Estimation localisée du chômage : une application des techniques d'estimation sur petits domaines

Pascal Ardilly
INSEE

Objectif, contexte

Effectif de chômeurs BIT au T1 de 2007, par zone d'emploi (France métropolitaine) ???

348 zones d'emploi (ZE) en France métropolitaine

Taille médiane ZE = 82 000 individus de 15 ans ou plus vivant en ménage ordinaire
(max = 1 850 000 ; min= 8 000)

Source : échantillon "Emploi" (T1 de 2007)

Méthode actuelle Insee : ventilation du chômage national proportionnellement aux DEFM par ZE (pour chaque croisement sexe X age).

Champ = personnes ayant 15 ans ou plus au 31 décembre 2007, résidant en ménage ordinaire.

Plan de sondage complexe

Taille échantillon **répondant** = 73 153 individus

90% des poids dans un rapport variant de 1 à 6.

Estimation du nombre total de chômeurs BIT = 2 416 000 (T1 de 2007)

10 ZE non couvertes par EEC
(21 700 <pop^t< 64 000).

Distribution des tailles d'échantillon répondants par ZE $(n > 0)$

Quantile		Taille d'échantillon répondant
100%	**Max**	2301
99%		1233
95%		693
90%		520
75%	**Q3**	227
50%	**Median**	**128**
25%	**Q1**	72
10%		38
5%		23
1%		12
0%	**Min**	1

Manifestement, l'estimation "classique" ne suffira pas (sauf pour les très grandes ZE ?)

Il va falloir formuler des hypothèses (modèles) "reliant" une ZE à un territoire plus grand

Les sources de données auxiliaires

A) *Le recensement de la population*

Un peu de variance, un peu de biais … on les ignorera !

Exploitation du fichier de cumul sur 5 ans ;

Restriction de champ : même champ que l'EEC;

⇒ **Niveau individu**

B) *Une base de données économiques / sociales ad-hoc*

329 variables de nature économique ou sociale par ZE.
Sources variées : RP, CAF, CLAP, Sirene, Lifi, DADS, données fiscales, …;

Atout : variables caractérisant la démographie d'entreprises et donc plutôt l'aspect "offre d'emploi".

⇒ **Niveau ZE**

C) *Les demandes d'emploi en fin de mois (DEFM)*

DEFM 1, 2, 3 : personnes sans emploi immédiatement disponibles, tenues d'accomplir des actes positifs de recherche d'emploi, à la recherche d'un emploi à durée indéterminé ;

Données individuelles anonymes disponibles ;

On dispose de : âge , sexe , niveau de formation.

⇒ **Niveau individu**

D) *La typologie "Tabard"*

Typologie socioprofessionnelle des quartiers et communes : profils de zonages en 27 postes.

Basée sur le RP 1999.

⇒ **Niveau ZE**

E) *Une base de données sur le RMI 2007*

effectifs allocataires + bénéficiaires par ZE.

⇒ **Niveau ZE**

F) *Une base relative aux ZUS*

Environ 850 Zones Urbaines Sensibles en France

Dénombrements par ZE : situation de l'année 2006.

⇒ **Niveau individu**

La base de données au niveau ZE

Variable d'intérêt (« expliquée ») par ZE :

$\hat{P}$ = rapport de l'estimateur pondéré du nombre de chômeurs à l'estimateur pondéré de la taille de ZE

ou

$\hat{N}_{cho} = N_{ZE} \cdot \hat{P}$;

France (métropole) : $\hat{P} = \frac{2\,416\,400}{49\,745\,000} \approx 4{,}86\%$

Distribution du ratio $\hat{P}$ par ZE

Quantile		Taux $\hat{P}$ (en %)
100%	**Max**	20.7
99%		14.2
95%		10.0
90%		8.2
75%	**Q3**	6.2
50%	**Median**	**4.2**
25%	**Q1**	2.4
10%		0.0
5%		0.0
1%		0.0
0%	**Min**	0.0

Base finale : 348 observations et 275 variables - essentiellement des proportions.

Modélisation au niveau ZE : la pré sélection des variables explicatives

Elimination de 2 ZE "aberrantes" ;

Elimination des ZE dont la taille d'échantillon est inférieure à 50 répondants (modèle à variable expliquée peu « fiable » !);

On cherche un processus pas trop compliqué, utilisant les logiciels existants.

- principe de sélection initiale basée sur les corrélations bivariées (non optimum…);

- on ignore toute structure complexe de la variance : variance = $\sigma^2 \cdot Id$

ETAPE 1 : Présélection des "meilleures" corrélations avec $\hat{P}$ (fixation de seuils pour ρ) :

- Recherche d'un emploi (déclaration au recensement) : $\rho = 42{,}7\,\%$
- Déclaration spontanée de chômage au recensement : $\rho = 42{,}2\,\%$
- Inscription DEFM, catégories 1,2,3 et HAR : $\rho = 39{,}5\,\%$
- Allocataire RMI : $\rho = 38{,}8\,\%$
- Inscription DEFM, catégories 1,2,3 et HAR et appartenir à la catégorie des hommes de 30 à 49 ans, peu diplômés : $\rho = 38{,}2\,\%$
- Ne pas vivre en couple : $\rho = 35{,}4\,\%$
- Part de la population ayant des bas revenus en 2005 : $\rho = 35{,}4\,\%$
- Part de la fonction "Administration publique" dans l'emploi total en 2006 : $\rho = 21{,}2\,\%$

- Etre marié : $\rho = -32{,}8\,\%$

- Taux d'activité dans la tranche d'âge 25-54 ans selon le RP 2006 : $\rho = -26{,}9\,\%$

- Avoir une profession d'indépendant : $\rho = -24{,}8\,\%$

- Part des agriculteurs dans la population de plus de 5 ans (RP 2006) : $\rho = -24{,}4\,\%$

- Taux d'activité dans la tranche d'âge 15-24 ans selon le RP 2006 : $\rho = -23{,}8\,\%$

- Avoir un niveau de diplôme égal au CAP / BEP : $\rho = -19{,}7\,\%$

ETAPE 2 : sélection de modèle avec **Proc GLMSELECT**

Choix d'une **stratégie STEPWISE**, où on a appliqué en parallèle plusieurs options :

a) Méthode basée sur la valeur du F de Fisher :

```
SELECTION = stepwise (select=SL
SLE=0.20 SLS=0.20 stop=ADJRSQ)
```

b) Méthode basée sur la valeur du R^2 ajusté :

```
SELECTION = stepwise (select=ADJRSQ
stop=CP)
```

c)Méthode de cross-validation :

```
CVMETHOD=BLOCK(5) CVDETAILS=ALL
SELECTION = STEPWISE (select=CV)
```

d)Méthode basée sur le critère d'Akaike

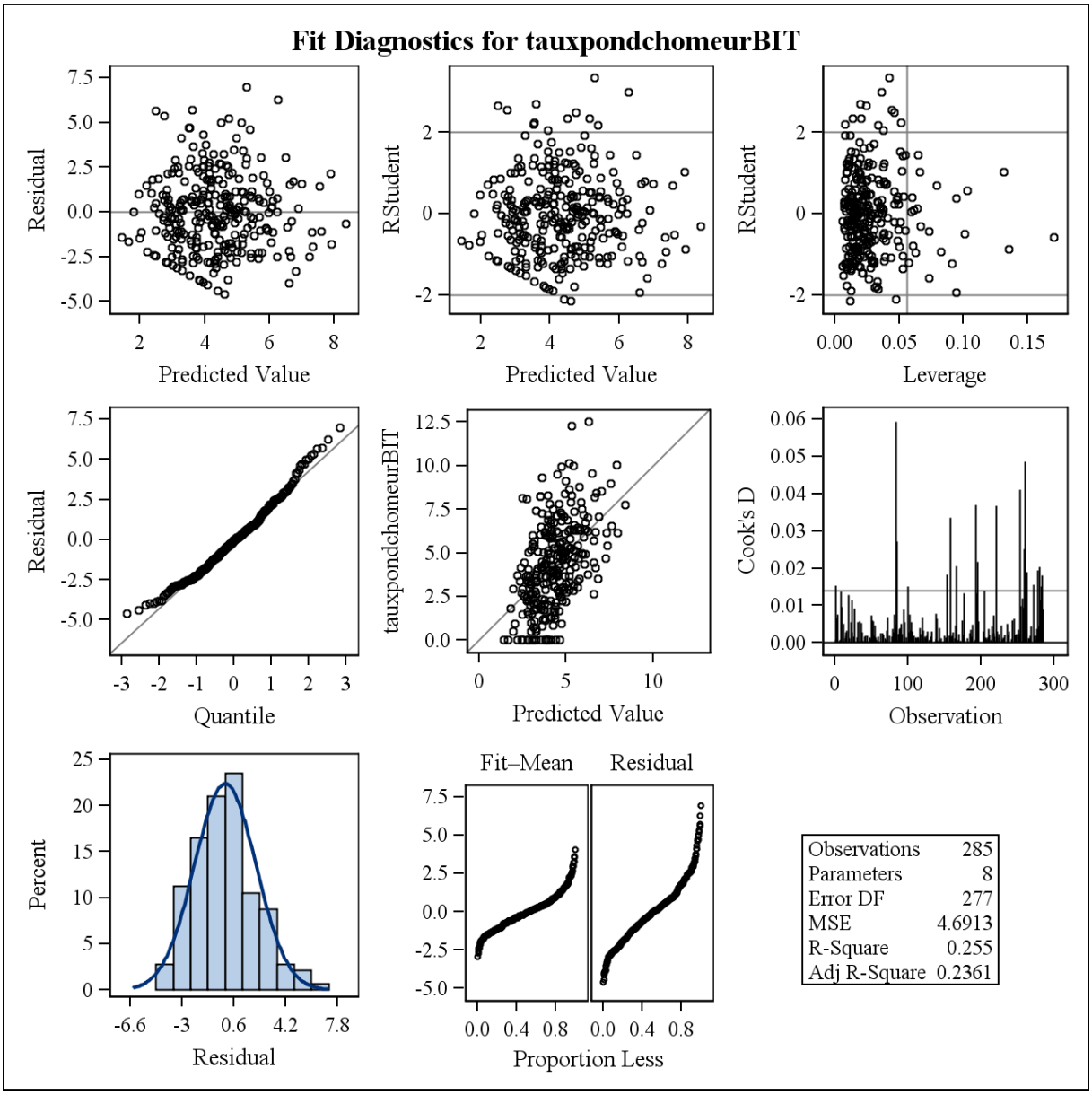
```
SELECTION = stepwise (select=AIC)
```

In fine, avec une dose d'empirisme, on arrive in fine à une sélection de 15 variables potentiellement explicatives.

- Taux de personnes qui recherchent un emploi (déclaration au recensement) → **t_rech_oui**
- Part de la population ayant des bas revenus en 2005 ;
- Taux d'allocataires RMI ;
- Taux de personnes DEFM catégories 1,2,3 et HAR et appartenant à la catégorie des hommes de 15 à 24 ans, diplômés → **t_age15_19HdiplBIT**;
- Taux de personnes DEFM catégories 1,2,3 et HAR et appartenant à la catégorie des hommes de 30 à 49 ans, peu diplômés → **t_age30_49HnondiplBIT**;
- Taux de personnes DEFM toutes catégories et appartenant à la catégorie des hommes de 50 à 64 ans, peu diplômés → **t_age50_64Hnondipl**;
- Taux de personnes ne vivant pas en couple (déclaration au recensement) → **t_couple_2**;
- Taux d'étrangers hors Europe
- Part de la population ayant des bas revenus en 2005 ;
- Part de la fonction "BTP" dans l'emploi total en 2006 ;
- Part de la fonction "Santé-action sociale" dans l'emploi total en 2006 ;
- Part de la fonction "Fabrication" dans l'emploi total en 2006 ;
- Part de la fonction "Gestion" dans l'emploi total en 2006 ;
- Part des agriculteurs dans la population des plus de 15 ans
- Taux de solde des établissements entre 2000 et 2006 = (arrivées - départs) divisé par stock d'établissements au 1/1/2006 → **c02_txsoldetab_0006**;
- Proportion de la population en catégorie Tabard dite "SEMAG02" (Hôtellerie, restauration) → **part_depcom_24**.

Ultime ajustement conduisant au modèle définitif (7 variables + la constante):

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	- 6.12	2.03	- 3.02	0.003
t_rech_oui	0.81	0.26	3.07	0.002
t_couple_2	0.18	0.06	2.99	0.003
t_age15_19HdiplBIT	137.70	53.26	2.59	0.010
t_age50_64Hnondipl	4.08	1.88	2.17	0.031
t_age30_49HnondiplBIT	- 7.65	2.91	- 2.63	0.009
c02_txsoldetab_0006	- 0.11	0.07	- 1.67	0.096
part_depcom_24	0.02	0.01	2.20	0.029



Le graphique central préfigure le "shrinkage" .

8

Estimation localisée par la technique de Fay et Herriot

Théorie (très) condensée

Nombre total de chômeurs BIT estimé en ZE d

$$\hat{Y}_d = \sum_{\substack{i \in s \\ i \in d}} w_i \cdot Y_i$$

Taille totale population du champ en ZE d : $\hat{N}_d = \sum_{\substack{i \in s \\ i \in d}} w_i$

Pour chaque ZE d, on dispose d'une information (pseudo) exacte $\overline{X}_d$ à p dimensions.

$\hat{\overline{Y}}_d = \dfrac{\hat{Y}_d}{\hat{N}_d} = \overline{Y}_d + \varepsilon_d$ où ε_d = **erreur d'échantillonnage**

$\overline{Y}_d = B^t \cdot \overline{X}_d + v_d$ où v_d **variable aléatoire = effet propre au domaine** d**.**

$$\boxed{\hat{\overline{Y}}_d = B^t \cdot \overline{X}_d + v_d + \varepsilon_d}$$

$Ev_d = E\varepsilon_d = 0$, $Var(v_d) = \sigma_v^2$ et $Var(\varepsilon_d) = \psi_d$.

La théorie de la prédiction linéaire sans biais optimale donne l'estimateur suivant de type composite :

$$\hat{\bar{Y}}_d^H = \hat{B}^t \cdot \bar{X}_d + \hat{v}_d$$

$$\boxed{\hat{\bar{Y}}_d^H = \hat{\gamma}_d \cdot \hat{\bar{Y}}_d + (1-\hat{\gamma}_d) \cdot \hat{B}^t \cdot \bar{X}_d}$$

où $\hat{B} = \left(\sum_d \frac{\bar{X}_d \cdot \bar{X}_d^t}{\hat{\sigma}_v^2 + \psi_d} \right)^{-1} \cdot \sum_d \frac{\bar{X}_d \cdot \hat{\bar{Y}}_d}{\hat{\sigma}_v^2 + \psi_d}$

et $\hat{\gamma}_d = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \psi_d}$

Hypothèse de normalité des aléas
$\Rightarrow \hat{\sigma}_v^2$ = maximum de vraisemblance.

$$E\left(\hat{\bar{Y}}_d^H - \bar{Y}_d\right)^2 = \gamma_d \cdot \psi_d + O\left(\frac{1}{m}\right)$$

La stabilisation de la variance est obtenue grâce à l'estimation synthétique $\hat{B}$.

Bien noter que la variable expliquée est par nature "mauvaise" $\Rightarrow$ **modèle à (très) forte erreur sur la variable expliquée : la théorie le gère !**

La délicate question des variances

Le calcul de ψ_d pose un sérieux problème, car :

- tailles d'échantillon égales à 1
  (+ plan complexe)
- surtout : très forte instabilité des estimations

Nécessité de simplifier et de lisser les variances d'échantillonnage

$deff_d$ = moyenne des estimations trimestrielles des *deff* **régionaux**, du T4 de 2005 au T2 de 2008.

