Extension de la méthode de Kokic et Bell au plan poissonnien

Thomas Deroyon, Insee, DG, DMS
Cyril Favre-Martinoz, Insee, DR de la Réunion-Mayotte, Criem

26 octobre 2018

Sommaire

Sommaire

Unité influente

- Dans les enquêtes, il est courant de mesurer des variables dont la distribution est asymétrique (revenu, CA,...).
- On peut également avoir une distribution de poids asymétrique.
- Il est fort probable qu'on observe des unités potentiellement influentes dans notre échantillon.
- Une unité influente fait partie intégrante de la population finie.
- Il s'agit d'une observation légitime.
- Problème induit par les unités influentes : elles rendent les estimateurs classiques très instables $\rightarrow$ variances élevées.
- Dans quelles situations, les estimateurs ponctuels sont instables ?

Unité influente : dans quelles situations, les estimateurs ponctuels sont instables ?

- Quand les poids de sondage sont très peu corrélés aux variables d'intérêt **et** les poids de sondage sont très dispersés → fréquent dans les enquêtes "Ménage".
- Quand la distribution de la variable d'intérêt est très asymétrique et/ou quand il y a des erreurs dans la base de sondage (i.e., mauvaises classifications) → problème des sauteurs de strate → fréquent dans les enquêtes "Entreprise".

Comment se prémunir contre les unités influentes à l'étape du plan de sondage ?

- Idéalement, on souhaiterait éliminer le problème des valeurs influentes dès le plan de sondage.
- Dans le cas des enquêtes "Entreprise", on peut se prémunir en construisant une ou plusieurs strates exhaustives → Les unités appartenant à ces strates ont une probabilité d'inclusion égale à 1 → Elles n'ont plus d'influence sur l'estimateur.
- Même avec un "bon plan", le problème des valeurs influentes n'est jamais complètement réglé :
  - on s'intéresse à beaucoup de variables d'intérêt et on dispose que d'un nombre limité de variables auxiliaires.
  - Problème des sauteurs de strate.
  - Etape de repondération : non réponse totale, calage

Sommaire

La winsorisation

- Méthode couramment utilisée : winsorisation
- Dans le cas d'un plan de sondage stratifié, elle consiste à associer à chaque partie de l'échantillon un seuil
- A chaque strate U_h, on définit un seuil K_h indépendant de l'échantillon S et la variable winsorisée $\tilde{X}$, pour $i \in S$, par :

$$\tilde{X}_{hi} = \begin{cases} X_{hi} \text{ si } X_{hi} < K_h \\ \frac{n_h}{N_h} X_{hi} + (1 - \frac{n_h}{N_h}) K_h \text{ si } X_{hi} \geq K_h \end{cases}$$

- L'estimateur winsorisé du total de X est alors l'estimateur par expansion du total de la variable winsorisée $\tilde{X}$: $\hat{T}(\tilde{X}) = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i \in S_h} \tilde{X}_{hi}$.

Détermination des seuils

- Pour déterminer les seuils :
  - Kokic et Bell (1994) pour les plans aléatoires simples stratifiés
  - Rivest et Hurtubise (1995)
  - Calcul de seuil basé sur les méthodes de biais conditionnel
- En pratique, à l'Insee, dans le cadre des entreprises, on mobilise la méthode de Kokic et Bell.

Sommaire

La méthode de Kokic et Bell

- Kokic et Bell (1994) ont déterminé les formules théoriques et des algorithmes de calcul des seuils qui conduisent à l'estimateur winsorisé ayant la plus faible erreur quadratique moyenne possible
  - sous l'hypothèse que les réalisations de la variable d'intérêt sont identiquement distribuées dans chaque strate, l'erreur quadratique moyenne étant calculée sous le plan de sondage et la loi de la variable d'intérêt
  - plan de sondage aléatoire simple stratifié
  - nécessite de disposer de données indépendantes de l'échantillon pour le calcul des seuils
- De plus en plus de tirages poissoniens dans le cadre des enquêtes entreprise
- La non-réponse est souvent modélisée comme une deuxième phase poissonienne
- Comment étendre K&B au cas poissonien ? Sous quelles hypothèses ? Robutesse à ces hypothèses ?

