Université de Montréal

# Inférence doublement robuste en présence de données imputées dans les enquêtes

par

**Frédéric Picard**

Département de mathématiques et de statistique
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en Statistique

février 2010

# Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé

## Inférence doublement robuste en présence de données imputées dans les enquêtes

présenté par

## Frédéric Picard

a été évalué par un jury composé des personnes suivantes :

*Pierre Duchesne*

(président-rapporteur)

*David Haziza*

(directeur de recherche)

*Mylène Bédard*

(membre du jury)

Mémoire accepté le:

*19 février 2010*

# SOMMAIRE

L'imputation est souvent utilisée dans les enquêtes pour traiter la non-réponse partielle. Il est bien connu que traiter les valeurs imputées comme des valeurs observées entraîne une sous-estimation importante de la variance des estimateurs ponctuels. Pour remédier à ce problème, plusieurs méthodes d'estimation de la variance ont été proposées dans la littérature, dont des méthodes adaptées de rééchantillonnage telles que le Bootstrap et le Jackknife. Nous définissons le concept de double-robustesse pour l'estimation ponctuelle et de variance sous l'approche par modèle de non-réponse et l'approche par modèle d'imputation. Nous mettons l'emphase sur l'estimation de la variance à l'aide du Jackknife qui est souvent utilisé dans la pratique. Nous étudions les propriétés de différents estimateurs de la variance à l'aide du Jackknife pour l'imputation par la régression déterministe ainsi qu'aléatoire. Nous nous penchons d'abord sur le cas de l'échantillon aléatoire simple. Les cas de l'échantillonnage stratifié et à probabilités inégales seront aussi étudiés. Une étude de simulation compare plusieurs méthodes d'estimation de variance à l'aide du Jackknife en terme de biais et de stabilité relative quand la fraction de sondage n'est pas négligeable. Finalement, nous établissons la normalité asymptotique des estimateurs imputés pour l'imputation par régression déterministe et aléatoire.

**Mots Clés** : Double robustesse ; Approche par modèle d'imputation ; Non-réponse partielle ; Estimateur de variance Jackknife ; Estimateur de variance linéarisé ; Approche par modèle de non-réponse ; Approche renversée ; Imputation par la régression ; Approche deux phases.

# SUMMARY

Imputation is often used in surveys to treat item nonresponse. It is well known that treating the imputed values as observed values may lead to substantial underestimation of the variance of the point estimators. To overcome the problem, a number of variance estimation methods have been proposed in the literature, including appropriate versions of resampling methods such as the jackknife and the bootstrap. We define the concept of doubly robust point and variance estimation under the so-called nonresponse and imputation model approaches. We focus on jackknife variance estimation, which is widely used in practice. We study the properties of several jackknife variance estimators under both deterministic and random regression imputation. We first consider the case of simple random sampling without replacement. The case of stratified simple random sampling and unequal probability sampling is also considered. A limited simulation study compares various jackknife variance estimators in terms of bias and relative stability when the sampling fraction is not negligible. Finally, the asymptotic normality of imputed estimator is established under both deterministic and random regression imputation.

**KEY WORDS** : Double robustness ; Imputation model approach ; Item nonresponse ; Jackknife variance estimator ; Linearization variance estimator ; Nonresponse model approach ; Reverse framework ; Regression imputation ; Two-phase framework.

# TABLE DES MATIÈRES

# LISTE DES FIGURES

# LISTE DES TABLEAUX

# REMERCIEMENTS

Je remercie David Haziza. En plus d'avoir eu la chance de suivre son excellent cours en échantillonnage, j'ai eu le privilège de l'avoir comme directeur de recherche. Il a partagé avec enthousiasme ses connaissances, son expertise et son intuition.

Je remercie les membres du jury pour leur lecture minutieuse de ce mémoire.

Je remercie mon épouse, Noriko, qui m'a autorisé à passer beaucoup de soirées devant l'ordinateur à travailler sur ce mémoire.

# INTRODUCTION

Dans toutes les branches de la statistique, nous devons faire face au problème des données manquantes. Ceci est particulièrement vrai dans le domaine des enquêtes. En effet, dans les enquêtes la non-réponse des unités sélectionnées dans l'échantillon est importante et ceci aura pour effet évidemment d'avoir des données manquantes. Les causes de la non-réponse sont multiples. Le refus de l'unité de répondre à l'enquête ou l'impossibilité de contacter l'unité en sont les principales.

On distingue généralement la non-réponse totale (aucune information recueillie sur l'unité) de la non-réponse partielle (réponse manquante pour certain items seulement). Dans ce mémoire, nous nous pencherons sur le cas de la non-réponse partielle qui est habituellement traitée par imputation. Nous supposerons que nous avons un vecteur de variables auxiliaires dont la valeur sera disponible pour toutes les unités sélectionnées (même si celles-ci ne sont pas répondantes). En plus de l'imputation nous mentionnons deux techniques de traitement de la non-réponse.

Une option qui est généralement peu recommandée est d'utiliser seulement les répondants complets. Le problème avec cette approche est qu'il risque d'y avoir perte d'information en excluant les répondants partiels. De plus, l'exclusion des répondants partiels risque d'entrainer un biais dans les estimations.

La repondération consiste à modifier les poids de sondages pour tenir compte de la non-réponse. En général, la repondération est plus souvent utilisée pour les unités pour lesquelles il y a non-réponse complète mais elle est aussi (plus rarement) utilisée pour la non-réponse partielle. Le problème est que si l'on étudie plus

d'une variable d'intérêt, cette méthode exige différents poids pour les différentes variables, ce qui peut entraîner une certaine confusion chez les utilisateurs.

L'imputation consiste à remplacer les valeurs manquantes par une ou plusieurs valeurs artificielles. Lorsque l'on impute plusieurs valeurs, on parle d'imputation multiple. Le cas de l'imputation multiple ne sera pas considéré ici. Le lecteur intéressé par l'imputation multiple pourra consulter Rubin (1987).

L'imputation a l'avantage de créer un fichier complet. Ce type de fichier est facile à utiliser. Les différentes analyses faites à partir d'un fichier imputé seront vraisemblablement cohérentes. Toutefois l'imputation comporte certains risques. L'imputation peut donner l'impression d'avoir des données complètes alors qu'en fait elles sont incomplètes. Ceci pourra créer des problèmes si les analyses à partir d'un fichier de données imputées ne sont faites avec aucune précaution prise à cet égard. L'imputation modifie généralement les relations entre les variables et un analyste pourrait ainsi obtenir de fausses conclusions.

Il est connu que traiter les valeurs imputées comme des valeurs observées entraîne une sous-estimation importante de la variance des estimateurs ponctuels. Un cas flagrant est le cas où l'on imputerait la moyenne des unités répondantes aux unités non-répondantes.

La non-réponse sera considérée comme un phénomène aléatoire. Ceci permet de faire de l'inférence en présence de données imputées si on émet des hypothèses supplémentaires.

En absence de non-réponse, il n'est pas nécessaire d'émettre d'hypothèses sur les variables d'intérêt pour faire de l'inférence dans les enquêtes. Les variables d'intérêt sont considérées comme fixées mais inconnues. La seule composante aléatoire est la sélection d'une unité dans l'échantillon qui est en général contrôlée par le plan de sondage. Toutefois, en présence de données imputées, il faut émettre des hypothèses supplémentaires pour faire une inférence. Il faut d'abord supposer que, conditionnellement aux variables auxiliaires, la probabilité de réponse ne dépende pas de la variable d'intérêt. Il faut également émettre des hypothèses supplémentaires parmi deux approches. La première est l'approche par modèle de non-réponse dans laquelle on suppose que la non-réponse est uniforme (les

unités répondent de façon indépendante avec la même probabilité) à l'intérieur de chaque classe d'imputation. La seconde est l'approche par modèle d'imputation. Sous cette approche on ne suppose plus que la variable d'intérêt est fixée. La population est alors considérée comme une réalisation d'un échantillon à partir d'une population infinie et on émet des hypothèses sur la distribution de la variable d'intérêt.

Dans ce mémoire nous commençons au chapitre 1 par un bref survol de l'univers des enquêtes, de l'échantillonnage, de la non-réponse et de l'imputation. Dans le chapitre 2, nous présentons l'article de David Haziza et Frédéric Picard intitulé *Doubly Robust Point and Variance Estimation in the Presence of Imputed Data*. Dans la section 2, nous discutons des approches par modèle de non-réponse et par modèle d'imputation. Ensuite nous discutons de la double robustesse de l'estimateur par la régression. Dans la section 3, nous dérivons un estimateur de variance linéarisé à l'aide de l'approche renversée. Nous traitons l'estimateur de variance Jackknife sous l'échantillon aléatoire simple sans remise dans la section 4. Dans la section 5, nous présentons les résultats d'une étude de simulation qui compare la performance de plusieurs estimateurs en terme de biais relatif et d'efficacité relative. Le cas de l'estimateur imputé par la régression aléatoire est traité dans la section 6. Dans la section 7, nous proposons un estimateur Jackknife de la variance dans le cadre de l'échantillonnage à probabilités inégales. Dans la section 8, nous concluons et discutons des généralisations possibles. Finalement, en annexe de l'article, nous démontrons la double robustesse ainsi que la normalité asymptotique de l'estimateur de régression.

# Chapitre 1

# PRÉLIMINAIRES

## 1.1. L'univers des enquêtes

Dans cette section, nous faisons un survol rapide des différents types d'erreurs dans les enquêtes. Nous commençons par rappeler le contexte. Supposons que nous avons une population finie $U$ composée de $N$ unités. L'objectif est d'estimer le total d'une variable d'intérêt $y$, $Y = \sum_{i \in U} y_i$, où $y_i$ désigne la valeur de la variable $y$ pour l'unité $i, i \in U$. On peut également s'intéresser à la moyenne de la variable $y$, $\bar{Y} = Y/N$. Pour des raisons de coût, d'efficacité, de faisabilité et même de précision il est souvent préférable de tirer un échantillon aléatoire $s \subset U$ duquel on observera la valeur $y$ uniquement pour les unités sélectionnées dans l'échantillon. L'estimation de $Y$ comportera une erreur, appelée erreur d'échantillonnage, causée par le fait que la variable $y$ n'est observée que pour les unité $i \in s$. Toutefois, l'erreur d'échantillonnage est contrôlée et peut être réduite en augmentant la taille de l'échantillon. Il existe d'autre types d'erreurs survenant dans les enquêtes que nous appelons erreurs non dues à l'échantillonnage. Nous classons ces types d'erreurs en quatre catégories :

- les erreurs de couverture ;
- les erreurs de mesure ;
- les erreurs de traitement ;
- les erreurs de non-réponse.

Lors des enquêtes, l'échantillon est tiré à partir d'une base de sondage qui idéalement contiendrait la liste exacte des unités de la population cible. Les erreurs

de couverture sont dues au fait que la base de sondage et la population cible ne coincident pas parfaitement. Parfois, certaines unités de la population cible ne font pas partie de la base de sondage et on parle alors de sous-couverture. Sinon, il se peut qu'il y ait des unités qui ne fassent pas parti de la population cible mais qui se trouvent sur la base de sondage. On parle alors de sur-couverture.

Les erreurs de mesure, qui surviennent lorsque la valeur observée n'est pas égale à la vraie valeur, peuvent avoir plusieurs sources : mauvaise interprétation du questionnaire, incapacité à répondre et parfois action délibérée du répondant à saboter l'enquête. Ces types d'erreurs, bien que possiblement très importantes ne seront pas considérées ici.

Les erreurs de traitement peuvent survenir à plusieurs endroits : saisie, transcription et codage. Nous supposerons aussi que ce type d'erreur est négligeable.

L'erreur due à la non réponse est en général causée par l'impossibilité de contacter le répondant ou par le refus de celui-ci d'y répondre partiellement ou complètement.

Les erreurs non dues à l'échantillonnage peuvent parfois être plus importantes que l'erreur d'échantillonnage. C'est pour cela qu'il est souvent préférable de sélectionner un échantillon car plus le nombre d'unités à observer est grand, plus les erreurs non dues à l'échantillonnage risquent d'être grandes. En effet, un nombre élevé d'unités rendra plus difficile le suivi des répondants et entraînera donc une augmentation de la non-réponse. De plus, le volume de données à traiter lorsque nous avons un nombre élevé de répondants risque d'augmenter les erreurs de mesure et de traitement.

Il est en général préférable de prévenir la non-réponse au cours de l'enquête plutôt que de la traiter par la suite. Pour prévenir la non-réponse, il est important de bien planifier la collecte des données, avoir une base de sondage à jour, un questionnaire simple et bien écrit. Beaucoup de facteurs influencent la non-réponse dont : la période de l'année durant laquelle l'enquête est faite, l'heure des entrevues, la compétence des interviewers, la méthode de collecte, le questionnaire, le fardeau de réponse (longueur du questionnaire), le suivi des répondants et les

mesures incitatives. Une enquête pilote permettra de détecter des problèmes potentiels d'une enquête et ces problèmes pourront être corrigés avant que la vraie enquête soit entreprise.

Toutefois, malgré toutes les mesures de prévention, il y aura toujours une non-réponse qu'il faudra traiter. En général, la non-réponse cause un biais dans les estimateurs et augmente leur variabilité. Il s'agira alors d'utiliser des méthodes de traitement permettant de réduire le biais.

## 1.2. Échantillonnage à partir d'une population finie

Nous faisons un bref survol des concepts de l'échantillonnage à partir d'une population finie. Le lecteur intéressé à un traitement plus détaillé est invité à consulter Särndal, Swensson et Wretman (1992).

Soit $U$ une population finie composée de $N$ unités. Dans le cadre des enquêtes, le problème consiste souvent à estimer le total $Y = \sum_{i \in U} y_i$ ou la moyenne $\bar{Y} = Y/N$ d'une variable d'intérêt $y$ à partir d'un échantillon aléatoire $s \subset U$. Cet échantillon est choisi selon un plan de sondage $p(.)$. Soit $\mathcal{S}$ l'ensemble des échantillons possibles de $U$. Pour chaque échantillon possible $s \in \mathcal{S}$, $p(s)$ dénote sa probabilité de sélection. Nous avons donc $\sum_{s \in \mathcal{S}} p(s) = 1$. Dans le cas où $p(s = U) = 1$, toutes les unités sont choisies avec probabilité 1 et nous sommes en présence d'un recensement.

Pour chaque unité $i \in U$, on désigne par $I_i$ la variable indicatrice de sélection définie par

$$I_i = \begin{cases} 1, & \text{si } i \in s, \\ 0, & \text{si } i \notin s. \end{cases}$$

De plus, pour chaque unité $i \in U$, on désigne par $\pi_i$ sa probabilité d'inclusion :

$$\pi_i = P[i \in s] = \sum_{s \in \mathcal{S}, i \in s} p(s).$$

On remarque que $E_p(I_i) = \pi_i$ et $V_p(I_i) = \pi_i(1 - \pi_i)$.

La taille de l'échantillon peut être aléatoire ou fixe selon le plan de sondage. Un exemple de plan à taille fixe est l'échantillonnage aléatoire simple sans remise. Soit $U$ une population de taille $N$ et $n$ un entier fixe tel que $n \in \{1, 2, ..., N\}$.

L'échantillonnage aléatoire simple sans remise de taille $n$ est le cas où n'importe quel sous-ensemble de $s \subset U$ de taille $n$ a la même probabilité d'être tiré que n'importe quel autre sous-ensemble de taille $n$. Puisqu'il y a $\binom{N}{n}$ sous-ensembles de $U$ de taille $n$, chaque échantillon $s$ de taille $n$ a une probabilité $\binom{N}{n}^{-1}$ d'être tiré. Sous ce plan, chaque unité $i \in U$ a une probabilité d'inclusion égale à $\pi_i = n/N$.

Un exemple de plan à taille aléatoire est l'échantillonnage de Bernoulli. Supposons que pour une population $U$ de taille $N$, chaque unité $i \in U$ est choisie de façon indépendante avec probabilité de sélection $\pi \in [0, 1]$. Alors, nous sommes en présence de l'échantillonnage de Bernouilli.

Un estimateur de $Y$ souvent utilisé est l'estimateur d'Horvitz-Thompson, $\hat{Y}_\pi$, défini par

$$\hat{Y}_\pi = \sum_{i \in s} d_i y_i = \sum_{i \in U} d_i I_i y_i,$$

où $d_i = \pi_i^{-1}$ désigne le poids de sondage de l'unité $i$. L'estimateur $\hat{Y}_\pi$ a l'avantage d'être sans biais pour le total $Y$ sous le plan $p$ pourvu que $\pi_i > 0$ pour tout $i \in U$. On a

$$E_p(\hat{Y}_\pi) = E_p\left(\sum_{i \in U} \frac{I_i}{\pi_i} y_i\right) = \sum_{i \in U} \frac{1}{\pi_i} y_i E(I_i) = \sum_{i \in U} \frac{1}{\pi_i} y_i \pi_i = Y.$$

Si on est intéressé à la moyenne de la population $\bar{Y} = Y/N$ où $N$ est la taille de la population, alors l'estimateur $\bar{y}_\pi = \hat{Y}_\pi/N$ est sans biais pour $\bar{Y}$. Si la taille de la population est inconnue, on peut l'estimer par $\hat{N}_\pi = \sum_{i \in s} d_i$, et on utilise l'estimateur $\tilde{y}_\pi = \hat{Y}_\pi/\hat{N}_\pi$ qui est asymptotiquement sans biais pour $\bar{Y}$.

Souvent de l'information auxiliaire est disponible pour toutes les unités échantillonnées. Plus précisément, pour chaque $i \in s$ on dispose d'un vecteur de $q$ variables auxiliaires

$$\mathbf{z}_i = (z_{1i}, z_{2i}, .., z_{qi})'.$$

De plus, on suppose que le vecteur des totaux correspondant aux variables auxiliaires, $\mathbf{Z} = (Z_1, Z_2, ..., Z_q)'$ est connu au niveau de la population, où $Z_k = \sum_{i \in U} z_{ki}, k = 1, 2, ..., q$.