$\tilde{p}_d$ = estimateur issu de la méthodologie actuelle de l'Insee (donc déjà « robuste »).

$$\hat{\psi}_d = deff_d \times \frac{\tilde{p}_d (1 - \tilde{p}_d)}{n_d}$$

Utiliser les estimations directes $\hat{\bar{Y}}_d$ dans $\hat{\psi}_d$ = catastrophe = variances nulles, et *in fine* $\hat{\sigma}_v^2 = 0$!

Dans ces conditions, on vérifie :

11,2% < CV estimés < 616,6 %,

avec $Q1 = 33{,}8\%$, médiane = $51{,}1\%$ et $Q3 = 72{,}5\%$

$\rightarrow$ **justifie bien la mise en œuvre de techniques spécifiques "petits domaines" ;**

Résolution informatique

PROC GLIMMIX de SAS

- Ajustement par maximum de vraisemblance restreint (correction réduisant le biais des estimateurs obtenus) ;

- Nombreuses options - mais extrêmement gourmande en mémoire vive :
  pour 338 observations, 15 Giga octets de RAM

Résultats

Les variables auxiliaires $\overline{X}_d$ sont les 7 proportions pré sélectionnées (+ la constante).

*A*justement du modèle restreint aux 287 ZE ayant au moins 50 individus répondants (justification : $\hat{\sigma}_v^2 = 0$ sinon);

On réduit successivement le nombre de regresseurs :

Critère de qualité utilisé	7 variables	5 variables	4 variables	3 variables	2 variables	1 variable
-2. Log vraisemblance restreinte	1280.2	**1268.1**	1273.3	1280.1	1292.9	1297.4
AIC	1282.2	**1270.1**	1275.3	1282.1	1294.9	1299.4
BIC	1285.9	**1273.1**	1279.0	1285.7	1298.6	1303.0
Chi-2 généralisé	**281.8**	284.1	286.9	288.7	292.2	295.6

Modèle définitif à 5 variables (+ la constante)

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	-4.70	1.64	134.7	**-2.86**	0.005
t_rech_oui	0.65	0.25	223.2	**2.64**	0.009
t_couple_2	0.12	0.049	126.3	**2.56**	0.012
t_age15_19 HdiplBIT	111.22	50.77	207.1	**2.19**	0.030
t_age50_64 Hnondipl	3.09	1.85	217.7	**1.67**	0.096
t_age30_49 HnondiplBIT	-3.27	2.76	190.9	**-1.19**	0.237

- deux variables non significatives mais les critères de choix de modèle incitent à les conserver.

- fort coefficient associé à *t_age15_19HdiplBIT* : les valeurs de cette variable sont très faibles.

Maximum de vraisemblance restreint $\hat{\sigma}_v^2$: estimateur relativement stable et σ_v^2 significatif - mais la qualité de $\hat{\sigma}_v^2$ reste médiocre (CV ≈ 30%).

Statistique	7 variables	5 variables	4 variables	3 variables	2 variables	1 variable
Estimateur $\hat{\sigma}_v^2$	1.137	**1.111**	1.090	1.088	1.085	1.197
Ecart-type estimé de $\hat{\sigma}_v^2$	0.33	**0.33**	0.33	0.33	0.33	0.34

L'estimateur de Fay et Herriot a la distribution suivante (rappel : $n_d \geq 50$ répondants) :

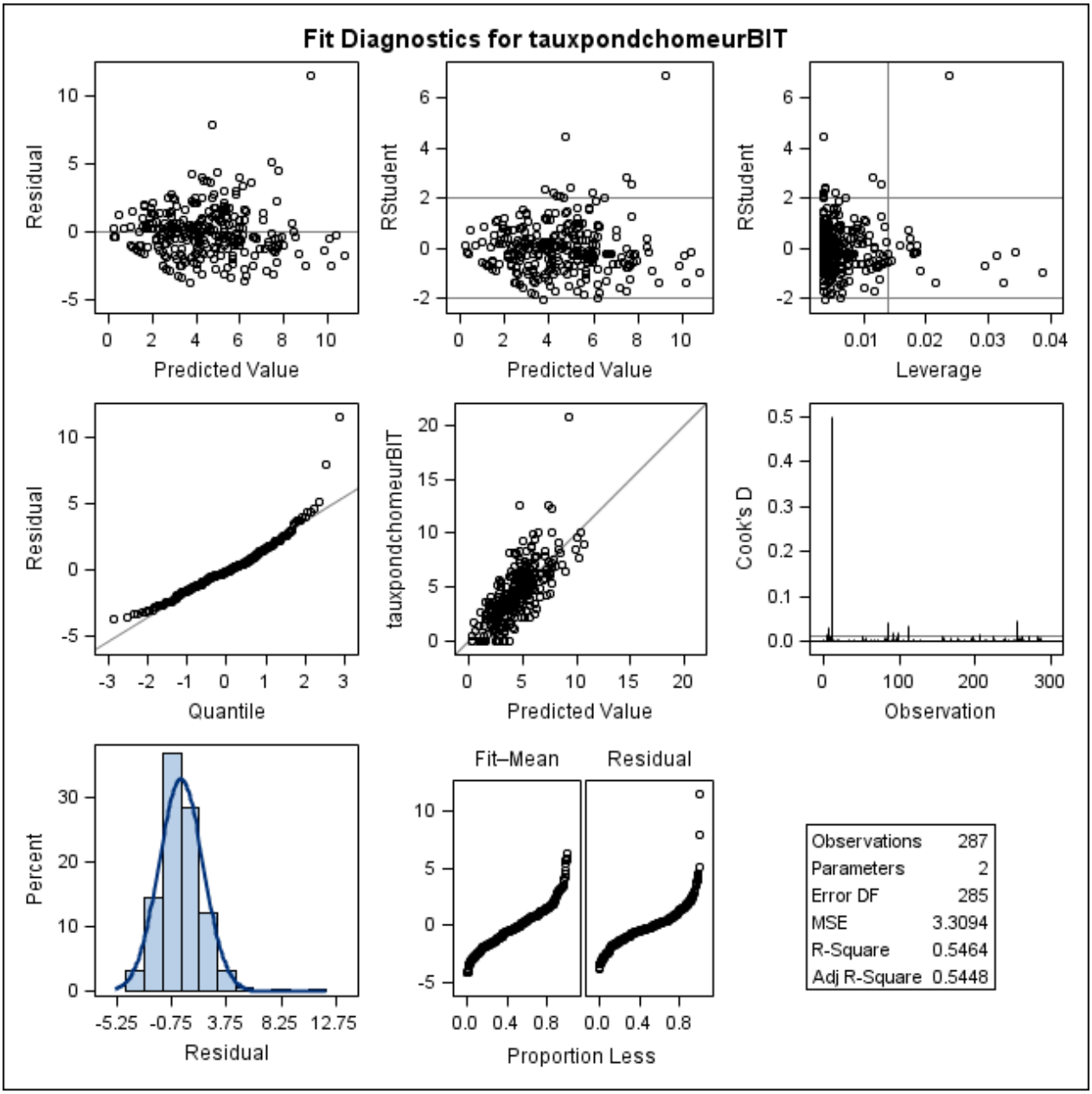
Quantile		Fay & Herriot	Direct $\hat{P}$
100%	Max	8.40	20.7
99%		8.02	14.2
95%		6.51	10.0
90%		6.16	8.2
75%	Q3	5.21	6.2
50%	Median	4.34	4.2
25%	Q1	3.47	2.4
10%		2.85	0.0
5%		2.58	0.0
1%		1.86	0.0
0%	Min	1.82	0.0

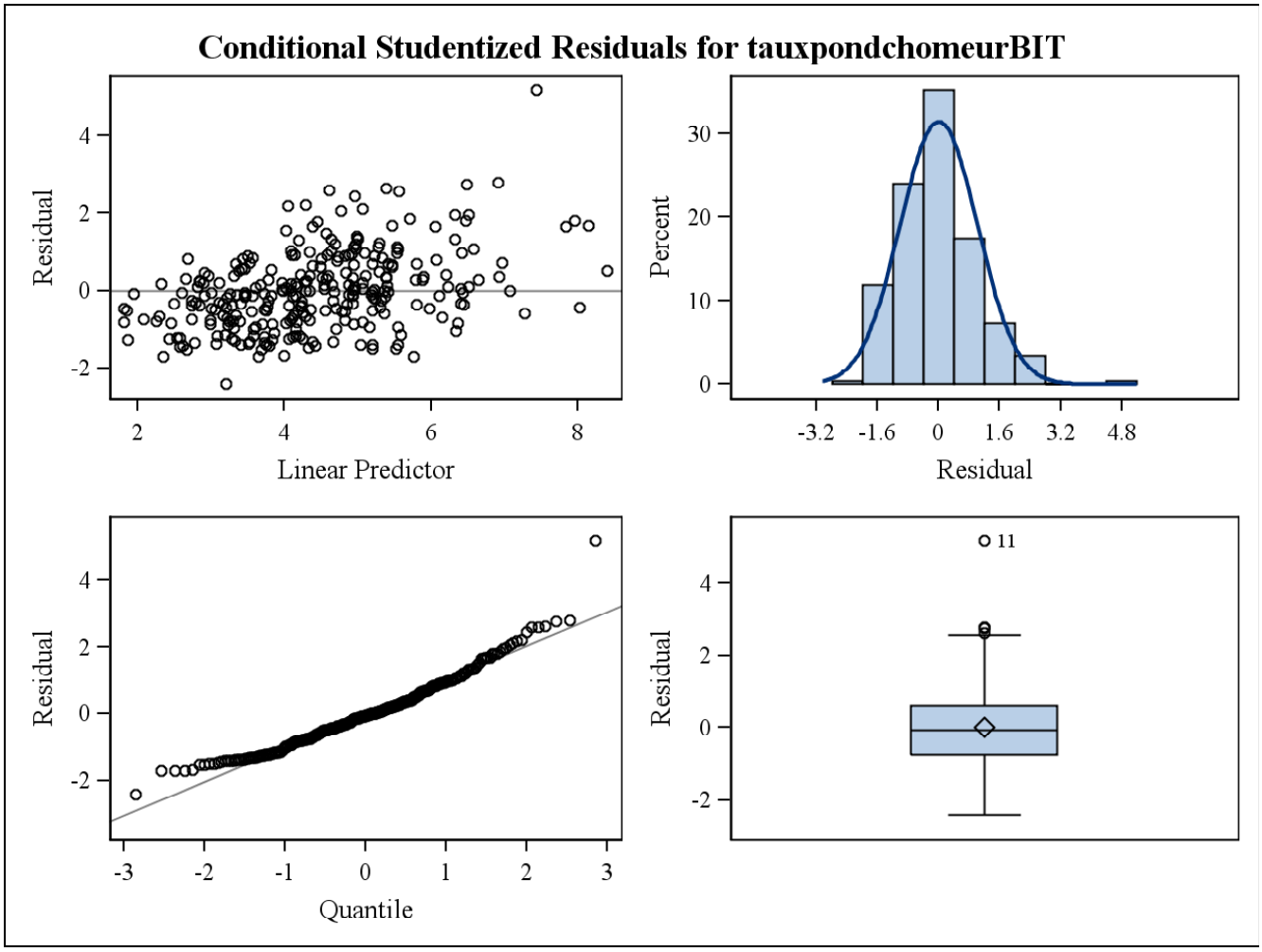
Qualité de l'ajustement :

résidu : $\hat{\bar{Y}}_d - \hat{B}^t \cdot \bar{X}_d$

résidu conditionnel studentizé :

$$\hat{U}_d = \frac{\hat{\bar{Y}}_d - \hat{B}^t \cdot \bar{X}_d - \hat{v}_d}{\sqrt{\hat{\psi}_d}}$$





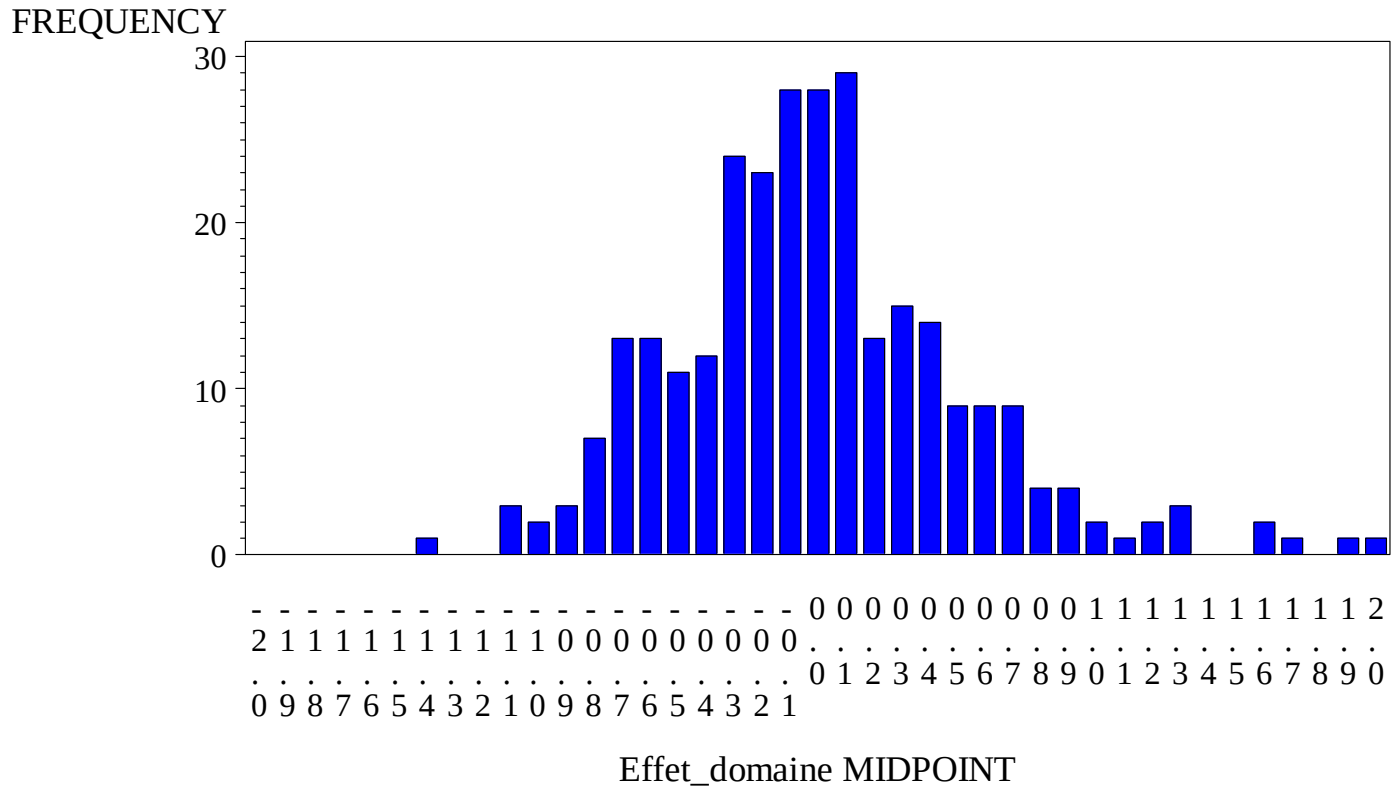
1 ZE très atypique :

$\hat{P}$ = 20,7% (CV estimé de 45,6%)

$\hat{\gamma}_d$ =0,13 $\Rightarrow$ estimation localisée finale égale à 7,4%

(estimation brute Insee = 6,1 %).

Distribution des effets domaines prédits $\hat{v}_d$



Les effets locaux ZE se situent entre -1,5 et 2 points de pourcentage.

Hypothèse de normalité de ces effets aléatoires acceptable.

Distribution du coefficient $\hat{\gamma}_d$ *modèle définitif*

Quantile		Gamma
100%	Max	0.74
99%		0.71
95%		**0.57**
90%		0.49
75%	Q3	0.33
50%	Median	**0.21**
25%	Q1	0.14
10%		0.10
5%		**0.09**
1%		0.07
0%	Min	0.05

Dans un peu plus de 90% des cas, priorité est donnée à l'estimateur synthétique.

