Recueil du symposium 2002 de Statistique Canada
Modélisation des données d'enquête pour la recherche économique et sociale

WESVAR : LOGICIEL D'ANALYSE DES DONNÉES D'ENQUÊTES COMPLEXES

G. Hussain Choudhry et Richard Valliant[1]

RÉSUMÉ

Dans presque toutes les enquêtes, on recourt à des plans d'échantillonnage complexes pour recueillir des données et celles-ci servent fréquemment à des analyses statistiques qui vont au-delà des estimations de simples paramètres descriptifs de la population visée. Nombreuses sont les techniques disponibles dans des progiciels statistiques en vogue qui ne se prêtent pas à un tel usage, parce que les analyses reposent sur l'hypothèse d'un échantillonnage aléatoire simple. Ainsi, les résultats des analyses effectuées à l'aide de ces progiciels ne sont pas valables si le plan de sondage comporte un échantillonnage à plusieurs degrés, une stratification ou une mise en grappes fondées sur la répétition. Il sera question du logiciel WesVar qui produit des estimations d'enquête et des estimations itératives de variance en tenant bien compte de l'application de méthodes complexes d'échantillonnage et d'estimation. Nous illustrerons aussi les caractéristiques WesVar à l'aide des données de deux enquêtes Westat à plan d'échantillonnage complexe, à savoir la Troisième étude internationale sur les mathématiques et les sciences (TEIMS) et la « National Health and Nutrition Examination Survey » (NHANES).

Mots clés : pondération des données d'enquête; correction de non-réponse; poststratification; méthode itérative du quotient; tableaux à deux entrées et à entrées multiples; estimation de quantiles; modèles logistiques et multinomiaux; estimation de variance fondée sur la répétition.

1. INTRODUCTION

Le recours aux techniques statistiques types ne convient pas à l'analyse de données recueillies dans le cadre d'enquêtes complexes. On a ordinairement besoin de logiciels spécialisés pour tenir compte de caractéristiques comme les probabilités variables de sélection et une sélection non indépendante. Les estimations de données d'enquêtes complexes sont elles-mêmes compliquées, qu'il s'agisse de moyennes de rapports, de coefficients de régression ou de rapports de cotes, et on aura besoin de méthodes d'estimation des erreurs-types pour prendre ces éléments de complexité en compte.

WesVar produit des estimations d'enquête et des estimations de variance fondées sur la répétition qui tiennent bien compte de l'emploi de méthodes complexes d'échantillonnage et d'estimation. L'estimation de variance fondée sur la répétition consiste à produire des estimations pour des sous-groupes de l'échantillon entier et à ensuite calculer la variance entre les estimations dues à chacune de ces répliques.

WesVar est un programme souple qui peut s'adapter à une grande diversité de plans d'échantillonnage complexes à degrés multiples, à stratification et à probabilités inégales d'échantillonnage. Les estimations de variance fondées sur la répétition peuvent également s'adapter à de nombreux types de plans d'estimation : correction de non-réponse, estimation de rapports (poststratification et méthode itérative du quotient, etc.), etc. Les puissantes caractéristiques de WesVar et son interface Windows[MD] conviviale facilitent la création de poids de répliques à mettre au service de l'analyse, ainsi que le transfert et l'analyse de fichiers qui possèdent déjà les poids de réplique.

WesVar peut produire des estimations de statistiques comme les totaux et les moyennes, tout comme les estimations correspondantes d'erreurs-types. On peut aussi l'utiliser aisément pour des estimations de variance dans le cas defonctions complexes d'estimation (rapports, différences de rapports, rapports de cotes en expression logarithmiqueetc.) fondées sur des données en tableaux. Il peut également permettre d'estimer les coefficients de

[1]Westat, Inc., 1650, boul. Research, Rockville, Maryland 20850, États-Unis.

modèles de régression linéaire et logistique et de vérifier la signification de combinaisons linéaires d'estimations paramétriques.

À la section 2, nous donnerons un aperçu des méthodes fondées sur la répétition que comporte WesVar. À la section 3, il sera question des types de pondération et, aux sections 4 et 5, des estimations fondées sur des données en tableaux et des modèles de régression. À la section 6, nous citerons des exemples d'analyses WesVar à l'aide de données d'enquêtes complexes. La version 4.2 de ce programme a récemment été rendue publique. On trouvera une description plus fine de ses caractéristiques à www.westat.com/wesvar. Une des mesures les plus utiles de mise à niveau dans la version 4.2 est la création d'une capacité de transfert direct de fichiers dans divers formats : entre autres, SAS[MD] (sd2, sas7bdat, ssd, transport), SPSS[MD], Stata[MD] et Excel[MD] et Access[MD] de Microsoft.