Sommaire

La méthode de Kokic et Bell : extension

- Nous nous intéressons à l'estimation du total dans la population $T(X) = \sum_{i \in U} X_i$ d'une variable X
- Plan de sondage P par lequel S est sélectionné : un plan de sondage poissonnien
- Chaque unité i de la population appartient à l'échantillon avec une probabilité $\pi_i > 0$.
- On suppose que X est une variable positive ou nulle ;

La méthode de Kokic et Bell : extension

Dans ce cadre, nous proposons comme dans la méthode originelle appliquée au sondage aléatoire simple stratifié d'associer un seuil K_h, $h = 1, \dots, H$ à chaque partie S_h, $h = 1, \dots, H$ et de définir :

- la variable winsorisée $\tilde{X}$ par

$$\tilde{X}_{hi} = \begin{cases} X_{hi} & \text{si } d_{hi} X_{hi} \leq K_h \\ \frac{X_{hi}}{d_{hi}} + \left(1 - \frac{1}{d_{hi}}\right) \frac{K_h}{d_{hi}} & \text{si } d_{hi} X_{hi} > K_h, \end{cases} \quad (1)$$

  où $d_{hi} = \frac{1}{\pi_i}$ est le poids de l'unité i dans la partie h.
- l'estimateur winsorisé du total de X comme l'estimateur par expansion usuel du total de $\tilde{X}$:

$$\hat{T}(\tilde{X}) = \sum_{h=1}^{H} \sum_{i \in S_h} d_{hi}\, \tilde{X}_{hi}. \quad (2)$$

La méthode de Kokic et Bell : extension

Dans ce cadre, nous proposons comme dans la méthode originelle appliquée au sondage aléatoire simple stratifié d'associer un seuil K_h, $h = 1, ..., H$ à chaque partie S_h, $h = 1, ..., H$ et de définir :

- la variable winsorisée $\tilde{X}$ par

$$\tilde{X}_{hi} = \begin{cases} X_{hi} \text{ si } X_{hi} < K_h \\ \frac{n_h}{N_h} X_{hi} + (1 - \frac{n_h}{N_h}) K_h \text{ si } X_{hi} \geq K_h \end{cases}$$

  où $d_{hi} = \frac{1}{\pi_i}$ est le poids de l'unité i dans la partie h.
- l'estimateur winsorisé du total de X comme l'estimateur par expansion usuel du total de $\tilde{X}$:

$$\hat{T}(\tilde{X}) = \sum_{h=1}^{H} \sum_{i \in S_h} d_{hi} \tilde{X}_{hi}. \tag{3}$$

La méthode de Kokic et Bell : extension

- On suppose qu'il est possible de partitionner la population et l'échantillon en sous-populations U_h et S_h dans lesquelles toutes les valeurs $d_{hi}X_{hi}$ sont des réalisations indépendantes issues d'un même modèle vérifiant :

$$\forall h = 1, ..., H, \, \forall i \in U_h, \, d_{hi}\, X_{hi} \; = \; \mu_h \, + \, \epsilon_{hi}, \qquad (4)$$

$$\text{avec} \left\{ \begin{array}{ll} E_m(\epsilon_{hi}) & = \; 0 \\ V_m(\epsilon_{hi}) & = \; \sigma_h^2 \; < \; +\infty \end{array} \right.$$

où E_m et V_m désigne l'espérance et la variance sous le modèle (4). Dans le cas SRS stratifié, on avait :

$$\forall h = 1, ..., H, \, \forall i \in U_h, \, X_{hi} \; = \; \mu_h \, + \, \epsilon_{hi}, \qquad (5)$$