Dans la moitié des ZE, l'estimateur direct issu de l'enquête Emploi contribue à plus de 20% dans la valeur de l'estimation composite finale de Fay et Herriot - ce qui est loin d'être négligeable.

La contribution maximale de l'estimateur direct = 74% (ZE de Paris), ce qui est logique ($n_d = 2301$), pour un CV de 11,3% ($\hat{\psi}_d = 0,4$).

La contribution minimale de l'estimation directe = 5% - également logique puisque $n_d = 50$ et CV = 71% ($\hat{\psi}_d = 19,7$).

Quid de l'erreur ?

Soit une ZE "médiane" :

- variance d'échantillonnage $\psi_d \approx$ 4 points de pourcentage ;
- $\hat{\gamma}_d$ = 20% ;

$\Rightarrow$ erreur de l'estimateur FH $\approx$ 0.8

$\Rightarrow$ écart-type $\approx$1 point de pourcentage.

$\Rightarrow$ incertitude $\approx \pm$2 points de pourcentage

$\Rightarrow$ on divise par 2 la largeur des intervalles de confiance par rapport à l'estimateur direct.

(Attention - rappel : erreur mélant 2 aléas !!!)

Aspect désagréable : la sensibilité de $\hat{\sigma}_v^2$ aux estimations $\hat{\psi}_d$ est forte $\Rightarrow$ danger et nécessité d'exclure d'emblée les ZE avec n_d trop petit. La limite de 50 observations est probante.

Autre critère qualité :

On compare l'estimation (pseudo) nationale FH et l'estimation directe issue de l'enquête Emploi, soit

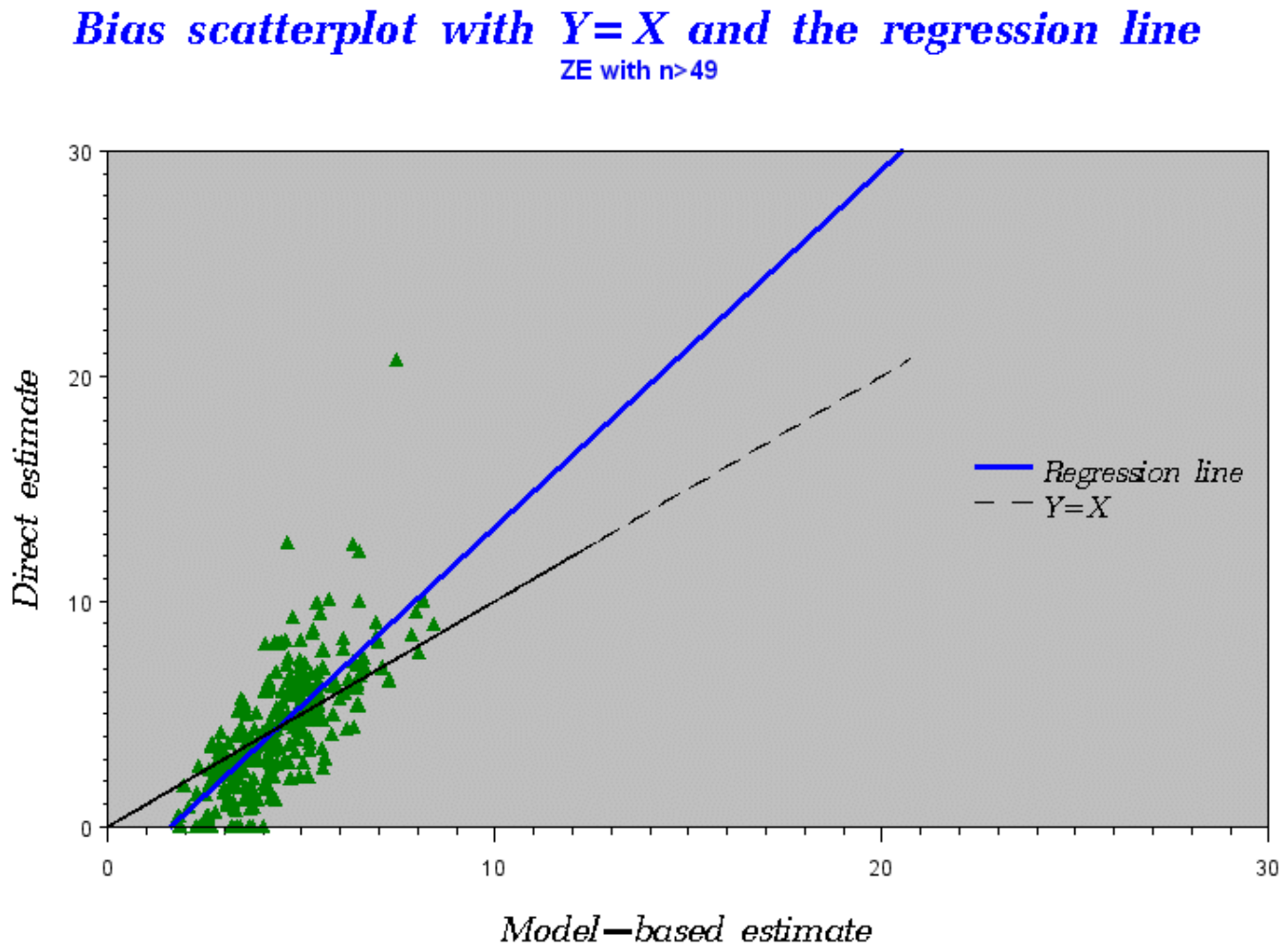
$$\sum_{ZE} N_{ZE} \cdot \frac{\hat{Y}_{ZE}}{\hat{N}_{ZE}}$$

Sur les 287 ZE concernées :

Estimateur (pseudo) national	Estimation totale
Fay et Herriot	2 339 000
Enquête Emploi	**2 340 000**
Méthodologie actuelle Insee	2 305 000

C'est un élément de validation fort de l'approche de Fay et Herriot - qui reste éloignée de l'inférence classique.

Pour juger du biais de l'estimateur *du seul point de vue de l'aléa d'échantillonnage* :
nuage réparti le long de la droite $(Y = X)$ = forte présomption d'absence de biais.



Différence significative entre la droite $Y = X$ et la droite de régression.

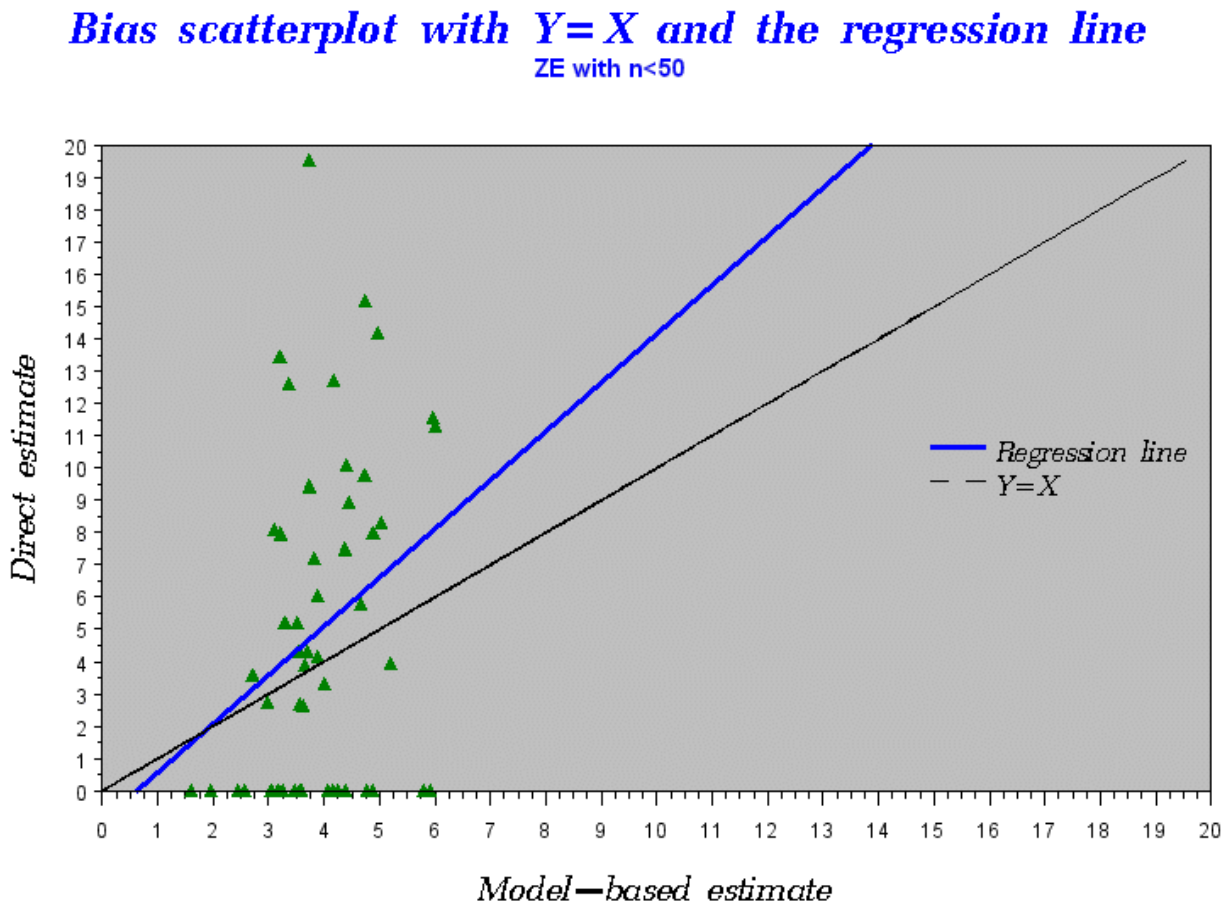
Phénomène - traditionnel - de resserrement de la distribution, dit "Shrinkage", classique dans les procédures utilisant une composante synthétique.

Ultime étape :
prise en compte des 61 ZE non traitées par le modèle : on impose l'estimateur **synthétique**. Pour toute ZE d avec $n_d < 50$, on choisit :

$$\hat{\bar{Y}}_d^{SYN} = \hat{B}^t \cdot \bar{X}_d$$

Cette stratégie concerne en particulier les 10 ZE où $n_d = 0$.

Phénomène de "shrinkage" (évidemment) très marqué



Critère qualité de l'opération d'ensemble :

Sur l'intégralité des 348 ZE de France métropolitaine :

Estimateur (pseudo) national	Estimation totale
Fay et Herriot (ou synthétique)	2 432 000
Enquête Emploi (post-stratifié)	**2 436 000**
Méthodologie actuelle Insee	2 408 000

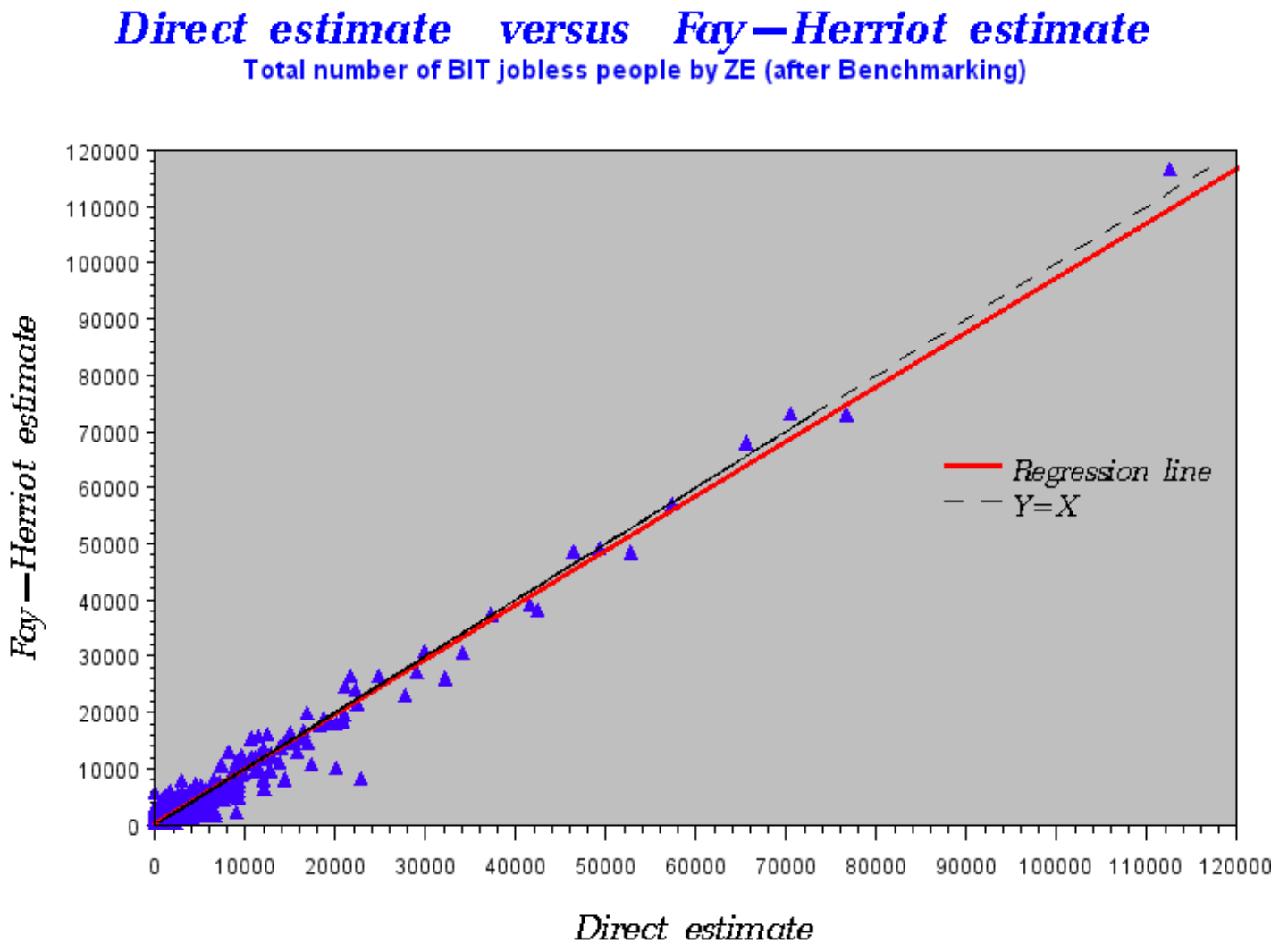
$\Rightarrow$ Proximité très satisfaisante avec l'estimation directe.

On peut vouloir ajouter un "calage" sur l'effectif "officiel" national :

- pour limiter encore la variance;
- pour assurer une cohérence dans la diffusion de l' effectif national;

D'où une technique dite de benchmarking :

$$\hat{\hat{Y}}_{ZE} = 2\,416\,000 \cdot \frac{\hat{Y}_{ZE}}{\sum_{ZE} \hat{Y}_{ZE}}$$



Biais final manifestement très modéré !

20

Estimation localisée utilisant le modèle de Poisson

Théorie (très) condensée

N_d^c = vrai nombre de chômeurs = variable aléatoire entière, strictement positive.

$\hat{N}_d^c$ = estimation pondérée issue de l'enquête Emploi, calculable ssi $n_d \geq 1$.

$$\hat{N}_d^c \rightarrow P(\mu_d)$$

$$g(\mu_d) = X_d^t \cdot \beta + v_d$$

$$E(v_d) = 0 \qquad Var(v_d) = \sigma_v^2$$

v_d Gaussien, indépendance mutuelle.

Modèle linéaire mixte généralisé (GLMM). Si $\sigma_v^2 = 0$: GLM (non mixte).