## 1.3. L'échantillonnage à deux phases

L'échantillonnage à deux phases est une technique utile lorsque la base de sondage contient peu d'information auxiliaire. En première phase, on sélectionne un échantillon $s_1$ suivant un plan de sondage $p_1(.)$. Une variable auxiliaire $z_i$ en général peu coûteuse est alors observée pour les unités $i \in s_1$. Il y a alors la deuxième phase, où l'on choisit un échantillon $s_2 \subset s_1$ suivant un plan de sondage $p_2(.|s_1)$ qui dépend de $s_1$. Pour les unités $i \in s_2$ on observe alors la variable d'intérêt $y_i$ qui est en général plus coûteuse à observer.

## 1.4. Le mécanisme de non-réponse

Nous rappelons que nous considérons la non-réponse comme un phénomène aléatoire. L'ensemble des répondants peut être vu comme l'échantillon de deuxième phase avec la différence que les probabilités de sélection (c'est-à-dire les probabilités de réponses) ne sont pas connues et contrôlées. Schématiquement on a :

$$U \to s \to (s_r, s_m),$$

où $s$ désigne l'échantillon, $s_r$ l'ensemble des unités sélectionnées répondantes et $s_m$ l'ensemble des unités sélectionnées non-répondantes. On a donc $s = s_r \cup s_m$.

Nous définissons la variable indicatrice de réponse $r_i$, $i \in U$, comme suit :

$$r_i = \begin{cases} 1, & \text{si } i \in s_r, \\ 0, & \text{si } i \notin s_r. \end{cases}$$

La distribution des variables indicatrices, $P[r_i|s]$, est appelée mécanisme de non-réponse. On désigne par $q(s_r)$ le mécanisme de non-réponse.

En général, la probabilité de réponse d'une unité dépend de l'échantillon $s$. Dans notre cas, nous supposerons que cette probabilité ne dépend pas de $s$. Nous supposerons aussi que les réponses des différentes unités sont indépendantes. Bien que cette supposition ne soit pas toujours vraie (par exemple, lorsque l'accès à un immeuble contenant plusieurs unités choisies est bloqué), dans la pratique elle est souvent vérifiée.

En général, nous disposons de vecteurs de variables auxiliaires $\mathbf{z}_i = (z_{1i}, ..., z_{qi})'$ pour toutes les unités échantillonnées. Souvent, on émet l'hypothèse que conditionnellement aux variables auxiliaires données, les probabilités de réponse des unités ne dépendent pas de la variable d'intérêt $y$. C'est-à-dire que

$$P(r_i = 1|\mathbf{z}_i, y) = P(r_i = 1|\mathbf{z}_i).$$

Dans ce cas, on dit que le mécanisme de non-réponse est ignorable (parfois dénotée MAR pour *missing at Random*).

Dans certains situations on suppose l'hypothèse encore plus forte d'uniformité des probabilités de réponse. Dans ce cas le mécanisme de non-réponse est dit uniforme (parfois dénoté MCAR pour *Missing completely at Random*).

Fay (1991) a proposé une approche alternative à l'approche deux phases. Celle-ci consiste à renverser l'ordre du mécanisme d'échantillonnage et du mécanisme de non-réponse. Cette approche est appelée "approche renversée". On suppose que le processus est le suivant : la population $U$ est divisée en une population $U_r$ de répondants et une population $U_m$ de non-répondants. Ensuite, à partir de $(U_r, U_m)$ est tiré l'échantillon qui contient les répondants et les non-répondants. Schématiquement, on a :

$$U \rightarrow (U_r, U_m) \rightarrow (s_r, s_m).$$

Cette approche sera utile lors de l'estimation de la variance. Elle permet, entre autres, de clarifier et de justifier certaines propriétés théorique d'estimateurs de variances obtenus à l'aide de méthodes de réplications en présences de données imputées.

## 1.5. L'IMPUTATION

L'imputation consiste à remplacer les valeurs manquantes par une ou plusieurs valeurs artificielles. Ici, nous ne considérerons que le cas où l'on impute qu'une seule valeur pour chaque valeur manquante. C'est le cas de l'imputation simple.

La majorité des méthodes d'imputation peuvent être représentées par le modèle suivant (Kalton et Kasprzyk, 1986),

$$m: \qquad y_i = f(\mathbf{z}_i) + \epsilon_i \tag{1.1}$$

où $E_m(\epsilon_i) = 0, E_m(\epsilon_i\epsilon_j) = 0, i \neq j, V_m(\epsilon_i) = \sigma_i^2$, et $\mathbf{z} = (z_1, ..., z_q)'$ est un vecteur de variables auxiliaires disponibles pour toutes les unités dans l'échantillon $s$. Dans le cas des méthodes déterministes, la valeur imputée $y_i^\star$ est obtenue en estimant la fonction $f$ par $\hat{f}$ au moyen des unités répondantes, et en posant $y_i^\star = \hat{f}(\mathbf{z}_i)$ . L'imputation aléatoire peut être vue comme une imputation déterministe à laquelle on a ajouté un résidu aléatoire. Ce résidu peut être tiré, par exemple, à partir d'une loi normale de moyenne 0 et de variance $\sigma^2$. En pratique, on préfère plutôt utiliser un résidu aléatoire tiré au hasard parmi les résidus observés dans l'ensemble des répondants. Soit $e_j = \frac{1}{\hat{\sigma}_j}\left[y_j - \hat{f}(\mathbf{z}_j)\right]$ le résidu standardisé pour le répondant $j$, où $\hat{\sigma}_j$ est un estimateur de $\sigma_j$, et soit $\bar{e}_r = (\sum_{k\in s_r} d_k e_k)/(\sum_{k\in s_r} d_k)$. La valeur manquante pour la $i$ème unité, est alors remplacée par

$$y_i^\star = \hat{f}(\mathbf{z}_i) + \hat{\sigma}_i \epsilon_i^\star,$$

où $\epsilon_i^\star = e_i^\star - \bar{e}_r$ est tel que $e_i^\star$ est tiré au hasard (habituellement avec remise), dans l'ensemble des résidus standardisés correspondant aux répondants, avec probabilité

$$P\left(e_i^* = e_j\right) = d_j / \left(\sum_{k\in s_r} d_k\right) \text{ pour } j \in s_r.$$

## 1.6. Quelques méthodes d'imputation déterministes

### 1.6.1. L'imputation par la régression

Supposons que l'on dispose d'un vecteur de variables auxiliaires $\mathbf{z}_i = (z_{1i}, ..., z_{qi})'$ pour chaque unité $i$ dans l'échantillon $s$. Supposons aussi que le modèle (1.1) est $f(\mathbf{z}_i) = \mathbf{z}_i'\boldsymbol{\beta}$, où $\boldsymbol{\beta}$ est un vecteur de paramètres inconnus et $\sigma_i^2 = \sigma^2\boldsymbol{\lambda}'\mathbf{z}_i$ où $\boldsymbol{\lambda}$ est un vecteur constant. La méthode d'imputation par la régression est très courante. La valeur imputée est donnée par

$$y_i^* = \mathbf{z}_i'\hat{\mathbf{B}}_r,$$

où

$$\hat{\mathbf{B}}_r = \left( \sum_{i \in s} d_i r_i \mathbf{z}_i \mathbf{z}_i' / \left( \boldsymbol{\lambda}' \mathbf{z}_i \right) \right)^{-1} \sum_{i \in s} d_i r_i \mathbf{z}_i y_i / \left( \boldsymbol{\lambda}' \mathbf{z}_i \right).$$

L'estimateur imputée de $Y$ est donc

$$\hat{Y}_I = \sum_{i \in s_r} d_i y_i + \sum_{i \in s_m} d_i y_i^{\star}.$$

On peut montrer que

$$\hat{Y}_I = \hat{\mathbf{Z}}_{\pi}' \hat{\mathbf{B}}_r,$$

où $\hat{\mathbf{Z}}_{\pi} = \sum_{i \in s} d_i \mathbf{z}_i$.

### 1.6.2. L'imputation par le ratio

L'imputation par le ratio est un cas particulier de l'imputation par la régression lorsque $\mathbf{z}_i$ est un scalaire et $\sigma_i^2 = \sigma^2 z_i$. On a :

$$y_i^* = \frac{\hat{Y}_r}{\hat{Z}_r} z_i,$$

où $\hat{Y}_r = \sum_{i \in s_r} d_i y_i$ et $\hat{Z}_r = \sum_{i \in s_r} d_i z_i$.

### 1.6.3. L'imputation par la moyenne

L'imputation par la moyenne est le cas particulier de l'imputation par la régression lorsque $z_i = 1$ (essentiellement nous sommes dans le cas où nous n'utilisons pas de variable auxiliaire). On a alors

$$y_i^* = \bar{y}_r := \frac{\sum_{i \in s_r} d_i y_i}{\sum_{i \in s_r} d_i}.$$

### 1.6.4. L'imputation par le plus proche voisin

Cette méthode consiste à remplacer la valeur manquante $y_i$ par la valeur $y_j$ de l'unité répondante $j$ qui est la plus proche de $i$ selon une distance $d(\mathbf{z}_i, \mathbf{z}_j)$ qui dépend seulement des variables auxiliaires.

## 1.7. Quelques méthodes d'imputation aléatoires

### 1.7.1. Imputation par hot-deck

Cette méthode est un exemple d'imputation par donneur. Pour une unité $i$ dont la valeur $y_i$ est manquante on choisit une unité $j$ au hasard parmi les unités répondantes $s_r$ et on remplace $y_i$ par $y_j$. Plus précisément $y_i^\star = y_j$ ou $j$ est choisie avec probabilité

$$P(y_i^\star = y_j) = d_j/(\sum_{k\in s_r} d_k).$$

On remarque que

$$y_i^\star = y_j = \bar{y}_r + (y_j - \bar{y}_r).$$

Essentiellement, on a donc imputé la moyenne des répondants à laquelle on a ajouté un résidu, $e_j = y_j - \bar{y}_r$.

### 1.7.2. L'imputation par la régression avec résidus

Cette méthode est équivalente à l'imputation par la régression déterministe à laquelle nous ajoutons un résidu aléatoire tiré avec remise parmi l'ensemble des résidus observés. La valeur imputée utilisée pour la valeur manquante $y_i$ pour une unité non-répondante est donc

$$y_i^* = \mathbf{z}_i'\hat{\mathbf{B}}_r + \left(\boldsymbol{\lambda}'\mathbf{z}_i\right)^{1/2}\epsilon_i^*,$$

où $\epsilon_i^\star = e_i^* - \bar{e}_r$ est tel que $P\left(e_i^* = e_j\right) = d_j/(\sum_{k\in s_r} d_k)$ pour $j \in s_r$, $e_j = \left(\boldsymbol{\lambda}'\mathbf{z}_j\right)^{-1/2}\left(y_j - \mathbf{z}_j'\hat{\mathbf{B}}_r\right)$ et $\bar{e}_r = (\sum_{k\in s_r} d_k e_k)/(\sum_{k\in s_r} d_k)$.

## 1.8. Inférence en présence d'imputation simple

L'inférence en présence de données imputées doit prendre en compte différents niveaux d'aléas : l'échantillonnage selon le plan $p(s)$, le mécanisme de non-réponse $q(s_r)$ ainsi que l'aléas due à la méthode d'imputation lorsque celle-ci est aléatoire. En plus des hypothèses émises à la section 1.4, des hypothèses supplémentaires sont nécessaires et il y a deux approches différentes généralement utilisées.

La première approche est l'approche par modèle de non-réponse. Sous cette approche on suppose que le mécanisme de non-réponse est uniforme. Cette hypothèse n'est pas réaliste. Toutefois on peut séparer les unités par classes d'imputation et alors l'hypothèse que le mécanisme de non-réponse est uniforme à l'intérieur de chaque classe est plausible et sera supposée vraie. Les méthodes de construction de classes ainsi que leurs justifications théoriques se trouvent dans Haziza et Beaumont (2007) et Little (1986).

La seconde approche est l'approche par modèle d'imputation. Sous cette approche, en plus de supposer que le mécanisme de non-réponse est ignorable, on suppose que le modèle suivant est valide,

$$m: \qquad y_i = \mathbf{z}_i'\boldsymbol{\beta} + \epsilon_i$$

où $E_m(\epsilon_i) = 0, V_m(\epsilon_i) = \sigma^2\boldsymbol{\lambda}'\mathbf{z}_i$ et $E_m(\epsilon_i\epsilon_j) = 0$ si $i \neq j$.

## 1.9. Asymptotique dans le contexte des enquêtes

L'étude des propriétés asymptotiques dans le cas des populations finies est différente du cas classique. En effet, lorsque nous avons une population finie $U$ de taille $N$, la taille de l'échantillon $n$ est limitée par la taille de la population $N$ ( c'est-à-dire $n \leq N$). Pour remédier à ce problème, nous considérons une suite de populations finies $\{U_j\}_{j\in\mathbb{N}}$, de taille, respectivement, $N_1, N_2, N_3, ...$ telles que $U_1 \subset U_2 \subset U_3...$, $N_1 < N_2 < N_3 < ...$ Pour chaque $\nu \in \mathbb{N}$, on choisit un échantillon $s_\nu \subset U_\nu$. Toutefois il n'est pas requis que $s_\nu \subset s_{\nu+1}$. Soit $n_\nu$ la taille de l'échantillon $s_\nu$. Si $n_\nu$ n'est pas aléatoire, c'est-à-dire $s_\nu$ est de taille fixe, on suppose que $n_\nu \to \infty$ lorsque $N \to \infty$. De façon plus générale, lorsque la taille de l'échantillon est aléatoire alors on supposera que $E(n_\nu) \to \infty$.

Nous rappelons quelques notations qui seront utilisées dans l'article :

**Définition 1.9.1.** *Soit $X_n$ une suite de variables aléatoires et $h_n$ une suite de nombres positifs. On écrit $X_n = o_p(h_n)$ si $\frac{X_n}{h_n} \to 0$ en probabilité. C'est-à-dire que pour tout $\epsilon > 0$, on a*

$$\lim_{n\to\infty} P\left(\left|\frac{X_n}{h_n}\right| > \epsilon\right) = 0.$$

**Définition 1.9.2.** *Soit $X_n$ une suite de variables aléatoires et $h_n$ une suite de nombres de positifs.On écrit $X_n = O_p(h_n)$ si pour tout $\epsilon > 0$, il existe un nombre $M_\epsilon > 0$ tel que*

$$P\left(\left|\frac{X_n}{h_n}\right| > M_\epsilon\right) \leq \epsilon$$

*pour tout $n = 1, 2, 3, ....$*

Les résultats suivants seront utiles dans la preuve des résultats asymptotiques dans l'annexe B de l'article.

**Lemme 1.9.1.** *(Théorème Central Limite) Supposons que pour tout entier naturel $T$, nous avons que $\{X_t\}_{t=1}^T$ sont des variables aléatoires indépendantes (n'ayant pas nécessairement la même distribution) telles que $E(X_t) = \mu_t$ et $V(X_t) = \sigma_t^2$. Si $\frac{1}{T}\sum_{t=1}^T \sigma_t^2 \to \sigma^2$ et $\frac{1}{T}\sum_{t=1}^T E|X_t - \mu_t|^{2+\delta} = O(1)$ quand $T \to \infty$ pour un certain $\delta > 0$, alors*

$$\sqrt{T}\left(\bar{X} - \bar{\mu}\right) \to_d N\left(0, \sigma^2\right)$$

*où $\bar{X} = \frac{1}{T}\sum_{t=1}^T X_t$ et $\bar{\mu} = \frac{1}{T}\sum_{t=1}^T \mu_t$.* (On remarque que nous avons un tableau triangulaire comme dans la plupart des versions du Théorème central limite de Lindeberg-Feller ; nous avons omis l'indice de rangée dans le but d'alléger la notation.)

**Lemme 1.9.2.** *(Théorème de Slutsky) Supposons que $\{X_n\}, \{Y_n\}$ et $X$ sont des variables aléatoires et que $c$ est une constante. Si $X_n \to_d X$ et $Y_n \to_p c$, alors*

*(i)* $X_n + Y_n \to_d X + c$ ;

*(ii)* $X_n Y_n \to_d cX$.

**Théorème 1.9.1.** *(Chen and Rao, 2007) Soient $\{U_n\}$ et $\{V_n\}$ deux suites de variables aléatoires et $\Im_n$ une suite de tribus. Supposons que*

1. *il existe $\sigma_{1n} > 0$ tel que*

$$\sigma_{1n}^{-1} V_n \to_d N(0, 1)$$

*quand $n \to \infty$, et $V_n$ est $\Im_n$-mesurable.*

2. *$E(U_n|\Im_n) = 0$ et $\sigma_{2n}^2 = \sigma_{2n}^2(\Im_n) > 0$ tels que*

$$\sup_t \left|P\left(\sigma_{2n}^{-1} U_n \leq t|\Im_n\right) - \Phi(t)\right| = o_p(1),$$

*où* $\Phi(.)$ *est la fonction de distribution d'une loi normale centrée réduite.*

3. $\gamma_n^2 = \sigma_{1n}^2/\sigma_{2n}^2 = \gamma^2 + o_p(1)$.

*Alors, quand* $n \to \infty$

$$\frac{U_n + V_n}{\sqrt{\sigma_{1n}^2 + \sigma_{2n}^2}} \to_d N(0,1).$$

# Chapitre 2

## L'ARTICLE

Voici l'article de David Haziza et Frédéric Picard ayant pour titre *Doubly Robust Point and Variance Estimation in the Presence of Imputed Data.*

**ABSTRACT**

Imputation is often used in surveys to treat item nonresponse. It is well known that treating the imputed values as observed values may lead to substantial underestimation of the variance of the point estimators. To overcome the problem, a number of variance estimation methods have been proposed in the literature, including appropriate versions of resampling methods such as the jackknife and the bootstrap. We define the concept of doubly robust point and variance estimation under the so-called nonresponse and imputation model approaches. We focus on jackknife variance estimation which is widely used in practice. We study the properties of several jackknife variance estimators under both deterministic and random regression imputation. We first consider the case of simple random sampling without replacement. The cases of stratified simple random sampling and unequal probability sampling are also considered. A limited simulation study compares various jackknife variance estimators in terms of bias and relative stability when the sampling fraction is not negligible. Finally, the asymptotic normality of the imputed estimator is established under both deterministic and random regression imputation.