La méthode de Kokic et Bell : extension

- L'hypothèse forte sous jacente au modèle (4) est que les valeurs X_{hi} multipliées par les poids d_{hi} sont supposées en espérance constantes dans chaque strate.
- Les probabilités d'inclusion au sein de chaque strate sont définies proportionnellement à la variable d'intérêt X.
- En pratique, il arrive fréquemment que ces probabilités d'inclusion soient définies proportionnellement à une variable auxiliaire connue et fortement corrélée à X, ce qui permet d'être proche de l'hypothèse sous jacente au modèle (4).
- Modèle (4) est celui sous lequel l'estimateur d'Horvitz-Thompson est optimal au sens de la minimisation de l'erreur quadratique moyenne anticipée sous le modèle.
- Dans la suite, les variables aléatoires $d_{hi}\, X_{hi}$ étant supposées indépendantes et identiquement distribuées au sein de chaque strate, nous noterons $Z_{hi} = d_{hi}\, X_{hi}$.

La méthode de Kokic et Bell : extension

- Nous nous plaçons de plus dans le même cadre asymptotique que Kokic et Bell 1994 en adaptant l'hypothèse portant sur les probabilités d'inclusion :
- $\forall \nu \in \mathbb{N}, \forall h = 1, ..., H, \; n_{h_\nu} > 1$;
- $N_\nu, n_\nu \underset{\nu \to +\infty}{\to} +\infty$;
- le nombre de strates H est fixe.

$$\forall h = 1, ..., H, \exists (\lambda_{1h}, \lambda_{2h}) \in]0,1[^2, \text{ tel que } \forall i \in U_h, \min(\pi_i) > \lambda_{1h} \text{ et } \max(\pi_i) < \lambda_{2h}.$$

La méthode de Kokic et Bell : extension

- Comme dans l'approche de K&B, les seuils K_h sont déterminés de manière à minimiser l'erreur quadratique moyenne de l'estimateur winsorisé $\hat{T}(\tilde{X})$ sous le modèle de la variable X et sous le plan de sondage P
- EQM minimisée en moyenne sur l'ensemble des populations possibles compte-tenu du modèle de super-population posé sur X et en moyenne sur l'ensemble des échantillons tirés dans ces populations compte-tenu du plan de sondage P :

$$(K_h^\star)_{h=1,\ldots,H} \in Argmin_{(K_h)_{h=1,\ldots,H}} E_m\, E_P \left\{ \left[\hat{T}(\tilde{X}) - T(X) \right]^2 \right\}$$

La méthode de Kokic et Bell : extension

- Il est possible de montrer qu'à l'optimum et asymptotiquement, en notant $J_{hi} = \mathbb{I}(Z_{hi} > K_h)$:

$$\forall\, h = 1, ..., H,\ K_h \sim -\frac{A_h}{C_h + D_h} B \qquad (6)$$

$$\text{avec} \begin{cases} A_h & = \sum_{i \in U_h} \frac{1}{d_{hi}} \left(1 - \frac{1}{d_{hi}}\right) \\ C_h & = \sum_{i \in U_h} \left(\frac{1}{d_{hi}}\right)^2 \left(1 - \frac{1}{d_{hi}}\right)^2 \\ D_h & = \sum_{i \in U_h} \frac{1}{d_{hi}} \left(1 - \frac{1}{d_{hi}}\right)^3 \end{cases}$$

$$\text{et } B = \sum_{h=1}^{H} A_h \left[K_h\, E_m(J_h) - E_m(J_h\, Z_h)\right]. \qquad (7)$$

- B est le biais de l'estimateur winsorisé optimal $\hat{T}(\tilde{X})$.
- A l'optimum, le seuil K_h est donc égal à un terme positif près, à l'opposé du biais multiplié par le terme $\frac{A_h}{C_h + D_h}$.