2. APERÇU DES MÉTHODES FONDÉES SUR LA RÉPÉTITION

L'idée qui se trouve à la base même de la répétition est de prélever à répétition des sous-échantillons sur l'échantillon entier, de calculer la statistique d'intérêt pour chaque sous-échantillon et de se servir ensuite des valeurs dégagées de sous-échantillon ou de réplique pour estimer la variance de la statistique de l'échantillon entier. Les méthodes fondées sur la répétition varient selon les façons de prélever les sous-échantillons. Ces sous-échantillons sont appelés *répliques* et les statistiques qui en sont tirées, *estimations itératives*. WesVar soutient les méthodes fondées sur la répétition suivantes d'estimation de variance :

- répétition compensée pour les plans d'échantillonnage à stratification comptant deux unités primaires d'échantillonnage (UPE) par strate en répétition compensée (RC);
- version RC de Fay (FAY);
- méthode jackknife pour les plans non stratifiés (JK1);
- méthode jackknife pour les plans stratifiés comptant deux UPE par strate (JK2);
- méthode jackknife pour les plans stratifiés comptant deux UPE et plus par strate (JKn).

D'autres méthodes fondées sur la répétition comme la méthode bootstrap peuvent être employées dans WesVar, mais il faut entrer dans le programme les facteurs et les poids de réplique qui conviennent.

Posons que $\hat{\theta}$ est l'estimation d'échantillon entier d'un paramètre de population quelconque θ. L'estimateur par répétition de variance $v\left(\hat{\theta}\right)$ calculé par WesVar prend la forme

$$v\left(\hat{\theta}\right) = c\sum_{g=1}^{G} f_g h_g \left(\hat{\theta}_{(g)} - \hat{\theta}\right)^2 , \qquad (2.1)$$

où

$\hat{\theta}_{(g)}$ est l'estimation de θ fondée sur les observations incluses dans la g^{e} réplique;
G est le nombre total de répliques formées;
c est une constante qui dépend de la méthode choisie.

Le facteur f_g est une correction de population finie qui peut s'employer avec les méthodes jackknife; h_g est un facteur d'échelle uniquement utilisé aux fins de l'application de ces mêmes méthodes.

Un des grands avantages de la répétition réside dans son utilité au stade de l'analyse. On applique la même méthode d'estimation pour l'échantillon entier et chacune de ses répliques. On calcule ensuite immédiatement les estimations de variance par une technique simple. Ajoutons que la méthode est applicable à la plupart des statistiques : moyennes, pourcentages, rapports, coefficients de régression, combinaisons comme les différences, etc. Il est aussi possible de produire de telles estimations pour des groupes d'analyse ou des sous-populations. L'utilisateur n'a pas à comprendre les méthodes d'échantillonnage ni d'estimation si les valeurs de pondération par répétition sont comprises avec les données fournies.

Un autre grand avantage de la répétition est qu'elle offre un moyen simple de tenir compte des corrections dans le cadre de la pondération. Souvent, on corrige la pondération d'échantillonnage en fonction de la non-réponse, de la

poststratification ou de l'alignement sur des totaux de contrôle. En calculant séparément les corrections de pondération pour chaque réplique, les estimations de variance peuvent traduire les effets de telles corrections. C'est ainsi que les estimations de variance fondées sur la répétition présentent des propriétés statistiques souhaitables sous le double angle du plan d'échantillonnage et du modèle. Shao (1996) passe en revue les méthodes d'estimation de population finie et leurs propriétés relatives au plan d'échantillonnage, tandis que Valliant, Dorfman et Royall (2000) décrivent les propriétés relatives au modèle. L'annexe D du guide d'utilisation WesVar (Westat, 2000) indique en détail comment construire des sous-groupes répliques pour certains plans d'échantillonnage populaires. Elle schématise aussi les questions à prendre en considération lorsqu'on forme des répliques.

2.1 Répétition compensée (RC) et méthode de Fay

La répétition compensée s'applique à des plans d'échantillonnage à degré unique ou à degrés multiples où la population d'UPE peut être regroupée en *L* strates de variance (ce qu'on appelle VarStrats dans WesVar) comptant chacune deux UPE (appelées VarUnits). Dans le cas de plans d'échantillonnage qui ne se prêtent pas à cette formation type, on peut souvent légitimement constituer des strates ou grouper les UPE – comme le décrit l'annexe D du guide WesVar – pour créer un plan de stratification « deux à deux ».

On établit chaque estimation de demi-échantillon répété en sélectionnant une des deux VarUnits de chaque VarStrat à l'aide d'une matrice de Hadamard (voir McCarthy, 1969). Seules les VarUnits sélectionnées servent ensuite à l'estimation du paramètre d'intérêt. On stocke des matrices de Hadamard de taille diverse (maximum de 512) et la même matrice s'applique à chaque fichier ayant le même nombre de VarStrats. Ce sont des matrices qui donnent des ensembles de répliques en compensation orthogonale (voir Wolter, 1985, p. 115). WesVar créera plus de 512 poids de répliques pour le RC (ou la version Fay de cette méthode) si vous pouvez fournir une matrice de Hadamard appropriée dans un fichier texte. La taille maximale de cette matrice doit être de 9 984 sur 9 984.

La méthode de Fay (Fay, 1989) est une version RC qui offre de meilleures propriétés dans certains cas. La méthode RC ordinaire s'expose à des problèmes lorsqu'on établit une estimation pour un petit domaine ou qu'on estime un rapport là où le dénominateur compte peu d'éléments de l'échantillon. La méthode de Fay résout ces problèmes en retenant toutes les unités de l'échantillon dans chaque réplique, tout en modifiant la pondération d'échantillonnage autrement que dans la méthode RC type.