**KEY WORDS** : Double robustness ; Imputation model approach ; Item non-response ; Jackknife variance estimator ; Linearization variance estimator ; Non-response model approach ; Reverse framework ; Regression imputation ; Two-phase framework.

---

[1] D. Haziza, Département de Mathématiques et de Statistique,Université de Montréal, Montréal, QC, Canada H37 3J7 ; F. Picard, Business Survey Methods Division, Statistics Canada, Ottawa, ON, Canada K1A 0T6.

## 1 Introduction

In recent years, doubly robust procedures have gained in popularity in several fields of statistics ; see Kang and Schafer (2008) and Robins etal. (2008), among others. In the survey sampling context and imputation for missing data, doubly robust estimation is discussed in Kott (1994), Rao (2000), Little and An (2004) and Haziza and Rao (2006), among others. In the presence of imputed data, two approaches may be used to study the properties of estimators and to derive valid variance estimators : the nonresponse model (NM) approach that requires the specification of a nonresponse model describing the nonresponse mechanism and the imputation model (IM) approach that requires the specification of an imputation model describing the distribution of the variable being imputed (see Section 2). An estimator is said to be doubly robust if it remains asymptotically unbiased and consistent if either model (nonresponse or imputation) is true (see Section 2). In our view, doubly robust procedures are attractive in the presence of missing data because in this context, point and variance estimators may be significantly biased if the underlying (nonresponse or imputation) model is not correctly specified. In other words, doubly robust procedures offer the survey statistician protection against misspecification of one model or the other. In this paper, we examine the problem of doubly robust point and variance estimation in the presence of deterministic regression imputation (DREGI), that includes ratio and mean imputation as special cases, and of random regression imputation (RREGI) that includes random hot-deck imputation (RHDI) as a special case (e.g., Haziza, 2009).

The problem of variance estimation in the presence of imputation has been widely treated in the literature. It is well known that treating the imputed values as if they were real observations will typically result in estimated variances (or coefficients of variation) that are usually too small because they fail to account for the variance due to nonresponse and imputation. As a result, inferences are generally invalid. For example, the coverage probability of 95% confidence intervals obtained by using naive variance estimation methods may be considerably

smaller than the nominal level, especially if the nonresponse rate is high. This led researchers to develop variance estimation methods that account for the variance due to nonresponse and imputation.

Traditionally, the total variance of the imputed estimator has been expressed as the sum of the sampling variance and the nonresponse variance. This decomposition of the total variance results from viewing nonresponse as a second-phase of selection. For this reason, this framework is often called the two-phase framework ; see Rao (1990), Särndal (1992) and Deville and Särndal (1994), among others. In this paper, we consider an alternative framework, which we call the *reverse framework* ; see Fay (1991) and Shao and Steel (1999). It consists of viewing the situation prevailing in the presence of nonresponse as follows : first, applying the nonresponse mechanism, the finite population $U$ is randomly divided into a population of respondents $U_r$ and a population of nonrespondents $U_m$. Then, given $(U_r, U_m)$, a sample $s$, containing both respondents and nonrespondents, is selected from $U$ according to the chosen sampling design. As we argue in section 3, the reverse framework is attractive because it facilitates the derivation of doubly robust variance estimators. Also, unlike the two-phase framework, it helps to clarify and to justify the theoretical properties of variance estimators obtained using replication methods in the presence of imputed data. The reader is referred to Rao and Shao (1992), Rao and Sitter (1995), Shao and Sitter (1996), Shao (2002) and Davison and Sardy (2007) for a description of replication methods in the presence of imputed data. In practice, replication methods are often used because, unlike Taylor linearization procedures, they do not require separate derivation for each particular estimator nor do they require second-order inclusion probabilities that may be difficult to obtain for complex designs.

In this paper, we focus on jackknife variance estimation, which is widely used in practice. It is well known that naive jackknife estimators (those that treat imputed values as if they were observed values) underestimate the true variance. To

overcome this problem, Rao and Shao (1992) proposed a jackknife variance estimator which is similar to the complete data jackknife variance estimator, except that whenever a responding unit is deleted, the imputed values are adjusted ; see also Rao and Sitter (1995). The reverse framework helps in clarifying the following points : (i) The Rao-Shao jackknife variance estimator is asymptotically unbiased and consistent for the total variance if the units are selected with replacement, or equivalently, if the (overall) sampling fraction is negligible. This property holds regardless of the validity of the assumed (nonresponse or imputation) model. As a result, it is doubly robust. Some authors (e.g. Rao and Shao, 1992 and Brick et al., 2005) showed/argued that the Rao-Shao jackknife variance estimator is consistent under a uniform nonresponse mechanism. Although this is true, the consistency property holds even when units have unequal response probabilities. As we argue in section 4, consistency follows from standard regularity conditions used in the complete data set-up. (ii) When the (overall) sampling fraction is not negligible, the Rao-Shao jackknife variance estimator tends to overestimate the the true variance. To overcome this difficulty, Lee, Rancourt and Särndal (1995) (henceforth, LRS) proposed an alternative variance estimator and, although this estimator was evaluated empirically in several papers (e.g., Brick etal., 2005, Hurtubise (2006)), its theoretical properties were not, to our knowledge, fully evaluated. The reverse framework facilitates the study of the properties of the LRS estimator and leads to alternative variance estimators. (iii) Under deterministic imputation, no adjustment of the imputed values is necessary if an appropriate standard jackknife procedure is applied. As a result, the Rao-Shao jackknife variance estimator can be obtained using software designed for complete data jackknife variance estimation. This is an attractive property from a practical point of view. (iv) Rao (1996) proposed a linearized jackknife variance estimator in the presence of imputed data which is obtained from the Rao-Shao jackknife variance estimator by performing a first-order Taylor expansion. As we discuss in section 4, the linearized jackknife variance estimator can be obtained by performing a complete data first-order Taylor expansion to approximate the first term of the variance under the reverse

framework.

The outline of the paper is as follows : in section 2, we first describe two approaches for inference : the NM and the IM approaches. Then, we define the imputed estimator of a total under DREGI and discuss the concept of double robustness. Finally, adapting a result in Chen and Rao (2007),the asymptotic normality of the imputed estimator is established. In section 3, we derive linearization variance estimators using the reverse framework. Jackknife variance estimation under simple random sampling without replacement (SRSWOR) is treated in section 4. First, the case of a negligible sampling fraction is considered. Then, several variance estimators in the case of nonnegligible sampling fractions are examined, including the LRS variance estimator. Section 4 ends with a brief discussion of jackknife variance estimation in the context of stratified sampling. Section 5 presents the results of a limited simulation study that compares the performance of several variance estimators in terms of relative bias and relative efficiency. The case of RREGI is considered in section 6. In section 7, extending results of Berger (2007), we propose a jackknife variance estimator in the context of unequal probability sampling designs. Finally, we conclude and discuss some possible extensions in section 8.

## 2 Theoretical set-up

Let $U = \{1, 2, ..., N\}$ be a population of $N$ identifiable elements. We consider the problem of estimating a population total $Y = \sum_{i \in U} y_i$, where $y_i$ denotes the $i$-th value of the variable of interest $y$, $i = 1, ..., N$. To that end, we select a sample, $s$, of size $n$, according to a given sampling design $p(s)$. Let $\pi_i$ denote the first-order inclusion probability of unit $i$ in the sample and let $d_i = 1/\pi_i$ denote its design weight. In the absence of nonresponse, a basic estimator of $Y$ is the expansion estimator GIVEN BY

$$\hat{Y}_\pi = \sum_{i \in s} d_i y_i. \tag{2.1}$$

The estimator $\hat{Y}_\pi$ in (2.1) is $p$-unbiased for $Y$ ; that is, $E_p\left(\hat{Y}_\pi\right) = Y$, where $E_p(.)$ denotes the expectation with respect to the sampling design $p(s)$. However, the

calculation of (2.1) requires the knowledge of the $y$-values for all the sample units. Let $y_i^*$ denote the imputed value for missing $y_i$. In the presence of nonresponse to item $y$, we define an imputed estimator $\hat{Y}_I$ by

$$\hat{Y}_I = \sum_{i \in s} d_i r_i y_i + \sum_{i \in s} d_i \left(1 - r_i\right) y_i^* = \sum_{i \in s} d_i \tilde{y}_i, \tag{2.2}$$

such that $r_i = 1$ if unit $i$ responds to item $y$ and $r_i = 0$, otherwise and $\tilde{y}_i = r_i y_i + \left(1 - r_i\right) y_i^*$.

In this paper, we consider regression imputation that can be motivated by the model

$$\begin{aligned} y_i &= \mathbf{z}_i' \boldsymbol{\beta} + \epsilon_i, \\ E_m\left(\epsilon_i\right) &= 0, Cov_m\left(\epsilon_i, \epsilon_j\right) = 0 \text{ if } i \neq j, V_m\left(\epsilon_i\right) \equiv \sigma_i^2 = \sigma^2\left(\boldsymbol{\lambda}' \mathbf{z}_i\right), \end{aligned} \tag{2.3}$$

where $\mathbf{z} = \left(z_1, ..., z_q\right)'$ is a $q$-vector of auxiliary variables available at the imputation stage for all the sample units, $\boldsymbol{\beta}$ is a $q$-vector of unknown parameters, $\sigma^2$ is an unknown parameter and $\boldsymbol{\lambda}$ is a vector of known constants. The subscript $m$ indicates that the expectations, variances and covariances are evaluated with respect to the model (2.3), which is often called an imputation model (e.g., Särndal, 1992). Under DREGI, the imputed values are given by

$$y_i^* = \mathbf{z}_i' \hat{\mathbf{B}}_r, \tag{2.4}$$

where

$$\hat{\mathbf{B}}_r = \left( \sum_{i \in s} d_i r_i \mathbf{z}_i \mathbf{z}_i' / \left(\boldsymbol{\lambda}' \mathbf{z}_i\right) \right)^{-1} \sum_{i \in s} d_i r_i \mathbf{z}_i y_i / \left(\boldsymbol{\lambda}' \mathbf{z}_i\right) \tag{2.5}$$

is the weighted least square estimator of $\boldsymbol{\beta}$ based on the responding units. Since $\sigma_i^2 = \sigma^2\left(\boldsymbol{\lambda}' \mathbf{z}_i\right)$, the imputed estimator (2.2) under DREGI reduces to

$$\hat{Y}_I = \sum_{i \in s} d_i \mathbf{z}_i' \hat{\mathbf{B}}_r. \tag{2.6}$$

Note that the form of the imputed estimator (2.6) is similar to that of a projection regression estimator in the context of two-phase sampling.

Different approaches may be used to evaluate the properties of the imputed estimator (2.6) and to derive corresponding variance estimators. To understand the nature of these approaches, we identify three sources of randomness : (i) the imputation model $m$ which generates the vector of $y$-values, $\mathbf{y} = (y_1, ..., y_N)'$; (ii) the sampling design $p(s)$, which generates the vector of sample selection indicators, $\mathbf{I} = (I_1, ..., I_N)'$, where $I_i = 1$ if unit $i$ is selected in the sample and $I_i = 0$, otherwise and (iii) the nonresponse mechanism $q(\mathbf{r}|\mathbf{I})$, which generates the vector of response indicators, $\mathbf{r} = (r_1, ..., r_N)'$. Let $p_i = P(r_i = 1|\mathbf{I}, i \in s)$ be the response probability of unit $i$ to item $y$. We assume that $p_i > 0$ for all $i$. Also, we assume that the units respond independently of one another; that is, $p_{ij} = P(r_i = 1, r_j = 1|\mathbf{I}, i \in s, j \in s, i \neq j) = p_i p_j$. In this paper, we consider two approaches for inference : the Nonresponse Model (NM) approach and the Imputation Model (IM) approach described below :

*The NM approach* : explicit assumptions are made about the nonresponse mechanism, called the nonresponse model. Inferences are made with respect to the joint distribution induced by the sampling design and the assumed nonresponse model, while the vector of $y$-values, $\mathbf{y}$, is treated as fixed. The NM approach has been studied by many including Rao (1990, 1996), Rao and Sitter (1995), Shao and Steel (1999), Beaumont (2005), Kim and Park (2006) and Haziza and Rao (2006). In this paper, we use the within-class uniform nonresponse model that assumes a constant probability of response within imputation classes. For simplicity, we consider the case of a single imputation class but the extension to the case of multiple classes is relatively straightforward.

*The IM approach* : explicit assumptions are made about the distributions of the values of the variables of interest, called the imputation model. Here, inference is with respect to the joint distribution induced by the imputation model, the sampling design and the nonresponse model. Unlike the NM approach, the underlying nonresponse mechanism is not explicitly specified, although it is assumed to be unconfounded; e.g., Rubin (1976). The IM approach has been studied

by Särndal (1992), Deville and Särndal (1994) and Shao and Steel (1999), among others. Under (deterministic and random) regression imputation, we assume that the imputation model (2.3) holds.

To study the bias of the imputed estimator $\hat{Y}_I$, we use the following decomposition of the total error, $\hat{Y}_I - Y$ :

$$\hat{Y}_I - Y = \left(\hat{Y}_\pi - Y\right) + \left(\hat{Y}_I - \hat{Y}_\pi\right). \tag{2.7}$$

The term $\hat{Y}_\pi - Y$ in (2.7) is called the sampling error, whereas the term $\hat{Y}_I - \hat{Y}_\pi$ is called the nonresponse error. Under the NM approach, it can be shown that the imputed estimator $\hat{Y}_I$ is asymptotically $pq$-unbiased provided the sample units have the same response probability to item $y$. That is, $E_{pq}\left(\hat{Y}_I\right) \equiv E_p E_q\left(\hat{Y}_I|\mathbf{I}\right) \approx Y$. Also, under the IM approach, $\hat{Y}_I$ is $mpq$-unbiased for $Y$ provided the imputation model (2.3) holds. That is, $E_{mpq}\left(\hat{Y}_I\right) \equiv E_m E_p E_q\left(\hat{Y}_I|\mathbf{I},\mathbf{r}\right) = Y$ . Therefore, the estimator $\hat{Y}_I$ is valid if one model or the other (nonresponse model or imputation model) is correctly specified. This is related to the concept of doubly robustness, which we define next.

**Definition 1** : An estimator $\hat{Y}_I$ is said to doubly robust if

(i) $E_{pq}(\hat{Y}_I - Y)/Y \to 0$ in probability under the imputation model (2.3) ; and

(ii) suppose that the imputation is performed according to (2.3) but that the true imputation model is not (2.3) but rather

$$m^* : y_i = \mu_i + \epsilon_i$$

such that $E_{m^*}(\epsilon_i) = 0$, $Cov_{m^*}(\epsilon_i, \epsilon_j) = 0$ if $i \neq j$, $V_{m^*}(\epsilon_i) = \sigma_i^2$ and $\mu_i \neq \mathbf{z}_i'\boldsymbol{\beta}$ . Then, we have

$$(\hat{Y}_I - Y)/N \to 0 \text{ in probability.}$$

under the nonresponse mechanism.

It is shown in Appendix A that the estimator $\hat{Y}_I$ given by (2.6) is doubly robust in the sense of the definition above. In the remainder of the paper, we assume that the imputed estimator (2.6) is (asymptotically) unbiased under the

chosen mode of inference.

We now turn to the variance of the imputed estimator $\hat{Y}_I$ under the reverse framework. Under the NM approach and DREGI, the total variance of the imputed estimator $\hat{Y}_I$ can be expressed as

$$\begin{aligned} V\left(\hat{Y}_I\right) &= E_q V_p\left(\hat{Y}_I|\mathbf{r}\right) + V_q E_p\left(\hat{Y}_I|\mathbf{r}\right) \\ &\equiv V_1^{NM} + V_2^{NM}. \end{aligned} \tag{2.8}$$

Note that, unlike the terms obtained using the two-phase framework, those on the right hand side of (2.8) do not represent the sampling and the nonresponse components and there is no natural interpretation of these terms. However, as we argue in section 4, the reverse framework provides a theoretical basis for studying the properties of the Rao-Shao jackknife variance estimator. Under the IM approach and DREGI, the total variance of the imputed estimator $\hat{Y}_I$ can be expressed as

$$\begin{aligned} V\left(\hat{Y}_I - Y\right) &= E_m E_q V_p\left(\hat{Y}_I|\mathbf{r}\right) + E_q V_m E_p\left(\hat{Y}_I - Y|\mathbf{r}\right) \\ &\equiv V_1^{IM} + V_2^{IM}. \end{aligned} \tag{2.9}$$

Note that since the imputed estimator $\hat{Y}_I$ is assumed to be $mpq$-unbiased for $Y$, the term $V_q E_m E_p\left(\hat{Y}_I - Y|\mathbf{r}\right)$ is equal to zero and was thus omitted from (2.9). In the remainder of the paper, we use the generic notation $V_1$ to denote $V_1^{NM}$ or $V_1^{IM}$ and $V_2$ to denote $V_2^{NM}$ or $V_2^{IM}$ when a statement applies for both the NM approach and the IM approach.

