusiness	0.02	0.04	0.54	0.59	0.02
secteurconstruc	-0.01	0.04	-0.34	0.73	-0.01
secteurenter	-0.11	0.06	-1.96	0.05	-0.11*
secteurfin	0.13	0.06	2.09	0.04	0.13*
secteurmanuf	0.06	0.03	1.89	0.06	0.06
secteurprorel	-0.04	0.04	-1.00	0.32	-0.04
secteurpublic	0.03	0.05	0.65	0.52	0.03
secteurtrading	-0.04	0.04	-1.16	0.25	-0.04
secteurtransport	0.01	0.04	0.33	0.74	0.01
wagepan$geo_zonenorth east	0.12	0.09	1.36	0.17	0.12
wagepan$geo_zoneother zones	0.05	0.06	0.87	0.38	0.05
union	0.07	0.02	3.77	0.00	0.07***
hours	0.00	0.00	-8.54	0.00	0.00***
wagepan$poorhlth	0.00	0.04	-0.12	0.91	0.00

Avec la régression within, on obtient peu de variables qui ont un effet statistiquement significatif. selon ce modèle, Cela veut donc dire que valorisation d'une année d'experience a augmenté de 7% dans la détermination des salaires. Le nombre d'heures travaillées n'a pas d'impact sur le salaire horaire. On peut donc supposer que les personnes ayant travaillées plus d'heures n'ont pas reçu de valorisation salariales. Les personnes syndiquées gagnent en moyenne mieux leur vie que lorsqu'elles ne le sont pas. De même le salaire horaire des personnes travaillant dans le secteur financier a lui bien plus augmenté que celui des agriculteurs de 1980 à 1987 soit 17 points de pourcentage en plus.

Régression first difference avec correction de white

Avec l'estimateur first difference (D1), étant donné que celui-ci implique la spécification d'un modèle à effet fixe (corrélation possible entre les effets fixes et les régresseurs), nous utilisons presque les mêmes régresseurs que pour le modèle estimé en within. En effet la principale différence entre un estimateur within et un estimateur "First-différence" est la façon dont ils contrôlent les effets fixes non observables et se concentrent sur les changements au fil du temps. L'estimation within soustrait les moyennes individuelles de chaque

	2.5 %	97.5 %
exper	0.0535004	0.0641445
secteurautre	-0.0624714	0.1494944
secteurbusiness	-0.0563410	0.0988912
secteurconstruc	-0.0971981	0.0681593
secteurenter	-0.2214270	-0.0001068
secteurfin	0.0083232	0.2536975
secteurmanuf	-0.0024387	0.1316284
secteurprorel	-0.1313083	0.0426713
secteurpublic	-0.0594434	0.1178701
secteurtrading	-0.1103903	0.0283685
secteurtransport	-0.0695036	0.0979938
wagepan$geo_zonenorth east	-0.0518964	0.2881466
wagepan$geo_zoneother zones	-0.0612322	0.1599066
union	0.0328677	0.1040979
hours	-0.0001690	-0.0001059
wagepan$poorhlth	-0.0827249	0.0734834

observation de la variable d'intérêt, tandis que l'estimation "First-différence" prend la différence entre les observations de deux périodes consécutives pour chaque unité individuelle dans les données. L'estimateur D1 nécessite une variation temporelle inconstante des variables explicatives pour chaque individu base son fonctionnement sur la variablité irrégulières Ainsi en vue de la méthode d'estimation, nous ne pouvons pas inclure la variable expérience qui a une évolution constante. Par définition elle augmente de une unité chaque année. Le modèle de différence première neutralise son effet et ne peut pas l'estimer correctement.

term	estimate	std.error	statistic	p.value	coefficients
(Intercept)	0.07	0.00	21.28	0.00	0.068***
secteurautre	0.03	0.04	0.83	0.41	0.031
secteurbusiness	-0.01	0.03	-0.44	0.66	-0.013
secteurconstruc	-0.02	0.04	-0.49	0.62	-0.020
secteurenter	-0.04	0.05	-0.69	0.49	-0.036
secteurfin	0.01	0.04	0.34	0.74	0.015
secteurmanuf	0.01	0.03	0.56	0.58	0.014
secteurprorel	-0.04	0.03	-1.29	0.20	-0.042
secteurpublic	0.00	0.04	-0.01	1.00	0.000
secteurtrading	-0.03	0.03	-1.05	0.29	-0.028
secteurtransport	-0.03	0.03	-0.75	0.45	-0.026
wagepan$geo_zonenorth east	0.11	0.08	1.40	0.16	0.115
wagepan$geo_zoneother zones	-0.02	0.05	-0.42	0.67	-0.021
union	0.04	0.02	2.30	0.02	0.037*
hours	0.00	0.00	-12.20	0.00	0.000***
wagepan$poorhlth	-0.02	0.04	-0.47	0.64	-0.020

	2.5 %	97.5 %
(Intercept)	0.0616143	0.0741192
secteurautre	-0.0424757	0.1044966
secteurbusiness	-0.0738142	0.0469707
secteurconstruc	-0.0996084	0.0595713
secteurenter	-0.1377153	0.0662370
secteurfin	-0.0722539	0.1022119
secteurmanuf	-0.0356227	0.0640123
secteurprorel	-0.1047763	0.0215873
secteurpublic	-0.0785907	0.0781880
secteurtrading	-0.0802527	0.0242319
secteurtransport	-0.0943483	0.0422051
wagepan$geo_zonenorth east	-0.0458255	0.2751098
wagepan$geo_zoneother zones	-0.1181548	0.0761312
union	0.0054976	0.0684064
hours	-0.0002565	-0.0001855
wagepan$poorhlth	-0.1031938	0.0632992

La variable à évolution constante étant omise comparé au modèle à estimation within, on retrouve avec une regression first difference à peu pres les mêmes résultats que pour notre première régression.