2.2 Méthodes jackknife

La méthode jackknife 1 (JK1) trouve sa place lorsqu'on n'a pas échantillonné avec une stratification explicite. Pour former les répliques JK1, on délimite *G* sous-ensembles de VarUnits. Si un sous-ensemble est formé d'une seule unité de l'échantillon (une seule VarUnit), la JK1 est la méthode jackknife type par retrait d'une unité. On forme les répliques en retirant une VarUnit à la fois et en multipliant les poids des autres VarUnits par $G/(G-1)$, où *G* est le nombre de répliques. Lorsque chaque sous-ensemble à soumettre au retrait est un groupe d'unités constitué au hasard, la JK1 correspond pour l'essentiel à la méthode des groupes aléatoires sans chevauchement dont parle Wolter (1985, chapitre 2). Le maximum d'itérations est de 9 999.

Le plan d'échantillonnage de base qui est posé pour l'application de la méthode jackknife 2 (JK2) est le même que dans une répétition compensée (RC), deux UPE (VarUnits) étant sélectionnées dans chacune de *L* strates (VarStrats). Avec un plan de stratification deux à deux, il y a simplification de la méthode jackknife pour les estimateurs linéaires. Quant à la méthode JKn, elle est plus générale et s'applique lorsque le nombre d'UPE (VarUnits) que compte une strate (VarStrat) est de deux et plus. Le nombre de répliques *G* est égal à $\sum_{h=1}^{L} n_h$, où *L* est le nombre de VarStrats et n_h , le nombre de VarUnits dans la VarStrat *h*.

2.3 Degrés de liberté des estimations de variance

Les bornes inférieures et supérieures des intervalles de confiance et la valeur *p* des statistiques de test sont fondées sur une statistique *t* dont les degrés de liberté (DL) sont déterminés par la méthode d'estimation de variance. Rust et Rao (1996) nous livrent la théorie des approximations DL. Une autre possibilité est de poser un nombre infini de degrés de liberté, auquel cas on recourra à l'approximation normale. Pour les méthodes d'estimation de variance que comporte le programme WesVar, les nombres (implicites) par défaut de degrés de liberté, qui sont fonction du nombre de VarStrats et de répliques, sont examinés à l'annexe A du guide WesVar.

2.4 Facteurs de correction de population finie

Dans la théorie des méthodes fondées sur la répétition, on suppose que les unités d'échantillonnage de premier degré ont été sélectionnées avec remise ou, si tel n'est pas le cas, que le plan d'échantillonnage peut être traité sans crainte comme s'il y avait eu remise. Notre capacité est restreinte d'apporter une correction de population finie (CPF) pour les méthodes jackknife, mais non pas pour la répétition compensée ni sa version Fay. Les CPF visent individuellement les répliques, comme on peut le voir à l'expression (2.1). On en discute en détail à l'annexe A du guide WesVar.

Les facteurs de correction de population finie peuvent se trouver dans un fichier séparé ou être spécifiés à l'écran *Attach Factors*.

3. PONDÉRATION WESVAR

La première étape de la démarche de pondération consiste à définir la méthode d'estimation de variance qui sera appliquée (RC, Fay, JK1, JK2 ou JKn). L'utilisateur spécifie les strates de variance (VarStrats) et les UPE (VarUnits) que comptera chacune de ces strates, ainsi que la pondération de base de l'échantillon entier. Le logiciel calcule ensuite une pondération par répétition de base convenant à la méthode choisie d'estimation de variance. WesVar permet en outre une correction de non-réponse et des corrections de poststratification et de méthode itérative du quotient.

3.1 Correction de non-réponse

WesVar calcule les corrections de non-réponse par la méthode des classes de pondération. Les unités admissibles répondantes et non répondantes sont rangées dans des cellules et la correction de non-réponse se calcule pour ces diverses cellules. WesVar établit les corrections de non-réponse pour l'échantillon entier et les poids de réplique. Il faut que les poids d'échantillon entier et les poids des répliques soient versés au fichier de données avant toute correction de non-réponse. Les produits de l'exercice sont les poids d'échantillon entier et les poids des répliques en correction de non-réponse dans un nouveau fichier de données WesVar qui compte le même nombre d'enregistrements que le fichier d'entrée. Les non-répondants figureront à ce fichier en pondération nulle. Toutes les unités inadmissibles demeureront au fichier avec leur pondération d'origine. On peut avoir à les éliminer en se servant des caractéristiques de sous-populations ou de sous-ensembles lorsqu'on procède aux totalisations.

Par la technique de correction de non-réponse, WesVar peut aussi modifier la pondération d'échantillonnage pour les cas d'admissibilité inconnue. Les facteurs de correction qui interviennent alors se calculent de la même manière que les corrections de non-réponse. Si on a recours à une telle correction, elle doit se faire préalablement à la correction de non-réponse. La correction d'admissibilité inconnue s'impose lorsqu'il est impossible de juger de l'admissibilité de toutes les unités de l'échantillon. En première étape, on met la pondération de base des unités de l'échantillon dont l'admissibilité est inconnue en distribution proportionnelle par rapport à celles dont l'admissibilité est connue. En seconde étape, on met les poids des non-répondants en correction d'admissibilité dans une distribution proportionnelle par rapport aux répondants. On doit déterminer séparément pour chaque correction les classes de pondération en correction d'admissibilité inconnue et en correction de non-réponse.

