Analyse de sensibilité de deux méthodes d'estimation de moyenne doublement robuste : Méthodes et plan de simulation

G. Santin[1], J. Bouyer[2], R. Sitta[3], A. Gueguen[3]

1/ Institut de veille sanitaire, Département santé travail, Saint-Maurice, France – 2/ CESP-INSERM 1018 - Équipe Épidémiologie de la reproduction et du développement de l'enfant, Le Kremlin-Bicêtre, France
3/ CESP-INSERM 1018 - Plateforme de recherche "Cohortes en population", Villejuif, France

Contexte

- Données manquantes inévitables dans les enquêtes
- Classification de Rubin [1976]
  - Missing Completly At Random $(Y \perp p)$: peu réaliste
  - Missing At Random $(Y \perp p)/X$: plus réaliste mais à quel point ?
  - Missing Not At Random $(Y \leftrightarrow p)/X$: probablement fréquent mais invérifiable
- Dans le cas MAR :

Nécessite d'avoir des données auxiliaires (covariables) disponibles chez les répondants et les non-répondants

- Méthodes classiques pour corriger les biais de non-réponse *a posteriori*
  . Modélisation de la propension à répondre (repondération)
  . Modélisation de la variable d'intérêt (imputation)
- Méthode récente qui combine repondération et imputation : estimation doublement robuste

- Développement de protocoles originaux en épidémiologie qui permettent l'accès à des données issues de systèmes d'information pour l'ensemble des personnes tirées au sort

➔ Application possible de ces méthodes

Objectifs

Analyse de sensibilité par simulation de méthodes d'estimation doublement robuste d'une moyenne et de la variance de l'estimateur de la moyenne applicables en surveillance épidémiologique (données d'enquête avec plan de sondage)

➔ **Quantification du biais de non-réponse pour l'estimation d'une moyenne quand on s'écarte des conditions MAR** en jouant sur 1) le taux de réponse 2) le lien entre la variable d'intérêt et la propension à répondre 3) la mauvaise modélisation de la propension à répondre ou/et de la variable d'intérêt en omettant des covariables explicatives du/des modèles.

Méthodes de correction de la non-réponse

Notations

Y : variable d'intérêt de la population finie P de taille N
μ : moyenne de Y de la population P
m : modèle de la variable d'intérêt Y
s : échantillon de taille n tiré au sort par sondage aléatoire simple dans P
$X=(X_1,\ldots,X_p)$ variables auxiliaires (covariables) disponibles sur (s)
p : propension à répondre à Y
R : variable indicatrice de non-réponse à Y (=1si Y mesuré / =0 sinon)

Correction de la non-réponse par repondération

Hypothèse : $p = P(R=1/Y=y,X) = P(R=1/X) \Leftrightarrow MAR$

Soit $p(X,\hat{\phi})$ un estimateur correctement spécifié de p

L'estimateur de la moyenne de Y $\hat{\mu}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i Y_i}{p_i(X,\hat{\phi})}$ est sans biais

(+) : modélisation explicite en utilisant les données sur l'échantillon complet
➔ possible de visualiser les problèmes potentiels liés à des $p_i(X,\hat{\phi})$ petits

(-) : Si les $p_i(X,\hat{\phi})$ sont très dispersés, qu'une petite quantité de $p_i(X,\hat{\phi})$ ont des valeurs très faibles et que Y et $p_i(X,\hat{\phi})$ ne sont pas corrélés, la variance de l'estimateur de la moyenne peut exploser.

Modèle d'imputation des Y manquants

Hypothèse : $m = P(Y=y/R=1,X) = P(Y=y/X) \Leftrightarrow MAR$

Soit $m(X,\hat{\beta})$ un estimateur correctement spécifié de m

L'estimateur de la moyenne de Y

$$\hat{\mu}_{OR} = \frac{1}{n}\sum_{i=1}^{n} m(X_i;\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(R_i Y_i + (1-R_i)m(X_i;\hat{\beta}))$$

est sans biais

(+) : ne pose pas de problème de variance de l'estimateur de la moyenne qui explose

(-) : modélisation implicite de Y en utilisant uniquement les données de Y observées

Estimation doublement robuste

Hypothèses : $\rho = P(R=1/Y=y,X) = P(R=1/X)$ et $m = P(Y=y/R=1,X) = P(Y=y/X) \Leftrightarrow MAR$

Sous l'hypothèse d'une bonne spécification du modèle $p(X,\hat{\phi})$ OU du modèle $m(X,\hat{\beta})$ l'estimateur de la moyenne de Y

$$\hat{\mu}_{DR} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{R_i Y_i}{p_i(X_i;\hat{\phi})} + \frac{p_i(X_i;\hat{\phi}) - R_i}{p_i(X_i;\hat{\phi})} m(X_i;\hat{\beta})\right\}$$

est sans biais

(+) : estimateur consistant si un des deux modèles est correctement spécifié

Beaucoup de propositions d'estimateurs liées à la variance de l'estimateur de la moyenne :

- statistique classique, statistique d'enquête ;
- prise en compte de poids dispersés et élevés.