Régression MCQG pour modèle à effet aléatoire avec correction de white

L'estimateur MCQG pour le modèle à effets aléatoires permet de tenir compte des caractéristiques invariables dans le temps dans sa méthode de calcul. La spécification d'un modèle à effets aléatoire suppose cependant que les effets fixes individuels inobservés ne sont pas corrélés avec les variables explicatives. Autrement dit les résidus ne doivent pas être corrélés avec les régresseurs. Ainsi les estimations pour ce modèle risquent d'être moins consistantes car les effets des variables non observées, tels que les compétences individuelles (motivation, détermination), ne sont pas pris en compte de manière appropriée car supposées non corrélées avec les régressuers. Ce modèle nous permettra d'estimer alors les effets des regresseurs constants dans le temps cette fois, tels que l'origine ethnique, le niveau d'étude, la situation maritale (nous considérons la variable comme constante étant donné qu'il y a peu de divorce sur les périodes observées). Nous utiliserons ensuite un test d'Hausman pour savoir si les résidus sont corrélés aux régresseurs. Les résultats du test nous permettront de savoir quelle spécification de modèle adopter. A priori en vue de la thématique abordée, le modèle à spécifier s'apparenterait plus à un modèle à effet fixe. Les résultats obtenus grâce au modèle à effet aléatoire risquent donc de ne pas être fiables.

term	estimate	std.error	statistic	p.value	coefficients
(Intercept)	0.13	0.11	1.19	0.24	0.130
exper	0.05	0.00	19.87	0.00	0.053***
secteurautre	0.09	0.05	1.81	0.07	0.091
secteurbusiness	0.06	0.04	1.66	0.10	0.060
secteurconstruc	0.04	0.04	1.07	0.28	0.041
secteurenter	-0.11	0.06	-1.92	0.05	-0.107
secteurfin	0.17	0.06	2.99	0.00	0.168**
secteurmanuf	0.12	0.03	3.81	0.00	0.116***
secteurprorel	-0.03	0.04	-0.65	0.52	-0.026
secteurpublic	0.06	0.04	1.56	0.12	0.063
secteurtrading	-0.01	0.03	-0.36	0.72	-0.011
secteurtransport	0.08	0.04	1.97	0.05	0.076*
married	0.06	0.02	3.79	0.00	0.059***
originhisp	0.15	0.05	2.86	0.00	0.148**
originother	0.14	0.05	3.16	0.00	0.142**
wagepan$geo_zonenorth east	0.11	0.04	2.83	0.00	0.108**
wagepan$geo_zoneother zones	0.04	0.03	1.25	0.21	0.040
union	0.08	0.02	4.89	0.00	0.083***
educ	0.10	0.01	13.77	0.00	0.104***
hours	0.00	0.00	-8.52	0.00	0.000***
wagepan$poorhlth	-0.01	0.04	-0.18	0.86	-0.007

	2.5 %	97.5 %
(Intercept)	-0.0847042	0.3445529
exper	0.0480666	0.0585897
secteurautre	-0.0075968	0.1887460
secteurbusiness	-0.0108398	0.1307144
secteurconstruc	-0.0342714	0.1169095
secteurenter	-0.2153897	0.0022046
secteurfin	0.0579245	0.2776407
secteurmanuf	0.0565321	0.1762306
secteurprorel	-0.1028730	0.0518458
secteurpublic	-0.0160838	0.1428995
secteurtrading	-0.0725233	0.0498011
secteurtransport	0.0003647	0.1518293
married	0.0283104	0.0888964
originhisp	0.0465439	0.2490064
originother	0.0540914	0.2307360
wagepan$geo_zonenorth east	0.0331236	0.1831201
wagepan$geo_zoneother zones	-0.0227592	0.1029612
union	0.0497223	0.1162156
educ	0.0894463	0.1191403
hours	-0.0001567	-0.0000981
wagepan$poorhlth	-0.0841582	0.0698847

Avec ce dernier, plus de resultats sont statistiquement significatifs. D'après ce modèle, une année d'experience supplémentaire permet une augmentation du salaire de 0,36 dollars. Les regresseurs constant dansle temps sont aussi plus ou moins significatifs. On remarque que d'après ce modèle, les hispaniques gagnent en

moyenne 0,75 dollars de l'heure de plus que les personnes noirs. Les personnes d'origine américaine ou autre gagnent en moyenne 0,79 dollars de plus que les personnes noires. On remarque également qu'une année d'étude supplémentaire créé une augmentation de salaire de à,68 dollars de l'heure en moyenne. Enfin, ce modèle nous dit également que le temps de travail n'a pas d'impact sur le salaire horaire.