3.2 Poststratification

Pour qu'il y ait poststratification, l'utilisateur doit fournir un fichier avec des codes de cellules de poststratification et des totaux de contrôle. À noter que le code de cellule peut être tiré d'une combinaison d'autres variables, si bien que la poststratification ne se limite pas nécessairement à une variable unique. WesVar calcule les corrections de poststratification de l'échantillon entier et des poids des répliques. Si le fichier contient déjà l'échantillon entier et les poids de réplique en poststratification, on doit faire entrer ces valeurs de pondération dans l'analyse des données. Il n'y a pas d'instructions particulières dont on ait besoin pour indiquer au programme que les poids ainsi entrés sont des valeurs de poststratification.

3.3 Méthode itérative du quotient

Pour utiliser la fonction « méthode itérative du quotient » de WesVar, il faut spécifier un fichier texte qui contient les totaux de contrôle de chaque dimension. Les zones de ce fichier doivent comprendre le niveau de la variable et le total de contrôle correspondant. WesVar permet une itération qui peut porter sur huit dimensions au maximum. La méthode étant itérative, l'utilisateur peut spécifier le maximum d'itérations ou de tolérances quant à la distance absolue ou relative qui sépare les estimations marginales des totaux de contrôle. Le traitement s'arrête au terme du nombre spécifié d'itérations ou une fois rempli un des critères d'arrêt que précise l'utilisateur. Le nombre d'itérations par défaut est de 4 et le maximum est de 100.

3.4 Unités autoreprésentatives

WesVar peut s'appliquer à des plans d'échantillonnage à unités autoreprésentatives (AR). Il convient de noter que, dans un échantillonnage à plusieurs degrés, les unités qui sont contenues dans une unité AR à un degré mais qui sont sélectionnées à un autre degré ne sont pas des unités AR. La méthode ne devrait s'appliquer qu'aux unités qui représentent une certitude absolue, et WesVar avertira l'utilisateur par un message que les unités AR reconnues sont traitées comme une certitude absolue dans le calcul des variances.

La spécification des unités AR dépend de la méthode d'estimation de variance. Si on emploie la méthode JK1, la variable de définition de VarUnit sert à reconnaître les unités AR. Pour les autres méthodes, la variable VarStrat sert à cette identification. Une fois les unités AR reconnues, on crée les poids itératifs liés aux unités « non-AR ». Le nombre d'itérations dépend du nombre d'unités « non-AR » pour la méthode JK1 ou du nombre de strates « non-AR » pour les autres méthodes.

4. ESTIMATIONS « TABLE REQUEST »

L'établissement d'estimations et de leurs erreurs-types dans des tableaux est largement commandé dans WesVar par la spécification d'options *Table Request* comme *Analysis Variables, Table Variables, Computed Statistics* et *Cell Function Statistics.* Une *Table Request* vous permet d'analyser les données d'enquêtes complexes en produisant des statistiques comme des totaux, des moyennes de rapports, des proportions, des rapports généraux ou d'autres fonctions de totaux. L'option *Analysis Variables* vous permet de spécifier les variables numériques pour lesquelles des agrégats de population doivent être estimés (comme le revenu). L'option *Computed Statistics* sert à l'établissement d'estimations qui sont des fonctions de totaux estimés.

Souvent, on a besoin de statistiques pour des sous-groupes (ou des domaines) et, dans bien des cas, l'analyse exigera qu'on utilise des totalisations croisées. On peut prendre l'option *Table Set* de *Table Request* pour spécifier des sous-groupes définis par une variable catégorique, tout comme des sous-groupes définis par la totalisation croisée de deux variables catégoriques et plus. Dans un tableau, l'option *Cell Function Statistics* sert à l'établissement d'estimations qui sont des fonctions des estimations de deux cellules et plus d'un tableau.

4.1 Estimations de totaux et de rapports

Par la fonction *Table Request*, on calcule les totaux pondérés des variables d'intérêt spécifiées. Le rapport général estimé est celui de deux totaux estimés. Les proportions et les moyennes de rapports sont des cas d'espèce du rapport général.

4.2 Médianes et quantiles

Par la fonction *Table Request*, on peut aussi estimer la médiane et la valeur à tout percentile entier. Le recours aux méthodes fondées sur la répétition pour l'estimation directe des variances de percentiles estimés, et notamment de la médiane, a constitué et constitue toujours un domaine de recherche dynamique. Les travaux de Kovar, Rao et Wu (1988) indiquent que la méthode jackknife donne de piètres résultats dans des estimations de quantiles, contrairement à la répétition compensée et à sa version Fay (Rao et Shao, 1999). WesVar peut calculer les variances de quantiles indirectement par la méthode de Woodruff (voir Särndal, Swensson et Wretman 1992) ou directement par répétition.