La méthode de Kokic et Bell : extension

- A l'optimum et asymptotiquement, B est l'opposé du point d'annulation de la fonction F définie par :

$$F(L) = L\left(1 + \sum_{h=1}^{H} \frac{A_h^2}{C_h + D_h} E_m(J_h^\star)\right) - \sum_{h=1}^{H} \frac{A_h^2}{C_h + D_h} E_m(J_h^\star\ X_h^\star).$$

- Supposons que nous disposons, pour chaque sous-population h, de p_h réalisations $\breve{X}_{hi}$ tirées dans la loi de X et indépendantes de l'échantillon S.
- On peut estimer F par :

$$\hat{F}(L) = L\left(1 + \sum_{h=1}^{H} \frac{A_h^2}{C_h + D_h} \frac{\sum_{i=1}^{p_h} \mathbb{I}(\breve{X}_{hi}^\star > L)}{p_h}\right) - \sum_{h=1}^{H} \frac{A_h^2}{C_h + D_h} \frac{\sum_{i=1}^{p_h} \breve{X}_{hi}^\star\, \mathbb{I}(\breve{X}_{hi}^\star > L)}{p_h}$$

et estimer B par l'opposé du point d'annulation de $\hat{F}$.

La méthode de Kokic et Bell : extension

- Pour déterminer le point d'annulation de $\hat{F}$, il faut donc opérer de manière analogue à la méthode proposée par K& B dans le cas du sondage aléatoire simple stratifié :
  - $\hat{F}$ est donc une fonction affine et croissante de L
  - calculer $\hat{F}(0)$, $\hat{F}\left(\breve{X}^{\star}_{(1)}\right)$, $\hat{F}\left(\breve{X}^{\star}_{(2)}\right)$, ..., $\hat{F}\left(\breve{X}^{\star}_{(p)}\right)$;
  - identifier la valeur j telle que $\hat{F}\left(\breve{X}^{\star}_{(j)}\right) \leq 0$ et $\hat{F}\left(\breve{X}^{\star}_{(j+1)}\right) \geq 0$, en posant que $\breve{X}^{\star}_{(0)} = 0$
  - B est alors estimé par interpolation :

$$\hat{B} = -\frac{\breve{X}^{\star}_{(j)}\,\hat{F}\left(\breve{X}^{\star}_{(j)}\right) - \breve{X}^{\star}_{(j+1)}\,\hat{F}\left(\breve{X}^{\star}_{(j+1)}\right)}{\hat{F}\left(\breve{X}^{\star}_{(j)}\right) - \hat{F}\left(\breve{X}^{\star}_{(j+1)}\right)}.$$

  - Puis $\hat{K}_h = -\frac{A_h}{C_h+D_h}\,\hat{B}$.

La méthode de Kokic et Bell : extension

- On va tester l'efficacité de ce nouvel estimateur en termes d'EQM
- On va aussi regarder sa robustesse à une mauvaise spécification du modèle imposé au $d_{hi}X_{hi}$
- Comparer avec les méthodes de biais conditionnel

Sommaire

La méthode de Kokic et Bell : extension

- Base d'apprentissage de taille $N = 5000$: $L = 1000$ réalisations d'un certain modèle ;
- pour chacune de ces réalisations, nous calculons le seuil optimal K_l selon la méthode présentée ;
- $M = 10000$ bases de sondages de test générées selon un (autre) modèle sur lesquelles nous sélectionnons un échantillon de taille espérée $n = 500$ suivant un tirage poissonnien et calculons l'estimateur robuste $\hat{\theta}_{(m)}$ avec le seuil K_l calculé.
- En guise de comparaison, nous calculons également l'estimateur robuste issu de la méthode basée sur le biais conditionnel.

La méthode de Kokic et Bell : extension

Les probabilités d'inclusion, ainsi que les valeurs de la variable X ont été générées selon le modèle suivant :

$$U_i \sim \mathcal{L}\text{og-}\mathcal{N}(1, 1.1),$$

$$\pi_i = n \times \frac{U_i}{\sum_{i=1}^{N} U_i},$$

$$X_i = 2000 \times \pi_i + \pi_i \epsilon_i + \delta_i V_i,$$

$$\epsilon_i \sim \mathcal{N}(0, 100), V_i \sim \mathcal{L}\text{og-}\mathcal{N}(log(500), 1.2), \delta_i \sim \mathcal{B}(\omega),$$

où ω est le paramètre de la Bernoulli, reflétant la proportion de valeurs influentes.