Regression en between

L'utilisation des estimateurs between et pooling fait l'hypothèse que les effets des variables explicatives sont les mêmes pour chaque individu. Ces hypothèses sont fortes et peu tenables dans un contexte réel, nous utiliserons les résultats des estimations en between et en pooling qu'à titre de comparaison avec des modèles plus appropriés.

term	estimate	std.error	statistic	p.value	coefficients
(Intercept)	0.51	0.18	2.81	0.01	0.51**
exper	0.01	0.01	1.19	0.24	0.01
married	0.12	0.04	3.12	0.00	0.12**
originhisp	0.14	0.05	2.63	0.01	0.14**
originother	0.13	0.04	2.99	0.00	0.13**
union	0.23	0.04	5.56	0.00	0.23***
educ	0.08	0.01	8.40	0.00	0.08***
hours	0.00	0.00	-0.91	0.36	0.00
wagepan$poorhlth	-0.20	0.21	-0.95	0.34	-0.20

	2.5 %	97.5 %
(Intercept)	0.1541269	0.8685906
exper	-0.0079021	0.0319766
married	0.0434924	0.1911996
originhisp	0.0349251	0.2422613
originother	0.0446827	0.2166308
union	0.1484631	0.3107245
educ	0.0621336	0.1000560
hours	-0.0001042	0.0000383
wagepan$poorhlth	-0.6257760	0.2186805

Regression en pooling

term	estimate	std.error	statistic	p.value	coefficients
(Intercept)	0.25	0.11	2.19	0.03	0.25*
exper	0.04	0.00	12.37	0.00	0.04***
married	0.10	0.02	4.18	0.00	0.10***
originhisp	0.14	0.05	2.68	0.01	0.14**
originother	0.13	0.05	2.87	0.00	0.13**
union	0.16	0.02	6.36	0.00	0.16***
educ	0.10	0.01	12.09	0.00	0.10***
hours	0.00	0.00	-4.09	0.00	0.00***
wagepan$poorhlth	-0.05	0.06	-0.79	0.43	-0.05

	2.5 %	97.5 %
(Intercept)	0.0257196	0.4687097
exper	0.0364354	0.0501599
married	0.0516675	0.1431248
originhisp	0.0382093	0.2469678
originother	0.0419369	0.2225811
union	0.1084913	0.2051258
educ	0.0821949	0.1140149
hours	-0.0001237	-0.0000436
wagepan$poorhlth	-0.1600263	0.0682610

Test d'effets individuels

```
##
##  Lagrange Multiplier Test - (Honda)
##
## data:  lwage ~ exper + married + origin + union + educ + hours + wagepan$poorhlth
## normal = 67.224, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

```
##
##  Lagrange Multiplier Test - (Breusch-Pagan)
##
## data:  lwage ~ exper + married + origin + union + educ + hours + wagepan$poorhlth
## chisq = 4519.1, df = 1, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

Pour notre thématique il est préférable de spécifier des modèles qui considère des effets individuels

Test d'Hausmann : Spécification d'un modèle à effet fixe ou effet aléatoire ?

```
##
##  Hausman Test
##
## data:  lwage ~ exper + secteur + wagepan$geo_zone + union + hours +  ...
## chisq = 78.823, df = 16, p-value = 2.71e-10
## alternative hypothesis: one model is inconsistent
```

Il n'y alors que le modèle à effet fixe qui peut être estimé de manière consistente. Ce résultat est assez intuitif étant donné la thématique abordée, en effet les effets fixes individuels sont fortement corrélés avec certaines variables explicatives comme on l'a déjà expliqué. Nous allons devoir alors déterminer quel estimateur choisir entre le within et le D1 pour spécifier notre modèle à effet fixe.

Within vs Différence Première

Dans un premier temps rappelons que l'estimateur différence première (D1) nécessite une variation temporelle des variables explicatives pour chaque individu. A l'inverse, l'estimateur à effet fixe within est lui souvent préférable à l'estimateur différence première dans les cas où il y a peu de variation temporelle des variables explicatives pour un individu donné. Cela peut arriver lorsque l'on utilise des données de panel où les individus sont suivis pendant une courte période de temps, comme pour le cas dans lequel nous nous trouvons. Nos variables explicatives varient peu également. Une partie de nos variables explicatives sont des dichotomiques constantes, et celles qui sont numériques continues ont une évolution assez constante. C'est le cas de l'expérience ou du nombre d'heure (voir statistiques descriptives). Nous penchons donc hypothétiquement pour l'estimateur within de préférence avant la validation par test.

Dans un second temps, rappelons que pour les deux estimateurs, nous considérons les mêmes modèles avec des résidus u_i corrélés avec nos régresseurs. Le calcul par estimateur within, pour un modèle à effet fixe, fait l'hypothèse que ces derniers ne sont pas corrélés entre eux au cours du temps. Il est de facto plus efficace que l'estimateur en différence première qui n'exige pas pour ses calculs la validation de cette hypothèse. Cependant dans de nombreux cas empiriques, ces résidus peuvent être de plus en plus corrélés au fil du temps, par exemple si les u_i suivent une marche aléatoire. Dans ce cas, leur valeur en période t dépend entièrement de leur valeur en t-1 plus un choc aléatoire. Cela implique qu'il y a un niveau de corrélation sérielle très fort et positif. Cependant les différentiels des u_i ne sont pas corrélés, dans ce cas il sera préférable d'utiliser l'estimateur différence première. Dans de nombreux cas il existe une corrélation sérielle entre les u_i mais bien moins forte que pour une marche aléatoire. Ainsi il sera difficile de faire un choix entre les 2 estimateurs.Il est difficile de tester si les u_i sont corrélés suite aux estimations within. Cependant il est possible de voir si les différentiels des résidus obtenus en différence première le sont. Si ce n'est pas le cas, D1 pourra être utilisé. Si ces derniers différentiels sont négativement et sériellement corrélés on optera pour within (recommandation Wooldridge, voir annexe 1). Si le test ne se présente pas en faveur d'un estimateur plutot qu'un autre, en vue de la longueur de notre dimension temportelle et de l'évolution de nos variables explicatives nos opterons pour l'estimateur within.

```
##
##  Wooldridge's first-difference test for serial correlation in panels
##
## data:  fd
## F = 47.839, df1 = 1, df2 = 3108, p-value = 5.594e-12
## alternative hypothesis: serial correlation in original errors
```

```
##
##  Wooldridge's first-difference test for serial correlation in panels
##
## data:  fd
## F = 134.04, df1 = 1, df2 = 3108, p-value < 2.2e-16
## alternative hypothesis: serial correlation in differenced errors
```