Les méthodes de calcul de variance de quantiles sont les méthodes avec et sans groupes. La méthode à groupes est le moyen de limiter les calculs portant sur de grands ensembles de données. Le nombre de groupes peut être de 3 à 500 (le nombre « implicite » par défaut est de 50).

4.3 Computed Statistics

Un certain nombre d'éléments sont disponibles pour la fonction *Computed Statistics* : moyenne, moyenne géométrique, médiane, quantiles, logarithmes, etc. D'autres fonctions *Computed Statistics* plus complexes peuvent également être spécifiées dans WesVar. À noter que, si une variable de tableau est spécifiée, l'expression est évaluée pour chaque recoupement de variables de tableau.

4.4 Valeurs plausibles ou imputation multiple

On doit la théorie de l'estimation par valeurs plausibles (VP) dans les évaluations de réussite scolaire à Mislevy et Sheehan (1989). Les travaux de ces auteurs sont fondés sur une technique plus générale d'imputation multiple décrite par Rubin (1987). Si on pose M valeurs plausibles et les estimations $\hat{\theta}_1$, $\hat{\theta}_2$, … $\hat{\theta}_M$ d'un paramètre θ selon ces VP, on emploiera un algorithme de combinaison des résultats des analyses répétées pour estimer le paramètre θ et sa variance.

L'estimateur du paramètre est la moyenne des estimations VP, c'est-à-dire $\hat{\theta}^* = M^{-1}\sum_{m=1}^{M}\hat{\theta}_m$. La variance de $\hat{\theta}^*$ se calcule par des formules propres à la méthode des valeurs plausibles ou à l'imputation multiple. Si nous désignons la variance par répétition de $\hat{\theta}_m$ par v_m, l'estimation finale de la variance se calculera par

$$v\left(\hat{\theta}^*\right) = \frac{1}{M}\sum_{m=1}^{M} v_m + \left(1+\frac{1}{M}\right)\times\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\theta}_m - \hat{\theta}^*\right)^2,$$

où le premier terme est la composante « intra » et le second, la composante « inter » de la variance.

4.5 Taux normalisés

WesVar peut calculer des taux normalisés par la méthode de normalisation directe. On corrige les taux par les totaux de contrôle (ou une distribution type) de manière à éliminer les effets de composition de la population au moment de procéder à des comparaisons entre groupes. Ainsi, les taux de mortalité de deux pays peuvent être en normalisation

d'âge si bien qu'un rapprochement de taux comparatifs nationaux ne subira pas l'influence des différences de structure par âge entre les pays en question.

4.6 Calcul de différences et autres estimations complexes

WesVar peut aussi servir à des analyses par fonctions complexes de totaux estimés par une spécification de *Computed Statistics*, de variables de tableau et d'une fonction des estimations de cellules à l'écran *Cell Functions*. La variable de tableau définit les cellules de la totalisation croisée. Il importe de distinguer la fonction *Computed Statistics* de la fonction *Cell Function Statistics*. La seconde permet de calculer les différences (ou des fonctions plus complexes) entre les cellules d'une totalisation croisée pour la même variable dans une diversité de sous-populations. Quant à la fonction *Computed Statistics*, elle sert au calcul des différences entre variables pour la population entière.

4.7 Effets de plan d'échantillonnage

L'effet de plan d'échantillonnage que calcule WesVar est le rapport entre la variance relative au plan d'échantillonnage effectif et la variance relative à un échantillon aléatoire simple avec remise. Cette définition diffère de celle de Kish (1965), qui porte au dénominateur la variance d'un échantillonnage aléatoire simple sans remise. De plus, la variance relative à l'échantillonnage aléatoire simple avec remise est conditionnelle à la taille d'échantillon réalisée pour le domaine d'intérêt. Toutefois, pour les tableaux à entrées multiples, la variance relative au plan d'échantillonnage effectif se calcule en fonction de la taille d'échantillon d'une totalisation marginale de deux variables.

4.8 Intervalles de confiance pour les proportions

Dans la méthode « implicite » par défaut d'élaboration d'un intervalle de confiance pour un pourcentage, on prend un intervalle symétrique de la forme $\hat{p} \pm t_\nu \sqrt{v(\hat{p})}$, où $\hat{p}$ est le pourcentage estimé, $v(\hat{p})$ la variance estimée et t_ν le multiplicateur de la distribution de t à ν degrés de liberté. Le défaut de cette méthode est que, dans le cas des pourcentages extrêmes, la borne supérieure ou inférieure de confiance peut sortir de la fourchette d'acceptabilité [0,100]. Dans un tel cas, la méthode d'évaluation de Wilson (Newcombe, 1998) peut servir à la délimitation d'intervalles de confiance qui ne quitteront jamais la fourchette [0,100]. À la différence des intervalles de *t* par approximation, des intervalles pas approximation *t*, les intervalles de Wilson ne seront pas disposés symétriquement autour de l'estimation ponctuelle du pourcentage.