## 3 Linearization variance estimators

In this section, we give expressions for a linearization variance estimator of $\hat{Y}_I$ under DREGI. As we argue below, linearization variance estimators are asymptotically unbiased and consistent for the total variance of the imputed estimator. For this reason, we use them as a reference to evaluate the properties of jackknife variance estimators (see Section 4). To obtain an estimator of the component

$V_1$, it suffices to estimate $V_p\left(\hat{Y}_I|\mathbf{r}\right)$, which represents the variance due to sampling conditional on the vector of response indicators $\mathbf{r}$. Thus, an estimator of $V_p\left(\hat{Y}_I|\mathbf{r}\right)$ can be obtained using standard variance estimation methods for the complete data case and standard software packages designed for complete data variance estimation. For example, we may use a first-order Taylor expansion to approximate the total error $\hat{Y}_I - Y$, which leads to

$$\hat{Y}_I - Y \approx \sum_{i\in s} d_i\hat{\xi}_i,$$

where

$$\hat{\xi}_i = r_i y_i + (1-r_i)\,\mathbf{z}_i'\hat{\mathbf{B}}_r + \left(\hat{\mathbf{Z}}_\pi - \hat{\mathbf{Z}}_r\right)'\hat{\mathbf{T}}_r^{-1}\frac{\mathbf{z}_i}{(\boldsymbol{\lambda}'\mathbf{z}_i)}r_i e_i$$

with $e_i = \left(y_i - \mathbf{z}_i'\hat{\mathbf{B}}_r\right)$, $\hat{\mathbf{Z}}_\pi = \sum_{i\in s} d_i\mathbf{z}_i$, $\hat{\mathbf{Z}}_r = \sum_{i\in s} d_i r_i\mathbf{z}_i$ and $\hat{\mathbf{T}}_r = \sum_{i\in s} d_i r_i\mathbf{z}_i\mathbf{z}_i'/(\boldsymbol{\lambda}'\mathbf{z}_i)$. Using the operator notation, an asymptotically $p$-unbiased estimator of $V_p\left(\hat{Y}_I|\mathbf{r}\right)$ is thus given by

$$v_1 = v\left(\hat{\xi}\right). \tag{2.10}$$

Under standard regularity conditions used in the complete data scenario, the estimator $v_1$ is asymptotically $pq$-unbiased and consistent for $V_1^{NM}$ as well as asymptotically $mpq$-unbiased and consistent for $V_1^{IM}$. In fact, the estimator $v_1$ is asymptotically unbiased and consistent regardless of the validity of the underlying model (nonresponse model or imputation model). Therefore, it is doubly robust. For example, in the case of SRSWOR, the estimator $v_1$ reduces to

$$v_1 = N^2\left(1-\frac{n}{N}\right)\frac{s_{\hat{\xi}}^2}{n}, \tag{2.11}$$

where $s_{\hat{\xi}}^2 = \frac{1}{n-1}\sum_{i\in s}\left(\hat{\xi}_i - \frac{\sum_{j\in s}\hat{\xi}_j}{n}\right)^2$.

Turning to the second component on the right hand side of (2.8) and (2.9), we use a first-order Taylor expansion to obtain

$$E_p\left(\hat{Y}_I - Y|\mathbf{r}\right) \approx \sum_{i\in U}\left[\mathbf{Z}'\mathbf{T}_r^{-1}\frac{r_i\mathbf{z}_i}{(\boldsymbol{\lambda}'\mathbf{z}_i)} - 1\right]y_i,$$

where $\mathbf{Z} = \sum_{i \in U} \mathbf{z}_i$ and $\mathbf{T}_r = \sum_{i \in U} r_i \mathbf{z}_i \mathbf{z}_i' / (\boldsymbol{\lambda}' \mathbf{z}_i)$. Neglecting higher order terms, we can approximate $V_2^{NM}$ by

$$V_2^{NM} \approx p\,(1-p)\,E_q \left[ \mathbf{Z}' \mathbf{T}_r^{-1} \left( \sum_{i \in U} r_i \mathbf{z}_i \mathbf{z}_i' \frac{E_i^2}{(\boldsymbol{\lambda}' \mathbf{z}_i)^2} \right) \mathbf{T}_r^{-1} \mathbf{Z} \right],$$

where $E_i = y_i - \mathbf{z}_i' \mathbf{B}_r$ with $\mathbf{B}_r = \mathbf{T}_r^{-1} \mathbf{t}_r$ and $\mathbf{t}_r = \sum_{i \in U} r_i \mathbf{z}_i y_i / (\boldsymbol{\lambda}' \mathbf{z}_i)$. Neglecting higher order terms, we can approximate $V_2^{IM}$ by

$$V_2^{IM} \approx \sigma^2 E_q \left[ \mathbf{Z}' \mathbf{T}_r^{-1} (\mathbf{Z} - \mathbf{Z}_r) \right]. \tag{2.12}$$

In order to estimate $V_2^{NM}$ and $V_2^{IM}$, we propose the following estimator :

$$v_2 = \hat{\sigma}^2 \hat{\mathbf{Z}}' \hat{\mathbf{T}}_r^{-1} \left( \hat{\mathbf{Z}} - \hat{\mathbf{Z}}_r \right), \tag{2.13}$$

where $\hat{\sigma}^2 = \frac{\sum_{i \in s} d_i r_i e_i^2}{\sum_{i \in s} d_i r_i \left( \boldsymbol{\lambda}' \mathbf{z}_i \right)}$ is an estimator of the model parameter $\sigma^2$. The estimator $v_2$ is asymptotically $pq$-unbiased and consistent for $V_2^{NM}$ if the nonresponse model is correctly specified as well as asymptotically $mpq$-unbiased and consistent for $V_2^{IM}$ if the imputation is correctly specified. Therefore, it is doubly robust. Rao (2000) discussed the double robustness property of (2.13) in the special case of deterministic ratio imputation. Finally, a doubly robust estimator of $V\left(\hat{Y}_I\right)$ is given by

$$v_t = v_1 + v_2. \tag{2.14}$$

Note that, unlike $v_1$, the component $v_2$ does not require the second-order inclusion probabilities $\pi_{ij}$ ; only the first-order inclusion probabilities are needed. Also, note that $v_2$ will tend to be small if the imputation model has good predictive power (which is the case if $\sigma^2$ is small) or if the response rate is low. Finally, under mild regularity conditions, it can be shown that the first component $v_1$ is $O_p\left(\frac{N^2}{n}\right)$, whereas the second component $v_2$ is $O_p(N)$. It follows that the contribution of $v_2$ to the total variance, $C(v_2) = \frac{v_2}{v_1 + v_2}$, is $O_p\left(\frac{n}{N}\right)$ and is thus negligible when the sampling fraction $n/N$ is negligible. In this case, we can omit this component from the calculations.

An 95% confidence interval for $Y$ is thus given by

$$\hat{Y}_I \pm 1.96 \sqrt{v_t}, \tag{2.15}$$

where $v_t$ is given by (2.14). The coverage probability of (2.15) is close to the nominal rate if the following criteria are met : (i) the asymptotic distribution of $\hat{Y}_I$ is normal ; (ii) the estimator $\hat{Y}_I$ is asymptotically unbiased for $Y$ and (iii) the variance estimator $v_t$ is consistent for the true variance of $\hat{Y}_I$. Instead of $v_t$, we could use any other consistent variance estimator ; for example a jackknife variance estimator (see section 4). In Appendix B, we establish the asymptotic normality of $\hat{Y}_I$ in the context of stratified multistage designs under both DREGI and RREGI. Our proof is different from the one provided in Rao and Shao (1992) in the special case of RHDI and is based on a result by Chen and Rao (2007). When both the point estimator, $\hat{Y}_I$, and the variance estimator $v_t$ are doubly robust, the confidence interval (2.15) is valid if either (nonresponse or imputation) model is true.

## 4 Jackknife variance estimation under DREGI

In this section, we first assume that the sample $s$ of size $n$ is selected according to SRSWOR. We first consider the case of negligible sampling fractions in section 4.1. The case of a nonnegligible sampling fraction is treated in section 4.2. Jackknife variance estimation in the context of stratified simple random sampling is briefly discussed in section 4.3.

### 4.1 The case of negligible sampling fractions

In the absence of nonresponse, a jackknife variance estimator is obtained as follows :

(i) remove the unit $j = 1$ from the sample ;

(ii) adjust the design weights $d_i$ to obtain the so-called jackknife weights $\tilde{d}_{i(j)}$, where $\tilde{d}_{i(j)}$ is given by

$$\tilde{d}_{i(j)} = \begin{cases} d_i \frac{n}{n-1} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

(iii) compute the estimator $\hat{Y}_{\pi(j)}$ which is calculated the same way as $\hat{Y}_\pi$ but using the adjusted weights $\tilde{d}_{i(j)}$ instead of the design weights $d_i$ ;

(iv) insert back unit $j = 1$ deleted in step (i) ;

(v) repeat the steps (i)-(iv) for $j = 2, ..., n$.

A jackknife variance estimator of $\hat{Y}_\pi$ is then given by

$$v_J = \left(\frac{n-1}{n}\right) \sum_{j=1}^{n} \left(\hat{Y}_{\pi(j)} - \hat{Y}_\pi\right)^2 . \tag{2.16}$$

In the special case of SRSWOR, it is well known that $v_J$ in (2.16) reduces to $v_J = s_y^2/n$, where $s_y^2 = (n-1)^{-1} \sum_{i \in s} (y_i - \bar{y})^2$ denotes the sample variability of the $y$-values with $\bar{y} = n^{-1} \sum_{i \in s} y_i$. That is, $v_J$ corresponds to the textbook variance estimator under simple random sampling with replacement (SRSWR). A trivial modification of $v_J$ consists of incorporating the finite population correction (fpc), $1 - n/N$, to obtain $v_J^* = \left(1 - \frac{n}{N}\right) v_J$ which corresponds to the textbook variance estimator under SRSWOR.

In the presence of nonresponse to item $y$, the use of (2.16) may lead to serious underestimation of the variance of the estimator, especially if the nonresponse rate is appreciable. Rao and Shao (1992) proposed an adjusted jackknife method that is calculated in a similar fashion as (2.16) except that, whenever a responding unit is deleted, the imputed values are adjusted. The imputed values are unchanged if a nonresponding unit is deleted. Let $y_{i(j)}^{a*}$ denote the adjusted imputed value for unit $i$ when unit $j$ was deleted. We have

$$y_{i(j)}^{a*} = \begin{cases} \mathbf{z}_i' \hat{\mathbf{B}}_{r(j)} & \text{if } r_j = 1 \\ \mathbf{z}_i' \hat{\mathbf{B}}_r & \text{if } r_j = 0 \end{cases}$$

where $\hat{\mathbf{B}}_{r(j)}$ is computed the same way as $\hat{\mathbf{B}}_r$ but replacing the design weights $d_i$ with the jackknife weights $\tilde{d}_{i(j)}$. Note that $\hat{\mathbf{B}}_{r(j)}$ is the estimated regression coefficient obtained by fitting a regression model using the set of respondents without unit $j$. Hence, in the case of DREGI, the Rao-Shao procedure is equivalent to re-imputing within each jackknife replicate.

The Rao-Shao jackknife variance estimator is then given by

$$v_{JRS} = \left(\frac{n-1}{n}\right) \sum_{j=1}^{n} \left(\hat{Y}_{I(j)}^{a} - \hat{Y}_I\right)^2, \tag{2.17}$$

where $\hat{Y}_{I(j)}^{a}$ is computed the same way as $\hat{Y}_I$ in (2.2) but with the adjusted imputed values $y_{i(j)}^{a*}$ instead of the imputed values $y_i^*$.

The variance estimator $v_{JRS}$ is an estimator of $V_p\left(\hat{Y}_I|\mathbf{r}\right)$ that we would have obtained had the sampling been performed with SRSWR. Hence, $v_{JRS}$ is asymptotically unbiased and consistent for $V_1$ under SRSWR, or equivalently, if the sampling fraction $n/N$ is negligible. This property is satisfied regardless of the validity of the nonresponse model or the imputation model. This does not mean that the validity of the underlying (nonresponse or imputation) model is not important. If the model is misspecified, the imputed estimator $\hat{Y}_I$ could be considerably biased. However, the estimator $v_{JRS}$ tracks $V_p\left(\hat{Y}_I|\mathbf{r}\right)$ be the imputed estimator $\hat{Y}_I$ biased or unbiased. Consistency of $v_{JRS}$ follows from standard regularity conditions used in the complete data case. Note that, unless an explicit adjustment is made, the second component, $V_2$ is not accounted for. Intuitively, this can be explained by the fact that once an unit is deleted, nonresponse is not generated in each jackknife replicate before the imputation process is performed. In other words, the Rao-Shao adjusted jackknife simulates the effect of (with replacement) sampling conditional only on the vector of response indicators $\mathbf{r}$.

The above discussion suggests that the estimator $v_{JRS}$ can alternatively be obtained by using a standard (complete-data) jackknife procedure. Let $\hat{Y}_{I(j)}$ denote the imputed estimator without unit $j$ which is computed the same way as $\hat{Y}_I$ but replacing the design weights $d_i$ with the jackknife weights $\tilde{d}_{i(j)}$. That is, $\hat{Y}_{I(j)} = \sum_{i\in s} d_{i(j)}\mathbf{z}_i'\hat{\mathbf{B}}_{r(j)}$. The variance estimator $v_{JRS}$ can be obtained as follows :

$$v_{JRS} = \left(\frac{n-1}{n}\right) \sum_{j=1}^{n} \left(\hat{Y}_{I(j)} - \hat{Y}_I\right)^2. \tag{2.18}$$

Thus, in the case of DREGI, adjusting the imputed values is not necessary if the above jackknife procedure is applied. Most importantly, note that obtaining

(2.18) does not require a specialized variance estimation software. It is readily obtained by using standard software packages designed for complete data jackknife variance estimation in the context of regression estimation.

We now turn to the asymptotic bias of $v_{JRS}$. Using the fact that $v_t$ given by (2.14) is asymptotically unbiased for the total variance, $V\left(\hat{Y}_I\right)$, we approximate the bias of $v_{JRS}$ by

$$\mathrm{B}\left(v_{JRS}\right) \approx E\left(v_{JRS} - v_t\right) = E\left[\left(v_{JRS} - v_1\right) - v_2\right], \tag{2.19}$$

where $v_1$ given by (2.10) and $v_2$ given by (2.13) are obtained under SRSWOR. Since both $v_1$ and $v_{JRS}$ are estimating the term $V_1$, the magnitude of the bias of $v_{JRS}$ depends on the average magnitude of $(v_{JRS} - v_1)$ and that of $v_2$.

In the case of DREGI and the NM approach, it can be shown that the asymptotic $pq$-bias of $v_{JRS}$ is given by

$$\mathrm{B}_{pq}\left(v_{JRS}\right) \equiv E_{pq}\left[\left(v_{JRS} - v_1\right) - v_2\right] \approx NS_y^2, \tag{2.20}$$

where $S_y^2 = (N-1)^{-1}\sum_{i\in U}\left(y_i - \bar{Y}\right)^2$ denotes the population variability of the $y$-values. Assuming that $S_y^2$ is $O(1)$ and is bounded away from 0, the term $\mathrm{B}_{pq}\left(v_{JRS}\right)$ is $O(N)$. On the other hand, under mild conditions, the total variance $V\left(\hat{Y}_I\right)$ is $O\left(\frac{N^2}{n}\right)$. It follows that the asymptotic relative bias of $v_{JRS}$, $RB_{pq}\left(v_{JRS}\right) = \frac{\mathrm{B}_{pq}(v_{JRS})}{V\left(\hat{Y}_I\right)}$ is $O\left(\frac{n}{N}\right)$ and is thus negligible when the sampling fraction $\frac{n}{N}$ is negligible. This can be explained by the fact that, in this case, the term $(v_{JRS} - v_1)$ is, on average, virtually equal to zero, whereas the contribution to the total variance of the term $v_2$ is negligible (see section 3.4).

In the case of DREGI and the IM approach, the asymptotic $mpq$-bias of $v_{JRS}$ is given by

$$\begin{aligned} \mathrm{B}_{mpq}\left(v_{JRS}\right) &\equiv E_{mpq}\left[\left(v_{JRS} - v_1\right) - v_2\right] \\ &\approx N\left[\sigma^2 \boldsymbol{\lambda}'\bar{\mathbf{Z}} + \boldsymbol{\beta}'\mathbf{S}_{\mathbf{zz}}\boldsymbol{\beta}\right], \end{aligned} \tag{2.21}$$

where $\mathbf{S_{zz}} = (N-1)^{-1}\sum_{i\in U}\left(\mathbf{z}_i - \bar{\mathbf{Z}}\right)\left(\mathbf{z}_i - \bar{\mathbf{Z}}\right)'$ with $\bar{\mathbf{Z}} = N^{-1}\sum_{i\in U}\mathbf{z}_i$. Once again, the asymptotic relative bias of $v_{JRS}$, $RB_{mpq}\left(v_{JRS}\right) = \frac{\mathrm{B}_{mpq}(v_{JRS})}{V\left(\hat{Y}_I - Y\right)}$ is $O\left(\frac{n}{N}\right)$ and is negligible if the sampling fraction $n/N$ is negligible.

In the complete data case, Yung and Rao (1996) proposed a linearization jackknife variance estimator, $v_{JL}$. Extension to missing data can be found in Rao (1996) and Yung and Rao (2000) who studied the properties of $v_{JL}$, which is obtained from $v_{JRS}$ by performing a first-order Taylor expansion. In fact, under the reverse framework it becomes clear that $v_{JL}$ is asymptotically equivalent to $v_1$ since both $v_{JRS}$ and $v_1$ estimate consistently the same term, $V_p\left(\hat{Y}_I|\mathbf{r}\right)$. Therefore, a jackknife linearization variance estimator can simply be obtained by performing a complete data first-order Taylor expansion instead of linearizing the jackknife variance estimator, which can involve tedious algebra. For example, one can use the method proposed by Demnati and Rao (2004) for Taylor linearization.