On observe que pour le modèle considéré, les résidus sont caractérisés par de la corrélation sérielle. Cependant on observe que pour l'estimation D1 les différenciels des erreurs à chaque période comportent eux aussi de la corrélation sérielle. Dans ce cas nous obtons par l'estimateur within.

Regression à modèle à effet fixe à estimateur within pour tester les effets de l'éducation sur l'évolution des salaires

Un problème se pose à nous maintenant : nous savons qu'en vue de la thématique abordée, un modèle à effet fixe avec une estimation within est plus approprié. Or, celui-ci ne permet pas de considérer des variables constantes dans le temps. Il est alors difficile pour nous de mesurer l'impact de ces dernières sur l'évolution des salaires. L'enjeu principal de la question d'étude étant de pouvoir quantifier les différentes dynamiques d'évolution des salaires entre différents groupes invariants dans le temps(origines ethniques ou niveau d'éducation)

S'il n'est pas possible de prendre en compte directement ces variables dans des modèles à effets fixes, il est tout de même possible de les considérer à travers leurs interractions avec d'autres variables inconstantes dans le temps et en particulier des variables dummy temporelles (effets combinés). C'est donc ce que nous avons spécifié pour la régression suivante. Nous avons introduit toutes les dummies temporelles exceptée celle pour notre période de référence. Il ne sera donc pas possible d'estimer les effets de l'éducation sur la détermination des salaires pour l'année 1980 qui est notre période de référence.

De même en introduisant ces dummies annuelles, les effets des variables dont l'évolution est constante dans le temps deviennent inestimables. C'est le cas pour les effets de l'expérience professionnelle qui par définition augmente de une unité chaque année.

term	estimate	std.error	statistic	p.value	coefficients
married	0.05	0.02	2.84	0.00	0.047**
exper	0.00	0.02	0.17	0.86	0.003
union	0.07	0.02	3.78	0.00	0.071***
hours	0.00	0.00	-8.42	0.00	0.000***
wagepan$poorhlth	0.00	0.04	-0.05	0.96	-0.002
d81	0.00	0.08	0.03	0.98	0.003
d82	-0.04	0.08	-0.44	0.66	-0.036
d83	-0.03	0.08	-0.32	0.75	-0.027
d84	-0.01	0.09	-0.07	0.94	-0.007
d85	-0.10	0.07	-1.45	0.15	-0.104
d86	-0.08	0.09	-0.94	0.35	-0.081
d81:educ	0.01	0.01	0.88	0.38	0.007
educ:d82	0.01	0.01	1.53	0.13	0.013
educ:d83	0.02	0.01	1.73	0.08	0.016
educ:d84	0.02	0.01	1.95	0.05	0.020
educ:d85	0.03	0.01	3.39	0.00	0.033***
educ:d86	0.04	0.01	3.48	0.00	0.036***
educ:d87	0.03	0.01	3.04	0.00	0.033**

	2.5 %	97.5 %
married	0.0144642	0.0790822
exper	-0.0323410	0.0386245
union	0.0342063	0.1079373
hours	-0.0001717	-0.0001068
wagepan$poorhlth	-0.0810799	0.0767958
d81	-0.1631000	0.1681216
d82	-0.1976926	0.1247592
d83	-0.1912391	0.1377336
d84	-0.1916211	0.1784116
d85	-0.2447873	0.0363992
d86	-0.2501835	0.0875865
d81:educ	-0.0084136	0.0221056
educ:d82	-0.0035594	0.0287927
educ:d83	-0.0020873	0.0335878
educ:d84	-0.0001113	0.0400134
educ:d85	0.0138562	0.0518345
educ:d86	0.0159231	0.0569886
educ:d87	0.0116968	0.0543229

Résultats du test de Fisher

```
## La significativité jointe de Fisher n'est pas atteinte. Le test de Fisher =  <
```

On voit bien que les effets estimés de l'éducation sont positifs et croissants au fil du temps. Les effets significatifs sont cependant ceux pour les dernières années d'observation avec par exemple une augmentation de la valorisation d'une année d'éducation d'environ 3 points de pourcentage en plus que pour l'année de référence, en ce qui concerne la détermination des salaires. En terme d'interval de confiance on voit tout de même que la valorisation de l'éducation est sensiblement la même pour la détermination du salaire des individues des individus (faible écart type) et qu'elle a été revu à la hause (coefficients positifs est croissants)

Cependant, en réalisant un test de Fisher de significativité jointe on voit qu'on ne peut pas rejeter l'hypothèse que les effets joints des dummies annuelles combinés à ceux de l'éducation sont différents de 0 bien que 3 de ces effets combinés se soient révélés significatif individuellement.