4.9 Statistiques chi-carré

Avec la fonction *Table Request*, les tests d'indépendance sont simples dans un tableau à deux entrées. On calcule à cette fin la statistique chi-carré de Pearson et deux statistiques chi-carré RS2 et RS3 qui ont été modifiées en fonction du plan d'échantillonnage complexe. Les statistiques chi-carré modifiées sont issues d'une correction de la statistique chi-carré de Pearson par estimation d'un « effet de plan d'échantillonnage », ainsi que l'ont proposé Rao et Scott (1981, 1984).

4.10 Techniques de traitement des données manquantes

Si l'ensemble de données d'entrée contient plus d'une représentation de données manquantes, toutes les représentations en question sont converties en une représentation unique pour WesVar et traitées comme une même valeur manquante dans toutes les techniques appliquées.

Si des données manquent, la fonction *Table Request* produira quand même les estimations et les erreurs-types dans la plupart des cas. Dans la délimitation des sous-groupes d'un tableau, l'opération « implicite » par défaut consiste à exclure de la sortie toute statistique relative à des sous-groupes définis par une valeur manquante pour une des

variables catégoriques de tableau. Si une variable d'analyse ou de définition d'une fonction *Computed Statistics* est manquante, l'opération par défaut consiste à retirer tout l'enregistrement de la demande. Ainsi, WesVar limite les données aux enregistrements exempts de valeurs manquantes pour toutes les variables d'analyse et toutes les variables de spécification de *Computed Statistics*. Les cas à valeurs manquantes sont aussi exclus des demandes de régression.

5. MODÈLES DE RÉGRESSION

L'utilisateur peut demander des modèles de régression linéaire, logistique ou multinomiale. Par les techniques de régression WesVar, on estime les paramètres d'un modèle de régression et fournit diverses autres statistiques, dont un test d'ajustement général du modèle et de chacun de ses paramètres, des mesures d'ajustement, des rapports de cotes, etc. À l'annexe C du guide WesVar, on examine en détail les méthodes de calcul servant à résoudre les équations relatives à des estimations paramétriques et à d'autres statistiques de modèle. Nous donnerons ici seulement un aperçu des techniques en usage.

5.1 Modèle de régression linéaire

WesVar ajuste des modèles de régression linéaire de la forme $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$, où $\mathbf{Y}$ est le $n \times 1$ vecteur colonne d'observations de l'échantillon, $\beta = \left(\beta_1, \ldots, \beta_p\right)$ le $p \times 1$ vecteur colonne de coefficients de régression, $\mathbf{X}$ la $n \times p$ matrice de variables indépendantes et $\mathbf{e}$ le $n \times 1$ vecteur colonne d'erreurs aléatoires. La i^e ligne de $\mathbf{X}$ est le vecteur des variables explicatives pour l'unité i, $\mathbf{x}_i' = \left(x_{i1}, x_{i2}, \ldots, x_{ip}\right)$. Le vecteur $\mathbf{x}_i$ peut contenir des variables continues ou discontinues. L'estimation par les moindres carrés pondérés du vecteur paramétrique est $\mathbf{b} = \left(\mathbf{X'WX}\right)^{-1}\mathbf{X'WY}$, où $\mathbf{W}$ est la $n \times n$ matrice diagonale tirée des n poids d'échantillon entier $w_1, w_2, \ldots, w_n$.

WesVar calcule aussi l'estimation selon chaque réplique par les moindres carrés pondérés. Ces estimations de répliques sont alors combinées à l'aide d'une formule matricielle analogue à (2.1) pour donner une estimation fondée sur la répétition de la matrice des covariances de **b**. On se sert ensuite des éléments de cette estimation de covariance pour élaborer des tests t et des intervalles de confiance pour les divers coefficients, ainsi que des tests spéciaux de combinaisons linéaires de coefficients de régression.

5.2 Modèle de régression logistique

Dans un modèle de régression logistique où $\mathbf{x}_i$ est défini comme ci-dessus pour un modèle de régression linéaire, l'espérance d'une variable dichotomique Y_i est posée comme $p\left(\mathbf{x}_i\right) = \dfrac{exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}{1 + exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}$. WesVar applique une technique de pseudomaximum de vraisemblance pour l'estimation paramétrique, où on trouve la valeur **b** qui maximise l'estimation logarithmique pondérée de vraisemblance de l'échantillon. C'est ce que l'on appelle l'estimation de pseudomaximum de vraisemblance (EMV). WesVar résout les équations de pseudomaximum à l'aide d'une version modifiée de la méthode de Newton-Raphson. Il y a aussi des contrôles de calcul tant pour la convergence que pour la divergence des estimations paramétriques.

WesVar établit trois mesures d'ajustement pour les modèles de régression logistique par comparaison de « logvraisemblance » entre le modèle ajusté et un modèle qui comprend seulement la valeur à l'origine. Ce sont les mesures de « logvraisemblance » négative (ou d'entropie) et des rapports de vraisemblances de Cox-Snell et d'Estrella dont parlent les études spécialisées. Toutes trois s'exécutent par défaut. Ces méthodes et plusieurs autres sont passées en revue dans Mittlböck et Schemper (1996) et Estrella (1998).