### 4.2 The case of non-negligible sampling fractions

We now turn to the case of non-negligible sampling fraction $n/N$. Under SRSWOR, we have

$$v_{JRS} - v_1 \geq 0$$

for all samples $s$. This difference tends to increases as the sampling fraction increases and the response rate decreases. Furthermore, the contribution to the total variance of the term $v_2$ can no longer be considered negligible and it tends to increase as the sampling fraction increases and the response rate decreases. As a result, $v_{JRS}$ is biased and alternative variance estimators are needed.

First, as in the complete data case, it would be tempting to use the following variance estimator :

$$v^*_{JRS} = \left(1 - \frac{n}{N}\right) v_{JRS}. \tag{2.22}$$

However, this estimator may be severely biased as its asymptotic relative bias is given by

$$\mathrm{RB}\left(v_{JRS}^{*}\right) \approx \frac{-V_2}{V\left(\hat{Y}_I\right)}. \tag{2.23}$$

It follows that the asymptotic relative bias of $v_{JRS}^{*}$ is always negative and its magnitude depends only on the contribution of $V_2$ to the total variance. In other words, the variance estimator $v_{JRS}^{*}$ correctly estimates $V_1$ but fails to estimate the component $V_2$. Therefore, the estimator $v_{JRS}^{*}$ should not be used since it can considerably underestimate the true variance, especially if the sampling fraction is large and the nonresponse rate is appreciable. Following the reverse approach of Shao and Steel (1999), a correct variance estimator in the case of nonnegligible sampling fraction is thus obtained as follows :

$$v_{SS} = \left(1 - \frac{n}{N}\right) v_{JRS} + v_2, \tag{2.24}$$

where $v_2$ is given by (2.13). Since the validity of $v_{JRS}$ does not depend on the validity of the underlying model and $v_2$ is doubly robust, it follows that $v_{SS}$ is doubly robust.

Finally, a third variance estimator was proposed by Lee, Rancourt and Särndal (1995). It is given by

$$v_{LRS} = v_{JRS} - N\hat{S}_y^2, \tag{2.25}$$

where $\hat{S}_y^2$ is an estimator of the population variability of the $y$-values. It follows from (2.20) that the LRS variance estimator can be seen as a bias-adjusted variance estimator under the NM approach. Under SRSWOR, LRS proposed to estimate $S_y^2$ by $s_{yr}^2 = (r-1)^{-1} \sum_{i \in s} r_i \left(y_i - \bar{y}_r\right)^2$, where $\bar{y}_r = r^{-1} \sum_{i \in s} r_i y_i$, which leads to

$$v_{LRS} = v_{JRS} - N s_{yr}^2. \tag{2.26}$$

Under the NM approach, $s_{yr}^2$ is asymptotically $pq$-unbiased for $S_y^2$. As a result, the estimator $v_{LRS}$ is asymptotically $pq$-unbiased for the total variance $V\left(\hat{Y}_I\right)$.

However, it is asymptotically *mpq*-biased ; to see this, we consider its conditional nonresponse bias under the IM approach given by

$$B_m\left(v_{LRS}|\mathbf{I},\mathbf{r}\right) \equiv E_m\left(v_{LRS}|\mathbf{I},\mathbf{r}\right) = N\left[\sigma^2\boldsymbol{\lambda}'\left(\bar{\mathbf{z}}-\bar{\mathbf{z}}_r\right)+\boldsymbol{\beta}'\left(\mathbf{s}_{\mathbf{zz}}-\mathbf{s}_{\mathbf{zz}r}\right)\boldsymbol{\beta}\right], \quad (2.27)$$

where $\mathbf{s}_{\mathbf{zz}r} = (r-1)^{-1}\sum_{i\in s} r_i\left(\mathbf{z}_i-\bar{\mathbf{z}}_r\right)\left(\mathbf{z}_i-\bar{\mathbf{z}}_r\right)'$ with $\bar{\mathbf{z}}_r = r^{-1}\sum_{i\in s} r_i\mathbf{z}_i$ and $\mathbf{s}_{\mathbf{zz}} = (n-1)^{-1}\sum_{i\in s}\left(\mathbf{z}_i-\bar{\mathbf{z}}\right)\left(\mathbf{z}_i-\bar{\mathbf{z}}\right)'$ with $\bar{\mathbf{z}} = n^{-1}\sum_{i\in s}\mathbf{z}_i$.

The bias given by (2.27) is not equal to zero, in general. Thus, the estimator $v_{LRS}$ is not doubly robust. Using the model (2.3), a doubly robust LRS type variance estimator can be obtained by first expressing $S_y^2$ as

$$S_y^2 = S_E^2 + \mathbf{B}'\mathbf{S}_{\mathbf{zz}}\mathbf{B}, \quad (2.28)$$

where $S_E^2 = (N-1)^{-1}\sum_{i\in U} E_i^2$ with $E_i = y_i - \mathbf{z}_i'\mathbf{B}$ and

$$\mathbf{B} = \left(\sum_{i\in U}\mathbf{z}_i\mathbf{z}_i'/\left(\boldsymbol{\lambda}'\mathbf{z}_i\right)\right)^{-1}\sum_{i\in U}\mathbf{z}_i y_i/\left(\boldsymbol{\lambda}'\mathbf{z}_i\right).$$

An alternative asymptotically *pq*-unbiased estimator of $S_y^2$ is given by

$$\tilde{\hat{S}}_y^2 = \left(\frac{\boldsymbol{\lambda}'\bar{\mathbf{z}}}{\boldsymbol{\lambda}'\bar{\mathbf{z}}_r}\right)s_{er}^2 + \hat{\mathbf{B}}_r'\mathbf{s}_{\mathbf{zz}}\hat{\mathbf{B}}_r, \quad (2.29)$$

where $s_{er}^2 = (r-1)^{-1}\sum_{i\in s} r_i e_i^2$. Using (2.29), an alternative LRS-type variance estimator is given by

$$v_{LRS}^* = v_{JRS} - N\tilde{\hat{S}}_y^2. \quad (2.30)$$

The variance estimator $v_{LRS}^*$ in (2.30) is asymptotically unbiased and consistent for $V\left(\hat{Y}_I\right)$ under either the NM approach or the IM approach. Therefore, it is doubly robust.

Both variance estimators (2.24) and (2.30) are doubly robust and can be used in the case of nonnegligible sampling fractions. However, the philosophy behind their construction is different. Obtaining (2.30) consists of determining and estimating the bias of the jackknife variance estimator, $v_{JRS}$. Then, a bias-adjusted variance estimator is obtained by subtracting the estimated bias from $v_{JRS}$. On the other hand, the estimator (2.24) is obtained by first noting that $v_{JRS}$ is an

estimator of the first component $V_1$. Thus, applying the finite population correction to $v_{JRS}$ provides an asymptotically unbiased estimator of $V_1$. Finally, adding the component $v_2$ leads to (2.24).

### 4.3 Stratified sampling

In sections 4.1 and 4.2, we have studied the properties of jackknife variance estimators in the context of SRSWOR. In practice, this design is seldom used. In this section, we consider the case of stratified sampling. The population $U$ is partitioned into $L$ strata, $U_1, U_2, ..., U_L$, of size $N_1, N_2, ..., N_L$, respectively. From stratum $h$, a sample $s_h$, of size $n_h$, is selected according to SRSWOR. The selection in one stratum is independent of the selection in any other stratum. We assume that imputation is performed independently within strata. That is, the strata correspond to imputation classes, which is a common situation in practice, especially in business surveys. Due to the independence feature, jackknife variance estimation is performed independently within each stratum. If the objective is to estimate the population total $Y$, the Rao-Shao jackknife variance estimator is asymptotically unbiased and consistent for $V\left(\hat{Y}_I\right)$ if the overall sampling fraction $\sum_{h=1}^{L} n_h / \sum_{h=1}^{L} N_h$ is negligible. This condition is often satisfied in practice. Note that we do not require the individual sampling fractions $n_h/N_h$ to be negligible, which is a much stronger condition. However, if the objective is to estimate a domain total, then the condition on the overall sampling fraction may not be sufficient for the jackknife to be valid. For example, suppose that the domains of interest are the individual stratum totals, $Y_h$, $h = 1, ..., L$. In this case, the stratum sampling fractions $n_h/N_h$ must be negligible.

## 5 Simulation study

We conducted a limited simulation study to investigate the performance of the variance estimators considered in section 3. We generated three populations of size $N = 500$ containing two variables : a variable of interest $y$ and an auxiliary variable $z$. We first generated the variable $z$ according to a gamma distribution with parameters $\alpha_0$ and $\alpha_1$ . The parameters $\alpha_0$ and $\alpha_1$ were chosen so that

$E(z) = 100$. Then, given the $z$-values, we generated the $y$-values according to the ratio model

$$y_i = 1.5z_i + \epsilon_i, \tag{2.31}$$

where the errors $\epsilon_i$ were generated from a normal distribution with mean 0 and variance $\sigma^2$. The value of $\sigma^2$ was chosen to give a model $R^2$ (coefficient of determination) approximately equal to 0.81. For population 1, the coefficient of variation of $z$, $CV(z)$, was set to 0.5, whereas it was set to 1 and 1.5 for populations 2 and 3, respectively. The goal is to estimate the population total of the $y$-values, $Y = \sum_{i \in U} y_i$.

From the generated populations, we selected $R = 10,000$ samples according to SRSWOR. The sampling fraction $n/N$ was set to 0.5 and 0.75. For each selected sample, nonresponse to item $y$ was generated according to two nonresponse mechanisms :

(i) Nonresponse mechanism 1 (uniform nonresponse) : the response probability $p_{1i}$ is constant for all the units in the population with probability 0.5.

(ii) Nonresponse mechanism 2 (non-uniform nonresponse based on $z$) : the response probability $p_{2i}$ for unit $i$ given by

$$p_i = 0.05 + 0.95\left[1 + exp\left(\lambda_0 + \lambda_1 z_i\right)\right]^{-1},$$

where $\lambda_0$ and $\lambda_1$ were chosen so that the overall response probability be equal to 0.5. Note that the minimum response probability is 0.05.

The response indicators $r_{1i}$ and $r_{2i}$ were then generated independently 10,000 times from a Bernoulli distribution with parameter $p_{1i}$ and $p_{2i}$. This led to 10,000 sets of respondents for each nonresponse mechanism.

To compensate for the missing $y$-values, deterministic ratio imputation was used. Deterministic ratio imputation is a special case of (2.4) with $\mathbf{z}_i = z_i$ and $\sigma_i^2 = \sigma^2 z_i$. Finally, in each sample, we computed the imputed estimator $\hat{Y}_I$ given

by (2.2) as well as the following variance estimators : $v_{JRS}$, $v^*_{JRS}$, $v_{LRS}$, $v^*_{LRS}$ and $v_{SS}$.

We define the Monte Carlo average of an estimator $\hat{\theta}$ by

$$E_{MC}\left(\hat{\theta}\right) = \frac{1}{R}\sum_{j=1}^{R}\hat{\theta}_{(r)}, \tag{2.32}$$

where $\hat{\theta}_{(r)}$ denotes the estimator $\hat{\theta}$ in the $r$-th simulated sample. We first calculated the Monte Carlo percent relative bias (RB) of $\hat{Y}_I$ as well as its Monte Carlo mean square error given respectively by

$$RB_{MC}\left(\hat{Y}_I\right) = 100 \times \frac{E_{MC}\left(\hat{Y}_I\right) - Y}{Y}, \tag{2.33}$$

where $E_{MC}\left(\hat{Y}_I\right)$ is obtained from (2.32) by letting $\hat{\theta} = \hat{Y}_I$ and

$$MSE_{MC}\left(\hat{Y}_I\right) = E_{MC}\left(\hat{Y}_I - Y\right)^2, \tag{2.34}$$

where $E_{MC}\left(\hat{Y}_I - Y\right)^2$ is obtained from (2.32) by letting $\hat{\theta} = \left(\hat{Y}_I - Y\right)^2$.

As a measure of bias of a variance estimator $v$, we used its Monte Carlo percent relative bias given by

$$RB_{MC}\left(v\right) = 100 \times \frac{E_{MC}\left[v - MSE_{MC}\left(\hat{Y}_I\right)\right]}{MSE_{MC}\left(\hat{Y}_I\right)}, \tag{2.35}$$

where $E_{MC}\left[v - MSE_{MC}\left(\hat{Y}_I\right)\right]$ is obtained from (2.32) by letting $\hat{\theta} = v - MSE_{MC}\left(\hat{Y}_I\right)$. As a measure of stability of a variance estimator, we used its Monte Carlo mean square error given by

$$MSE_{MC}\left(v\right) = E_{MC}\left[v - MSE_{MC}\left(\hat{Y}_I\right)\right]^2, \tag{2.36}$$

where $E_{MC}\left[v - MSE_{MC}\left(\hat{Y}_I\right)\right]^2$ is obtained from (2.32) by letting $\hat{\theta} = \left[v - MSE_{MC}\left(\hat{Y}_I\right)\right]^2$. To compare the relative stability of the variance estimators, using $v_{SS}$ as the reference, we used the following measure of relative efficiency :

$$RE\left(v\right) = \frac{MSE_{MC}\left(v\right)}{MSE_{MC}\left(v_{SS}\right)}. \tag{2.37}$$

We first note that the RB of the imputed estimator $\hat{Y}_I$ under deterministic ratio imputation was negligible (less than 0.1%) in all the scenarios, as expected. Table 1 shows the percent Monte Carlo RB of the variance estimators. It is clear from Table 1 that $v_{JRS}$ overestimates the total variance of the imputed estimator under both nonresponse mechanisms, as expected. For a given value of $CV(z)$, the RB of $v_{JRS}$ increases as the sampling fraction $n/N$ increases. Also, for a given value of $n/N$, the RB of $v_{JRS}$ increases as $CV(z)$ increases. The variance estimator $v^*_{JRS}$ always underestimates the total variance under both nonresponse mechanisms, as expected. The RB increases as the sampling fraction increases but seems to decrease as $CV(z)$ increases. Under the nonresponse mechanism 1, the three variance estimators $v_{LRS}$, $v^*_{LRS}$ and $v_{SS}$ show a small RB (less than 5%), as expected. Under the nonresponse mechanism 2, both $v^*_{LRS}$ and $v_{SS}$ perform well in terms of RB. The variance estimator $v_{LRS}$ performs well when $CV(z)$ is low but is considerably biased when $CV(z)$ is high. This result is not surprising as $v_{LRS}$ is not doubly robust.

Table 2 shows the Monte Carlo relative efficiency given by (2.37). It is clear that in all the scenarios (except one), the variance estimator $v_{SS}$ was more stable than its competitors. In particular, it is more stable than the doubly robust variance estimator, $v^*_{LRS}$.

## 6 Jackknife variance estimation under RREGI

In this section, we consider the problem of jackknife variance estimation under RREGI, which uses the imputed values

$$y^*_i = \mathbf{z}'_i\hat{\mathbf{B}}_r + (\boldsymbol{\lambda}'\mathbf{z}_i)^{1/2}\,\epsilon^*_i, \tag{2.38}$$

where $\epsilon^*_i = e^*_i - \bar{e}_r$ such that $P(e^*_i = e_j) = d_j/\sum_{l\in s} d_l r_l$ with $e_j = (\boldsymbol{\lambda}'\mathbf{z}_j)^{-1/2}\left(y_j - \mathbf{z}'_j\hat{\mathbf{B}}_r\right)$ and $\bar{e}_r = \sum_{l\in s} d_l r_l e_l / \sum_{l\in s} d_l r_l$. The imputed value $y^*_i$ in (2.38) is the sum of a deterministic component, $\mathbf{z}'_i\hat{\mathbf{B}}_r$, and a random component $\epsilon^*_i$. The deterministic component corresponds to DREGI. RHDI is a special case of (2.38) with $\mathbf{z}_i = 1$ for all $i$.

TABLE 2.1. Monte Carlo percent relative bias of the variance estimators

| Nonresponse mechanism | Variance estimators | Population 1 | | Population 2 | | Population 3 | |
|---|---|---|---|---|---|---|---|
| | | $f = 1/2$ | $f = 3/4$ | $f = 1/2$ | $f = 3/4$ | $f = 1/2$ | $f = 3/4$ |
| 1 | $v_{JRS}$ | 66.6 | 174.8 | 86.3 | 225.2 | 175.8 | 230.3 |
| | $v^*_{JRS}$ | -16.7 | -31.3 | -6.9 | -18.7 | -8.0 | -17.5 |
| | $v_{LRS}$ | -3.7 | 0.4 | 0.5 | 1.0 | -1.7 | -0.3 |
| | $v^*_{LRS}$ | -5.0 | -2.5 | -1.2 | -3.9 | -2.7 | -3.4 |
| | $v_{SS}$ | -3.9 | 0.5 | 0.6 | 0.8 | -1.7 | -0.4 |
| 2 | $v_{JRS}$ | 85.9 | 238.5 | 93.1 | 283.0 | 97.9 | 298.1 |
| | $v^*_{JRS}$ | -7.0 | -16.2 | -5.5 | -8.0 | -3.0 | -5.0 |
| | $v_{LRS}$ | 0.3 | 4.1 | -11.0 | -26.7 | -24.8 | -71.1 |
| | $v^*_{LRS}$ | 0.0 | 3.9 | -4.2 | -7.0 | -2.5 | -3.5 |
| | $v_{SS}$ | 0.5 | 4.7 | -1.8 | 0.4 | -1.0 | 0.9 |

To study the properties of the imputed estimator $\hat{Y}_I$ under RREGI and to derive corresponding variance estimators, we need to account for the random imputation mechanism. The total error of the imputed estimator can now be expressed as

$$\hat{Y}_I - Y = \left(\hat{Y}_\pi - Y\right) + \left[E_I\left(\hat{Y}_I\right) - \hat{Y}_\pi\right] + \left[\hat{Y}_I - E_I\left(\hat{Y}_I\right)\right], \tag{2.39}$$

where $E_I(.)$ denotes the expectation with respect to the random imputation mechanism. As before, the first term on the right hand side of (2.39) represents the sampling error, the second term represents is called the nonresponse error, whereas the third term represents the imputation error. Since $E_I(\epsilon^*_i) = 0$, we have $E_I(y^*_i) = \mathbf{z}'_i\hat{\mathbf{B}}_r$, which corresponds to the imputed value had DREGI been used. It follows that the imputed estimator $\hat{Y}_I$ under RREGI is asymptotically $pqI$-unbiased for $Y$ provided the underlying nonresponse model is correctly specified and asymptotically $mpqI$-unbiased for $Y$ provided the underlying imputation is correctly specified.