Régression pour déterminer les effets de l'origine ethnique sur l'évolution des salaires

La méthode précédente nous ayant permis de surmonter le problème d'estimation de variables constantes dans le temps au sein d'un modèle à effet fixe, nous réitérons la méthodes pour estimer les effets de l'origine ethnique.

term	estimate	std.error	statistic	p.value	coefficients
married	0.05	0.02	2.91	0.00	0.048**
exper	0.06	0.02	3.41	0.00	0.062***
union	0.07	0.02	3.95	0.00	0.074***
hours	0.00	0.00	-8.32	0.00	0.000***
wagepan$poorhlth	0.00	0.04	-0.01	0.99	0.000
d81	0.03	0.08	0.32	0.75	0.027
d82	0.00	0.08	0.01	0.99	0.001
d83	-0.01	0.08	-0.08	0.94	-0.007
d84	0.01	0.09	0.12	0.90	0.012
d85	0.01	0.07	0.10	0.92	0.007
d86	0.02	0.09	0.20	0.84	0.017

	2.5 %	97.5 %
married	0.0150428	0.0809986
exper	0.0546241	0.0689444
union	0.0373145	0.1111660
hours	-0.0001698	-0.0001057
wagepan$poorhlth	-0.0804518	0.0795862
d81	-0.0063264	0.0604248
d82	-0.0316339	0.0328103
d83	-0.0391685	0.0256424
d84	-0.0181695	0.0415939
d85	-0.0181601	0.0329603
d86	-0.0080287	0.0421972
d81:black	-0.0707813	0.0827409
black:d82	-0.1373656	0.0474425
black:d83	-0.1267344	0.0683451
black:d84	-0.1836310	0.0262481
black:d85	-0.2305921	-0.0209509
black:d86	-0.1763870	0.0348324
black:d87	-0.2510514	-0.0317322
d81:hisp	-0.0871942	0.0432930
d82:hisp	-0.0760937	0.0670377
d83:hisp	-0.1283967	0.0349174
d84:hisp	-0.1559481	0.0362825
d85:hisp	-0.1216047	0.0675307
d86:hisp	-0.1966607	0.0284147
d87:hisp	-0.1414943	0.0628392

La régression ne veut pas afficher les effets combinés des années aux origines ethnique. On a cependant une idée de l'interval de confiance de ces coefficients. On ne sait pas pourquoi la régression ne les affiche pas.

Ainsi en se basant sur les intervals de confiance on peut tout de même se rendre compte que contrairement à l'éducation, l'effet de l'origine éthnique a été beaucoup plus disparatre. En effet au fil des années on voit que certaines personnes afro-américaines, et hispaniques ont été tres désavantagées au niveau des salaires alors que d'autres ont été tres avantagées par rapport aux autres orgines ethniques présupposées subir moins de discrimination. En effet, il y a grand écart type avec des effets de l'origines ethnique fortement postif et fortement négatif concernant l'évolution des salaires par rapport aux autres originies ethniques).

Il reste cependant compliqué pour nous de tester la significativé de cette variable origine ethnique car n'ayant pas de résultats affichés pour la régression.

Spécification d'un Modèle Hausman-Taylor.

Le modèle Hausman-Taylor est un modèle de panel pour lequel on prend soin de préciser quels régresseurs sont corrélés avec effet fixes et lesquels sont invariants dans le temps. Pour celui-ci, l'instrumentalisation est utilisée pour résoudre ce problème d'endogénéité, ce qui peut conduire à une estimation biaisée des coefficients. Pour résoudre ce problème, on utilise des variables instrumentales. Cette instrumentalisation se fait par la transformation within des variables endogènes. Elle consiste a soustraire la moyenne temporelle des valeurs de la variable prises pour l'individu i à la valeur de la variable pour ce même individu i à chaque période. Or, si ces variables endogènes sont initialement corrélées avec les effets fixes, leur transformation ne l'est pas puisque la corrélation avec ces effets fixes ne peut être qu'avec la partie de la valeur du régresseur invariante dans le temps. Les variables variantes dans le temps et exogènes peuvent elles aussi subir une transformation within afin de les séparer de leur partie constante dans le temps. Apres que ces variables exogènes soient instrumentalisées, leur partie constante,elles, peuvent servir pour l'instrumentalisation des variables endogènes constantes dans le temps.

Tout l'enjeu de la spécification de ce modèle reste de pouvoir estimer les regresseurs constant dans le temps. Or, apres toutes les instrumentalisations réalisées, si le nombre de régresseurs endogènes et variants dans le temps sont au moins égaux au nombre de régresseurs endogènes et constant dans le temps , il sera possible d'estimer ces régresseurs constant dans le temps.

Dans notre cas nous supposons :

- Variables endogènes :
  - educ : éducation
  - exper : années d'expériences
  - union : syndicalisation
  - married : marié
  - secteur : secteurs d'activités
- Variables endogènes constantes :
  - educ : éducation
- Variables exogènes inconstantes :
  - South : habite dans le sud
  - Nrthcen : habite dans le nord centrale
  - Nrtheast : habite dans le nord est
  - Poorhlth : est en mauvaise santé
- Variables exogènes constantes :
  - Black : est afro-américain
  - Hisp : est d'origine hispanique

On se rend compte qu'on possède en effet plus de variables exogènes inconstantes que de variables endogènes constantes. On a donc assez d'instruments pour estimer pour estimer les régresseurs endogènes et exogènes constant.