Maximum de vraisemblance restreint sur **modèle approché** (GLM) :

$$P_d = X_d^t \cdot \beta + v_d + \varepsilon_d$$

avec $E\varepsilon_d = 0$

En théorie : $Var\left(\hat{N}_d^c | v_d\right) = \mu_d$

En réalité : $Var\left(\hat{N}_d^c | v_d\right) = \phi \cdot \mu_d$

Phénomène **d'overdispersion** : nouveau paramètre ϕ

En outre $Var(\varepsilon) = \phi \cdot M(\beta)$

β, σ_v^2 et ϕ estimés par maximum de vraisemblance

$$\hat{\hat{N}}_d^c = \hat{\mu}_d = g^{-1}\left(X_d^t \cdot \hat{\beta} + \hat{v}_d\right)$$

.

La stabilisation de la variance est obtenue grâce à l'estimation synthétique $\hat{\beta}$.

Paramétrage du modèle

On choisit $g(\mu_d) = Log(\mu_d) \Rightarrow \hat{\hat{N}}_d^c = \exp\left(X_d^t \cdot \hat{\beta} + \hat{v}_d\right)$.

On part de la présélection du GLM - méthode discutable mais pratique (absence d'outil informatique) !

Variables explicatives X_d^k : ce sont des taux. On pose

$$Log(\hat{N}_d^c) = Log(poptot2007) + \sum_{k=1}^{p} \beta_k \cdot Log\left(X_d^k\right)$$

Aménagements ultimes :

- "le" code Tabard : supprimé (trop souvent nul);
- *t_age15_19HdiplBIT* et *c02_txsoldetab_0006* : pas de logarithme.

Participation à l'ajustement possible SSI $n_d \geq 1$.

Sinon $\hat{\hat{N}}_d^c = \exp\left(X_d^t \cdot \hat{\beta}\right)$ = estimation synthétique

Doute : ajustement incluant ou non les ZE où $\hat{N}_d^c = 0$?

Proc GLIMMIX de SAS

Il n'y a aucun problème de temps ni de mémoire.

Résultats

Différents scénarii d'exclusion des ZE

	Somme des $\hat{N}_d^c$	$\hat{\phi}$	R^2	$\hat{b}$	Conclusion test
$n_d > 0$	2 431 000	1751	0.91	1.01	égalité
$n_d > 49$	2 463 000	1711	0.91	1.01	égalité
$n_d > 49$ et $\hat{N}_d^c > 0$	2 513 000	1598	0.91	1.03	égalité
$n_d > 49$ et $0 < \hat{N}_d^c < 25000$	2 322 000	1463	0.90	1.23	écart significatif

Exclusion $\uparrow$ $\Leftrightarrow$ Biais $\uparrow$ $\Leftrightarrow$ Variance $\downarrow$

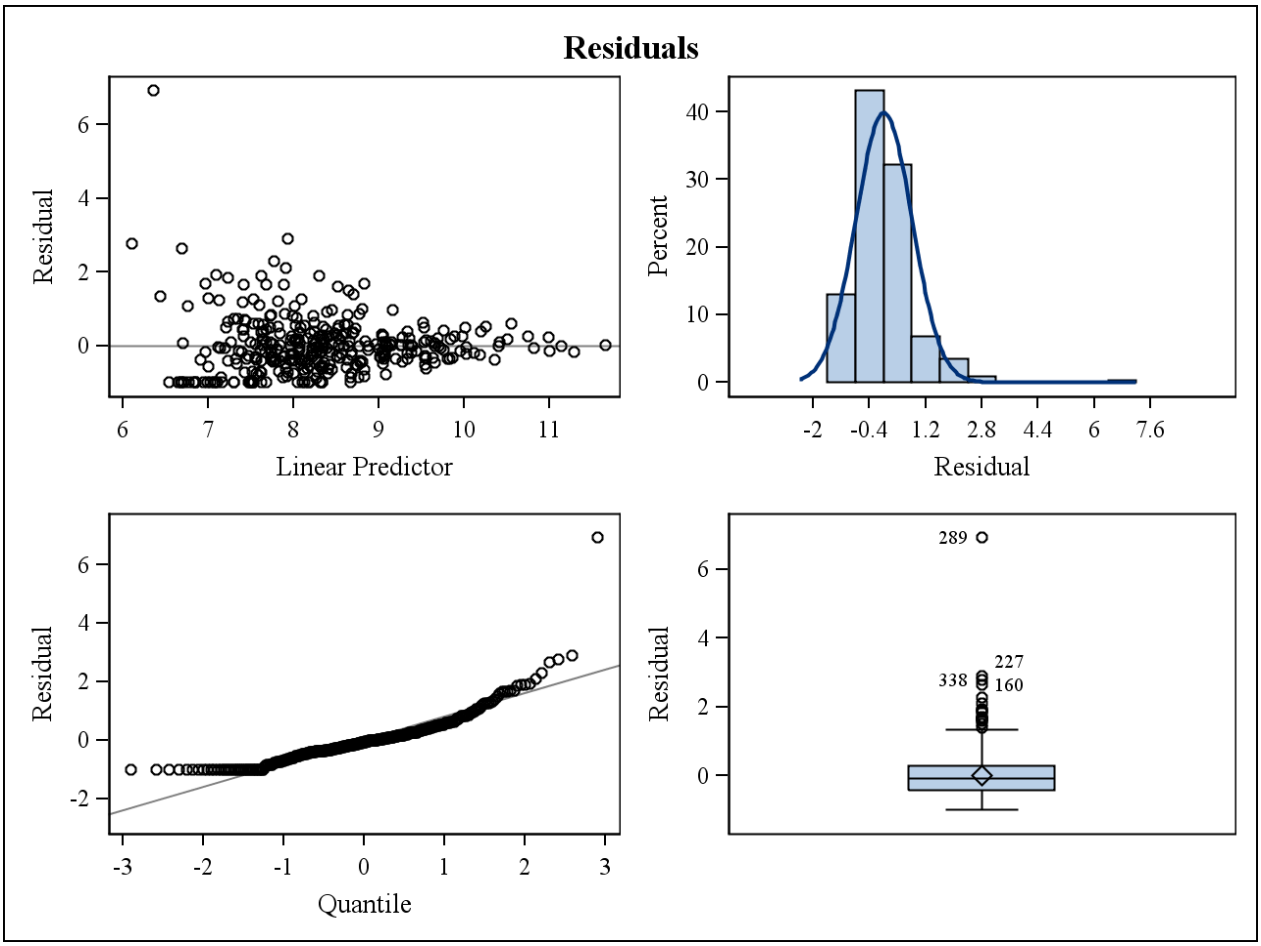
Finalement : sélection $n_d > 0$ = élimination de 10 ZE

a) *Modélisation classique*

Parameter Estimates - definitive model - classic					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	-2.8401	0.07061	325	-40.22	<.0001
log_t_rech_oui	1.2594	0.01202	325	104.79	<.0001
log_t_couple_2	-0.1865	0.01658	325	-11.25	<.0001
log_t_age30_49Hnondi	-0.2398	0.009844	325	-24.37	<.0001
log_t_age50_64Hnondi	0.1705	0.006534	325	26.10	<.0001
log_b05_partbasrev	0.01324	0.007453	325	1.78	0.0765
log_a06_partagri06	0.01781	0.000877	325	20.31	<.0001
log_c08_partslbtp06	-0.1746	0.005822	325	-29.99	<.0001
log_c08_partslsante0	0.01255	0.005127	325	2.45	0.0149
log_c08_partslfabri0	-0.2042	0.002581	325	-79.12	<.0001
log_c08_partslgest06	-0.3828	0.005696	325	-67.20	<.0001
c02_txsoldetab_0006	-0.02422	0.000396	325	-61.16	<.0001
t_age15_19HdiplBIT	-3.3256	0.2740	325	-12.14	<.0001

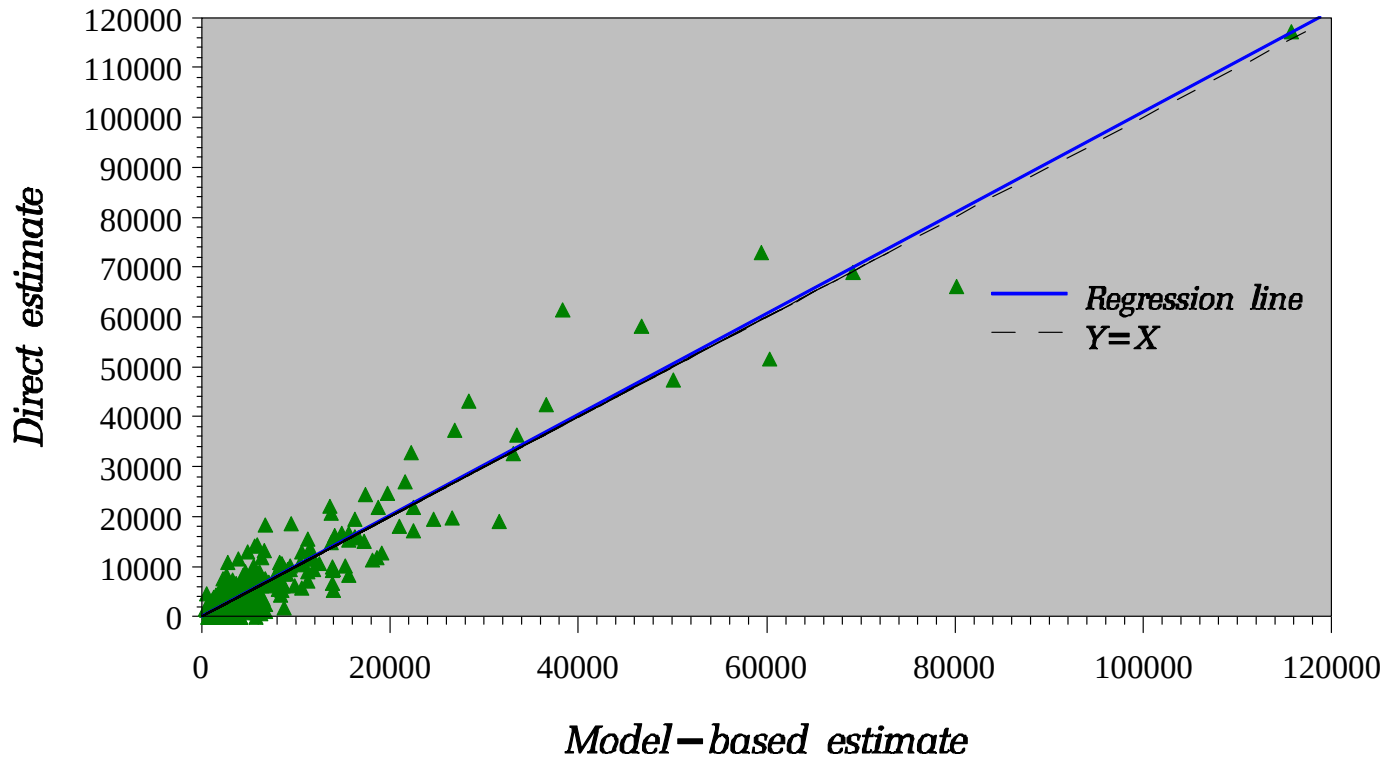
⇒ **Liste très "riche" - contrairement au modèle linéaire FH !!!**

Nota : on obtient les mêmes $\hat{\beta}$ avec et sans overdispersion. Justification : variance-covariance du pseudo modèle = $\phi \cdot \Delta$ où Δ matrice diagonale (propriété fausse si GLMM).



Bias scatterplot with Y=X and the regression line

Poisson GLM



Pas de shrinkage, pas de biais !

Estimation nationale : 2 431 000 chômeurs
Rappel EEC : 2 416 000 chômeurs

.

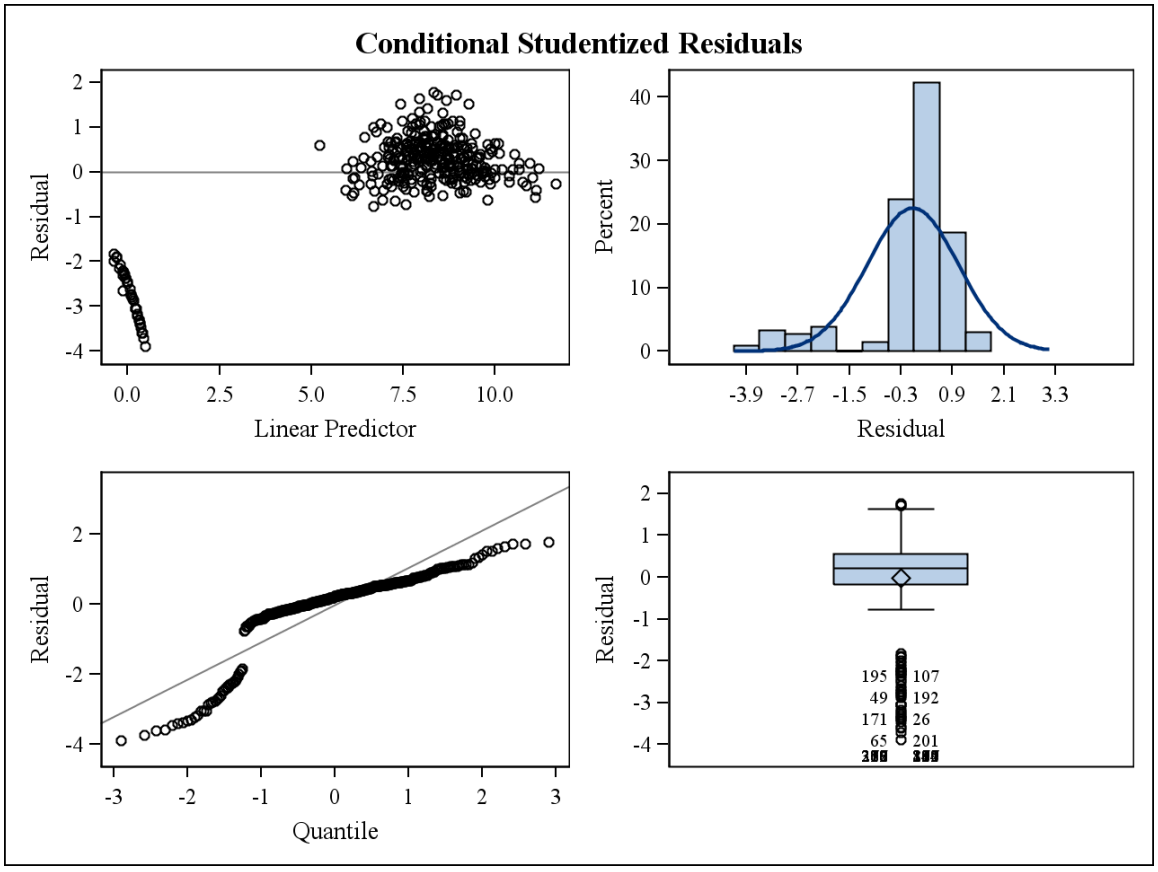
b) Modélisation mixte

L'ajustement SANS overdispersion : $\hat{\sigma}_v^2 = 4{,}96$

(pour v_d compris entre 3 et 4 $\Leftrightarrow$ effets locaux sur $\hat{N}_d^c \approx e^3$ ou e^4).

Estimation nationale = 4 124 000 chômeurs BIT !!!

Résidus d'allure inquiétante :



Conclusion : c'est inacceptable.

AVEC un paramètre d'overdispersion : $\hat{\sigma}_v^2 = 0{,}039$ (écart-type 0,024) et $\hat{\phi} = 1434$ (écart-type $169{,}9$)

$\Rightarrow$ $\hat{\sigma}_v^2$ **non significativement différent de zéro**.

$\hat{\phi} >> \hat{\sigma}_v^2$: la variance des $\hat{N}_d^c$ est attribuée à l'overdispersion ("R-effect") et non à l'effet local ("G-effect").