TABLE 2.2. Relative efficiency (RE) of the variance estimators

| Nonresponse mechanism | Variance estimators | Population 1 | | Population 2 | | Population 3 | |
|---|---|---|---|---|---|---|---|
| | | $f = 1/2$ | $f = 3/4$ | $f = 1/2$ | $f = 3/4$ | $f = 1/2$ | $f = 3/4$ |
| 1 | $v_{JRS}$ | 39.3 | 431.4 | 68.3 | 868.2 | 22.5 | 390.4 |
| | $v^*_{JRS}$ | 2.9 | 13.9 | 1.3 | 6.4 | 1.1 | 2.9 |
| | $v_{LRS}$ | 2.0 | 6.7 | 1.8 | 7.0 | 2.0 | 12.0 |
| | $v^*_{LRS}$ | 1.5 | 2.4 | 1.5 | 2.9 | 1.2 | 2.3 |
| | $v_{SS}$ | 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | $v_{JRS}$ | 97.9 | 973.2 | 94.0 | 2658.1 | 29.1 | 680.2 |
| | $v^*_{JRS}$ | 1.6 | 4.4 | 1.1 | 1.6 | 1.0 | 1.0 |
| | $v_{LRS}$ | 0.8 | 1.9 | 2.0 | 24.6 | 2.0 | 38.2 |
| | $v^*_{LRS}$ | 1.5 | 2.2 | 1.5 | 5.6 | 1.1 | 1.7 |
| | $v_{SS}$ | 1 | 1 | 1 | 1 | 1 | 1 |

We now turn to the variance of $\hat{Y}_I$. Under the NM approach and RREGI, the total variance of the imputed estimator $\hat{Y}_I$ can be expressed as

$$\begin{aligned} V\left(\hat{Y}_I\right) &= E_q V_p E_I\left(\hat{Y}_I|\mathbf{r}\right) + E_q E_p V_I\left(\hat{Y}_I|\mathbf{r}\right) + V_q E_p E_I\left(\hat{Y}_I|\mathbf{r}\right) \\ &\equiv \tilde{V}_1^{NM} + \tilde{V}_I^{NM} + \tilde{V}_2^{NM}, \end{aligned} \tag{2.40}$$

where $V_I(.)$ denotes the variance with respect to the random imputation mechanism. Under the IM approach and RREGI, the total variance of the imputed estimator $\hat{Y}_I$ can be expressed as

$$\begin{aligned} & V\left(\hat{Y}_I - Y\right) \\ = \; & E_m E_q V_p E_I\left(\hat{Y}_I - Y|\mathbf{r}\right) + E_q E_m E_p V_I\left(\hat{Y}_I - Y|\mathbf{r}\right) + E_q V_m E_p E_I\left(\hat{Y}_I - Y|\mathbf{r}\right) \\ \equiv \; & \tilde{V}_1^{IM} + \tilde{V}_I^{IM} + \tilde{V}_2^{IM}. \end{aligned} \tag{2.41}$$

When a statement applies for both the NM approach and the IM approach, we use the generic notation $\tilde{V}_1$ to denote $\tilde{V}_1^{NM}$ or $\tilde{V}_1^{IM}$, $\tilde{V}_2$ to denote $\tilde{V}_2^{NM}$ or $\tilde{V}_2^{IM}$ and $\tilde{V}_I$ to denote $\tilde{V}_I^{NM}$ or $\tilde{V}_I^{IM}$. Note that the term $\tilde{V}_I$ represents the imputation

variance due to random imputation.

### 6.1 Linearization variance estimators

As in the case of DREGI, we first derive variance estimators using a first-order Taylor expansion. Noting that $E_I\left(\hat{Y}_I\right) = \sum_{i\in s} d_i \mathbf{z}_i' \hat{\mathbf{B}}_r$, an estimator of $\tilde{V}_1$ is obtained by estimating $V_p E_I\left(\hat{Y}_I - Y|\mathbf{r}\right)$ and is given by $v_1$ in (2.10). Similarly, an estimator of $\tilde{V}_2$ is given by $v_2$ in (2.13). Finally, to estimate the variance due to imputation, $\tilde{V}_I$, it suffices to determine $V_I\left(\hat{Y}_I|\mathbf{r}\right)$, which we denote by $v_I$. We obtain

$$v_I = \left[\sum_{i\in s} d_i^2 \left(1 - r_i\right) \mathbf{z}_i \mathbf{z}_i'\right] s_{er}^2, \tag{2.42}$$

where $s_{er}^2 = \frac{1}{\sum_s d_i r_i} \sum_s d_i r_i e_i^2$. Note that $v_I$ remains valid regardless of the approach (NM or IM) used as the basis for inference. Also, note that $v_I$ is simple to obtain since it does not require the second-order inclusion probabilities. Finally, a doubly robust estimator of the total variance under RREGI is given by

$$v_t = v_1 + v_I + v_2. \tag{2.43}$$

### 6.2 Jackknife variance estimation

In this section, we consider two jackknife estimators under RREGI. Once again, we confine ourselves to the case of SRSWOR. The first variance estimator assumes that the deterministic and the random components of (2.38) are reported in two separate columns in the imputed data file. In this case, we can apply a standard jackknife procedure to the deterministic component (which corresponds to DREGI) such as described in section 3.2. An obvious jackknife variance estimator under RREGI is then given by

$$\tilde{v}_{JRS}^* = \left(1 - \frac{n}{N}\right) v_{JRS} + v_I + v_2, \tag{2.44}$$

where $v_{JRS}$ is given by (2.18), $v_2$ is given by (2.13) and $v_I$ is given by (2.42). The estimator $\tilde{v}_{JRS}^*$ is asymptotically unbiased and consistent for the total variance of $\hat{Y}_I$ under either the NM approach or the IM approach.

In practice, the imputed values are typically reported in a single column. In this case, the data user cannot distinguish the deterministic component from the

random component. Rao and Shao (1992) provide a consistent variance estimator in this context under SRSWR. Let $y_{i(j)}^{a*}$ denote the adjusted imputed value for unit $i$ when unit $j$ was deleted. We have

$$y_{i(j)}^{a*} = \begin{cases} y_i^* + \mathbf{z}_i'\hat{\mathbf{B}}_{r(j)} - \mathbf{z}_i'\hat{\mathbf{B}}_r & \text{if } j \in s_r \\ \mathbf{z}_i'\hat{\mathbf{B}}_r & \text{if } j \in s_m \end{cases}$$

Using these imputed values, an adjusted jackknife variance estimator under RREGI, denoted by $\tilde{v}_{JRS}$, is obtained by using (2.17). The estimator $\tilde{v}_{JRS}$ is asymptotically unbiased and consistent for $\tilde{V}_1 + \tilde{V}_I$ if the sample is selected with replacement, or equivalently, if the sampling fraction $n/N$ is negligible. In the case of nonnegligible $n/N$, it would be tempting to use an estimator of the form $\left(1 - \frac{n}{N}\right)\tilde{v}_{JRS} + v_2$, as we proposed in section 3.2 (see expression (2.24)). However, this estimator is not appropriate because the fpc, $1 - n/N$, is applied on an estimator of $\tilde{V}_1 + \tilde{V}_I$, while it should only be applied to the part estimating $\tilde{V}_1$. This suggests an alternative estimator of the form

$$\tilde{v}_{JRS}^{**} = \left(1 - \frac{n}{N}\right)\tilde{v}_{JRS} + \frac{n}{N}v_I + v_2, \tag{2.45}$$

where $v_2$ is given by (2.13) and $v_I$ is given by (2.42). As $\tilde{v}_{JRS}^*$, the estimator $\tilde{v}_{JRS}^{**}$ is asymptotically unbiased and consistent for the total variance of $\hat{Y}_I$ under either the NM approach or the IM approach.

We now examine the difference between $\tilde{v}_{JRS}^*$ and $\tilde{v}_{JRS}^{**}$ for the special case of RHDI and consider the case of $n/N$ negligible so we can omit the term $v_2$ in (2.44) and (2.45). Le $r$ and $m = m - r$ denote the number of respondents and nonrespondents, respectively. Assuming that $n/(n-1) \approx 1$ and $r/(r-1) \approx 1$, the variance estimators $\tilde{v}_{JRS}^*$ and $\tilde{v}_{JRS}^{**}$ reduce respectively to

$$\tilde{v}_{JRS}^* = N^2\left[\frac{1}{r} + \frac{m}{n^2}\right]s_{yr}^2 \tag{2.46}$$

and

$$\tilde{v}_{JRS}^{**} = N^2\left[\frac{s_{yI}^2}{n} + \left(\frac{1}{r} - \frac{1}{n} + \frac{m}{n^2}\right)s_{yr}^2\right], \tag{2.47}$$

where $s_{yI}^2 = (n-1)^{-1}\sum_{i\in s}(\tilde{y}_i - \bar{y}_I)^2$ with $\bar{y}_I = n^{-1}\sum_{i\in s}\tilde{y}_i$. Noting that $E_I\left(s_{yI}^2\right) = \left(1 - \frac{m}{n^2}\right)s_{yr}^2$, we have $E_I\left(\tilde{v}_{JRS}^{**}\right) \approx \tilde{v}_{JRS}^*$.

## 7 Unequal probability sampling designs without replacement

In this section, we consider the problem of jackknife variance estimation under unequal probability sampling without replacement designs that include Inclusion Probability Proportional to Size (IPPS) sampling designs as special cases. Suppose we want to estimate a parameter $\theta$ that can be expressed as a smooth function of means of $q$ survey variables; i.e., $\theta = g\left(\bar{Y}_1, ..., \bar{Y}_q\right)$, where $\bar{Y}_j = N^{-1} \sum_{i \in U} y_{ji}$. An estimator of $\theta$ is the so-called plug-in estimator given by $\hat{\theta} = g\left(\bar{y}_1, ..., \bar{y}_q\right)$, where $\bar{y}_j = \sum_{i \in s} d_i y_{ji} / \hat{N}_\pi$ and $\hat{N}_\pi = \sum_{i \in s} d_i$.

Let $\pi_{ij}$ denote the joint inclusion probability in the sample for unit $i$ and $j$. In the complete data situation, Campbell (1980) proposed a jackknife estimator analogue to a standard linearization variance estimator, which is given by

$$v_{JC} = \sum_{i \in s} \sum_{j \in s} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}} u_i u_j, \tag{2.48}$$

where $u_j = (1 - w_j)\left(\hat{\theta} - \hat{\theta}_{(j)}\right)$ with $w_j = \frac{d_j}{\sum_{k \in s} d_k}$ and $\hat{\theta}_{(j)}$ is calculated the same way as $\hat{\theta}$ but using the adjusted weights $\tilde{d}_{i(j)}$ instead of the design weights $d_i$. Here, the adjusted weights $\tilde{d}_{i(j)}$ are given by

$$\tilde{d}_{i(j)} = \begin{cases} d_i & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

Berger and Skinner (2005) established the consistency of $v_{JC}$. However, it requires the second-order inclusion probabilities which may not be easy to compute for complex designs. To overcome this difficulty, Berger (2007) proposed a jackknife variance estimator in the absence of nonresponse that requires the first-order inclusion probabilities only. Under some regularity conditions, he showed that the proposed estimator is consistent in the case of unequal probability sampling designs. One important condition for consistency is that the sampling design is required to have high entropy. High entropy designs include the maximum entropy design often called Conditional Poisson Sampling (Hajek, 1981), the Rao-Sampford design (Rao, 1965; Sampford, 1967)and Chao's procedure (Chao,

1982) ; see, e.g., Berger (1998). Berger's jackknife variance estimator is given by

$$v_{JB} = \left(\frac{n}{n-1}\right) \sum_{j \in s} c_j \left(u_j - \sum_{k \in s} \phi_k u_k\right)^2, \tag{2.49}$$

where $c_j = (1 - \pi_j)$ and $\phi_k = c_k / \sum_{l \in s} c_l$. The quantity $c_j$ can be seen as a finite population correction for unequal probability sampling designs.

Following Campbell (1980), Berger and Rao (2006) proposed a variance estimator in the presence of imputed data when $\theta = Y$, and established its consistency. Once again, the proposed estimator requires the second-order inclusion probabilities. We propose to extend Berger's estimator to the case of imputation for missing values. Noting that $\hat{Y}_I$ in (2.6) can be expressed as $\hat{Y}_I = g\left(\hat{\bar{\mathbf{Z}}}_\pi, \hat{\bar{\mathbf{T}}}_r, \hat{\bar{\mathbf{t}}}_r\right)$, where $\hat{\bar{\mathbf{Z}}}_\pi = \hat{\mathbf{Z}}_\pi / \hat{N}_\pi$, $\hat{\bar{\mathbf{T}}}_r = \hat{\mathbf{T}}_r / \hat{N}_\pi$ and $\hat{\bar{\mathbf{t}}}_r = \left[\sum_{i \in s} d_i r_i \mathbf{z}_i y_i / \left(\boldsymbol{\lambda}' \mathbf{z}_i\right)\right] / \hat{N}_\pi$, a jackknife variance estimator is readily obtained from (2.49) by replacing $y_{1i}$ with $\mathbf{z}_i$, $y_{2i}$ with $r_i \mathbf{z}_i \mathbf{z}_i' / \left(\boldsymbol{\lambda}' \mathbf{z}_i\right)$ and $y_{3i}$ with $r_i \mathbf{z}_i y_i / \left(\boldsymbol{\lambda}' \mathbf{z}_i\right)$. Under regularity conditions similar to the ones provided in Berger (2007), the estimator $v_{JB}$ is consistent for $V_1$. Under DREGI, an estimator of the total variance of $\hat{Y}_I$ is given by

$$v_{JB}^* = v_{JB} + v_2, \tag{2.50}$$

where $v_2$ is given by (2.13). Under RREGI, an estimator of the total variance is given by

$$\tilde{v}_{JB}^* = v_{JB} + v_I + v_2, \tag{2.51}$$

where $v_I$ is given by (2.42). Note that both $v_{JB}^*$ and $\tilde{v}_{JB}^*$ are doubly robust because they are asymptotically unbiased and consistent under either the NM or the IM approach. Also, note that both variance estimators do not require the second-order inclusion probabilities.

## 8 Discussion

In this paper, we studied the problem of doubly robust inference in the presence of imputed data. Using the reverse framework, we have examined the theoretical properties of linearization and jackknife variance estimators under both

DREGI and RREGI.

Shao and Sitter (1996) proposed a bootstrap variance estimator under imputation for missing data. Their method consists of re-imputing the missing values in each bootstrap sample by the same imputation method used in the original sample. Like the Rao-Shao jackknife variance estimator, the Shao-Sitter variance estimator is an estimator of $V_p\left(\hat{Y}_I|\mathbf{r}\right)$ that we would have obtained had the sampling been performed with SRSWR. Hence, their estimator is asymptotically unbiased and consistent for $V\left(\hat{Y}_I\right)$ if the overall sampling fraction is negligible. Therefore, like the Rao-Shao jackknife variance estimator, the Shao-Sitter variance estimator is doubly robust when $n/N$ is negligible. In the case of SRSWOR (or stratified simple random sampling), one can use (2.24) and (2.30) if the sampling fraction is not negligible, where $v_{JRS}$ is replaced by the Shao-Sitter estimator. The problem of bootstrap variance estimation in the presence of imputed data and unequal probability sampling without replacement requires further research.

**A Appendix : Double robustness of the imputed estimator**

We first show part (i) of Definition 1. To that end, we write the imputed estimator $\hat{Y}_I$ given by (2.6) as

$$\hat{Y}_I = \sum_{i\in s} d_i \mathbf{z}'_i \hat{\mathbf{B}}_r = \sum_{i\in s} \tilde{\omega}_i r_i y_i,$$

where

$$\tilde{\omega}_i = d_i \hat{\mathbf{Z}}'_\pi \hat{\mathbf{T}}_r^{-1} \mathbf{z}_i / (\boldsymbol{\lambda}' \mathbf{z}_i).$$

Suppose that the true model is given by (2.3). Further, we assume that both the sampling design and the nonresponse mechanism are ignorable with respect to the model (2.3). Showing (i) is equivalent to showing that

$$E_{pq}(\hat{Y}_I/Y) \to 1 \text{ in probability .} \tag{2.52}$$

To show (2.52), notice that it suffices to show

(a) $Y/\mathbf{Z}'\boldsymbol{\beta} \to 1$ as $n \to \infty$.

(b) $E_{pq}(\hat{Y}_I/\mathbf{Z}'\boldsymbol{\beta}) \to 1$.