term	estimate	std.error	statistic	p.value	coefficients
(Intercept)	-3.08	2.02	-1.53	0.13	-3.085
married	0.04	0.01	3.19	0.00	0.042**
union	0.07	0.01	5.27	0.00	0.074***
exper	0.05	0.00	25.91	0.00	0.049***
wagepan$secteurautre	0.06	0.04	1.26	0.21	0.056
wagepan$secteurbusiness	0.03	0.04	0.80	0.42	0.031
wagepan$secteurconstruc	0.01	0.04	0.27	0.78	0.011
wagepan$secteurenter	-0.10	0.05	-1.83	0.07	-0.097
wagepan$secteurfin	0.14	0.05	2.88	0.00	0.137**
wagepan$secteurmanuf	0.08	0.03	2.20	0.03	0.076*
wagepan$secteurprorel	-0.02	0.04	-0.59	0.55	-0.023
wagepan$secteurpublic	0.04	0.04	0.93	0.35	0.040
wagepan$secteurtrading	-0.03	0.04	-0.76	0.44	-0.027
wagepan$secteurtransport	0.03	0.04	0.83	0.40	0.034
black	-0.05	0.10	-0.48	0.63	-0.049
hisp	0.29	0.19	1.53	0.13	0.290
wagepan$nrthcen	0.01	0.04	0.22	0.83	0.008
wagepan$nrtheast	0.12	0.05	2.41	0.02	0.117*
wagepan$south	0.07	0.04	1.55	0.12	0.067
wagepan$poorhlth	0.00	0.03	-0.07	0.94	-0.002
educ	0.37	0.17	2.18	0.03	0.366*

	2.5 %	97.5 %
(Intercept)	-7.0404267	0.8714030
married	0.0162213	0.0677564
union	0.0466980	0.1019744
wagepan$exper	0.0455711	0.0530284
wagepan$secteurautre	-0.0305088	0.1415341
wagepan$secteurbusiness	-0.0439312	0.1049824
wagepan$secteurconstruc	-0.0665059	0.0880329
wagepan$secteurenter	-0.2000960	0.0066257
wagepan$secteurfin	0.0437839	0.2305022
wagepan$secteurmanuf	0.0081812	0.1439167
wagepan$secteurprorel	-0.1014018	0.0544292
wagepan$secteurpublic	-0.0448613	0.1255780
wagepan$secteurtrading	-0.0966572	0.0424410
wagepan$secteurtransport	-0.0458268	0.1135960
black	-0.2499892	0.1521493
hisp	-0.0813153	0.6611766
wagepan$nrthcen	-0.0674251	0.0842951
wagepan$nrtheast	0.0219872	0.2121240
wagepan$south	-0.0176531	0.1524536
wagepan$poorhlth	-0.0693432	0.0644699
educ	0.0370431	0.6948875

Pour ce modèle on peut donc supposer que sur la période d'année éttudiée, la valorisation de l'éducation aurait augmenté de de 37 % de manière significative (95 %). De même l'impact de l'origine n'aurait pas été significative sur l'évolution des salaires. Les afro-américains auraient leur salaire évoluer 5 % de moins que

la catégories autres autres origines ethniques. les hispaniques eux auraient vu leur salaire augmenter de 30% de plus que la catégorie autres origines ethniques, cependant il n'y a rien de significatif à ces estimations.

Conclusion.

Pour résumer notre travail, nous avons dans un premier temps estimé des modèles classiques (within, différence première, effet aléatoire, between, pooling). Nous supposions que pour notre thématique d'étude, un modèle où les effets fixes et individuels étaientt corrélés avec les régresseurs était plus adapté. Nous avions alors du tester entre un modèle à estimer en within ou différence première. Nous avons donc opté pour un estiâteur within selon les conseils méthodologiques de Woolridge.

Within une fois choisi, il nous était impossible d'estimer des effets de régresseurs constant dans le temps. Pour résoudre ce problème nous avons utilisé une méthode permettant de considérer les effets des régresseurs constants en les combinant avec des dummies temporelles.

Dans un second temps nous avons aussi utilisé un modèle dynamique sous la spécification Hausman-Taylor. Les conditions pour pouvoir estimers les régresseurs constants étaient d'autant plus satisfaites.

Il reste compliqué cependant de choisir un modèle en particulier. Nos 2 modèles les plus fiables restent le within et le Hausman-Taylor. Cependant en termes d'estimation, ces modèles fournissent des coefficients assez similaires (ex : la valorisation d'une année d'expérience aurait augmenté de 6% envrion par estimation within et de 5 % avec modèle Hausman-Taylor). Cependant en spécifiant un modèle Hausman Taylor et en comparant les résultat au modèle effets fixes within avec effets combinés avec dummies, on a des résultats assez différents. L'effet de l'éxpérience est assez différent 0,3 % d'augmentation de valorisation dans la détermination des salaires. Cet effet n'est pas significatif. De même les deux modèles permettent d'évaluer les effets de l'éducation (régresseur constant). on retrouve des résultats également très différents : la valorisation de l'éducation aurait augmenté de de 37 % de manière significative selon les estimations du modèle Hausman-Taylor. La valorisation d'une année d'éducation aurait elle augmenté d'environ 3 points de pourcentage par rapport à 1980 en 1987 selon les estimations du modèle within avec effets combinés. Les résultats de ce dernier modèle sont d'autant plus différents des autres modèles. Cependant sa spécification a été faite suite au instruction pédagogique de wooldridge (voir annexe 2).

Ainsi face à ces résultats assez différents pour 2 modèles dont la pertinence théorique semble justifiée, il nous est bel et bien compliqué d'opter pour un en particulier.