Par défaut, WesVar calcule des rapports de cotes seulement pour les principaux effets qui ne sont pas liés aux interactions. Il calcule aussi un intervalle de confiance bilatéral pour chacun de ces rapports. Le logarithme de la cote

(ou la « logcote ») que la réponse soit 1 pour l'unité *i* de l'échantillon est $log\left[\frac{p(\mathbf{x}_i)}{1-p(\mathbf{x}_i)}\right] = \mathbf{x}_i'\boldsymbol{\beta}$, ce que l'on appelle aussi le logit de $p(\mathbf{x}_i)$. Dans le cas d'une variable x_{ik} continue, un paramètre β_k est le logarithme du rapport de cotes d'un changement unitaire de x_{ik}, toutes les autres variables étant constantes. La quantité $exp(\beta_k)$ est donc le rapports de cotes d'un changement unitaire de x_{ik} pour toute unité *i*. On trouve l'intervalle de confiance bilatéral pour une cote en fixant un tel intervalle au paramètre β_k et en le transformant à l'échelle du rapport de cotes. On estime l'erreur-type de l'estimation de β_k par la méthode fondée sur la répétition que spécifie l'utilisateur.

Dans un modèle comportant des interactions, le calcul d'un rapport de cotes significatif est plus complexe, mais WesVar offre un outil par lequel l'utilisateur peut calculer un raport spécial en combinant des estimations des paramètres du modèle.

5.3 Modèle de régression logistique multinomiale

Le modèle de régression logistique multinomiale ou le modèle des logits généralisés est un développement du modèle de régression logistique. De tels modèles sont décrits en détail dans Agresti (1990). Dans une régression logistique multinomiale, la variable de réponse *Y* peut correspondre à une réponse catégorique à *K* catégories. Comme dans une régression logistique dichotomique, WesVar recourt à la méthode de pseudomaximum de vraisemblance pour l'estimation paramétrique, et ce, tant pour l'échantillon entier que pour les répliques.

Les trois mesures d'ajustement précitées – entropie, Cox-Snell et Estrella – s'établissent dans une régression logistique multinomiale comme dans une régression logistique dichotomique. Les raports de cotes et les intervalles de confiance se calculent de la même manière. Dans la régression logistique multinomiale, le rapport de cotes « implicite » par défaut est le rapport des cotes d'appartenance à la catégorie *k* du modèle multinomial par rapport à la catégorie de référence *K* pour un changement unitaire d'une des variables explicatives (les autres variables prévisionnelles étant constantes). WesVar offre également une fonction de calcul de rapports de cotes spéciaux qui ne s'exécute pas par défaut.

6. EXEMPLES D'UTILISATION DE WESVAR

Nous citerons deux exemples d'analyses de données, celles de la Troisième étude internationale sur les mathématiques et les sciences (TEIMS) et de la National Health and Nutrition Examination Survey (NHANES).

6.1 Troisième étude internationale sur les mathématiques et les sciences (TEIMS)

Le plan d'échantillonnage de base de l'étude TEIMS est généralement à deux degrés avec strates et grappes. Le premier degré est un échantillon stratifié d'établissements scolaires et le deuxième, des échantillons de classes des établissements échantillonnés pour chaque année d'études admissible et visée. Dans certains pays, on ajoute un troisième degré d'échantillons d'élèves des classes. Le tableau 6.1 compare les programmes non pondéré et pondéré SAS et le programme WesVar dans leurs résultats. Sans programmation spéciale, la fonction SAS PROC MEANS peut traiter une VP à la fois. Les erreurs-types que dégage le programme pondéré SAS ne tiennent pas compte des grappes et se révèlent bien inférieures à celles du programme WesVar où seule la première VP ou, d'une manière plus appropriée, les cinq entrent dans la formule d'imputation multiple. À noter que la composante « inter » est faible par rapport à la composante « intra » dans ce cas pour l'analyse par valeurs plausibles.

Tableau 6.1 : Comparaison des résultats des programmes non pondéré et pondéré SAS et du programme WesVar

Pays	Test	SAS non pondéré		SAS pondéré		WesVar			
		Première VP		Première VP		Première VP		5 VP	
		Moyenne	ET	Moyenne	ET	Moyenne	ET	Moyenne	ET
Pays-Bas	Sciences	551,7	1,3	543,5	1,3	543,5	6,6	544,8	7,0
	Math.	544,4	1,2	537,9	1,3	537,9	6,8	539,9	7,2
Belgique	Sciences	547,1	0,9	534,0	1,0	534,0	2,7	534,9	2,9
	Math.	573,9	0,9	556,5	1,1	556,5	3,1	558,0	3,5

Note : Dans l'analyse des données de la TEIMS, on utilise la méthode de répétition JK2. Le nombre de répliques est de 74.

6.2 « National Health and Nutrition Examination Survey » (NHANES)

L'échantillon de l'enquête NHANES représente toute la population civile hors établissement des 50 États américains et du district de Columbia. On applique un plan d'échantillonnage à quatre degrés. Pour réduire le nombre de déplacements, on définit les UPE comme les comtés ou les groupes de comtés voisins. Le deuxième degré est celui des secteurs aréolaires formés d'îlots ou de combinaisons d'îlots de recensement. Le troisième degré d'échantillonnage vise les ménages et les logements collectifs qui ne sont pas des établissements. Au quatrième degré, on trouve les membres des ménages ou les occupants des logements collectifs. On sélectionne les UPE et les secteurs aréolaires de l'échantillon par échantillonnage PPT (probabilités proportionnelles à la taille). L'échantillon est conçu pour donner une taille d'échantillon à peu près égale par UPE.