Proof of (a) : first, note that $E_m(\hat{Y}/\mathbf{Z}'\boldsymbol{\beta}) = 1$. Now ,

$$V_m(\hat{Y}/\mathbf{Z}'\boldsymbol{\beta}) = \frac{1}{(\mathbf{Z}'\boldsymbol{\beta})^2}\sum_{i\in U}\sigma_i^2 = \frac{1}{N(\overline{\mathbf{Z}}'\boldsymbol{\beta})^2}\left(\sum_{i\in U}\sigma_i^2/N\right).$$

If $(\overline{\mathbf{Z}}'\boldsymbol{\beta})^{-1} = O(1)$ and $\sum_{i\in U}\sigma_i^2/N = O(1)$ we have $V_m(Y/\mathbf{Z}'\boldsymbol{\beta}) \to 0$ as $N \to \infty$. Hence,

$$Y/\mathbf{Z}'\boldsymbol{\beta} \to 1 \text{ in probability as } n \to \infty.$$

Proof of (b) : first, note that $E_{mpq}(\hat{Y}_I/\mathbf{Z}'\boldsymbol{\beta}) = E_{qpm}(\hat{Y}_I/\mathbf{Z}'\boldsymbol{\beta}) = 1$. Interchanging the order of expectations is correct because both the sampling design and the nonresponse mechanism are assumed to be ignorable. Now, we need to show that $V_m E_{pq}(\hat{Y}_I/\mathbf{Z}'\boldsymbol{\beta}) \to 0$ as $n, N \to \infty$. To show this, it suffices to show that $V_{mpq}(\hat{Y}_I/\mathbf{Z}'\boldsymbol{\beta}) \to 0$ since

$$\begin{aligned} V_{mpq}(\hat{Y}_I/\mathbf{Z}'\boldsymbol{\beta}) &= V_m E_{pq}(\hat{Y}_I/\mathbf{Z}'\boldsymbol{\beta}) + E_m V_p E_q(\hat{Y}_I/\mathbf{Z}'\boldsymbol{\beta}) + E_{mp}V_q(\hat{Y}_I/\mathbf{Z}'\boldsymbol{\beta}) \\ &\geq V_m E_{pq}(\hat{Y}_I/\mathbf{Z}'\boldsymbol{\beta}). \end{aligned}$$

Now,

$$\begin{aligned} V_{mpq}(\hat{Y}_I/\mathbf{Z}'\boldsymbol{\beta}) &= \frac{1}{(\mathbf{Z}'\boldsymbol{\beta})^2}E_{mpq}(\hat{Y}_I - E_{mpq}(\hat{Y}_I))^2 \\ &= \frac{1}{(\mathbf{Z}'\boldsymbol{\beta})^2}E_{pqm}(\hat{Y}_I - \mathbf{Z}'\boldsymbol{\beta})^2 \\ &= \frac{1}{N}\frac{1}{(\overline{\mathbf{Z}}'\boldsymbol{\beta})^2}\left[\sum_{i\in U}E_{pq}(\tilde{\omega}_i r_i I_i)\sigma_i^2/N\right]. \end{aligned}$$

Hence, $V_{mpq}(\hat{Y}_I/\mathbf{Z}'\boldsymbol{\beta}) \to 0$ as $N \to \infty$ if $(\overline{\mathbf{Z}}'\boldsymbol{\beta})^{-1} = O(1)$ and $\sum_{i\in U}E_{pq}(\tilde{\omega}_i r_i I_i)\sigma_i^2/N = O(1)$. The latter condition essentially means that $\max_{i\in U} d_i = O(\frac{n}{N})$ and $p_i$ is bounded away from 0 ; i.e. there exists $p_{\min} > 0$ such that $p_{\min} < p_i$ for all $i \in U$.

We now show part (ii) of Definition 1. Suppose the true model is not given by (2.3) but rather :

$$m^\star : y_i = \mu_i + \epsilon_i,$$

such that $E_{m^*}(\epsilon_i) = 0$, $E_{m^*}(\epsilon_i \epsilon_j) = 0, i \neq j,$ and $V_{m^*}(\epsilon_i) = \sigma_i^2$. We assume that we have uniform nonresponse. Further, we assume that

$$\frac{\sum_{i=1}^{N} \mu_i}{N} \to \mu,$$

where $\mu$ is a constant and

$$\sum_{k=1}^{\infty} \frac{\sigma_k^2}{k^2} < \infty.$$

From the strong law of large numbers, we have

$$\frac{\sum_{k=1}^{N} Y_k}{N} \to \mu \text{ almost surely.}$$

Hence, to show part (ii) of Definition 1 it remains to show that

$$\frac{\hat{Y}_I}{N} \to \mu \text{ in probability.} \tag{2.53}$$

To that end, we also that the usual assumptions on the variables $\{\mathbf{z}_k\}$ and $\{y_k\}$ (see expressions (3.1)-(3.3) in Deville (1999)) . Under the assumptions, the proof of (2.53) is similar to that of the asymptotic normality of $U_n$ in appendix B.1 and hence, is skipped.

**B Appendix : Asymptotic normality of $\hat{Y}_I$**

We show the asymptotic normality of $\hat{Y}_I$ in the context of a stratified multistage sampling design. That is, the population under consideration is stratified into $L$ strata with $N_h$ primary sampling units (PSU's) or clusters in the $h^{th}$ stratum. Within each stratum, $n_h \geq 2$ clusters are selected from stratum $h$, independently across strata. The first-stage clusters are usually selected without replacement to avoid the selection of the same cluster more than once. Within the $(hi)^{th}$ sampled first-stage cluster, $m_{hi}$ ultimate units (elements) are sampled according to some probability sampling method, $i = 1, ..., n_h$ ; $h = 1, ..., L$. Note that we do not need to specify the number of stages or the sampling methods beyond the first stage. We simply assume that subsampling within sampled clusters is performed to ensure unbiased estimation of cluster totals, $Y_{hi}$, $i = 1, ..., n_h$ ; $h = 1, ..., L$. Associated with the $j^{th}$ sampled element in the $i^{th}$ sampled cluster belonging to the $h^{th}$ stratum is the variable of interest, $y_{hij}$, and

the basic survey weights, $d_{hij}$. When working on asymptotic properties, it is convenient to work with $\bar{y}_I = \hat{Y}_I/N$, where $\hat{Y}_I = \sum_{(hij)\in s} d_{hij}\tilde{y}_{hij}$.

We use the asymptotic framework described in Krewski and Rao (1981). We have a sequence of nested populations with $L$ strata. We assume that the number of strata $L \to \infty$ and so $n = \sum_{h=1}^{L} n_h \to \infty$ and $N = \sum_{h=1}^{L} N_h \to \infty$. We will use $n$ as the index of the sequence of population. We also assume that the auxiliary variables and the variable of interests are such that we can use the Taylor linearization (see equations (3.1)-(3.3) in Deville(1999)).

**B.1 Asymptotic Normality under DREGI**

Recall that under DREGI, the imputed values are given by $y^*_{hij} = \mathbf{z}'_{hij}\hat{\mathbf{B}}_r$, where $\hat{\mathbf{B}}_r$ is given by (2.5).

Let $p_{hij}$ be the probability of response of unit $(hij)$, $(hij) \in U$. We assume that that the units respond independently of one another and that they are independent of the realized sample. We also assume also that the $p_{hij}$'s are bounded away from zero (that is, there exists a constant $p_{min} > 0$ such that $p_{hij} \geq p_{min}$ for all $(hij) \in U$).

Under uniform response, (when all the $p_j$'s are equal), the imputed estimator $\bar{y}_I$ is consistent for $\bar{Y}$. However, when the $p_j$'s are unequal, this is not necessarily true and, as $n \to \infty$, the estimator $\bar{y}_I$ converges to $\bar{Y}_p = \bar{\mathbf{Z}}'\mathbf{B}_p$, where

$$\mathbf{B}_p = \left( \sum_{(hij)\in U} d_{hij} p_{hij} \mathbf{z}_{hij}\mathbf{z}'_{hij} / \left(\boldsymbol{\lambda}'\mathbf{z}_{hij}\right) \right)^{-1} \sum_{(hij)\in U} d_{hij} p_{hij} \mathbf{z}_{hij} y_{hij} / \left(\boldsymbol{\lambda}'\mathbf{z}_{hij}\right).$$

The total error of $\bar{y}_I$ can be expressed as

$$\begin{aligned} \bar{y}_I - \bar{Y}_p &= (\bar{y}_I - \bar{\mathbf{Z}}'\mathbf{B}_r) + (\bar{\mathbf{Z}}'\mathbf{B}_r - \bar{Y}_p) \\ &= U_n + V_n. \end{aligned}$$

Let $\Im_K = \sigma(r_{hij}, (hij) \in U)$ be the $\sigma$-field generated by the response indicators, $r_{hij}$. We first establish the asymptotic normality of $U_n$ conditional to the vector of response indicators, $\mathbf{r}$. Using a first-order Taylor expansion, we obtain

$$U_n = \bar{y}_I - \bar{\mathbf{Z}}'\mathbf{B}_r = \left( \sum_{(hij)\in U} \tilde{d}_{hij} e_{hij} \right) - \bar{\mathbf{Z}}'\mathbf{B}_r + O_p\left(n^{-1}\right), \tag{2.54}$$

where

$$e_{hij} = \left( \mathbf{Z}'\mathbf{T}_r^{-1} r_{hij}\mathbf{z}_{hij} \frac{\left(y_{hij} - \mathbf{z}'_{hij}\mathbf{B}_r\right)}{\left(\boldsymbol{\lambda}'\mathbf{z}_{hij}\right)} + \mathbf{z}'_{hij}\mathbf{B}_r \right)$$

and $\tilde{d}_{hij} = N^{-1} d_{hij} I_{hij}$. Conditionally on $\mathbf{r}$, we have $E_p(\bar{e}) = \bar{\mathbf{Z}}'\mathbf{B}_r$, where $\bar{e} = \sum_{hij \in s} \tilde{d}_{hij} e_{hij}$.

We assume that the following regularity conditions hold :

C1 : $n^{1+\delta} \sum_h \sum_i E_p |e_{hi} - E_p(e_{hi})|^{2+\delta} = O(1)$ for some $\delta > 0$ and $e_{hi} = \sum_j \tilde{d}_{hij} e_{hij}$.

C2 : $nV_p(\bar{e}) \to \sigma_e^2$, say.

Condition C1 is a standard Liapunov-type condition on the $2 + \delta$ moments used in establishing a central limit Theorem for independent nonidentically distributed random variables. Condition C2 assumes that the limit of the variance of $\bar{e}$ exists when multiplied by the normalized factor $n$. Let

$$X_{hi} = n\left(e_{hi} - E_p\left(e_{hi}\right)\right),$$

where $e_{hi} = \sum_j \tilde{d}_{hij} e_{hij}$. Notice that the variables $X_{hi}$ are independent. Under condition C2, we have

$$\frac{1}{n} \sum_h \sum_i V_p\left(X_{hi}\right) = n \sum_h \sum_i V_p\left(e_{hi}\right) = nV_p\left(\bar{e}\right) \to \sigma_e^2.$$

Also, under condition C1, we have

$$\frac{1}{n} \sum_h \sum_i E_p \left|X_{hi}\right|^{2+\delta} = n^{1+\delta} \sum_h \sum_i E_p \left|e_{hi} - E_p\left(e_{hi}\right)\right|^{2+\delta} = O\left(1\right),$$

satisfying the conditions of the Central Limit Theorem (see Lemma 3.1 in Krewski and Rao, 1981). Therefore, applying Slutzky's Theorem to (2.54), we obtain conditionally given $\Im_K$,

$$\frac{1}{\sigma_{2n}} U_n \to_d N\left(0, 1\right),$$

where $\sigma_{2n}^2 = \sigma_e^2 / n$.

Now, we need to show the normality of $V_n = \overline{\mathbf{Z}}'\mathbf{B}_r - \bar{Y}_p$. Note that $\mathbf{B}_r$ can be written as

$$\mathbf{B}_r = \left( \sum_{(hij) \in U} \frac{r_{hij}}{p_{hij}} c_{hij}\mathbf{z}_{hij}\mathbf{z}'_{hij} \right)^{-1} \sum_{(hij) \in U} \frac{r_{hij}}{p_{hij}} c_{hij}\mathbf{z}_{hij} y_{hij},$$

where $c_{hij} = p_{hij}/(\boldsymbol{\lambda}'\mathbf{z}_{hij})$. Under assumptions similar to that of Deville (1999), we write

$$V_n = \overline{\mathbf{Z}}'\mathbf{B}_r - \bar{Y}_p = \overline{\mathbf{Z}}' \sum_{(hij)\in U} \mathbf{T}_p^{-1}\mathbf{z}_{hij} \frac{(y_{hij} - \mathbf{z}'_{hij}\mathbf{B}_p)}{(\boldsymbol{\lambda}'\mathbf{z}_{hij})} r_{hij} + O_p(N^{-1}).$$

Since the random variables $\overline{\mathbf{Z}}'\mathbf{T}_p^{-1}\mathbf{z}_{hij}\frac{(y_{hij}-\mathbf{z}'_{hij}\mathbf{B}_p)}{(\boldsymbol{\lambda}'\mathbf{z}_{hij})}r_{hij}$ are independent, it follows from the Central Limit Theorem (see Lemma 3.1 in Krewski and Rao, 1981) that

$$\frac{V_n}{\sigma_{1n}} \to_d N(0,1),$$

where

$$\sigma_{1n}^2 = \sum_{(hij)\in U} \left( \overline{\mathbf{Z}}'\mathbf{T}_p^{-1}\mathbf{z}_{hij} \frac{(y_{hij} - \mathbf{z}'_{hij}\mathbf{B}_p)}{(\boldsymbol{\lambda}'\mathbf{z}_{hij})} \right)^2 p_{hij}(1-p_{hij}),$$

as long the following conditions hold :

C3 : $\sigma_{1n}^2/N$ converges to a positive number as $n \to \infty$;

C4 : $\frac{1}{N}\sum_{(hij)\in U} \left| \bar{\mathbf{Z}}'\mathbf{T}_p^{-1}\mathbf{z}_{hij} \frac{(y_{hij}-\mathbf{z}'_{hij}\mathbf{B}_p)}{(\boldsymbol{\lambda}'\mathbf{z}_{hij})} \right|^{2+\delta} = O(1)$.

Applying Theorem 2 of Chen and Rao (2007), we conclude that

$$\frac{U_n + V_n}{\sqrt{\sigma_{1n}^2 + \sigma_{2n}^2}} \to_d N(0,1).$$

**B.2 Asymptotic normality under RREGI**

Under RREGI, we have

$$\bar{y}_I = \sum_{(hij)\in s} \tilde{d}_{hij} r_{hij} y_{hij} + \sum_{(hij)\in s} \tilde{d}_{hij}\left(1 - r_{hij}\right)\left(\mathbf{z}'_{hij}\hat{\mathbf{B}}_r + (\lambda'\mathbf{z}_{hij})^{1/2}\epsilon^*_{hij}\right).$$

In this case, the total error of $\bar{y}_I$, $\bar{y}_I - \bar{Y}$, can be decomposed as

$$\bar{y}_I - \bar{Y} = \left[\bar{y}_I - E_I\left(\bar{y}_I\right)\right] + \left[E_I\left(\bar{y}_I\right) - \bar{\mathbf{Z}}'\mathbf{B}_r\right] + \left[\bar{\mathbf{Z}}'\mathbf{B}_r - \bar{Y}\right]. \tag{2.55}$$

Let $\Im_n = \sigma\left(\left(I_{hij}, r_{hij}\right), (hij) \in s\right)$, where $I_{hij}$ denotes the sample selection indicator for unit $(hij)$. Note that since $E_I\left(\epsilon^*_{hij}\right) = 0$, we have $E_I\left(\bar{y}_I\right) = \hat{\mathbf{Z}}'_\pi\hat{\mathbf{B}}_r$, which coincides with the imputed estimators one would obtain by performing DREGI. Hence, we can use the result in Appendix B.1 and conclude that the sum of the last two terms on the right hand side of (2.55) is asymptotically normally distributed.

Now, consider the term $\bar{y}_I - E_I(\bar{y}_I)$ on the right hand side of (2.55). First, note that

$$\begin{aligned}\bar{y}_I - E_I(\bar{y}_I) &= \sum_{(hij)\in s} \tilde{d}_{hij}(1-r_{hij})(\lambda'\mathbf{z}_{hij})^{1/2}\epsilon^*_{hij} \\ &= \sum_{(hij)\in s} \xi^*_{hij},\end{aligned}$$

where $\xi^*_{hij} = \tilde{d}_{hij}(1-r_{hij})(\lambda'\mathbf{z}_{hij})^{1/2}\epsilon^*_{hij}, (hij) \in s$. In addition to C1-C4, we assume the following regularity conditions :

C5 : $n \max_{h,i} \sum_j \tilde{d}_{hij}(\lambda'\mathbf{z}_{hij})^{1/2} = O_p(1)$.

C6 : $\sum_{(hij)\in s} \tilde{d}_{hij} |e_{hij} - \bar{e}_r|^{2+\delta} = O_p(1)$.