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	-2.25	3.6216	207.4	-0.62	0.5339
log_t_rech_oui	1.25	0.5823	249.9	2.14	0.0331
log_t_couple_2	-0.24	0.8199	166	-0.29	0.7723
log_t_age30_49Hnondi	-0.29	0.4419	276	-0.65	0.5192
log_t_age50_64Hnondi	0.27	0.2921	187.9	0.94	0.3504
log_b05_partbasrev	0.032	0.3976	254.7	0.08	0.9359
log_t_allocRMI	-0.015	0.2435	292.3	-0.06	0.9518
log_t_natc_afri	0.0023	0.05515	299.9	0.04	0.9664
log_a06_partagri06	0.0008	0.04370	122.8	0.02	0.9850
log_c08_partslbtp06	-0.184	0.2676	221	-0.69	0.4937
log_c08_partslsante0	-0.005	0.2263	301	-0.02	0.9815
log_c08_partslfabri0	-0.22	0.1128	283.2	-1.93	0.0552
log_c08_partslgest06	-0.46	0.2558	258.7	-1.81	0.0711
c02_txsoldetab_0006	-0.024	0.01693	295.1	-1.43	0.1544
t_age15_19HdiplBIT	-3.88	12.0684	209	-0.32	0.7480

- L'introduction des $\hat{v}_d$ impacte peu les $\hat{\beta}$ mais les t-values s'effondrent.

- Estimation nationale = 2 449 000 chômeurs BIT ;

- Résidus d'apparence très corrects;

- Absence de biais d'échantillonnage (après test).

Réduction de la population pour l'ajustement du modèle

Avec 14 variables explicatives, si on restreint ($n_d > 49$ et $0 < \hat{N}_d^c < 25000$) alors σ_v^2 devient significatif et $\hat{\phi}$ s'effondre ($\hat{\phi} = 650$; écart-type 228).

MAIS *log_t_rech_oui* perd très nettement sa significativité (p-value de 0.20), et le test habituel conclut à un biais des estimateurs "petits domaines" ...

Processus de réduction du nombre de variables

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	-4.4401	0.5224	176.5	-8.50	<.0001
log_t_rech_oui	1.1907	0.1506	144.9	7.91	<.0001
log_c08_partslfabri0	-0.1219	0.08028	240.9	-1.52	0.1302
log_c08_partslgest06	-0.2258	0.1398	85.6	-1.62	0.1098

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	-5.2258	0.2740	117.5	-19.07	<.0001
log_t_rech_oui	1.1643	0.1433	95.45	8.12	<.0001

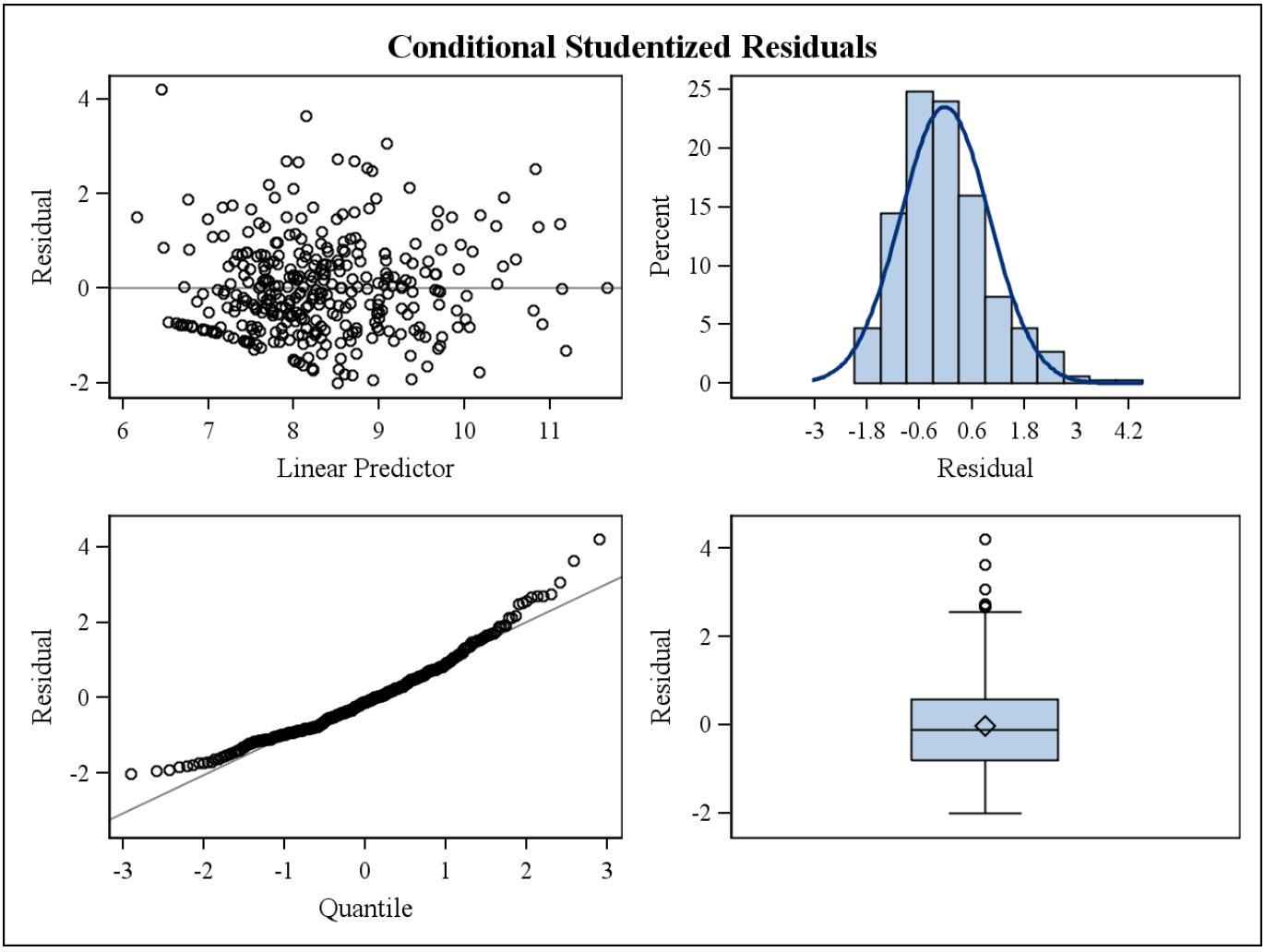
Que choisir ?

	14 variables	3 variables	1 variable
-2 Res Log Pseudo-Likelihood	666.52	669.52	684.38
Generalized Chi-Square	463178.9	516992.4	568416.1
Estimation nationale	2 449 000	2 447 000	2 432 000
Biais selon test ?	NON	NON	NON
Overdispersion	1434	1548	1692

Conclusion 1 : résultats un peu déroutants …

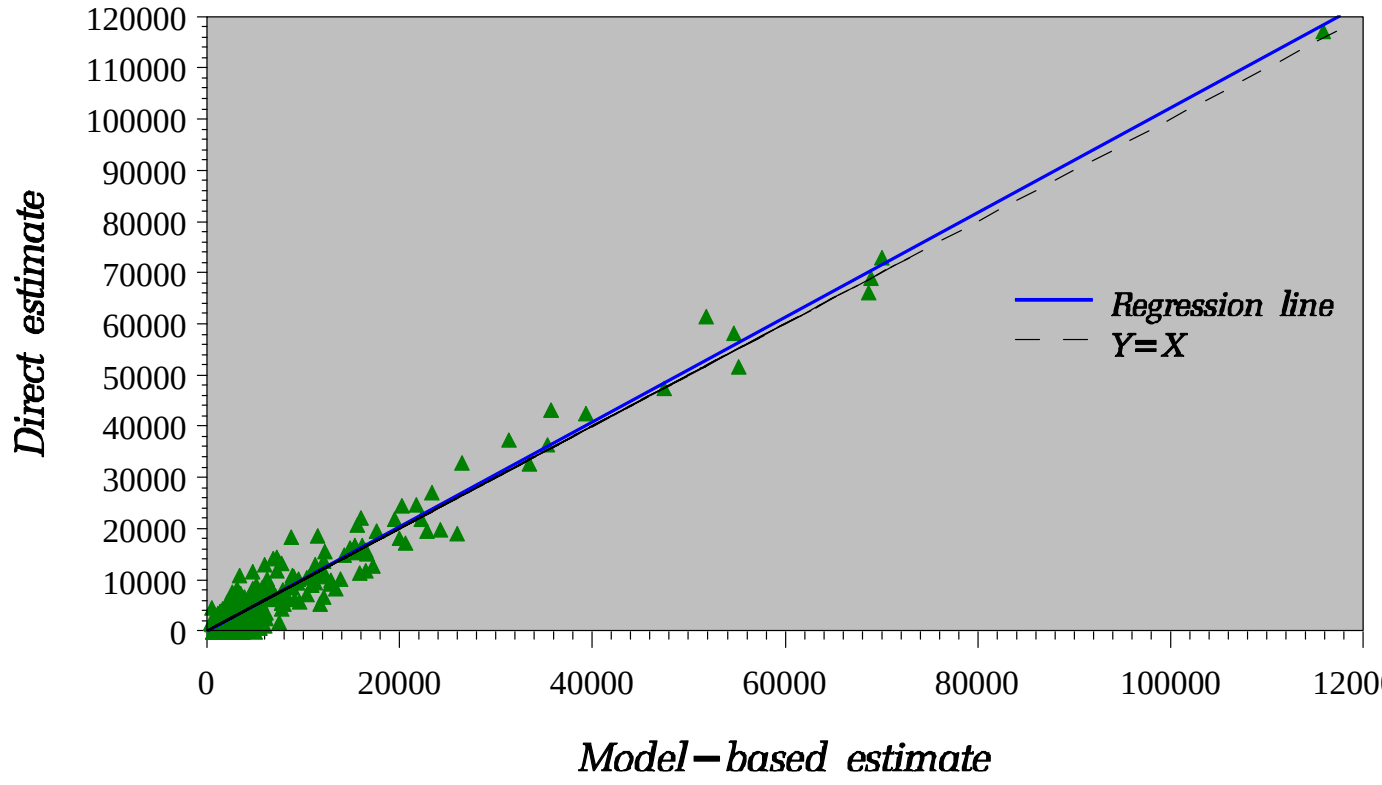
Conclusion 2 : on reste proche de l'estimation nationale tout en utilisant une méthode très éloignée de l'inférence classique.

Finalement, si on opte pour le modèle complet :



Bias scatterplot with Y=X and the regression line

Poisson GLMM



Remarque : l'utilisation du modèle simplifié donne des graphiques semblables.

Appréciation de la qualité par ZEAT, avant benchmarking
(*nota* : estimateur Insee déjà calé …) :

ERREURS par ZEAT
Comparaison avec l'estimation directe EE
- Modèle complet (14 variables) -

ZEAT	Ecart Poisson classique	Ecart Poisson mixte	Ecart Insee
1	0,1 %	2,8 %	10,6 %
2	3,2 %	3,6 %	2,4 %
3	7,1 %	1,7 %	8,2 %
4	11,0 %	9,5 %	10,6 %
5	1,0 %	1,8 %	8,1 %
7	2,2 %	1,6 %	0,4 %
8	3,0 %	2,0 %	3,5 %
9	1,9 %	1,6 %	1,2 %
Total	**Σ = 29.5**	**Σ = 24,6**	**Σ = 45,0**

<u>Estimations nationales</u> :

Direct :	**2 416 000**
Classique :	2 431 000
Mixte :	2 449 000
Insee :	2 408 000

28

Estimation localisée par la technique de l'estimation logistique

Théorie (très) condensée

Il s'agit ici d'un modèle **individuel**.

$Y_{d,i}$ = 1 si l'individu d,i est chômeur BIT et $Y_{d,i}$ = 0 sinon.

$$Y_{d,i} \rightarrow B(1, P_{d,i})$$

$P_{d,i}$ variable aléatoire liée aux variables auxiliaires $X_{d,i}$

$$g(P_{d,i}) = \beta^t \cdot X_{d,i} + \nu_d$$

On estime (cas standard) ou on prédit (cas mixte) $P_{d,i}$ par

$$\hat{P}_{d,i} = g^{-1}\left(\hat{\beta}^t \cdot X_{d,i} + \hat{\nu}_d\right)$$

Puis *in fine* : $\hat{N}_d^c = \sum_{i \in s_d} w_{d,i} \cdot \hat{Y}_{d,i} = \sum_{i \in s_d} w_{d,i} \cdot \hat{P}_{d,i}$

s_d = individus recensés

Fonction de lien utilisée $g(x) = Log \frac{x}{1-x}$

$$\hat{P}_{d,i} = \frac{e^{\left(\hat{\beta}^t \cdot X_{d,i} + \hat{v}_d\right)}}{1 + e^{\left(\hat{\beta}^t \cdot X_{d,i} + \hat{v}_d\right)}}$$

Choix des variables auxiliaires

Le choix est fort restreint par 2 conditions :

1) Une information $X_{d,i}$ présente dans l'EE (ajustement du modèle);

2) Calcul des $\hat{P}_{d,i} \Rightarrow$ obtenir $X_{d,i}$ sur une base exhaustive (u extrapolable) ;

Seul le recensement le permet !

(sauf si $X_{d,i}$ est l'indicatrice d'appartenance à la base - mais c'est pauvre et on en revient aux modèles agrégés …)

On est donc très contraint, car le recensement est limité en variables potentiellement explicatives (présentes dans l'EE).

En pratique, on doit hélas se limiter à :

- La situation déclarée pour le mois en cours (déclaration spontanée de l'état de chômage ou non)

- La recherche ou non d'un emploi

- Le sexe
- L'âge
- La nationalité
- Le diplôme le plus élevé
- L'indicateur de vie en couple ou non
- Le statut matrimonial
- Le statut d'occupation du logement

… sauf à définir des variables au niveau commune (ou ZE).

A) *Modélisation standard non pondérée :*

Estimation nationale : 3 062 000 chômeurs BIT

⇒ **pas acceptable**

Justification probable : " *Recherchez vous un emploi ?* " est **hétérogène** entre les sources (EE ≠ RP)

Effectif de personnes recherchant un emploi :

- EE : 2 512 000
- Recensement 2007 : 3 296 000 (+ 31,2 %)

EE 2007 T1	Recherche un emploi	Ne recherche pas d'emploi	*% marginal*
Chômeur BIT	3 254	**382**	4,97%
N'est pas chômeur BIT	**534**	68 983	95,03%
% marginal	5,18 %	94,82 %	73 153

La variable sexe n'est pas significative.

Effect	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-6.1625	0.3845	-16.03	<.0001
stoc_loc	-0.7554	0.03843	-19.66	<.0001
stoc_loc	0	.	.	.
AGE	3.5158	0.3888	9.04	<.0001
AGE	5.3603	0.3828	14.00	<.0001
AGE	5.2693	0.3829	13.76	<.0001
AGE	5.0095	0.3798	13.19	<.0001
AGE	4.5242	0.3804	11.89	<.0001
AGE	0	.	.	.
dipl_binaire	-0.5705	0.03813	-14.96	<.0001
dipl_binaire	0	.	.	.
nat	-0.5125	0.05949	-8.62	<.0001
nat	0.1269	0.09095	1.40	0.1628
nat	0	.	.	.
matr_celib	-0.5323	0.04400	-12.10	<.0001
matr_celib	0	.	.	.