Bibliographie :

F. Vella and M. Verbeek (1998), "Whose Wages Do Unions Raise? A Dynamic Model of Unionism and Wage Rate Determination for Young Men," Journal of Applied Econometrics 13, 163-183.

Wooldridge, J. M. (2003). Introductory Econometrics: A Modern Approach. South Western Educational Publishing.

Bound, J., & Johnson, G. (1992). Changes in the structure of wages in the 1980s: An evaluation of alternative explanations. American economic review, 82(3), 371-392.

Katz, L. F., & Murphy, K. M. (1992). Changes in relative wages, 1963-1987: Supply and demand factors. The Quarterly Journal of Economics, 107(1), 35-78.

Annexes :

- Annexe 1 : Choix entre estimateur Within et différences premières
- Annexe 2 : Spécification d'un modèle effets fixes avec utilisations d'effets combinés pour déterminer les effets des variables constantes

EXAMPLE 14.2 HAS THE RETURN TO EDUCATION CHANGED OVER TIME?

The data in WAGEPAN.RAW are from Vella and Verbeek (1998). Each of the 545 men in the sample worked in every year from 1980 through 1987. Some variables in the data set change over time: experience, marital status, and union status are the three important ones. Other variables do not change: race and education are the key examples. If we use fixed effects (or first differencing), we cannot include race, education, or experience in the equation. However, we can include interactions of *educ* with year dummies for 1981 through 1987 to test whether the return to education was constant over this time period. We use log(*wage*) as the dependent variable, dummy variables for marital and union status, a full set of year dummies, and the interaction terms *d81*·*educ*, *d82*·*educ*, …, *d87*·*educ*.

The estimates on these interaction terms are all positive, and they generally get larger for more recent years. The largest coefficient of .030 is on *d87*·*educ*, with $t = 2.48$. In other words, the return to education is estimated to be about 3 percentage points larger in 1987 than in the base year, 1980. (We do not have an estimate of the return to education in the base year for the reasons given earlier.) The other significant interaction term is *d86*·*educ* (coefficient = .027, $t = 2.23$). The estimates on the earlier years are smaller and insignificant at the 5% level against a two-sided alternative. If we do a joint F test for significance of all seven interaction terms, we get p-value = .28: this gives an example where a set of variables is jointly insignificant even though some variables are individually significant. [The *df* for the F test are 7 and 3,799; the second of these comes from $N(T - 1) - k = 545(8 - 1) - 16 = 3{,}799$.] Generally, the results are consistent with an increase in the return to education over this period.

The Dummy Variable Regression

A traditional view of the fixed effects approach is to assume that the unobserved effect, a_i, is a parameter to be estimated for each i. Thus, in equation (14.4), a_i is the intercept for person i (or firm i, city i, and so on) that is to be estimated along with the β_j. (Clearly, we cannot do this with a single cross section: there would be $N + k$ parameters to estimate with only N observations. We need at least two time periods.) The way we estimate an intercept for each i is to put in a dummy variable for each cross-sectional observation, along with the explanatory variables (and probably dummy variables for each time period). This method is usually called the **dummy variable regression**. Even when N is not very large (say, $N = 54$ as in Example 14.1), this results in many explanatory variables—in most cases, too many to explicitly carry out the regression. Thus, the dummy variable method is not very practical for panel data sets with many cross-sectional observations.

Nevertheless, the dummy variable regression has some interesting features. Most importantly, it gives us *exactly* the same estimates of the β_j that we would obtain from the regression on time-demeaned data, and the standard errors and other major statistics are identical. Therefore, the fixed effects estimator can be obtained by the dummy variable regression. One benefit of the dummy variable regression is that it properly computes the degrees of freedom directly. This is a minor advantage now that many econometrics packages have programmed fixed effects options.

The R-squared from the dummy variable regression is usually rather high. This occurs because we are including a dummy variable for each cross-sectional unit, which explains much of the variation in the data. For example, if we estimate the unobserved effects

model in Example 13.8 by fixed effects using the dummy variable regression (which is possible with $N = 22$), then $R^2 = .933$. We should not get too excited about this large R-squared: it is not surprising that we can explain much of the variation in unemployment claims using both year and city dummies. Just as in Example 13.8, the estimate on the EZ dummy variable is more important than R^2.

The R-squared from the dummy variable regression can be used to compute F tests in the usual way, assuming, of course, that the classical linear model assumptions hold (see the chapter appendix). In particular, we can test the joint significance of all of the cross-sectional dummies ($N - 1$, since one unit is chosen as the base group). The unrestricted R-squared is obtained from the regression with all of the cross-sectional dummies; the restricted R-squared omits these. In the vast majority of applications, the dummy variables will be jointly significant.

Occasionally, the estimated intercepts, say $\hat{a}_i$, are of interest. This is the case if we want to study the distribution of the $\hat{a}_i$ across i, or if we want to pick a particular firm or city to see whether its $\hat{a}_i$ is above or below the average value in the sample. These estimates are directly available from the dummy variable regression, but they are rarely reported by packages that have fixed effects routines (for the practical reason that there are so many $\hat{a}_i$). After fixed effects estimation with N of any size, the $\hat{a}_i$ are pretty easy to compute:

$$\hat{a}_i = \bar{y}_i - \hat{\beta}_1\bar{x}_{i1} - \ldots - \hat{\beta}_k\bar{x}_{ik}, \; i = 1, \ldots, N, \qquad [14.6]$$

where the overbar refers to the time averages and the $\hat{\beta}_j$ are the fixed effects estimates. For example, if we have estimated a model of crime while controlling for various time-varying factors, we can obtain $\hat{a}_i$ for a city to see whether the unobserved fixed effects that contribute to crime are above or below average.