Pour illustrer l'application de la méthode de régression logistique, nous utilisons un sous-ensemble de l'échantillon réuni en 1994 pour modéliser la présence d'asthme (ASTHMA) en fonction de l'éventualité que quelqu'un ait fumé 100 cigarettes et plus dans sa vie (CIG100[2]), d'une variable de race-ethnicité à quatre niveaux (RACETHN[4]), de la présence ou de l'absence de fièvre des foins (HAYFEVER[2]) et du sexe (SEX[2]). WesVar note par [*n*] une variable catégorique à *n* niveaux. Il crée automatiquement des variables fictives pour chaque variable catégorique et fixe la solution paramétrique du plus haut niveau de chacune à zéro pour dégager un train de solutions.

Il estime les variances par la méthode de Fay avec 24 répliques. Par défaut, il exécute les estimations paramétriques et divers tests d'hypothèse, dont nous ne parlerons pas par souci de concision. La valeur F d'ajustement global était de 30,35 avec 6 et 18 degrés de liberté, ce qui représente un haut niveau de signification. Le tableau 6.2 présente les rapports de cotes des principaux effets. Ainsi, CIG100.1 désigne le premier niveau de CIG100, celui de la personne qui a fumé 100 cigarettes et plus. Le coefficient correspondant de l'asthme pour quelqu'un qui a autant fumé par rapport aux autres est de 1,65 pour un intervalle de confiance à 95 % de [1,27, 2,13]. Nous présentons aussi un rapport de cotes spécifié par l'utilisateur qui compare les fumeuses ayant la fièvre des foins aux non-fumeurs (de sexe masculin) qui ne l'ont pas (sans égard à la race-ethnicité). À l'échelle des logits, la différence d'« espérance » pour les deux groupes s'établit à CIG100.1 + HAYFEVER.1-SEX.1. Le rapport de cotes, qui est désigné par OR1 au tableau 6.2, est de 9,63 pour un intervalle de confiance à 95 % de [5,52, 16,81].

Tableau 6.2 Rapports de cotes obtenus

Paramètre	Estimation	Valeur inférieure à 95 %	Valeur supérieure à 95 %
CIG100.1	1,65	1,27	2,13
RACETHN.1	0,72	0,47	1,11
RACETHN.2	0,90	0,59	1,38
RACETHN.3	0,53	0,37	0,77
HAYFEVER.1	4,45	3,24	6,11
SEX.1	0,76	0,59	0,98
OR1	9,63	5,52	16,81

RÉFÉRENCES

Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley & Sons.

Estrella, A. (1998). A New Measure of Fit for Equations with Dichotomous Dependent Variables. *Journal of Business and Economic Statistics*, **16**, pp. 198-205.

Fay, R.E. (1989). Theoretical Application of Weighting for Variance Calculation. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, pp. 212-217.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.

Kovar, J.G., Rao, J.N.K., et Wu, C.F.J. (1988). Bootstrap and Other Methods to Measure Errors in Survey Estimates. *Canadian Journal of Statistics*, **16**, pp. 25-45.

McCarthy, P.J. (1969). Pseudo-replication: Half-samples. *Review of the International Statistical Institute*, **37**, pp. 239-264.

Mislevy, R.J., et Sheehan, K.M. (1989). The Role of Collateral Information about Examinees in Item Parameter Estimation. *Psychometrika*, **54**, pp. 661-679.

Mittlböck, M., et Schemper, M. (1996). Explained Variation for Logistic Regression. *Statistics in Medicine*, **15**, pp. 1987-1997.

Newcombe, R.G. (1998). Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods. *Statistics in Medicine*, **17**, pp. 857-872.

Rao, J.N.K., et Scott, A.J. (1981). The Analysis of Categorical Data from Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables. *Journal of the American Statistical Association*, **76**, pp. 221-230.

Rao, J.N.K., et Scott, A.J. (1984). On Chi-Squared Tests for Multiway Contingency Tables with Cell Proportions Estimated from Survey Data. *The Annals of Statistics*, **12**, pp. 46-60.

Rao, J.N.K., et Shao, J. (1999). Modified Balanced Repeated Replication for Complex Survey Data. *Biometrika*, **86**, pp. 403-415.

Rubin, D. (1987). *Multiple Imputations for Nonresponse in Sample Surveys*. New York: John Wiley & Sons.

Rust, K., et Rao, J.N.K. (1996). Variance estimation for complex estimators in sample surveys. *Statistics in Medical Research*, **5**, pp. 381-397.

Särndal, C.E., Swensson, B., et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Shao, J. (1996). Resampling Methods in Sample Surveys, (with Discussion). *Statistics*, **27**, pp. 203-254.

Valliant, R., Dorfman, A.H., et Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley & Sons.

Westat (2000). *WesVar 4.0 User's Guide*. Rockville MD: Westat.

Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.