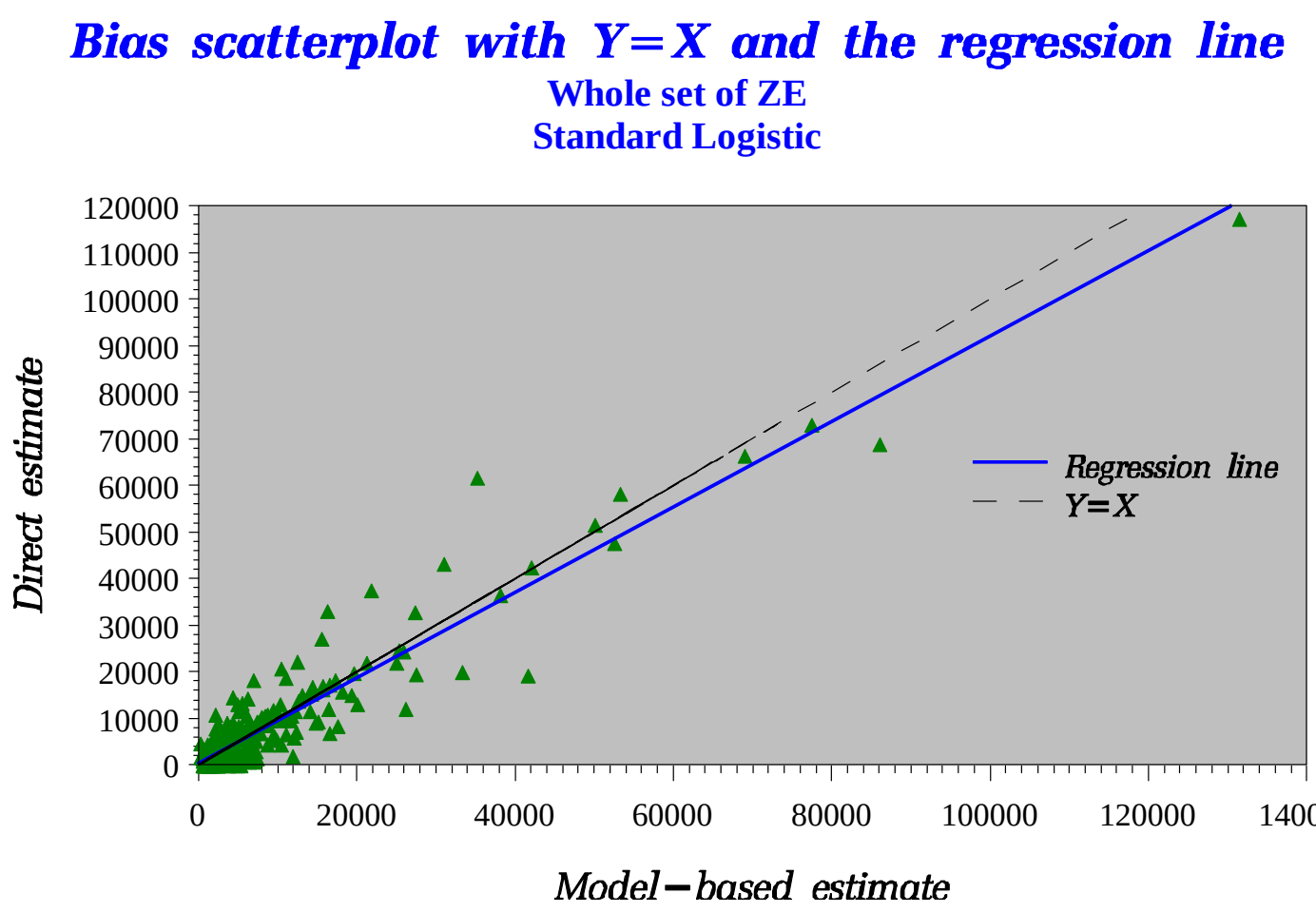
Critères de qualité

Fit Statistics	Modèle initial	Modèle réduit
-2 Res Log Pseudo-Likelihood	7 518	25 136
Critère d'Akaike (AIC)	7 556	25 158
Generalized Chi-Square	52 084	75 904

Modèle réduit = sans la variable de recherche d'emploi

Estimation nationale : 2 504 000 chômeurs en France métropolitaine

Présence d'un biais (test) + shrinkage "à l'envers" !



B) Modélisation mixte non pondérée :

Souci de convergence (réglé, mais troublant)
($\Delta\, objectif\ :\ 10^{-8} \rightarrow 10^{-5}$)

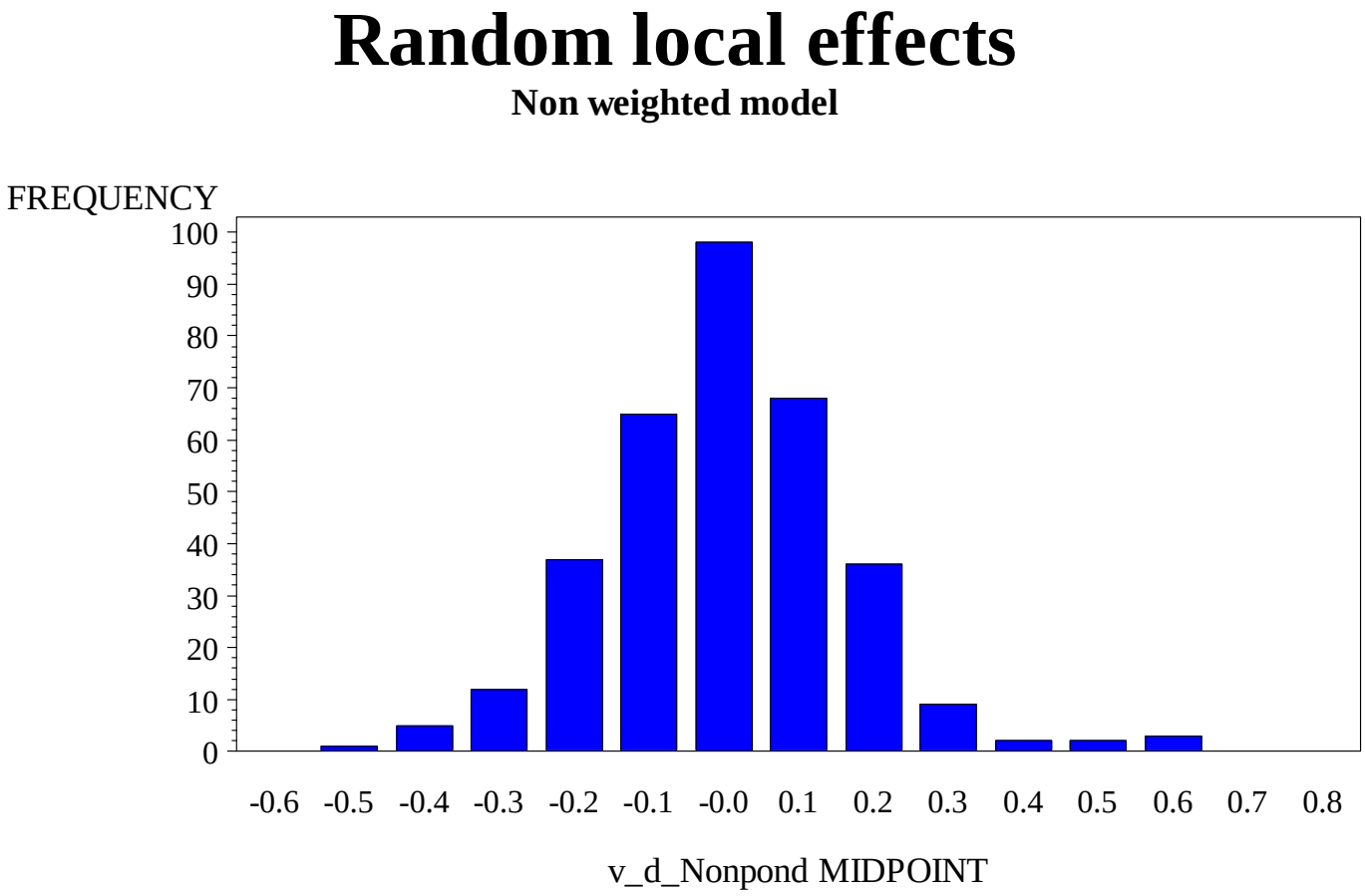
Estimation nationale : 2 484 000 chômeurs

$\hat{\sigma}_v^2$ = 0.082 (écart-type = 0.017)
$\rightarrow$ il y a des effets ZE.

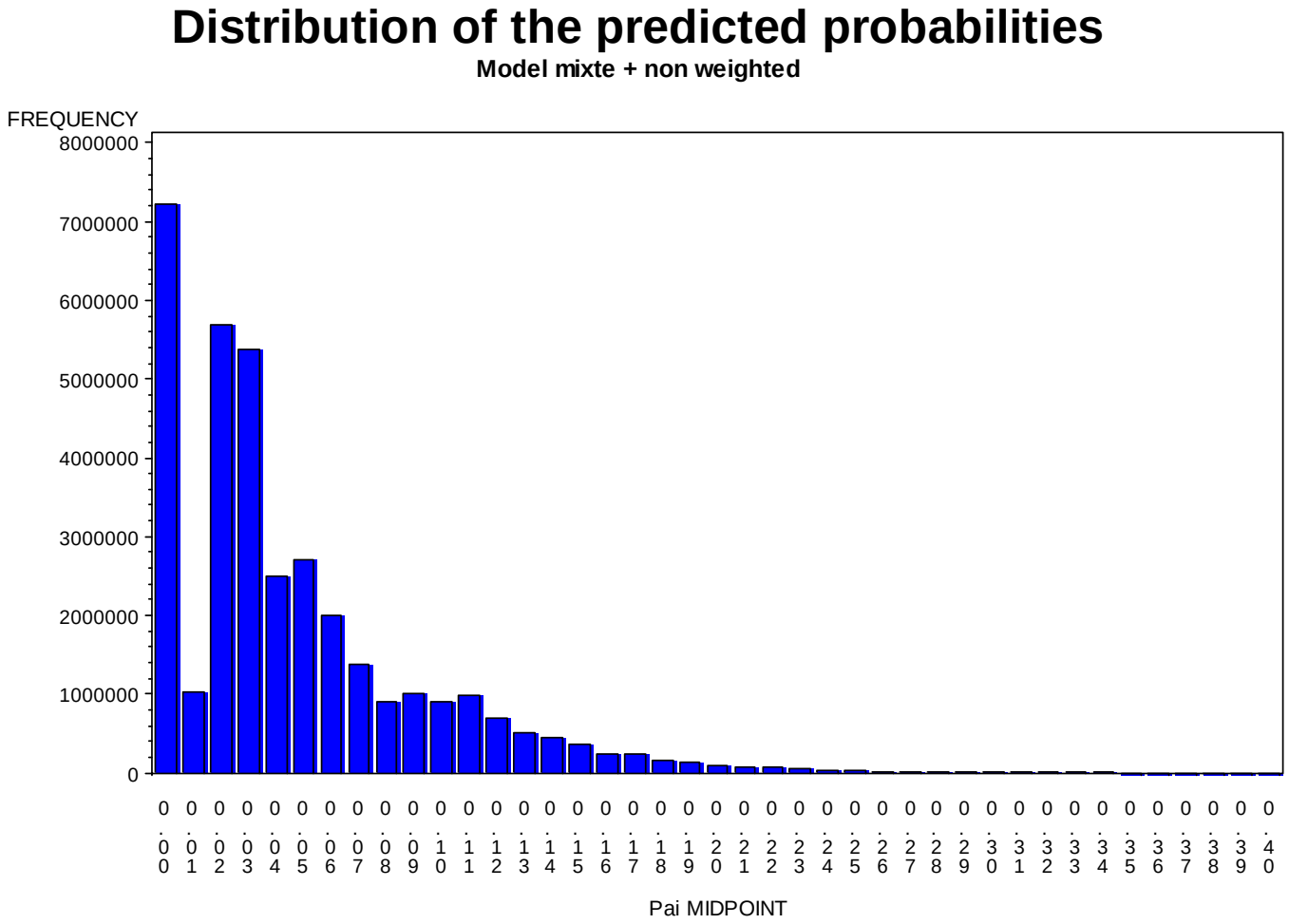
Toutes les variables retenues sont significatives (mais *n* très grand …) :

Effect	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-6.1460	0.3855	-15.94	<.0001
stoc_loc	-0.7615	0.03919	-19.43	<.0001
stoc_loc	0	.	.	.
AGE	3.4938	0.3889	8.98	<.0001
AGE	5.3303	0.3829	13.92	<.0001
AGE	5.2553	0.3830	13.72	<.0001
AGE	4.9974	0.3798	13.16	<.0001
AGE	4.5184	0.3805	11.88	<.0001
AGE	0	.	.	.
dipl_binair	-0.5320	0.03863	-13.77	<.0001
dipl_binair	0	.	.	.
nat	-0.5434	0.06131	-8.86	<.0001
nat	0.1085	0.09149	1.19	0.2356
nat	0	.	.	.
matr_celib	-0.5539	0.04428	-12.51	<.0001
matr_celib	0	.	.	.

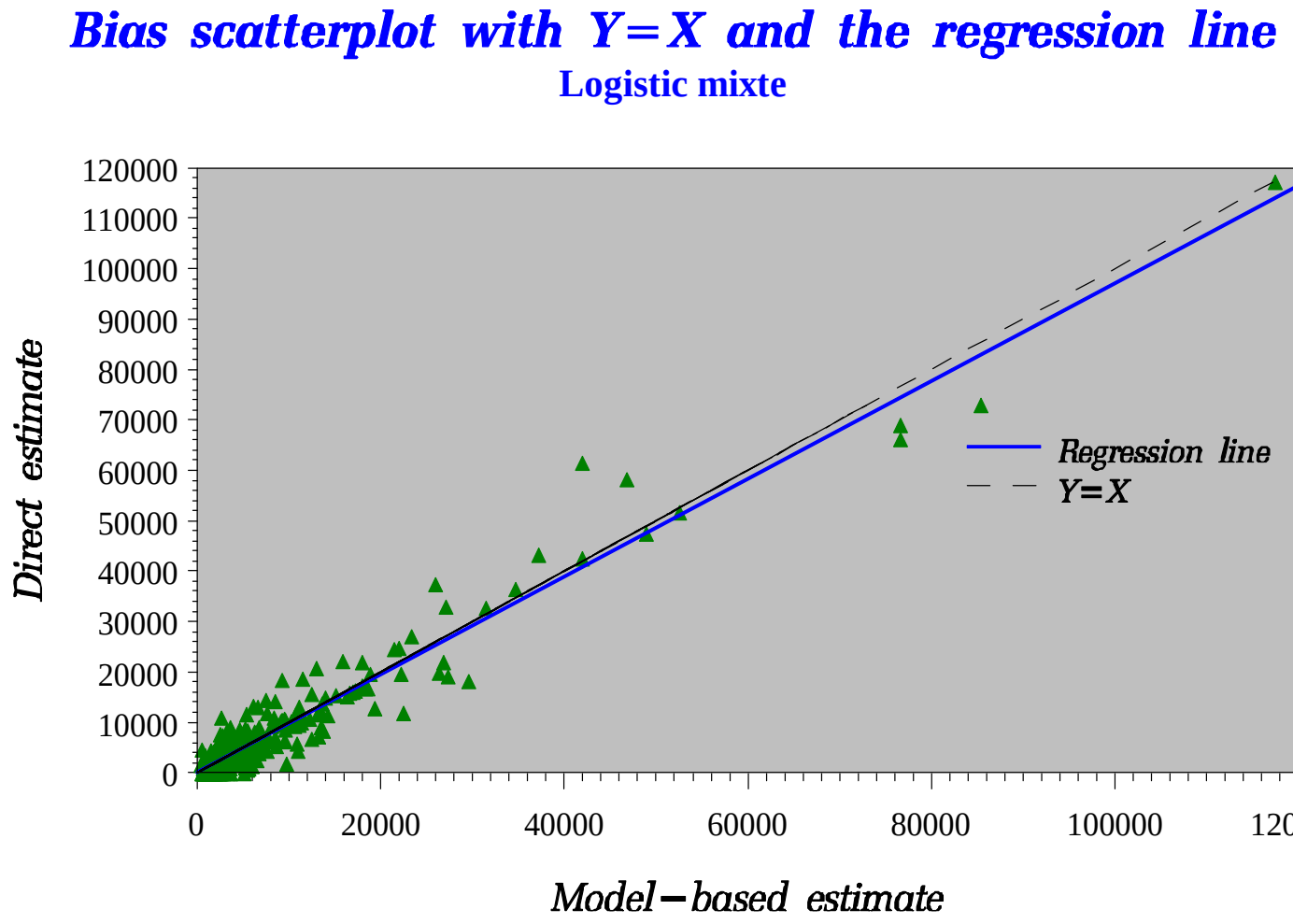
Distribution des effets locaux $\hat{v}_d$:



Distribution des probabilités individuelles $\hat{P}_{d,i}$ prédites par le modèle sur l'ensemble des individus recensés



Absence de biais (test) :



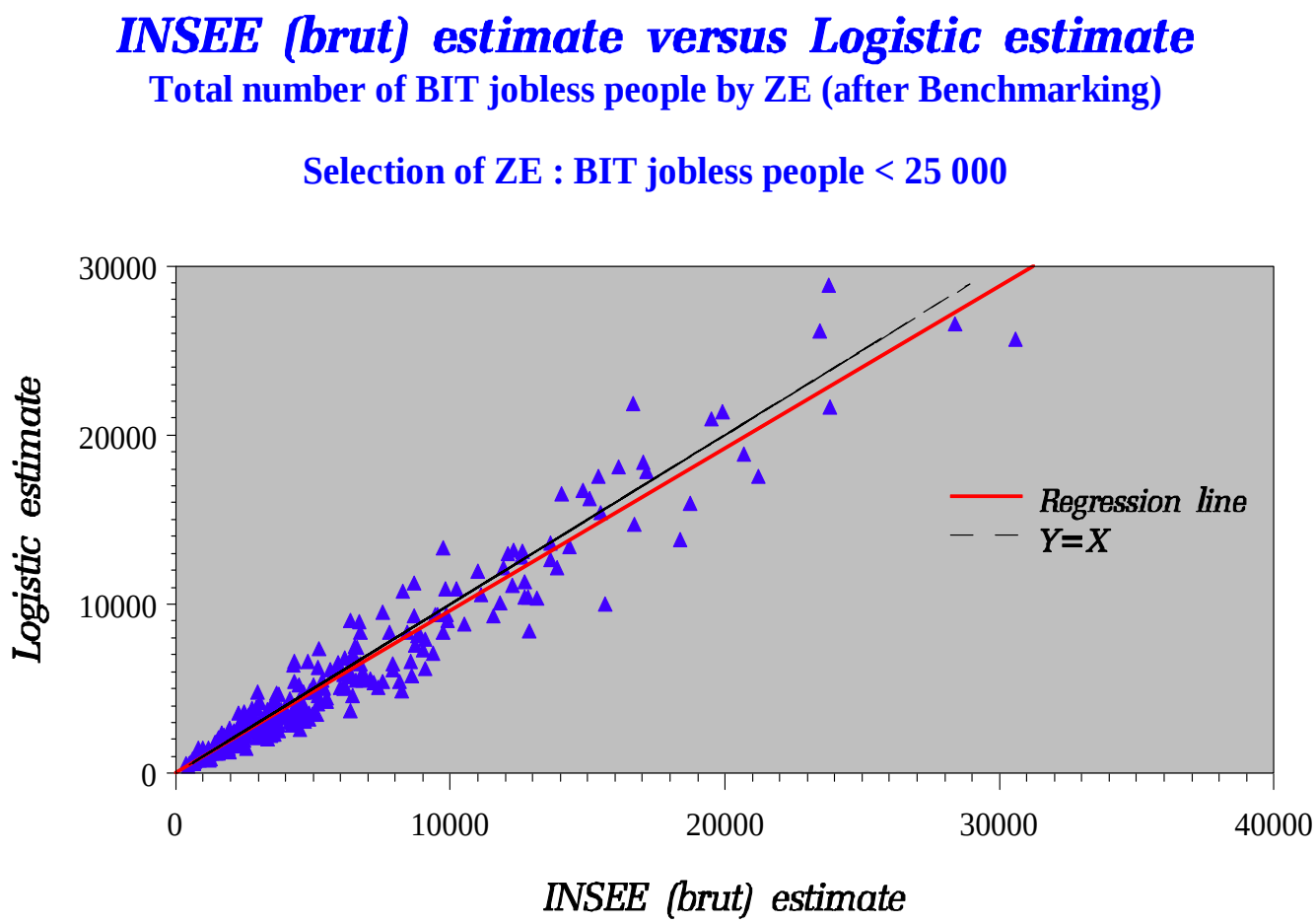
ERREURS par ZEAT
Comparaison avec l'estimation directe EE

ZEAT	Classique	Mixte	Insee
1	+12.8 %	+ 4.3 %	- 10.6 %
2	+ 3.4 %	+ 2.5 %	+ 2.4 %
3	- 26.1 %	- 9.0 %	- 8.2 %
4	+ 16.1 %	+ 9.2 %	+ 10.6 %
5	+ 12.5 %	+ 8.7 %	+ 8.1 %
7	+ 0.9 %	- 3.3 %	+ 0.4 %
8	+ 21.1 %	+ 17.2 %	+ 3.5 %
9	- 13.5 %	- 4.7 %	+ 1.2 %
TOTAL	**Sigma = 106,4**	**Sigma = 58,9**	**Sigma = 45,0**

Estimations nationales (avant benchmarking) :

Direct :	2 416 000
Classique :	2 504 000
Mixte :	2 484 000
Insee :	2 408 000

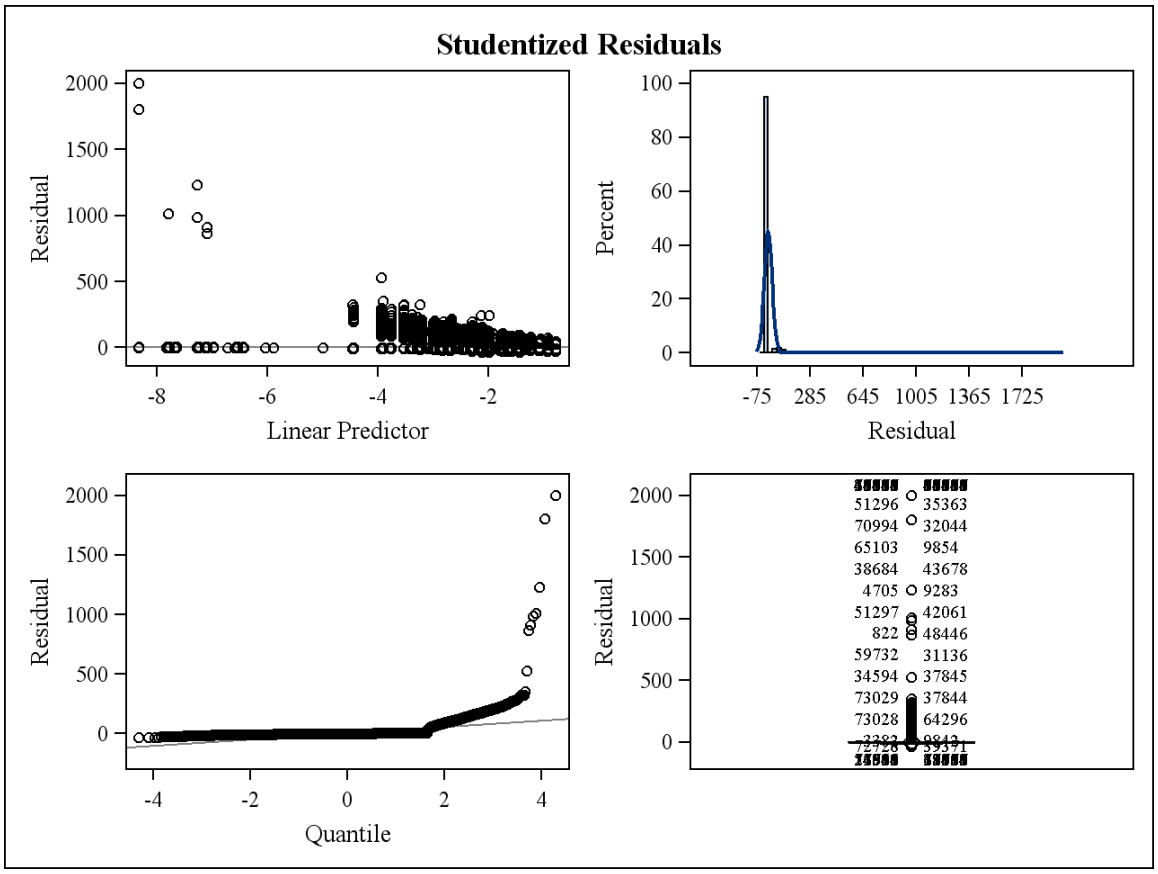
Comparaison estimateur INSEE / estimateur logistique mixte après benchmarking



C) Modélisation standard pondérée

Maximum de vraisemblance pondéré

Effect	Estimateur pondéré	Rappel estim^{t} NON pondéré	Standard Error	Pr > |t|
Intercept	-6.02	-6.16	0.01408	<.0001
stoc_loc	-0.69	-0.76	0.00147	<.0001
stoc_loc	0	0	.	.
AGE	3.33	3.52	0.01427	<.0001
AGE	5.10	5.36	0.01401	<.0001
AGE	4.97	5.27	0.01402	<.0001
AGE	4.80	5.01	0.01389	<.0001
AGE	4.39	4.52	0.01391	<.0001
AGE	0	0	.	.
dipl_binair	-0.54	-0.57	0.00148	<.0001
dipl_binair	0	0	.	.
nat	-0.52	-0.51	0.00231	<.0001
nat	0.14	0.13	0.00353	<.0001
nat	0	0	.	.
matr_celib	-0.55	-0.53	0.00168	<.0001
matr_celib	0	0	.	.

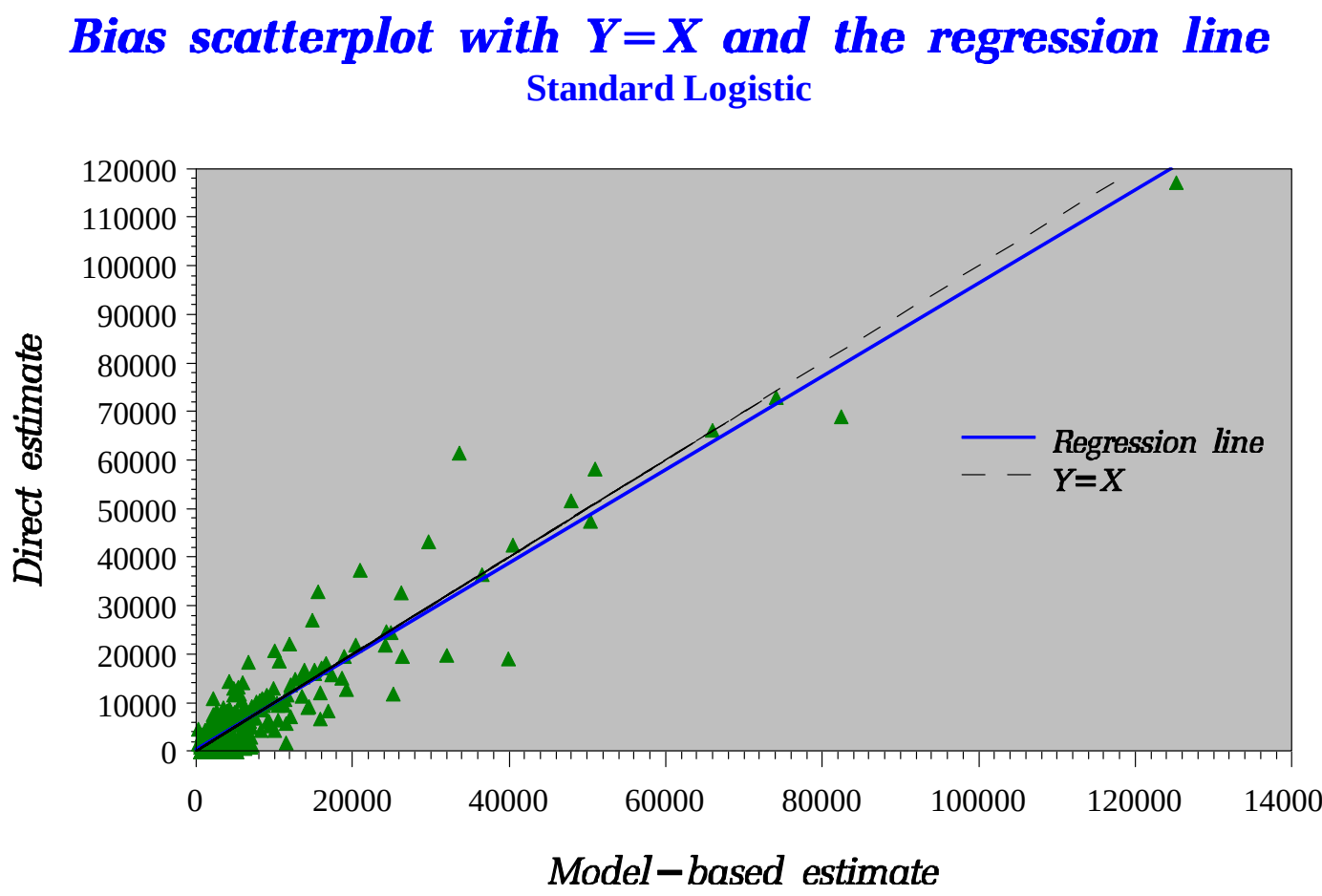


Estimation nationale = 2 409 000 chômeurs

Proximité peu surprenante avec l'EE car

$$\sum_{d,i \in emp} w_{d,i} \cdot \hat{P}_{d,i} = \sum_{d,i \in emp} w_{d,i} \cdot Y_{d,i}$$

Absence de biais (test)



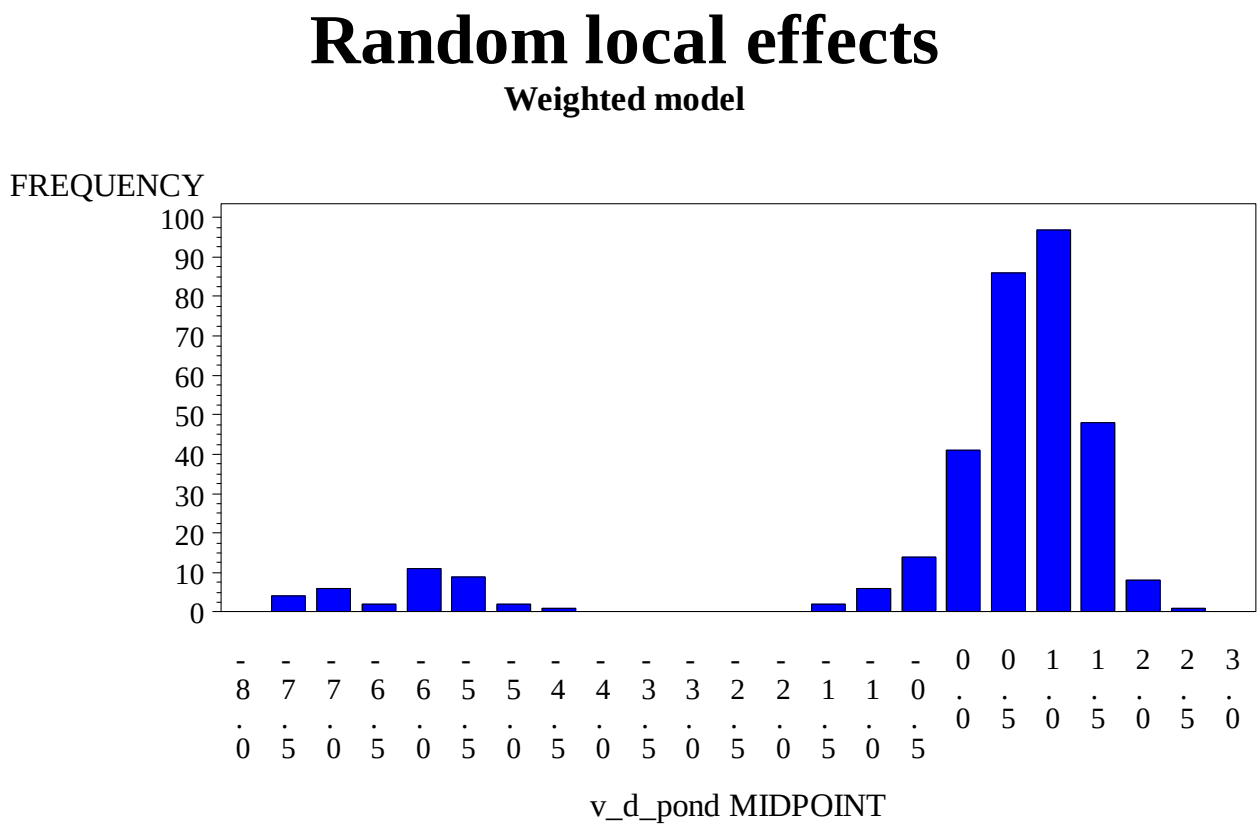
D) Modélisation mixte pondérée

Effect	Estimateur	Rappel estima[t] NON pondéré	Std	t Value	Pr > |t|
Intercept	-6.76	-6.15	0.1215	-55.62	<.0001
stoc_loc	-0.69	-0.76	0.00155	-444.74	<.0001
stoc_loc	0	0	.	.	.
AGE	3.31	3.49	0.01429	231.48	<.0001
AGE	5.10	5.33	0.01403	363.31	<.0001
AGE	4.99	5.26	0.01403	355.68	<.0001
AGE	4.80	5.00	0.01389	345.50	<.0001
AGE	4.40	4.52	0.01391	316.02	<.0001
AGE	0	0	.	.	.
dipl_binair	-0.50	-0.53	0.00153	-324.93	<.0001
dipl_binair	0	0	.	.	.
nat	-0.53	-0.54	0.00243	-219.97	<.0001
nat	0.12	0.11	0.00358	34.50	<.0001
nat	0	0	.	.	.
matr_celib	-0.58	-0.55	0.00171	-338.37	<.0001
matr_celib	0	0	.	.	.