Some econometrics packages that support fixed effects estimation report an "intercept," which can cause confusion in light of our earlier claim that the time-demeaning eliminates all time-constant variables, including an overall intercept. [See equation (14.5).] Reporting an overall intercept in fixed effects (FE) estimation arises from viewing the a_i as parameters to estimate. Typically, the intercept reported is the average across i of the $\hat{a}_i$. In other words, the overall intercept is actually the average of the individual-specific intercepts, which is an unbiased, consistent estimator of $\alpha = \mathrm{E}(a_i)$.

In most studies, the $\hat{\beta}_j$ are of interest, and so the time-demeaned equations are used to obtain these estimates. Further, it is usually best to view the a_i as omitted variables that we control for through the within transformation. The sense in which the a_i can be estimated is generally weak. In fact, even though $\hat{a}_i$ is unbiased (under Assumptions FE.1 through FE.4 in the chapter appendix), it is not consistent with a fixed T as $N \to \infty$. The reason is that, as we add each additional cross-sectional observation, we add a new a_i. No information accumulates on each a_i when T is fixed. With larger T, we can get better estimates of the a_i, but most panel data sets are of the large N and small T variety.

Fixed Effects or First Differencing?

So far, setting aside pooled OLS, we have seen two competing methods for estimating unobserved effects models. One involves differencing the data, and the other involves time-demeaning. How do we know which one to use?

We can eliminate one case immediately: when $T = 2$, the FE and FD estimates, as well as all test statistics, are *identical*, and so it does not matter which we use. Of course, the equivalance between the FE and FD estimates requires that we estimate the same model in each case. In particular, as we discussed in Chapter 13, it is natural to include an intercept in the FD equation; this intercept is actually the intercept for the second time period in the original model written for the two time periods. Therefore, FE estimation must include a dummy variable for the second time period in order to be identical to the FD estimates that include an intercept.

With $T = 2$, FD has the advantage of being straightforward to implement in any econometrics or statistical package that supports basic data manipulation, and it is easy to compute heteroskedasticity-robust statistics after FD estimation (because when $T = 2$, FD estimation is just a cross-sectional regression).

When $T \geq 3$, the FE and FD estimators are not the same. Since both are unbiased under Assumptions FE.1 through FE.4, we cannot use unbiasedness as a criterion. Further, both are consistent (with T fixed as $N \rightarrow \infty$) under FE.1 through FE.4. For large N and small T, the choice between FE and FD hinges on the relative efficiency of the estimators, and this is determined by the serial correlation in the idiosyncratic errors, u_{it}. (We will assume homoskedasticity of the u_{it}, since efficiency comparisons require homoskedastic errors.)

When the u_{it} are serially uncorrelated, fixed effects is more efficient than first differencing (and the standard errors reported from fixed effects are valid). Since the unobserved effects model is typically stated (sometimes only implicitly) with serially uncorrelated idiosyncratic errors, the FE estimator is used more than the FD estimator. But we should remember that this assumption can be false. In many applications, we can expect the unobserved factors that change over time to be serially correlated. If u_{it} follows a random walk—which means that there is very substantial, positive serial correlation—then the difference Δu_{it} is serially uncorrelated, and first differencing is better. In many cases, the u_{it} exhibit some positive serial correlation, but perhaps not as much as a random walk. Then, we cannot easily compare the efficiency of the FE and FD estimators.

It is difficult to test whether the u_{it} are serially uncorrelated after FE estimation: we can estimate the time-demeaned errors, $\ddot{u}_{it}$, but not the u_{it}. However, in Section 13.3, we showed how to test whether the differenced errors, Δu_{it}, are serially uncorrelated. If this seems to be the case, FD can be used. If there is substantial negative serial correlation in the Δu_{it}, FE is probably better. It is often a good idea to try both: if the results are not sensitive, so much the better.

When T is large, and especially when N is not very large (for example, $N = 20$ and $T = 30$), we must exercise caution in using the fixed effects estimator. Although exact distributional results hold for any N and T under the classical fixed effects assumptions, inference can be very sensitive to violations of the assumptions when N is small and T is large. In particular, if we are using unit root processes—see Chapter 11—the spurious regression problem can arise. First differencing has the advantage of turning an integrated time series process into a weakly dependent process. Therefore, if we apply first differencing, we can appeal to the central limit theorem even in cases where T is larger than N. Normality in the idiosyncratic errors is not needed, and heteroskedasticity and serial correlation can be dealt with as we touched on in Chapter 13. Inference with the fixed effects estimator is potentially more sensitive to nonnormality, heteroskedasticity, and serial correlation in the idiosyncratic errors.

Like the first difference estimator, the fixed effects estimator can be very sensitive to classical measurement error in one or more explanatory variables. However, if each x_{itj} is uncorrelated with u_{it}, but the strict exogeneity assumption is otherwise violated—for example, a lagged dependent variable is included among the regressors or there is feedback between u_{it} and future outcomes of the explanatory variable—then the FE estimator likely has substantially less bias than the FD estimator (unless $T = 2$). The important theoretical fact is that the bias in the FD estimator does not depend on T, while that for the FE estimator tends to zero at the rate $1/T$. See Wooldridge (2010, Section 10.7) for details.

Generally, it is difficult to choose between FE and FD when they give substantively different results. It makes sense to report both sets of results and to try to determine why they differ.