C7 : $nV_I\left(\sum_h \sum_i \sum_j \xi^*_{hij}\right) \to \sigma^2_\xi$, say, as $n \to \infty$,

Let

$$t^*_{hi} = \sum_j \xi^*_{hij} = \sum_j \tilde{d}_{hij}(1-r_{hij})(\lambda'\mathbf{z}_{hij})^{1/2}\epsilon^*_{hij}.$$

We have

$$\begin{aligned}& E_I|t^*_{hi}|^{2+\delta} \\ = & \left[\sum_{(glk)\in s} \tilde{d}_{glk} r_{glk} \left|\sum_j \tilde{d}_{hij}(1-r_{hij})(\boldsymbol{\lambda}'\mathbf{z}_{hij})^{1/2}(e_{glk}-\bar{e}_r)\right|^{2+\delta}\right]\left[\sum_{(glk)\in s} \tilde{d}_{glk} r_{glk}\right]^{-1} \\ = & \left[\sum_{(glk)\in s} \tilde{d}_{glk} r_{glk} |(e_{glk}-\bar{e}_r)|^{2+\delta} \left|\sum_j \tilde{d}_{hij}(1-r_{hij})(\boldsymbol{\lambda}'\mathbf{z}_{hij})^{1/2}\right|^{2+\delta}\right]\left[\sum_{(glk)\in s} \tilde{d}_{glk} r_{glk}\right]^{-1} \\ \leq & \left[\sum_{(glk)\in s} \tilde{d}_{hij} |e_{glk}-\bar{e}_r|^{2+\delta} \left|\sum_j \tilde{d}_{hij}(\boldsymbol{\lambda}'\mathbf{z}_{hij})^{1/2}\right|^{2+\delta}\right]\left[\sum_{(glk)\in s} \tilde{d}_{glk} r_{glk}\right]^{-1} \qquad (2.56)\end{aligned}$$

since $r_{hij} = 0$ or 1 and $\tilde{d}_{hij} \geq 0$. Now, noting that the $p_{hij}$'s are bounded away from zero, we have

$$\left[\sum_{(glk)\in s} \tilde{d}_{glk} r_{glk}\right]^{-1} = O_p(1).$$

We obtain, using the conditions C5 and C6 in (2.56),

$$E_I|t^*_{hi}|^{2+\delta} \leq \left(\sum_j \tilde{d}_{hij}(\boldsymbol{\lambda}'\mathbf{z}_{hij})^{1/2}\right)^{2+\delta} O_p(1).$$

Combining with condition C5, we get

$$\begin{aligned}
\frac{1}{n}\sum_h\sum_i E_I\left|nt_{hi}^*\right|^{2+\delta} &\leq n^{1+\delta}\sum_h\sum_i\left(\sum_j \tilde{d}_{hij}(\boldsymbol{\lambda}'\mathbf{z}_{hij})^{1/2}\right)^{2+\delta} O_p\left(1\right)\\
&\leq n^{2+\delta}\max_{h,i}\left(\sum_j \tilde{d}_{hij}(\boldsymbol{\lambda}'\mathbf{z}_{hij})^{1/2}\right)^{2+\delta} O_p\left(1\right)\\
&= \left(n\max_{h,i}\sum_j \tilde{d}_{hij}(\boldsymbol{\lambda}'\mathbf{z}_{hij})^{1/2}\right)^{2+\delta} O_p\left(1\right)\\
&= O_p\left(1\right).
\end{aligned}$$

Also by C7,

$$\frac{1}{n}\sum_h\sum_i V_I\left(nt_{hi}^*\right) = nV_I\left(\sum_h\sum_i\sum_j \xi_{hij}^*\right) \to \sigma_\xi^2, \text{ as } n\to\infty.$$

Using the fact that $\frac{1}{n}\sum_h\sum_i E_I\left(nt_{hi}^*\right) = \sum_h\sum_i\sum_j E_I\left(\xi_{hij}^*\right) = 0$ and the Central Limit Theorem (see Lemma 3.1 in Krewski and Rao, 1981), we obtain

$$\sqrt{n}\left(\frac{1}{n}\sum_h\sum_i nt_{hi}^*\right) \to_d N\left(0,\sigma_\xi^2\right).$$

Finally, applying Theorem 2 of Chen and Rao (2007) with $V_n = \left[E_I\left(\bar{y}_I\right) - \bar{\mathbf{Z}}'\mathbf{B}_r\right] + \left[\bar{\mathbf{Z}}'\mathbf{B}_r - \bar{Y}\right]$ and $U_n = \left[\bar{y}_I - E_I\left(\bar{y}_I\right)\right]$, we obtain

$$\frac{1}{\sqrt{V\left(\bar{y}_I\right)}}\left(\bar{y}_I - \bar{Y}\right) \to_d N\left(0,1\right),$$

where $V\left(\bar{y}_I\right) = \sigma_{1n}^2 + \sigma_{2n}^2$ with $\sigma_{1n}^2$ as the asymptotic variance of $E_I\left(\bar{y}_I\right)$ and $\sigma_{2n}^2 = V_I\left(\bar{y}_I\right)$.

**Acknowledgment**

David Haziza's research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. We would like to thank Professor J.N.K. Rao for helpful comments and suggestions.

**REFERENCES**

Beaumont, J.-F. (2005). Calibrated Imputation in Surveys under a Quasi-Model-Assisted Approach. *Journal of the Royal Statistical Society B*, 67, 445-458.

Berger, Y.G. (2007). A Jackknife Variance Estimator for Unistage Stratified Samples with Unequal Probabilities. *Biometrika*, 94, 4, 953-964.

Berger, Y.G. (1998). Rate of Convergence to Asymptotic Variance for the Horvitz-Thompson Estimator. *Journal of Statistical Planning and Inference*, 74, 149-168.

Berger, Y.G. and Rao, J.N.K. (2006). Adjusted Jackknife for Imputation under Unequal Probability Sampling Without Replacement. *Journal of the Royal Statistical Society B*, 68, 531-547.

Berger, Y.G. and Skinner, C.J. (2005). A Jackknife Variance Estimator for Unequal Probability Sampling. *Journal of the Royal Statistical Society B*, 67, 79-89.

Brick, J. M., Jones, E., Kalton, G. and Vaillant, R. (2005). Variance Estimation with Hot Deck Imputation : a Simulation Study of Three Methods. *Survey Methodology*, 31, 151-159.

Campbell, C. (1980). A Different View of Finite Population Estimation. Proceedings of the Survey Research Methods, Section of the American Statistical Association, 319-324.

Chao, M.T. (1982). A General Purpose Unequal Probability Sampling Plan. *Biometrika*, 69, 653-656.

Chen, J. and Rao, J. N. K. (2007). Asymptotic Normality under Two-phase Sampling Designs. *Statistica Sinica*, 17, 1047-1064.

Davison, A. C. and Sardy, S. (2007). Resampling Variance Estimation in Surveys with Missing Data. *Journal of Official Statistics*, 23, 371-386.

Demnati, A. and Rao, J. N. K. (2004). Linearization Variance Estimators for Survey Data. *Survey Methodology*, Vol 30, No. 1, 17-26.

Deville, J. C. (1999). Variance Estimation for Complex Statistics and Estimators : Linearization and Residual Techniques. *Survey Methodology*, Vol 25, No. 2, 193-203.

Deville, J. C. and Särndal, C. E. (1994). Variance Estimation for the Regression Imputed Horvitz-Thompson Estimator. *Journal of Official Statistics*, 23, 33-40.

Fay, R. E. (1991). A Design-Based Perspective on Missing Data Variance. *Proceedings of the 1991 Annual Research Conference, US Bureau of the Census*, 429-440.

Hajek, J. (1981). *Sampling from a finite population.* Marcel Dekker Inc : Bassel.

Haziza, D. and Rao, J. N. K. (2006). A Nonresponse Model Approach to Inference under Imputation for Missing Survey Data. *Survey Methodology*, 32, 53-64.

Haziza, D. (2009). Imputation and Inference in the Presence of Missing Data. *To appear in the Handbook of Statistics, vol. 29, Sample Surveys : Theory, Methods and Inference*

Hurtubise, D. (2006). Variance Due à l'Imputation à l'Aide d'un Modèle par le Ratio. *Working Paper no. BSMD-2006-002F, Statistics Canada, Ottawa.*

Kang, J.D.Y. and Schafer, J.L. (2008). Demystifying Double Robustness : a Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science*, 22, 523-539.

Kim, J.K. and Park, H. (2006). Imputation using Response Probability. *The Canadian Journal of Statistics*, 34, 171-182.

Kott, P.S. (1994). A Note on Handling Nonresponse in Sample Surveys. *Journal of American Statistical Association*, 89, 693-696.

Krewski, D. and Rao, J.N.K. (1981). Inference from Stratified Samples : Properties of Linearization, Jackknife and Balanced Repeated Replication Methods. *Annals of Statistics*, 9, 1010-1019.

Lee, E., Rancourt, E. and Särndal, C. E. (1995). Jackknife Variance Estimation for Data with Imputed Values. *Proceedings of the Section on Survey Research Methods, Statistical Society of Canada*, 111-115.

Little, R. J. A. and An, H. (2004). Robust Likelihood-based Analysis for Multivariate Data with Missing Values. *Statistica Sinica*, 14, 949-968.

Rao, J.N.K. (1965). On Two Simple Schemes of unequal Probability Sampling Without Replacement. *Journal of the Indian Statistical Association*, 3, 173-180.

Rao, J. N. K. (1990). Variance Estimation under Imputation for Missing Data. *Technical report*, Statistics Canada, Ottawa.

Rao, J. N. K. (1996). On Variance Estimation with Imputed Survey Data. *Journal of American Statistical Association*, 91, 499-506.

Rao, J. N. K. (2000). Variance Estimation in the Presence of Imputation for Missing Data. *Proceedings of the Second International Conference on Establishment Surveys*, 599-608.

Rao, J. N. K. and Shao, J. (1992). On Variance Estimation under Imputation for Missing Data. *Biometrika*, 79, 811-822.

Rao, J. N. K. and Sitter, R. R. (1995). Variance Estimation under Two-phase Sampling with Application to Imputation for Missing Data. *Biometrika*, 82, 453-460.

Robins, J.M., Sued, M., Lei-Gomez, Q. and Rotnitzky, A. (2008). Performance of Double-robust Estimators when "Inverse Probability" Weights are Highly Variable. *Statistical Science*, 22, 544-559.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, 63, 581-590.

Sampford, M.R. (1967). On Sampling Without Replacement with Unequal Probabilities of Selection. *Biometrika*, 54, 499-513.

Särndal, C. E. (1992). Method for Estimating the Precision of Survey Estimates when Imputation has Been Used. *Survey Methodology*, 18, 241-252.

Shao, J. (2002). Replication Methods for Variance Estimation in Complex Surveys with Imputed Data. *In Survey Nonresponse edited by R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little*, New York : John Wiley and Sons, 303-314.

Shao, J. and Sitter, R. R. (1996). Bootstrap for Imputed Survey Data. *Journal of the American Statistical Association*, 93, 819-831.

Shao, J. and Steel, P. (1999). Variance Estimation for Survey Data With Composite Imputation and Nonnegligible Sampling Fractions. *Journal of the American Statistical Association*, 94, 254-265.

Yung, W. and Rao, J. N. K. (1996). Jackknife Linearization Variance Estimators under Stratified Multi-stage Sampling. *Survey Methodology*, 22, 23-31.

Yung, W. and Rao, J. N. K. (2000). Jackknife Variance Estimation under Imputation for Estimators using Poststratification Information. *Journal of the American Statistical Association*, 95, 903-915.

# CONCLUSION

Dans ce mémoire nous avons discuté d'inférence dans les enquêtes en présence données imputées. Nous avons étudiés le problème d'inférence doublement robuste dans ce contexte. En utilisant l'approche renversée, nous avons étudié les propriétés théoriques des estimateurs de variance par linéarisation et des estimateurs de variance Jacknife pour le cas de l'imputation par la régression. Nous avons également établi la normalité asymptotique des estimateurs imputés pour le cas de l'imputation par régression.

Shao and Sitter (1996) ont proposé l'estimateur de variance Bootstrap en présence de données imputées. Leur méthode consiste à réimputer les valeurs manquantes pour chaque échantillon bootstrap par la même méthode d'imputation que dans l'échantillon original. Comme c'est le cas pour l'estimateur de variance Jacknife de Rao-Shao, l'estimateur de variance de Shao-Sitter est un estimateur de $V_p\left(\hat{Y}_I|\mathbf{r}\right)$ que nous aurions obtenu dans le cas d'un échantillon aléatoire simple sans remise. En conséquence leur estimateur est consistent pour $V\left(\hat{Y}_I\right)$ si la fraction de sondage est négligeable. Il s'ensuit que comme dans le cas de l'estimateur de variance de Rao-Shao, l'estimateur de variance de Shao-Sitter est doublement robuste quand $n/N$ est négligeable.

Dans le cas de l'échantillonnage aléatoire simple sans remise (ou stratifié avec échantillon aléatoire simple sans remise dans chaque strate), on utilisera (2.24) et (2.30) si la fraction de sondage n'est pas négligeable, où $v_{JRS}$ est remplacé par l'estimateur de Shao-Sitter. L'étude du problème de l'estimation de la variance en présence de données imputées pour le cas de l'échantillonnage sans remise à probabilités inégales devra être approfondie.

# BIBLIOGRAPHIE


[1] Beaumont, J.-F. (2005). Calibrated Imputation in Surveys under a Quasi-Model-Assisted Approach. *Journal of the Royal Statistical Society B*, 67, 445-458.

[2] Berger, Y.G. (2007). A jackknife variance estimator for unistage stratified samples with unequal probabilities. *Biometrika*, 94, 4, 953-964.

[3] Berger, Y.G. (1998). Rate of Convergence to Asymptotic Variance for the Horvitz-Thompson Estimator. *Journal of Statistical Planning and Inference*, 74, 149-168.

[4] Berger, Y.G. et Rao, J.N.K. (2006). Adjusted Jackknife for Imputation under Unequal Probability Sampling without Replacement. *Journal of the Royal Statistical Society B*, 68, 531-547.

[5] Berger, Y.G. et Skinner, C.J. (2005) A Jackknife Variance Estimator for Unequal Probability Sampling. *Journal of the Royal Statistical Society B*, 67, 79-89.

[6] Brick, J. M., Jones, E., Kalton, G. et Vaillant, R. (2005). Variance estimation with hot deck imputation : a simulation study of three methods. *Survey Methodology*, 31, 151-159.

[7] Campbell, C. (1980). A different view of finite population estimation. Proceedings of the Survey Research Methods, Section of the American Statistical Association, 319-324.

[8] Chao, M.T. (1982). A General Purpose Unequal Probability Sampling Plan. *Biometrika*, 69, 653-656.

[9] Chen, J. et Rao, J. N. K. (2007). Asymptotic normality under two-phase sampling designs. *Statistica Sinica*, 17, 1047-1064.

[10] Davison, A. C. et Sardy, S. (2007). Resampling variance estimation in surveys with missing data. *Journal of Official Statistics*, 23, 371-386.

[11] Demnati, A. et Rao, J. N. K. (2004). Linearization variance estimators for survey data *Survey Methodology*, Vol 30, No. 1, 17-26.

[12] Deville, J. C. (1999). Variance estimation for Complex Statistics and Estimators : Linearization and Residual Techniques *Survey Methodology*, Vol 25, No. 2, 193-203.

[13] Deville, J. C. et Särndal, C. E. (1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics*, 23, 33-40.

[14] Fay, R. E. (1991). A Design-Based Perspective on Missing Data Variance. *Proceedings of the 1991 Annual Research Conference, US Bureau of the Census*, 429-440.

[15] Hajek, J. (1981). *Sampling from a finite population.* Marcel Dekker Inc : Bassel.

[16] Haziza, D. et Rao, J. N. K. (2006). A nonresponse model approach to inference under imputation for missing survey data. *Survey Methodology*, 32, 53-64.

[17] Haziza, D. (2009). Imputation and inference in the presence of missing data. *To appear in the Handbook of Statistics, vol. 29, Sample Surveys : Theory, Methods and Inference*

[18] Haziza, D. et Beaumont, J-F (2007). On the construction of imputation classes in surveys. *International Statistical Review*, 75, 1, 25-43.

[19] Hurtubise, D. (2006). Variance due à l'imputation à l'aide d'un modèle par le ratio. *Working Paper no. BSMD-2006-002F, Statistics Canada, Ottawa.*

[20] Kang, J.D.Y. et Schafer, J.L. (2008). Demystifying double robustness : a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22 , 523-539.

[21] Kim, J.K. et Park, H. (2006). Imputation using response probability. *The Canadian Journal of Statistics*, 34, 171-182.

[22] Kott, P.S. (1994). A note on handling nonresponse in sample surveys. *Journal of American Statistical Association*, 89, 693-696.

[23] Krewski, D. et Rao, J.N.K. (1981). Inference from stratified samples : properties of linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.

[24] Lee, E., Rancourt, E. et Särndal, C. E. (1995). Jackknife variance estimation for data with imputed values. *Proceedings of the Section on Survey Research Methods, Statistical Society of Canada*, 111-115.

[25] Little, R. J. A. (1986). Survey nonreponse adjustements, *International Statistical Review*, 54, 139-157.

[26] Little, R. J. A. et An, H. (2004). Robust Likelihood-based analysis for multivariate data with missing values. *Statistica Sinica*, 14, 949-968.

[27] Rao, J.N.K. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal of the Indian Statistical Association*, 3, 173-180.

[28] Rao, J. N. K. (1990). Variance estimation under imputation for missing data. *Technical report*, Statistics Canada, Ottawa.

[29] Rao, J. N. K. (1996). On variance estimation with imputed survey data. *Journal of American Statistical Association*, 91, 499-506.

[30] Rao, J. N. K. (2000). Variance estimation in the presence of imputation for missing data. *Proceedings of the Second International Conference on Establishment Surveys*, 599-608. .

[31] Rao, J. N. K. et Shao, J. (1992). On variance estimation under imputation for missing data. *Biometrika*, 79, 811-822.

[32] Rao, J. N. K. et Sitter, R. R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.

[33] Robins, J.M., Sued, M., Lei-Gomez, Q. et Rotnitzky, A. (2008). Performance of double-robust estimators when "inverse probability" weights are highly variable. *Statistical Science*, 22, 544-559.

[34] Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, 63, 581-590.

[35] Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*, New York : John Wiley & Sons, Inc.

[36] Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection.*Biometrika*, 54, 499-513.

[37] Särndal, C. E. (1992). Method for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18, 241-252.

[38] Särndal, C. E., Swensson, B. et Wretman, J.H. (1992). *Model Assisted Survey Sampling*, New York, Springer Verlag.

[39] Shao, J. (2002). Replication Methods for Variance Estimation in Complex Surveys with Imputed Data. *In Survey Nonresponse edited by R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little*,New York : John Wiley and Sons, 303-314.

[40] Shao, J. et Sitter, R. R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 93, 819-831.

[41] Shao, J. et Steel, P. (1999). Variance Estimation for Survey Data With Composite Imputation and Nonnegligible Sampling Fractions. *Journal of the American Statistical Association*, 94, 254-265.

[42] Yung, W. et Rao, J. N. K. (1996). Jackknife linearization variance estimators under stratified multi-stage sampling. *Survey Methodology*, 22, 23-31.

[43] Yung, W. et Rao, J. N. K. (2000). Jackknife variance estimation under imputation for estimators using poststratification information. *Journal of the American Statistical Association*, 95, 903-915.