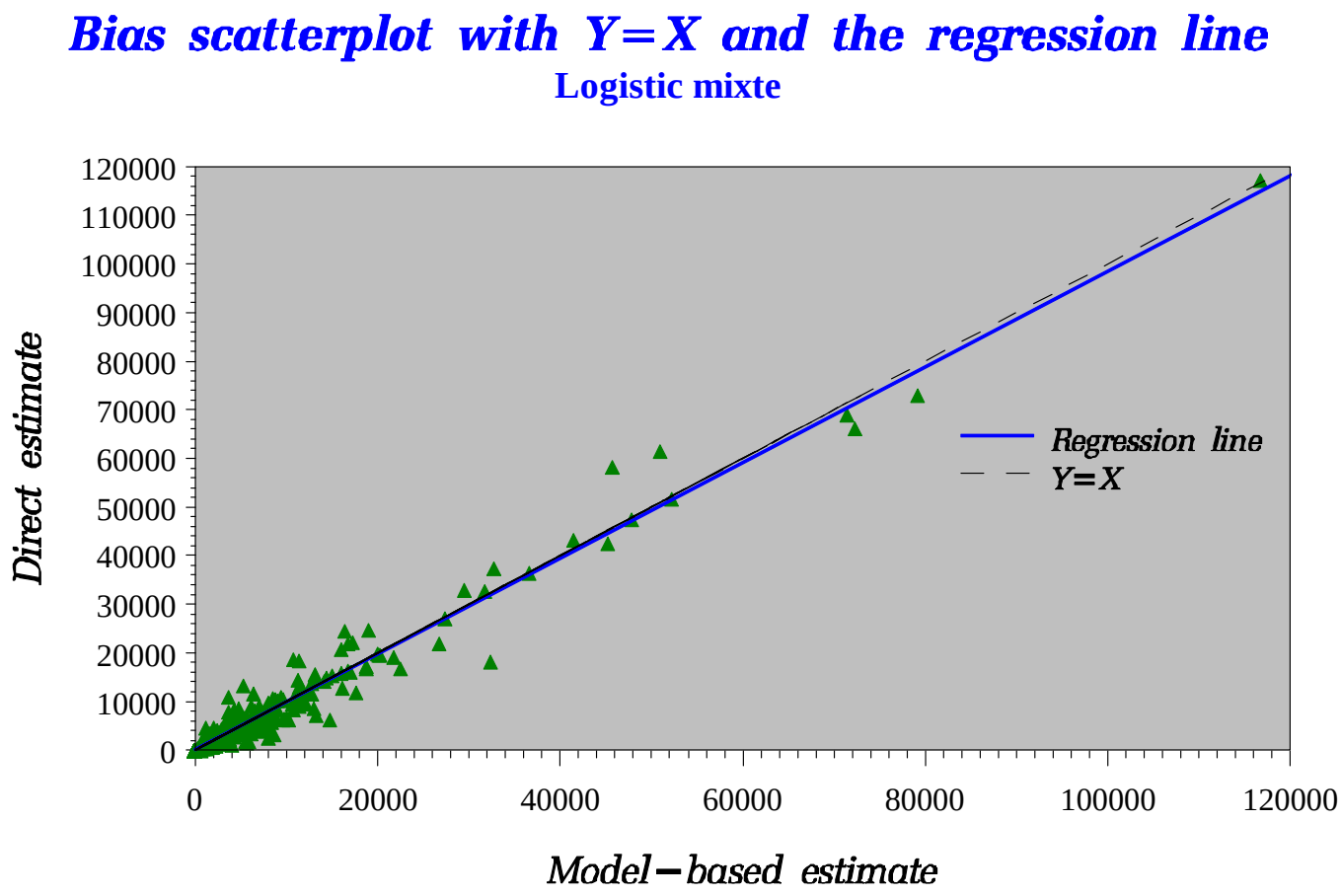
$\hat{\sigma}_v^2$ = 4.839 (écart-type = 0.426)

$\rightarrow$ il y a des effets ZE

Distribution des effets locaux $\hat{v}_d$:



Absence de biais.

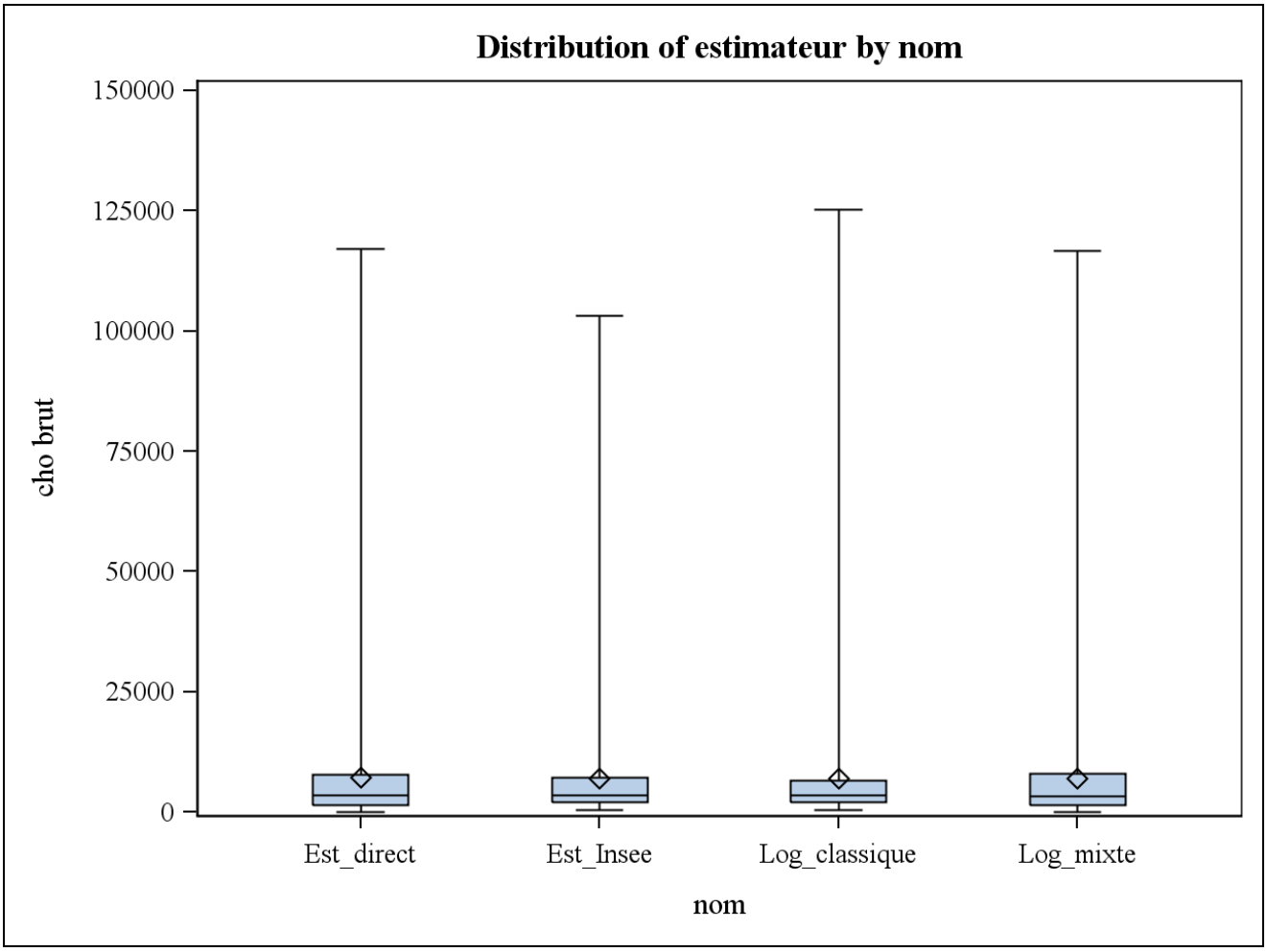


Estimations par ZEAT avant benchmarking

ZEAT	Classique	Mixte	Insee
1	+ 8.1 %	+ 0.7 %	- 10.6 %
2	- 0.4 %	+ 0.9 %	+ 2.4 %
3	- 29.1 %	- 0.5 %	- 8.2 %
4	+ 11.8 %	- 3.0 %	+ 10.6 %
5	+ 8.6 %	+ 0.5 %	+ 8.1 %
7	- 2.6 %	- 6.4 %	+ 0.4 %
8	+ 16.6 %	+ 2.8 %	+ 3.5 %
9	- 16.8 %	- 0.1 %	+ 1.2 %
TOTAL	**Sigma = 94,0**	**Sigma = 14,9**	**Sigma = 45,0**
RAPPEL ***Cas non pondéré***	**Sigma = 106,4**	**Sigma = 58,9**	**Sigma = 45,0**

Estimations nationales (avant benchmarking) :

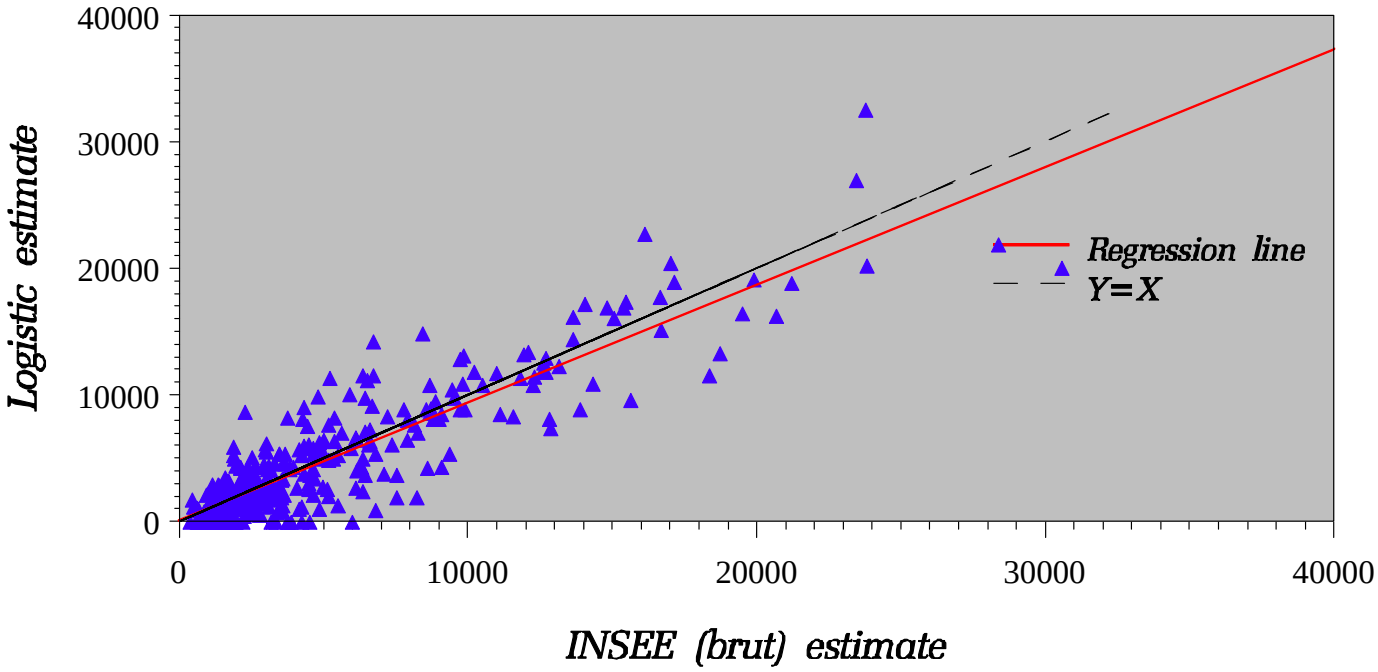
Direct :	2 416 000
Classique :	2 409 000
Mixte :	2 409 000
Insee :	2 408 000


Distribution of estimateur by nom
cho brut
150000
125000
100000
75000
50000
25000
0
Est_direct
Est_Insee
Log_classique
Log_mixte
nom

INSEE (brut) estimate versus Logistic estimate

Total number of BIT jobless people by ZE (after Benchmarking)

Selection of ZE : BIT jobless people < 25 000



Conclusion

- L'objectif est de faire "mieux" que l'estimateur direct - avant de faire "bien". *Il faut se contenter de l'à-peu-près :*

1. *ONS* : CV $\leq$ 20% ;

2. *ISTAT* : niveau province ($\approx$département) :

   CV $\leq$ 17,5% (moyennes annuelles)

   CV $\leq$ 25% (moyennes trimestrielles);

3. *Statistique Canada* :

   - CV $\leq$16,5% : aucun avertissement,

   - 16,5%$\leq$ CV$\leq$ 33,3% : mise en garde de l'utilisateur,

   - CV>33,3% : diffusion déconseillée

     (signature d'un "avis de non-respon-

     -sabilité si l'utilisateur passe outre).

$\Rightarrow$ Garder des ambitions "modestes" !

- L'examen du biais prime sur la variance : toutes ces méthodes vont (fortement) réduire la variance $\Rightarrow$ *le vrai enjeu est le biais*;

- La pratique = mélange savant de théorie très compliquée et d'empirisme;

- Le non-linéaire est très compliqué : plus d'intuition possible, des cathédrales … se laisser porter !

- Finalement, tout ça marche bien … et on peut en être surpris, compte tenu de la distance séparant ces techniques de l'inférence classique.

- Il faudrait regarder la stabilité des résultats dans le temps;

- Finalement, le classement de ces méthodes n'est pas immédiat - multiples considérations; il faudrait positionner la méthode officielle.

- Les perspectives d'amélioration sont quasi innombrables …

    - corrélation spatiale
    - corrélation temporelle
    - modèles optimisant des propriétés d'ensemble
    - modèles intégrant des benchmarking
    - domaine Bayésien
    - etc…