Simulations

Méthodes testées (méthodes applicables en surveillance épidémiologique)

- **Deux méthodes classiques** :
  - repondération (IPW)

$\hat{p}_w$: probabilité de réponse prédite par régression logistique

  - imputation (OR)

$\hat{m}_w$: modèle d'imputation de Y par régression linéaire

- **Deux méthodes de double robustesse** :
  - Wirth (Epidemiology, 2010)

$$\hat{\mu}_{DR-W} = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i Y_i}{\hat{p}_{i,w}} + \frac{\hat{p}_{i,w} - R_i}{\hat{p}_{i,w}}\hat{m}_{i,w}$$

  - Kim-Haziza – cas où Y est imputé par régression linéaire (Proceedings of the Survey Research Methods, American Statistical Association, 2010)

$\hat{p}_{KH}$: probabilité de réponse prédite par calage

$\hat{m}_{KH}$: modèle d'imputation de Y par régression linéaire pondérée par $\frac{1-\hat{p}_{KH}}{\hat{p}_{KH}}$

$$\hat{\mu}_{DR-KH} = \frac{1}{n}\sum_{i=1}^{n} R_i Y_i + (1-R_i)\hat{m}_{i,KH}$$

Constitution de la population

Population de taille N=1 000 000

Covariables :

X_i (i=1,…,5)~N(0,1)

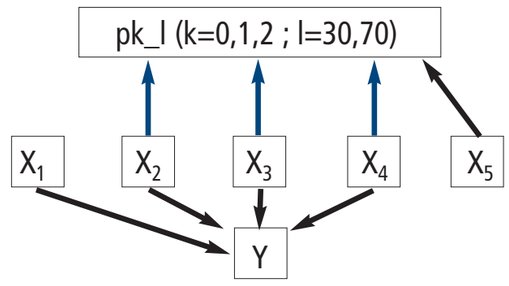
Variable d'intérêt :

$Y=f(X_1, X_2, X_3, X_4)$

Propension à répondre : 6 scenarii

		Lien entre Y et p		
		Faible	**Moyen**	**Fort**
Taux de réponse	**70 %**	p0_70	p1_70	p2_70
	30 %	p0_30	p1_30	p2_30

Génération de la non-réponse



Flèches bleues : modulation du lien entre Y et p

- lien fort p_2~ $f(X_2, X_3, X_4, X_5)$:
  X_2 et X_3 et X_4 observées ➔ MAR ; X_2 ou X_3 ou X_4 non observées ➔ MNAR
- lien moyen p_1~ $f(X_2, X_3, X_5)$:
  X_2 et X_3 observées ➔ MAR ; X_2 ou X_3 non observées ➔ MNAR
- lien faible p_0~ $f(X_2, X_5)$:
  X_2 observée ➔ MAR ; X_2 non observée ➔ MNAR

Constitution des échantillons : tirage au sort de 5 000 échantillons

Paramètres estimés et critères de comparaison retenus pour chaque scénario

- Moyenne et variance de l'estimateur de la moyenne pour chaque méthode *(meth=IPW, OR, DR-KW, DR-KH)*
- Biais relatif des moyennes et des variances des estimateurs des moyennes

$$BR_{MC}(\hat{\mu}_{meth}) = \frac{\frac{1}{5000}\sum_{s=1}^{5000}\hat{\mu}^{s}_{meth} - \mu}{\mu} \times 100$$

- Erreur quadratique moyenne des moyennes et des variances des estimateurs de moyenne pour chaque méthode
- Taux de couverture d'un intervalle de confiance à 95 %

$$TC_{MC} = \frac{1}{5000}\sum_{s=1}^{5000} A_s \times 100 \text{ où } A_s = \begin{cases} 1 \text{ si } \mu \in \text{IC95\% de } \hat{\mu}^{s}_{meth} \\ 0 \text{ si } \mu \notin \text{IC95\% de } \hat{\mu}^{s}_{meth} \end{cases}$$

Types de résultats futurs

		Méthodes utilisées pour les estimations			
Combinaison de variables omises	**Critères étudiés**	OR	IPW	DR-W	DR-KH
MAR : [aucune]	BR moyenne				
	BR variance				
	EQM moyenne				
	EQM variance				
	Taux de couverture				
MNAR : X_1, X_2, X_3	BR moyenne				
	BR variance				
	EQM moyenne				
	EQM variance				
	Taux de couverture				
MNAR : …					

Septembre 2012 - Adelf, Bruxelles

INSTITUT DE VEILLE SANITAIRE

Contact : Gaëlle Santin (g.santin@invs.sante.fr) - Institut de veille sanitaire - 12, rue du Val d'Osne 94415 Saint-Maurice Cedex - Tél : +33 (0)1 55 12 54 28 - Fax : +33 (0)1 41 79 67 88