La méthode de Kokic et Bell : extension

	Valeurs du paramètre ω	
Scénario	Modèle d'apprentissage	Modèle de test
1	0	0
2	0.01	0.01
3	0.01	0.1
4	0.1	0.01

Table 1: Valeurs du paramètre ω utilisées afin de générer les populations

La méthode de Kokic et Bell : extension

Comme mesure du biais d'un estimateur $\hat{\theta}$ d'un total T, nous avons calculé le biais relatif Monte Carlo (en %) :

$$BR_{MC}(\hat{\theta}) = \frac{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{\theta}_{(m)} - T\right)}{T} \times 100,$$

où $\hat{\theta}_{(m)}$ désigne l'estimateur $\hat{\theta}$ dans l'échantillon m, $m = 1, \ldots, M$. Nous avons également calculé l'efficacité relative des estimateurs robustes relativement à l'estimateur par dilatation, $\hat{t}$:

$$RE_{MC}(\hat{\theta}) = \frac{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{\theta}_{(m)} - T\right)^2}{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{t}_{(m)} - T\right)^2} \times 100.$$

La méthode de Kokic et Bell : extension

Statistique descriptive	Scénario 1				Scénario 2			
	K&B		BHR		K&B		BHR	
	BR	RE	BR	RE	BR	RE	BR	RE
min	-0.2	100	-0.43	100	-9.0	1	-4.3	26
$Q1$	-0.1	100	-0.32	100	-2.9	35	-1.9	51
Médiane	0.0	100	-0.27	100	-1.8	50	-1.5	62
Moyenne	0.0	100	-0.27	100	-2.0	50	-1.6	62
$Q3$	0.0	100	-0.23	100	-1.0	64	-1.3	73
max	0.0	100	-0.14	100	-0.1	109	-0.6	91

Table 2: Statistiques descriptives pour les scénarios 1 et 2 sur les 1000 simulations pour $n = 500$

La méthode de Kokic et Bell : extension

Statistique descriptive	Scénario 3				Scénario 4			
	K&B		BHR		K&B		BHR	
	BR	RE	BR	RE	BR	RE	BR	RE
min	-32.2	2	-7.8	27	-4.5	1	-4.3	26
$Q1$	-18.9	50	-5.1	59	-1.8	48	-1.9	51
Médiane	-13.9	82	-4.6	66	-1.5	70	-1.5	62
Moyenne	-14.2	89	-4.7	65	-1.5	68	-1.6	62
$Q3$	-9.3	138	-4.2	72	-1.2	91	-1.3	73
max	-0.01	537	-2.7	89	-0.6	100	-0.6	91

Table 3: Statistiques descriptives pour les scénarios 3 et 4 sur les 1000 simulations pour $n = 500$

Quelques remarques

Ces simulations montrent donc :

- qu'en l'absence d'unités influentes, les deux méthodes d'estimation robuste n'entraînent pas de perte d'efficacité d'estimation ;
- que quand elle est appliquée dans ses hypothèses, la méthode de Kokic et Bell conduit à des estimateurs plus précis que la méthode du biais conditionnel ;
- que la méthode de Kokic et Bell est cependant sensible aux données utilisées pour calculer les seuils ; si ces données ne sont pas générées suivant le même modèle que les données auxquelles les seuils sont appliqués, la méthode peut conduire à une perte de précision ;
- que la méthode du biais conditionnel permet de gagner toujours en précision sur ces simulations, même si ce gain n'est pas optimal.

Sommaire

Conclusion

- Une alternative à K& B pour le cas poissonien
- Elle repose sur la possibilité de disposer de données historiques ou auxiliaires
- Efficace dans le cas où le modèle est bien spécifié : le modèle ayant généré les données historiques est proche du modèle ayant généré nos données d'enquête
- Alternative non paramétrique efficace : méthode de biais conditionnel qui ne nécessite pas d'information complémentaire.

Merci pour votre attention !